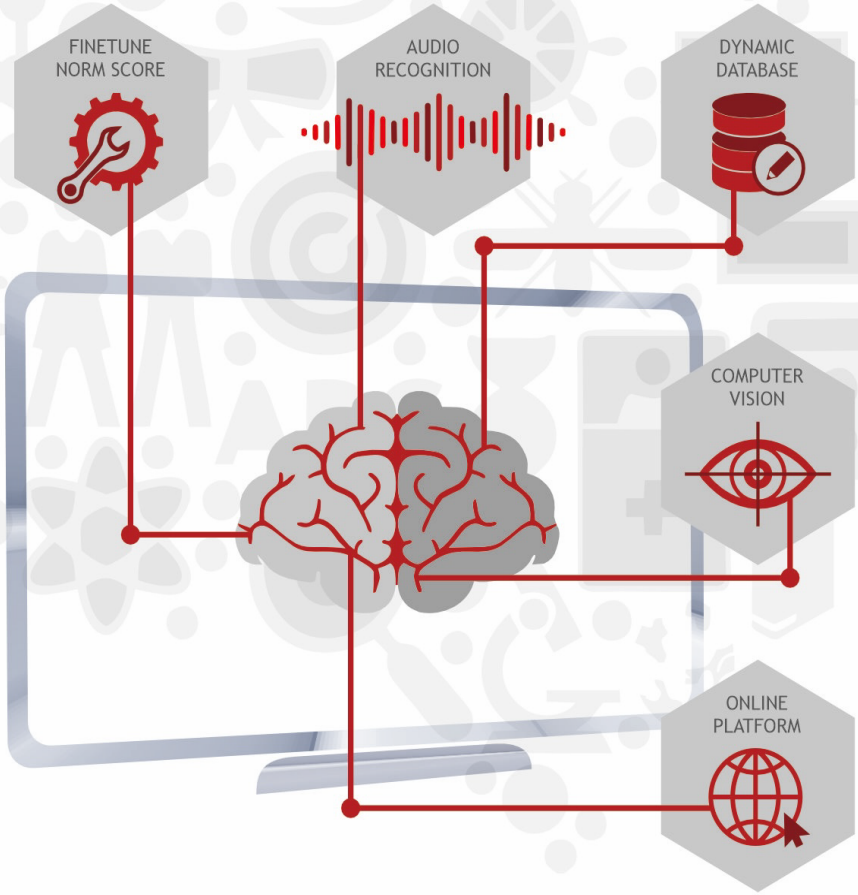


ADVANCING
INDONESIAN NEUROPSYCHOLOGICAL ASSESSMENT WITH TOOLS FROM
COMPUTER SCIENCE

DONDERS SERIES

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Shinta Estri Wahyuningrum

Advancing Indonesian Neuropsychological Assessment with Tools from Computer Science

Shinta Estri Wahyuningrum

Advancing Indonesian Neuropsychological Assessment with Tools from Computer Science
Shinta Wahyuningrum

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Proefschrift AIO | Annelies Lips
Cover: Arwin Purnama Jati
Printing: DPN Rikken/Pumbo

ISBN: 9789465150956
DOI: 10.54195/9789465150956
Free download at: https://doi.org/10.54195/9789465150956

RADBOUD
UNIVERSITY
PRESS

Advancing Indonesian Neuropsychological Assessment with Tools from Computer Science

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag, 18 Juni 2025
om 12.30 uur precies

door

Shinta Estri Wahyuningrum
geboren op 27 september 1982
te Semarang (Indonesië)

Promotoren:
Prof. dr. E.L.J.M. van Luijtelaar
Prof. dr. T.M. Heskes

Copromotoren:
Dr. M.P.H. Hendriks
Dr. A. Sulastri (Universitas Katolik Soegijapranata, Indonesië)

Manuscriptcommissie:
Prof. dr. J.M. Oosterman
Prof. dr. E.O. Postma (Tilburg University)
Prof. dr. ing. R. Pulungan (Universitas Gadjah Mada, Indonesië)

Advancing Indonesian Neuropsychological Assessment with Tools from Computer Science

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Wednesday, June 18, 2025
at 12.30 pm

by

Shinta Estri Wahyuningrum
born on September 27, 1982
in Semarang (Indonesia)

PhD supervisors:
Prof. dr. E.L.J.M. van Luijtelaar
Prof. dr. T.M. Heskes

PhD co-supervisors:
Dr. M.P.H. Hendriks
Dr. A. Sulastri (Soegijapranata Catholic University, Indonesia)

Manuscript Committee:
Prof. dr. J.M. Oosterman
Prof. dr. E.O. Postma (Tilburg University)
Prof. dr. ing. R. Pulungan (Gadjah Mada University, Indonesia)

Table of contents

Chapter 1

Introduction

Cognitive function

Neuropsychological evaluation, used to assess a person's cognitive abilities, started to develop in part, after the Second World War. At that time, psychological methods were employed to evaluate the cognitive condition of individuals suffering from brain damage caused by World War II. However, at that time, psychological tests were not sensitive enough to detect cognitive problems caused by Traumatic Brain Injury (TBI). Some people with TBI occasionally exhibit no apparent or obvious symptoms of cognitive change, but they still might have cognitive problems that can worsen over time (Dean & Sterr, 2013). In the long run, TBI could result in changes in cognitive performance (e.g. working memory and processing speed) next to an increased risk of mental health issues (e.g. anxiety and post-traumatic stress disorder) (Clarke et al., 2012; Wilson et al., 2021). Therefore, more sensitive neuropsychology tests were needed to accurately detect cognitive problems associated with TBI.

Neuropsychology focuses on the relationships between the brain and behaviour, evaluating the characteristics of both behavioural and cognitive functions or dysfunctions (Zucchella et al., 2018). These comprehensive assessments not only identify cognitive weaknesses but also reveal an individual's strengths, providing a more holistic picture of their cognitive and emotional, behavioural profile (Mitrushina et al., 2005; Strauss et al., 2006). Over several decades, there has been a notable evolution in neuropsychological evaluations, driven by advancements in neuroscience, psychology, and technology (Da Motta et al., 2021). Today, neuropsychological assessment plays a crucial role in the diagnosis, understanding, and treatment planning of many neurological and psychiatric disorders (Bilder et al., 2023; Germine et al., 2019; Parsey & Schmitter-Edgecombe, 2013; Philip & Jeffrey, 2002).

Cognitive assessments like the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) serve as starting points due to their quick and easy way of administration, making them ideal for screening purposes (Damian et al., 2011). Fortunately, a comprehensive neuropsychological assessment delves deeper than these two short cognitive tests. Alexander Luria's post-World War II research in Russia marked a pivotal moment in the development of neuropsychological assessment (Casaletto & Heaton, 2017). Alongside these developments in Russia, neuropsychological assessment was also gaining ground in the United States. Arthur Benton decried the absence of standardized measures for neurological disorders such as aphasia and agnosia during the 1950s. Next, the fixed battery approach was pioneered by Ward Halstead and Ralph Reitan who developed the Halstead-Reitan

Neuropsychological Battery (HRNB). It is a set of neuropsychological tests designed to assess the condition and functioning of the brain (Morlett Paredes et al., 2021).

A comprehensive approach is often needed for a complete overview or for diagnosing different kinds of brain dysfunctions, such as in TBI, stroke, and neurodegenerative diseases (e.g. different types of dementia, Parkinson disease, Amyotrophic Lateral Sclerosis /ALS, and Huntington's disease) (Arevalo-Rodriguez et al., 2021; Cumming et al., 2013; Schroeder et al., 2019). Neuropsychology utilizes standardized test batteries to evaluate a broad range of cognitive domains memory, attention, language, problem-solving, and executive functioning (Hoelzle et al., 2011; Kim & Park, 2018). This allows neuropsychologists to construct a detailed cognitive profile for patients, providing valuable insights into their brain function and dysfunction (Lezak et al., 2012; Strauss et al., 2006).

Neuropsychological Assessment

Neuropsychological assessments offer a significant advantage due to their flexibility. Neuropsychologists can tailor the assessment by selecting specific tests from a comprehensive battery or the literature (Bigler, 2007; Casaletto & Heaton, 2017; Durisko et al., 2016). This customization considers factors like the individual's suspected area of brain damage (e.g. stroke), their reported symptoms (e.g. memory problem, difficulty planning), age group (children or adults), and educational background. This flexibility translates into several advantages for individuals and neuropsychologists (Bilder et al., 2023; Casaletto & Heaton, 2017; Schaefer et al., 2023). Firstly, it allows for a more accurate diagnosis of the type and extent of brain damage. Secondly, by focusing on the relevant cognitive domains, the assessment provides a deeper understanding of how brain damage impacts daily life. This information is crucial for developing personalized treatment plans that target specific cognitive impairments for individuals with or without existing cognitive decline (Eling, 2019; Wang et al., 2023). Thirdly, it might save time and available resources.

The reach of neuropsychology extends beyond the clinical realm. Establishing cognitive baselines in healthy individuals may play a vital role in pre-surgical evaluations, college entrance exams, and even employee selection (Lezak et al., 2012). In addition, cognitive assessments can serve as a valuable tool for career and academic planning by pinpointing both strengths and weaknesses, enabling individuals to make informed decisions about their future trajectories (Mitrushina

et al., 2005). Furthermore, neuropsychological tests are invaluable research instruments for investigating the intricate relationship between the brain and behaviour, monitoring the cognitive changes associated with aging, and promoting brain health through targeted interventions (Mancioppi et al., 2019; Perry-Young et al., 2018; Pressler et al., 2018).

The accuracy of neuropsychological evaluations depends on the establishment of robust reference parameters. These parameters provide a standardized framework for interpreting an individual's performance in the context of their age, education, and other relevant demographic factors (Mitrushina et al., 2005; Ruan et al., 2020). Many tests require a certain level of education, which can lead to misinterpretations in populations with limited formal education (Aoki et al., 2023; Da Motta et al., 2021; Kiselica et al., 2020). Scores obtained from healthy control groups (normative data) serve as a critical benchmark against which patient performance is compared (Jaffe et al., 2022; Strauss et al., 2006). Cognitive assessments use cut-off scores to identify individuals who may have cognitive impairment (Kessels & Hendriks, 2023; Lyu et al., 2024; McCarthy et al., 2024; Zhang et al., 2021). These scores provide a benchmark for performance, and scoring below the cut-off indicates a need for further assessment (Conca et al., 2023; Kasten et al., 2021; Weintraub et al., 2009; Woods et al., 2016).

Besides age, education, and sex, the quality of neuropsychological assessments is also affected by cultural and linguistic factors (Davies et al., 2014; Fernandez, 2019; Franzen et al., 2023; Manly, 2008). Most tests are not adapted for different cultural contexts and may yield misleading results. Language-based tasks and assessments for cognition may require adaptations to ensure cultural sensitivity (Pesau et al., 2023; Zucchella et al., 2018). A study conducted across 12 countries found that the participants' nationality accounted for more than 20% of the variance in mentalizing and emotion recognition measures, even after controlling for age, sex, and level of education (Quesque et al., 2022). Also, cultural factors have a determining influence on an individual's behavior regardless of the neuropsychological status of the brain (Manly, 2008). While there has been research on how culture affects cognitive functioning and measurement within the field of neuropsychology, this research has primarily focused on the United States. Consequently, the influence of culture on these factors in other countries is largely unknown and requires further investigation (Pérez-Arce, 1999).

Research has consistently shown that using standardized neuropsychological tests and norms developed in Western Educated Industrialized Rich Democratic (WEIRD)

countries can result in biased assessments of individuals from non-Western cultures (Franzen et al., 2023; Nguyen et al., 2024). For instance, Norman et al. (2011) reported that age significantly impacts test scores differently for African Americans than Caucasians. This underscores the importance of developing new race-specific scores to address the bias. Accurate norms are necessary to prevent misdiagnosis (Fujii & Wong, 2006; Thaler & Fujii, 2014; Werry et al., 2019). The challenge is particularly acute in Indonesia, a nation with over 700 regional languages spread across its 7,000 islands, fostering a multitude of distinct cultural communities (Nur & Shi, 2022). Addressing these issues is paramount to ensuring accurate neuropsychological assessments for individuals from diverse backgrounds.

Empowering Neuropsychology with Computer Science

Computer science has revolutionized neuropsychology by introducing innovative tools and techniques that enhance the efficiency, accuracy, and accessibility of neuropsychological tests (Bilder & Reise, 2019; Miller, 2019). These advancements impact both clinical practice and the technical aspects of testing (Bilder, 2011). Clinical advanced, computerized administration streamlines the process with electronic delivery and scoring, saving valuable time. It also promotes standardization through features like item shuffling and randomization, ensuring consistency across assessments. Additionally, test design has become more flexible and engaging with multimedia elements and adaptive questioning, personalizing the experience for each patient. Technical advancement is equally impressive as clinical advances such as electronic data management simplify tracking patient progress, identifying trends, and generating reports (Bilder & Reise, 2019; De Vent et al., 2016). Securing data delivery is ensured through encryption, while large question banks with algorithmic selection enable customized testing. Finally, automated scoring and real-time feedback enhance both accuracy and patient understanding.

The integration of computerization also enhances data visualization and representation, supporting the analysis and interpretation of test results (Langer et al., 2022; Vogt et al., 2019). This process includes the automatic generation of normative scores and facilitates the assessment procedure (Aoki et al., 2023; De Vent et al., 2016; Kiselica et al., 2020; Ruan et al., 2020; Shirk et al., 2011). It is also crucial to recognize that advancements in computer science extend beyond neuropsychological testing. Neuroimaging, for instance, relies extensively on

computing power to process the vast datasets produced by techniques such as structural and functional imaging techniques like magnetic resonance imaging (sMRI and fMRI), computerized tomography (CT), and positron emission tomography (PET) (Davenport & Kalakota, 2019), as well as multichannel Electroencephalogram (EEG) and Magnetoencephalography (MEG). These techniques are vital for brain visualization, aiding clinicians in identifying the precise locations of brain damage and used to detect early signs of neurodegenerative diseases like Alzheimer's and Parkinson's (Balestri et al., 2024; Pathak et al., 2022).

In healthcare, Artificial intelligence (AI) is emerging as a powerful tool, leveraging the power of computer science. AI can significantly reduce inefficiency in healthcare, improve patient flow and experience, and enhance caregiver experience and patient safety through the care pathway (Bajwa et al., 2021). For example, AI can be applied to the remote monitoring of patients (e.g., intelligent telehealth through wearables/sensors) to identify and provide timely care to patients at risk of deterioration (Leff, 2005; Shaik et al., 2023). As another example, the automated scoring of assessments with the use of AI further improves efficiency, leads to better standardization, and reduces the risk of human error (Cavedoni et al., 2020; Chen et al., 2020; Kang et al., 2019; Zhang et al., 2021).

Advancing Neuropsychological Assessment with Computer Science

Applications of technology in neuropsychological assessment have been developed since the 1980s (Parsey & Schmitter-Edgecombe, 2013). One of the most famous computer assessments was CANTAB (Picanco Diniz et al., 2014). CANTAB is a computer-based cognitive assessment system. It was originally developed at the University of Cambridge and is now commercially available from Cambridge Cognition. The program CANTAB has been developed to meet the demand for customized and flexible battery testing in response to the current technological breakthroughs (Schaefer et al., 2023). CANTAB aims to be culture and language independent by using non-verbal stimuli in most of its tests. This version of the CANTAB assessment is best suited for individuals familiar with English and Western cultural norms.

In the field of neuropsychological assessment, more computer applications are emerging as valuable tools for automating the evaluation of participant responses, particularly for tests that involve qualitative scoring based on written or drawn

responses. An example of such a test is the Five-Point Test: participants connect dots within a grid, creating unique patterns. Traditionally, scoring involves the manual evaluation of both the number of designs and any repetitive patterns (so-called perseveration errors) made (Elderson et al., 2016; Tucha et al., 2012). Similarly, the Trail Making Test requires participants sequentially to connect numbered circles as quickly and accurately as possible without lifting the pen (Reitan & Wolfson, 1995). This automation can reduce the time required for manual scoring. The Figural Reproduction Test and The Rey-Osterrieth Complex Figure (ROCF) also rely on drawn responses. These tests present different geometric stimuli that participants are asked to reproduce from memory. A challenge in the manual scoring of this type of test is the subjective nature of interpretation in the evaluation process (Brito-Marques et al., 2012).

In addition to tests based on handwritten responses including drawings, there are various tests based on verbal responses. Scoring verbal responses in neuro-psychological scoring procedures can also be time-consuming and require significant attention, so can also benefit from automation with the use of ICT. The Digit Span is a prototypical example of such a verbal response test. In the Digit Span, derived from the Wechsler Adult Intelligence Scale (WAIS), the examiner reads a sequence of digits, and the participant repeats them back in the same or reversed order (Lezak, 2004). In the Rey Auditory Verbal Learning Test (RAVLT), a list of 15 words is read aloud by the tester, after which the participant recalls as many as possible (Strauss et al., 2006). In the Boston Naming Test (BNT), the participant names 60 pictures presented within a time limit (Sulastri et al., 2018). In the Phonetic Verbal Fluency Test (PVFT), the participant generates words within a time limit based on a phonemic cue or a starting letter. Similar to the challenges with scoring drawn responses, evaluating spoken responses can lead to inconsistencies, especially in large studies.

Manual vs Automated Scoring: Transforming Neuropsychological Assessments

Manual scoring has several advantages, including expert judgement, flexibility, adaptability, accountability, and transparency (Schatz & Browndyke, 2002). However, it also has notable limitations. The subjective nature of manual scoring can lead to inconsistencies due to factors such as scorer experience, fatigue, or expectations occurring both between different scorers and within the same scorer over time (Awad et al., 2004; Elderson et al., 2016; Fleuret et al., 2011; Kenda et

al., 2022). Furthermore, manual scoring is labor-intensive and time-consuming, especially when dealing with large datasets. The limited availability of trained human scorers further restricts scalability, and maintaining consistent scoring requires rigorous training and ongoing monitoring, adding to the complexity and cost (Diaz-Orueta et al., 2022; Moetesum et al., 2022).

Automated scoring particularly for tasks involving hand-drawn responses like the ROCF test, a widely used tool for measuring visuospatial cognitive abilities (Canham et al., 2000, 2005; Frank & Landeira-Fernandez, 2008; Webb et al., 2021). Different (so-called parallel) versions of the ROCF have been developed because of test-retest learning effects (Langer et al., 2022; Park et al., 2023; Taleb et al., 2019). However, transitioning from manual to automated scoring presents challenges (Porter & Johnson, 2020). One key issue is the variability of human drawings, which makes it difficult for algorithms to consistently identify key features for accurate scoring. In addition, the quality of the input data, such as poor lighting or scanning issues, can further affect the accuracy. Another challenge is the need for detailed and explicit guidelines in test manuals to ensure accurate and consistent evaluations (Diaz-Orueta et al., 2022).

A particular challenge arises when adapting new versions of visuospatial tests for automated scoring. These tests often provide a limited number of drawing samples available, which hinders the development of robust automated scoring systems (Di Febbo et al., 2023). To address these challenges, a flexible and adaptable approach to the development of hand-drawn image processing systems is crucial. This approach should focus on strategies to overcome the limitations of small sample sizes.

Automatic speech recognition (ASR), also known as speech-to-text technology, offers a promising approach for scoring verbal response tests in neuropsychology. Ziman et al. (2018) highlighted the potential of speech-to-text engines as a cost-effective, reliable, and fast method for automated transcription of speech data in psychological experiments. This technology goes beyond simple transcription, with applications in several areas: analyzing speech content to assess meaning and language use (Abad et al., 2013; Iglesias et al., 2022), assisting in the treatment of language disorders such as aphasia (Favaro et al., 2023), and distinguishing speech patterns associated with different conditions of cognitive decline (Holmlund et al., 2020; Jamal et al., 2017).

1

Enhancing Accessibility and Efficiency in Indonesian Neuropsychology with Computer Science

Indonesia, an archipelagic nation with a diverse population, rich cultural landscape, and numerous local languages, faces unique challenges in adapting neuropsychological test batteries. The lack of a centralized framework hinders the standardization of testing procedures, data collection, and the development of normative scores. This fragmented system leads to inconsistencies across assessment practices, making it difficult to accurately measure cognitive functioning across the vast archipelago. Neuropsychologists across the islands lack a central storage system and a shared framework for integrated references. This is further complicated by the absence of a standardized system for interpreting test results within the Indonesian context.

Ensuring consistency in scoring is another critical area that requires attention. Scoring the neuropsychological tests, particularly those involving visuospatial skills is inherently susceptible to inconsistencies due to the variability in human responses. This variability can stem from differences in how the subject can memorize the stimulus or from variations in how scorers evaluate the responses. Furthermore, the limitations of human scorers, such as fatigue, inattention, and even unconscious biases, can further worsen this inconsistency.

In neuropsychological assessment besides the cultures, language barriers remain a significant challenge. Most ASR systems are currently optimized for English. Adapting them for accurate performance in other languages requires significant resources (Alharbi et al., 2021; Baevski et al., 2020; Schneider et al., 2019). This limitation hinders the widespread adoption of ASR for multilingual neuropsychological assessments and this is certainly an issue in Indonesia.

This research aims to explore how advances in computer science can contribute to the development of normative scores for neuropsychological battery tests specifically tailored to the Indonesian population. In addition, the study will investigate the potential of automated scoring systems to improve the objectivity and consistency of neuropsychological test evaluation.

Thesis Outline

The research presented in this thesis aims to advance neuropsychological assessment in Indonesia through the application of computer science. In Chapter 2, we propose the development of a dynamic database and an online platform for Indonesian psychologists, allowing them to choose a reference group with its normative score of the neuropsychological tests, mimicking the client's demographics as close as possible. Traditionally, researchers and clinicians have relied on static normative data, which limits the ability to account for individual variations in cognitive performance. However, the rise of dynamic databases and online platforms has the potential to revolutionize access to these reference scores. This shift enables a more nuanced understanding of cognitive function by allowing adjustments based on individual demographic characteristics such as sex, age, education level, and perhaps ethnic background (Casaletto & Heaton, 2017; Mitrushina et al., 2005; Ruan et al., 2020). Next, cognitive abilities naturally change throughout life. Dynamic databases will offer the flexibility to incorporate the outcomes of longitudinal studies as well.

Indonesia's geographically dispersed nature makes managing and accessing neuropsychological test score data from remote regions challenging. However, the development of a centralized online platform can offer a transformative solution. This platform can enhance neuropsychological assessment in Indonesia by improving data management serving as a central repository for storing and managing test score data collected from diverse locations across the country.

Furthermore, the online platform has the potential to facilitate collaboration and resource sharing among researchers and healthcare professionals throughout Indonesia (Smith et al., 2011). Sharing anonymized test data and best practices can accelerate advancements in Indonesian neuropsychology. Researchers can leverage the platform to identify regional trends, develop culturally sensitive assessments, and refine existing tests to better reflect the unique characteristics of the Indonesian population (De Vent et al., 2016; Shirk et al., 2011).

Chapter 3a presents an in-depth analysis of the effects of age and education on performance in the Indonesia Neuropsychological Test Battery (INTB). The INTB is a comprehensive neuropsychological test battery administered in Bahasa Indonesia, designed to assess a wide range of cognitive domains through a series of ten tests. These domains include learning and memory, language production and comprehension, various types of attention, and executive functions. The specific

tests included in the INTB are Digit Span, Stroop Test, Rey Auditory Verbal Learning Test (RAVLT), Figural Reproduction Test, Trail Making Test (TMT), Five Point Test, phonetic Verbal Fluency Test (pVFT), Boston Naming Test (BNT), Token Test (TT), and Bourdon Wiersma Test (BWT).

This chapter emphasizes the evaluation of the reliability and validity of the INTB, as well as the effect of age and education effects on test scores. The analysis will be conducted at the level of individual subtests to provide detailed insights into the cognitive functions assessed by each test (Strauss et al., 2006). The reliability of the subtest will be assessed among others by using methods such as the test-retest approach, ensuring consistent results over time (Jacobsen et al., 2003; Karlsen et al., 2022; Sim & Wright, 2005). The validity of the tests will be evaluated here by comparing the influence of education and age on INTB performance with findings from similar studies conducted in other countries.

Chapter 3b will provide a comprehensive analysis of the preliminary normative scores for the Javanese population. This analysis will examine the reliability of the tests of the INTB and the underlying cognitive constructs measured by the INTB. Cognition is a multifaceted concept that involves various mental processes and activities, and by examining the interrelations between the test scores, this study aims to gain a more comprehensive understanding of cognition and its constructs. Principal Component Analysis (PCA) will be employed to identify these underlying cognitive constructs. PCA is a statistical technique that reduces data complexity by identifying a smaller number of independent latent variables (Lee et al., 2006; Testa et al., 2012). This analysis will help to clarify the distinct cognitive domains assessed by the INTB.

In addition to identifying these constructs, Chapter 3b will investigate the effects of demographic variables, specifically age and education, on these cognitive constructs. This involves analyzing how these factors influence performance on the INTB and determining the extent to which they affect different cognitive domains. It is expected that there are differences between the cognitive constructs regarding how they are affected by age and education. The results of this investigation will be used to establish construct validity for the INTB by comparing the findings with those from other studies. This comparative approach will help to confirm that the INTB effectively measures the intended cognitive abilities and that its constructs are consistent with established cognitive theories and findings from other populations.

In Chapter 4 we will describe the development of an automated scoring system for the Figural Reproduction Test (FRT) based on AI technology and computer

vision. Traditionally, scoring relies on trained professionals using the scoring system outlined in the test manual. However, this manual scoring process is susceptible to human subjectivity and inconsistencies potentially impacting the accuracy and reliability of test results. By leveraging sophisticated computer vision techniques, this system offers a highly precise, objective, and standardized approach to scoring (Gao et al., 2018; Nevatia, 2000; Pereira et al., 2015).

The automated scoring system will be developed by first digitizing a hand-drawn picture using a scanning device. Utilizing computer vision methodology, the image will be analyze and separated into distinct elements based on the test's manual scoring rules. It will identify the presence, shape, position, and orientation of each scored element (Canham et al., 2000, 2005; Troyer & Wishart, 1997).

To ensure the system's trustworthiness and effectiveness, a rigorous validation process will be undertaken. This process involves comparing the automated scores generated by the AI system with those obtained through manual assessment by two independent, trained professionals. Robust statistical measures, such as Cohen's Kappa statistic, will quantify the level of agreement between the automated and manual scoring methods (Awad et al., 2004; Kenda et al., 2022). Additionally, a confusion matrix will provide a breakdown of specific areas where the automated system agrees or disagrees with the manual scores. This analysis helps identify potential areas for further refinement of the AI model, ensuring its accuracy and reliability in real-world applications (Chen et al., 2020; Vogt et al., 2019).

Chapter 5 will investigate the feasibility of leveraging computer science methods to automate the scoring of verbal speech-based tests. Traditionally, these tests require manual scoring by trained professionals, a process that can be time-consuming and susceptible to human error. This chapter explores the potential of utilizing computer science techniques to automate the scoring process for verbal speech-based tests within the framework of the INTB.

Five out of ten verbal response tests (I-BNT, VFT, Digit Span, RAVLT, Stroop) in the INTB employ distinct stimulus word lists. These tests involve comparing client verbal responses to word stimuli, necessitating automated speech-to-text conversion for accurate scoring. This chapter will initially focus on transcribing spoken words from the I-BNT as an example. However, achieving accurate speech-to-text conversion in Indonesian requires addressing several complexities, such as accent variability, pronunciation nuances, background noise, and speaker independence

across diverse demographics like age and sex (Kim et al., 2020; Kitzing et al., 2009; Koenecke et al., 2020).

Our research aims to identify opportunities for developing accurate speech-to-text transcription models in Indonesia by investigating parameters that impact transcription quality and enhancing their integration into automated assessment systems. This work will contribute to the advancement of neuropsychological assessment by improving the efficiency and reliability of scoring verbal speech-based tests.

Chapter 6 is the General Discussion with conclusions and outlooks on the further usage of computer science in the field of neuropsychology.

References

Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., Farrajota, L., & Martins, I. P. (2013). Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, *27*(6), 1235–1248. https://doi.org/10.1016/j.csl.2012.10.003

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, *9*, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Aoki, S., Nagatani, F., Kagitani-Shimono, K., Ohno, Y., Taniike, M., & Mohri, I. (2023). Examining normative values using the Cambridge neuropsychological test automated battery and developmental traits of executive functions among elementary school-aged children in Japan. *Frontiers in Psychology*, *14*, 1141628. https://doi.org/10.3389/fpsyg.2023.1141628

Arevalo-Rodriguez, I., Smailagic, N., Roqué-Figuls, M., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., Pedraza, O. L., Bonfill Cosp, X., & Cullum, S. (2021). Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*, *2021*(7). https://doi.org/10.1002/14651858.CD010783.pub3

Awad, N., Tsiakas, M., Gagnon, M., Mertens, V. B., Hill, E., & Messier, C. (2004). Explicit and Objective Scoring Criteria for the Taylor Complex Figure Test. *Journal of Clinical and Experimental Neuropsychology*, *26*(3), 405–415. https://doi.org/10.1080/13803390490510112

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (arXiv:2006.11477). arXiv. http://arxiv.org/abs/2006.11477

Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal*, *8*(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

Balestri, W., Sharma, R., Da Silva, V. A., Bobotis, B. C., Curle, A. J., Kothakota, V., Kalantarnia, F., Hangad, M. V., Hoorfar, M., Jones, J. L., Tremblay, M.-È., El-Jawhari, J. J., Willerth, S. M., & Reinwald, Y. (2024). Modeling the neuroimmune system in Alzheimer's and Parkinson's diseases. *Journal of Neuroinflammation*, *21*(1), 32. https://doi.org/10.1186/s12974-024-03024-8

Bigler, E. (2007). A motion to exclude and the 'fixed' versus 'flexible' battery in 'forensic' neuropsychology: Challenges to the practice of clinical neuropsychology. *Archives of Clinical Neuropsychology*, *22*(1), 45–51. https://doi.org/10.1016/j.acn.2006.06.019

Bilder, R. M. (2011). Neuropsychology 3.0: Evidence-Based Science and Practice. *Journal of the International Neuropsychological Society*, *17*(01), 7–13. https://doi.org/10.1017/S1355617710001396

Bilder, R. M., & Reise, S. P. (2019). Neuropsychological tests of the future: How do we get there from here?. *The Clinical Neuropsychologist*, *33*(2), 220–245. https://doi.org/10.1080/13854046.2018.1521993

Bilder, R. M., Widaman, K. F., Bauer, R. M., Drane, D., Loring, D. W., Umfleet, L. G., Reise, S. P., Vannier, L. C., Wahlstrom, D., Fossum, J. L., Wong, E., Enriquez, K., Whelan, F., & Shih, S. (2023). Construct identification in the neuropsychological battery: What are we measuring?. *Neuropsychology*, *37*(4), 351–372. https://doi.org/10.1037/neu0000832

Brito-Marques, P. R. D., Cabral-Filho, J. E., & Miranda, R. M. (2012). Visual reproduction test in normal elderly: Influence of schooling and visual task complexity. *Dementia & Neuropsychologia*, *6*(2), 91–96. https://doi.org/10.1590/S1980-57642012DN06020005

Canham, R. O., Smith, S. L., & Tyrrell, A. M. (2000). Recognition and Grading of Severely Distorted Geometric Shapes from within a Complex Figure. *Pattern Analysis & Applications*, *3*(4), 335–347. https://doi.org/10.1007/s100440070005

Canham, R. O., Smith, S. L., & Tyrrell, A. M. (2005). Location of structural sections from within a highly distorted complex line drawing. *IEEE Proceedings - Vision, Image, and Signal Processing*, *152*(6), 741. https://doi.org/10.1049/ip-vis:20045166

Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological Assessment: Past and Future. *Journal of the International Neuropsychological Society*, *23*(9–10), 778–790. https://doi.org/10.1017/S1355617717001060

Cavedoni, S., Chirico, A., Pedroli, E., Cipresso, P., & Riva, G. (2020). Digital Biomarkers for the Early Detection of Mild Cognitive Impairment: Artificial Intelligence Meets Virtual Reality. *Frontiers in Human Neuroscience*, *14*, 245. https://doi.org/10.3389/fnhum.2020.00245

Chen, S., Stromer, D., Alabdalrahim, H. A., Schwab, S., Weih, M., & Maier, A. (2020). Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, *10*(1), 20854. https://doi.org/10.1038/s41598-020-74710-9

Clarke, L. A., Genat, R. C., & Anderson, J. F. I. (2012). Long-term cognitive complaint and post-concussive symptoms following mild traumatic brain injury: The role of cognitive and affective factors. *Brain Injury*, *26*(3), 298–307. https://doi.org/10.3109/02699052.2012.654588

Conca, F., Esposito, V., Rundo, F., Quaranta, D., Muscio, C., Manenti, R., Caruso, G., Lucca, U., Galbussera, A. A., Di Tella, S., Baglio, F., L'Abbate, F., Canu, E., Catania, V., Filippi, M., Mattavelli, G., Poletti, B., Silani, V., Lodi, R., … Cappa, S. F. (2023). Correction: Italian adaptation of the Uniform Data Set Neuropsychological Test Battery (I-UDSNB 1.0): development and normative data. *Alzheimer's Research & Therapy*, *15*(1), 104. https://doi.org/10.1186/s13195-023-01247-0

Cumming, T. B., Churilov, L., Linden, T., & Bernhardt, J. (2013). Montreal Cognitive Assessment and Mini-Mental State Examination are both valid cognitive tools in stroke. *Acta Neurologica Scandinavica*, *128*(2), 122–129. https://doi.org/10.1111/ane.12084

Da Motta, C., Carvalho, C. B., Castilho, P., & Pato, M. T. (2021). Assessment of neurocognitive function and social cognition with computerized batteries: Psychometric properties of the Portuguese PennCNB in healthy controls. *Current Psychology*, *40*(10), 4851–4862. https://doi.org/10.1007/s12144-019-00419-2

Damian, A. M., Jacobson, S. A., Hentz, J. G., Belden, C. M., Shill, H. A., Sabbagh, M. N., Caviness, J. N., & Adler, C. H. (2011). The Montreal Cognitive Assessment and the Mini-Mental State Examination as Screening Instruments for Cognitive Impairment: Item Analyses and Threshold Scores. *Dementia and Geriatric Cognitive Disorders*, *31*(2), 126–131. https://doi.org/10.1159/000323867

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthc J*. 6(2). 94-98. Doi: 10.7861/ futurehosp.6-2-94.

Davies, M. S., Strickland, T. L., & Cao, M. (2014). Neuropsychological evaluation of culturally diverse populations. In F. T. L. Leong, L. Comas-Díaz, G. C. Nagayama Hall, V. C. McLoyd, & J. E. Trimble (Eds.), *APA handbook of multicultural psychology, Vol. 2: Applications and training.* (pp. 231–251). American Psychological Association. https://doi.org/10.1037/14187-014

De Vent, N. R., Agelink Van Rentergem, J. A., Schmand, B. A., Murre, J. M. J., Huizenga, H. M., & ANDI Consortium. (2016). Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A Normative Database Created from Control Datasets. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01601

Dean, P. J. A., & Sterr, A. (2013). Long-term effects of mild traumatic brain injury on cognitive performance. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00030

Di Febbo, D., Ferrante, S., Baratta, M., Luperto, M., Abbate, C., Trimarchi, P. D., Giunco, F., & Matteucci, M. (2023). A decision support system for Rey–Osterrieth complex figure evaluation. *Expert Systems with Applications*, *213*, 119226. https://doi.org/10.1016/j.eswa.2022.119226

Diaz-Orueta, U., Rogers, B. M., Blanco-Campal, A., & Burke, T. (2022). The challenge of neuropsychological assessment of visual/visuo-spatial memory: A critical, historical review, and lessons for the present and future. *Frontiers in Psychology*, *13*, 962025. https://doi.org/10.3389/fpsyg.2022.962025

Durisko, C., McCue, M., Doyle, P. J., Dickey, M. W., & Fiez, J. A. (2016). A Flexible and Integrated System for the Remote Acquisition of Neuropsychological Data in Stroke Research. *Telemedicine and E-Health*, *22*(12), 1032–1040. https://doi.org/10.1089/tmj.2015.0235

Elderson, M. F., Pham, S., Van Eersel, M. E. A., LifeLines Cohort Study, Wolffenbuttel, B. H. R., Kok, J., Gansevoort, R. T., Tucha, O., Van Der Klauw, M. M., Slaets, J. P. J., & Izaks, G. J. (2016). Agreement between Computerized and Human Assessment of Performance on the Ruff Figural Fluency Test. *PLOS ONE*, *11*(9), e0163286. https://doi.org/10.1371/journal.pone.0163286

Eling, P. (2019). History of Neuropsychological Assessment. In J. Bogousslavsky, F. Boller, & M. Iwata (Eds.), *Frontiers of Neurology and Neuroscience* (Vol. 44, pp. 164–178). S. Karger AG. https://doi.org/10.1159/000494963

Favaro, A., Tsai, Y.-T., Butala, A., Thebaud, T., Villalba, J., Dehak, N., & Moro-Velázquez, L. (2023). Interpretable speech features vs. DNN embeddings: What to use in the automatic assessment of Parkinson's disease in multi-lingual scenarios. *Computers in Biology and Medicine*, *166*, 107559. https://doi.org/10.1016/j.compbiomed.2023.107559

Fernandez, A. L. (2019). Modern neuropsychological tests for a diversity of cultural contexts. *The Clinical Neuropsychologist*, *33*(2), 438–445. https://doi.org/10.1080/13854046.2018.1560501

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, *108*(43), 17621–17625. https://doi.org/10.1073/pnas.1109168108

Frank, J., & Landeira-Fernandez, J. (2008). Comparison between two scoring systems of the Rey–Osterrieth Complex Figure in left and right temporal lobe epileptic patients. *Archives of Clinical Neuropsychology*, *23*(7–8), 839–845. https://doi.org/10.1016/j.acn.2008.06.001

Franzen, S., Van Den Berg, E., Bossenbroek, W., Kranenburg, J., Scheffers, E. A., Van Hout, M., Van De Wiel, L., Goudsmit, M., Van Bruchem-Visser, R. L., Van Hemmen, J., Jiskoot, L. C., & Papma, J. M. (2023). Neuropsychological assessment in the multicultural memory clinic: Development and feasibility of the TULIPA battery. *The Clinical Neuropsychologist*, *37*(1), 60–80. https://doi.org/10.1080/13854046.2022.2043447

Fujii, D. E., & Wong, T. M. (2006). Neuropsychological Assessment with Asian-American Immigrants: Recommendations for Meeting *Daubert* Standards. *Journal of Forensic Neuropsychology*, *4*(4), 3–31. https://doi.org/10.1300/J151v04n04_02

Gao, J., Yang, Y., Lin, P., & Park, D. S. (2018). Computer Vision in Healthcare Applications. *Journal of Healthcare Engineering*, *2018*, 1–4. https://doi.org/10.1155/2018/5157020

Germine, L., Reinecke, K., & Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist*, *33*(2), 271–286. https://doi.org/10.1080/13854046.2018.1535662

Hoelzle, J. B., Nelson, N. W., & Smith, C. A. (2011). Comparison of Wechsler Memory Scale–Fourth Edition (WMS–IV) and Third Edition (WMS–III) dimensional structures: Improved ability to evaluate auditory and visual constructs. *Journal of Clinical and Experimental Neuropsychology*, *33*(3), 283–291. https://doi.org/10.1080/13803395.2010.511603

Holmlund, T. B., Chandler, C., Foltz, P. W., Cohen, A. S., Cheng, J., Bernstein, J. C., Rosenfeld, E. P., & Elvevåg, B. (2020). Applying speech technologies to assess verbal memory in patients with serious mental illness. *Npj Digital Medicine*, *3*(1), 33. https://doi.org/10.1038/s41746-020-0241-7

Iglesias, M., Favaro, A., Motley, C., Oh, E. S., Stevens, R. D., Butala, A., Moro-Velazquez, L., & Dehak, N. (2022). Cognitive and Acoustic Speech and Language Patterns Occurring in Different Neurodegenerative Disorders while Performing Neuropsychological Tests. *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1–6. https://doi.org/10.1109/SPMB55497.2022.10014965

Jacobsen, S. E., Sprenger, T., Andersson, S., & Krogstad, J.-M. (2003). Neuropsychological assessment and telemedicine: A preliminary study examining the reliability of neuropsychology services performed via telecommunication. *Journal of the International Neuropsychological Society*, *9*(3), 472–478. https://doi.org/10.1017/S1355617703930128

Jaffe, P. I., Kaluszka, A., Ng, N. F., & Schafer, R. J. (2022). A massive dataset of the NeuroCognitive Performance Test, a web-based cognitive assessment. *Scientific Data*, *9*(1), 758. https://doi.org/10.1038/s41597-022-01872-8

Jamal, N., Shanta, S., Mahmud, F., & Sha'abani, M. (2017). Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review. *AIP Conf. Proc. 1883, 020028*. https://doi.org/10.1063/1.5002046

Kang, M. J., Kim, S. Y., Na, D. L., Kim, B. C., Yang, D. W., Kim, E.-J., Na, H. R., Han, H. J., Lee, J.-H., Kim, J. H., Park, K. H., Park, K. W., Han, S.-H., Kim, S. Y., Yoon, S. J., Yoon, B., Seo, S. W., Moon, S. Y., Yang, Y., Shim, Y.S., Baek, M.J., Jeong, J.H., Choi, S.H., & Youn, Y. C. (2019). Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data. *BMC Medical Informatics and Decision Making*, *19*(1), 231. https://doi.org/10.1186/s12911-019-0974-x

Karlsen, R. H., Karr, J. E., Saksvik, S. B., Lundervold, A. J., Hjemdal, O., Olsen, A., Iverson, G. L., & Skandsen, T. (2022). Examining 3-month test-retest reliability and reliable change using the Cambridge Neuropsychological Test Automated Battery. *Applied Neuropsychology: Adult*, *29*(2), 146–154. https://doi.org/10.1080/23279095.2020.1722126

Kasten, E., Barbosa, F., Kosmidis, M. H., Persson, B. A., Constantinou, M., Baker, G. A., Lettner, S., Hokkanen, L., Ponchel, A., Mondini, S., Jonsdottir, M. K., Varako, N., Nikolai, T., Pranckeviciene, A., Harper, L., & Hessen, E. (2021). European Clinical Neuropsychology: Role in Healthcare and Access to Neuropsychological Services. *Healthcare*, *9*(6), 734. https://doi.org/10.3390/healthcare9060734

Kenda, M., Cheng, Z., Guettler, C., Storm, C., Ploner, C. J., Leithner, C., & Scheel, M. (2022). Inter-rater agreement between humans and computer in quantitative assessment of computed tomography after cardiac arrest. *Frontiers in Neurology*, *13*, 990208. https://doi.org/10.3389/fneur.2022.990208

Kessels, R. P. C., & Hendriks, M. P. H. (2023). Neuropsychological assessment. In *Encyclopedia of Mental Health (Third Edition)* (pp. 622–628). https://doi.org/10.1016/B978-0-323-91497-0.00017-5

Kim, K. W., Lee, S. Y., Choi, J., Chin, J., Lee, B. H., Na, D. L., & Choi, J. H. (2020). A Comprehensive Evaluation of the Process of Copying a Complex Figure in Early- and Late-Onset Alzheimer Disease: A Quantitative Analysis of Digital Pen Data. *Journal of Medical Internet Research*, *22*(8), e18136. https://doi.org/10.2196/18136

Kim, S.-J., & Park, E. H. (2018). Relationship of Working Memory, Processing Speed, and Fluid Reasoning in Psychiatric Patients. *Psychiatry Investigation*, *15*(12), 1154–1161. https://doi.org/10.30773/pi.2018.10.10.2

Kiselica, A. M., Webber, T. A., & Benge, J. F. (2020). The Uniform Dataset 3.0 Neuropsychological Battery: Factor Structure, Invariance Testing, and Demographically Adjusted Factor Score Calculation. *Journal of the International Neuropsychological Society*, *26*(6), 576–586. https://doi.org/10.1017/S135561772000003X

Kitzing, P., Maier, A., & Åhlander, V. L. (2009). Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phoniatrics Vocology*, *34*(2), 91–96. https://doi.org/10.1080/14015430802657216

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Langer, N., Weber, M., Vieira, B. H., Strzelczyk, D., Wolf, L., Pedroni, A., Heitz, J., Müller, S., Schultheiss, C., Tröndle, M., Lasprilla, C. A., Rivera, D., Scarpina, F., Zhao, Q., Leuthold, R., Jenni, O. G., Brugger, P., Zaehle, T., Lorenz, R., & Zhang, C. (2022). The AI Neuropsychologist: Automatic scoring of memory deficits with deep learning. *Https://Www.Biorxiv.Org/Content/10.1101/2022.06.15.496291v4*. https://doi.org/10.1101/2022.06.15.496291

Lee, Y.-S., Koo, H.-S., & Jeong, C.-S. (2006). A straight line detection using principal component analysis. *Pattern Recognition Letters*, *27*(14), 1744–1754. https://doi.org/10.1016/j.patrec.2006.04.016

Leff, B. (2005). Hospital at Home: Feasibility and Outcomes of a Program to Provide Hospital-Level Care at Home for Acutely Ill Older Patients. *Annals of Internal Medicine*, *143*(11), 798. https://doi.org/10.7326/0003-4819-143-11-200512060-00008

Lezak, M. D. (Ed.). (2004). *Neuropsychological assessment* (4th ed). Oxford University Press.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (Fifth edition). Oxford University Press.

Lyu, W.-J., Chiu, P.-Y., Liu, C.-H., Liao, Y.-C., & Chang, H.-T. (2024). Determining optimal cutoff scores of Cognitive Abilities Screening Instrument to identify dementia and mild cognitive impairment in Taiwan. *BMC Geriatrics*, *24*(1), 216. https://doi.org/10.1186/s12877-024-04810-y

Mancioppi, G., Fiorini, L., Timpano Sportiello, M., & Cavallo, F. (2019). Novel Technological Solutions for Assessment, Treatment, and Assistance in Mild Cognitive Impairment. *Frontiers in Neuroinformatics*, *13*, 58. https://doi.org/10.3389/fninf.2019.00058

Manly, J. J. (2008). Critical Issues in Cultural Neuropsychology: Profit from Diversity. *Neuropsychology Review*, *18*(3), 179–183. https://doi.org/10.1007/s11065-008-9068-8

McCarthy, L., Rubinsztein, J., Lowry, E., Flanagan, E., Menon, V., Vearncombe, S., Mioshi, E., & Hornberger, M. (2024). Cut-off scores for mild and moderate dementia on the Addenbrooke's Cognitive Examination-III and the Mini-Addenbrooke's Cognitive Examination compared with the Mini-Mental State Examination. *BJPsych Bulletin*, *48*(1), 12–18. https://doi.org/10.1192/bjb.2023.27

Miller, J. B. (2019). Big data and biomedical informatics: Preparing for the modernization of clinical neuropsychology. *The Clinical Neuropsychologist*, *33*(2), 287–304. https://doi.org/10.1080/13854046.2018.1523466

Mitrushina, M., Boone, K. B., Razani, J., & D'elia, L. F. (2005). Handbook of normative data for neuropsychological assessment (2nd ed). *Oxford University Press*.

Moetesum, M., Diaz, M., Masroor, U., Siddiqi, I., & Vessio, G. (2022). A survey of visual and procedural handwriting analysis for neuropsychological assessment. *Neural Computing and Applications*, *34*(12), 9561–9578. https://doi.org/10.1007/s00521-022-07185-6

Morlett Paredes, A., Carrasco, J., Kamalyan, L., Cherner, M., Umlauf, A., Rivera Mindt, M., Suarez, P., Artiola I Fortuny, L., Franklin, D., Heaton, R. K., & Marquine, M. J. (2021). Demographically adjusted normative data for the Halstead category test in a Spanish-speaking adult population: Results from the Neuropsychological Norms for the U.S.-Mexico Border Region in Spanish (NP-NUMBRS). *The Clinical Neuropsychologist*, *35*(2), 356–373. https://doi.org/10.1080/13854046.2019.1709660

Nevatia, R. (2000). Perceptual Organization for Generic Object Descriptions. In K. L. Boyer & S. Sarkar (Eds.). *Perceptual Organization for Artificial Vision Systems*, 546, 173–189. Springer US. https://doi.org/10.1007/978-1-4615-4413-5_10

Nguyen, C. M., Rampa, S., Staios, M., Nielsen, T. R., Zapparoli, B., Zhou, X. E., Mbakile-Mahlanza, L., Colon, J., Hammond, A., Hendriks, M., Kgolo, T., Serrano, Y., Marquine, M. J., Dutt, A., Evans, J., & Judd, T. (2024). Neuropsychological application of the International Test Commission Guidelines for Translation and Adapting of Tests. *Journal of the International Neuropsychological Society*, 1–14. https://doi.org/10.1017/S1355617724000286

Norman, M. A., Moore, D. J., Taylor, M., Franklin, D., Cysique, L., Ake, C., Lazarretto, D., Vaida, F., Heaton, R. K., & The Hnrc Group. (2011). Demographically corrected norms for African Americans and Caucasians on the Hopkins Verbal Learning Test–Revised, Brief Visuospatial Memory Test–Revised, Stroop Color and Word Test, and Wisconsin Card Sorting Test 64-Card Version. *Journal of Clinical and Experimental Neuropsychology*, *33*(7), 793–804. https://doi.org/10.1080/13803395.2011.559157

Nur, A. C., & Shi, Z. (2022). *Indonesian Culture* (1st ed.). Culture Independently Published. ISBN: 979-8838118691.

Park, J. Y., Seo, E. H., Yoon, H.-J., Won, S., & Lee, K. H. (2023). Automating Rey Complex Figure Test scoring using a deep learning-based approach: A potential large-scale screening tool for cognitive decline. *Alzheimer's Research & Therapy*, *15*(1), 145. https://doi.org/10.1186/s13195-023-01283-w

Parsey, C. M., & Schmitter-Edgecombe, M. (2013). Applications of Technology in Neuropsychological Assessment. *The Clinical Neuropsychologist*, *27*(8), 1328–1361. https://doi.org/10.1080/13854046.2013.834971

Pathak, N., Vimal, S. K., Tandon, I., Agrawal, L., Hongyi, C., & Bhattacharyya, S. (2022). Neurodegenerative Disorders of Alzheimer, Parkinsonism, Amyotrophic Lateral Sclerosis and Multiple Sclerosis: An Early Diagnostic Approach for Precision Treatment. *Metabolic Brain Disease*, *37*(1), 67–104. https://doi.org/10.1007/s11011-021-00800-w

Pereira, C. R., Pereira, D. R., Silva, F. A. D., Hook, C., Weber, S. A. T., Pereira, L. A. M., & Papa, J. P. (2015). A Step Towards the Automated Diagnosis of Parkinson's Disease: Analyzing Handwriting Movements. *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, 171–176. https://doi.org/10.1109/CBMS.2015.34

Pérez-Arce, P. (1999). *The influence of culture on cognition. Archives of clinical neuropsychology: The official journal of the National Academy of Neuropsychologists*. *14*(7), 581–592.

Perry-Young, L., Owen, G., Kelly, S., & Owens, C. (2018). How people come to recognise a problem and seek medical help for a person showing early signs of dementia: A systematic review and meta-ethnography. *Dementia*, *17*(1), 34–60. https://doi.org/10.1177/1471301215626889

Pesau, H. G., Immanuel, A. S., Sulastri, A., & Van Luijtelaar, G. (2023). The role of daily spoken language on the performance of language tests: The Indonesian experience. *Bilingualism: Language and Cognition*, *26*(3), 538–549. https://doi.org/10.1017/S136672892200075X

Philip, S., & Jeffrey, B. (2002). Applications of Computer-based Neuropsychological Assessment. *Journal of Head Trauma Rehabilitation*, *17*(5), 395–410.

Picanco Diniz, C., Cabral Soares, F., Galdino De Oliveira, T. C., Dias E Dias Macedo, L., Wanderley Picanco Diniz, D. L., Valim Oliver Bento-Torres, N., Bento-Torres, J., & Tomás, A. (2014). CANTAB object recognition and language tests to detect aging cognitive decline: An exploratory comparative study. *Clinical Interventions in Aging*, 37. https://doi.org/10.2147/CIA.S68186

Porter, S. J., & Johnson, D. E. (2020). Clinical Use of the Automated Neuropsychological Assessment Metrics TBI-Mil Expanded Battery in Evaluating Concussion Recovery: A Retrospective Study. *Military Medicine*, *185*(9–10), e1722–e1727. https://doi.org/10.1093/milmed/usaa075

Pressler, S. J., Giordani, B., Titler, M., Gradus-Pizlo, I., Smith, D., Dorsey, S. G., Gao, S., & Jung, M. (2018). Design and Rationale of the Cognitive Intervention to Improve Memory in Heart Failure Patients Study. *Journal of Cardiovascular Nursing*, *33*(4), 344–355. https://doi.org/10.1097/JCN.0000000000000463

Quesque, F., Coutrot, A., Cox, S., De Souza, L. C., Baez, S., Cardona, J. F., Mulet-Perreault, H., Flanagan, E., Neely-Prado, A., Clarens, M. F., Cassimiro, L., Musa, G., Kemp, J., Botzung, A., Philippi, N., Cosseddu, M., Trujillo-Llano, C., Grisales-Cardenas, J. S., Fittipaldi, S., et al. (2022). Does culture shape our understanding of others' thoughts and emotions? An investigation across 12 countries. *Neuropsychology*, *36*(7), 664–682. https://doi.org/10.1037/neu0000817

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

Ruan, Q., Xiao, F., Gong, K., Zhang, W., Zhang, M., Ruan, J., Zhang, X., Chen, Q., & Yu, Z. (2020). Demographically Corrected Normative *Z* Scores on the Neuropsychological Test Battery in Cognitively Normal Older Chinese Adults. *Dementia and Geriatric Cognitive Disorders*, *49*(4), 375–383. https://doi.org/10.1159/000505618

Schaefer, L. A., Thakur, T., & Meager, M. R. (2023). Neuropsychological Assessment. In StatPearls. *StatPearls Publishing*. http://www.ncbi.nlm.nih.gov/books/NBK513310/

Schatz, P., & Browndyke, J. (2002). Applications of Computer-based Neuropsychological Assessment. *Journal of Head Trauma Rehabilitation*, *17*(5), 395–410. https://doi.org/10.1097/00001199-200210000-00003

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). *wav2vec: Unsupervised Pre-training for Speech Recognition* (arXiv:1904.05862). arXiv. http://arxiv.org/abs/1904.05862

Schroeder, R. W., Martin, P. K., & Walling, A. (2019). Neuropsychological Evaluations in Adults. American family Physician, *99*(2), 101-108.

Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., & Acharya, U. R. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *Arxiv*. https://doi.org/10.48550/ARXIV.2301.10009

Shirk, S. D., Mitchell, M. B., Shaughnessy, L. W., Sherman, J. C., Locascio, J. J., Weintraub, S., & Atri, A. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimer's Research & Therapy*, *3*(6), 32. https://doi.org/10.1186/alzrt94

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Smith, A. K., Ayanian, J. Z., Covinsky, K. E., Landon, B. E., McCarthy, E. P., Wee, C. C., & Steinman, M. A. (2011). Conducting High-Value Secondary Dataset Analysis: An Introductory Guide and Resources. *Journal of General Internal Medicine*, *26*(8), 920–929. https://doi.org/10.1007/s11606-010-1621-5

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary, Third Edition (3 rd). *Oxford University Press*.

Sulastri, A., Utami, M. S. S., Jongsma, M., Hendriks, M., & van Luijtelaar, G. (2018). The Indonesian Boston Naming Test: Normative Data among Healthy Adults and Effects of Age and Education on Naming Ability. *International Journal of Science and Research (IJSR)*, *8*(11), 134-139.

Taleb, C., Khachab, M., Mokbel, C., & Likforman-Sulem, L. (2019). Visual Representation of Online Handwriting Time Series for Deep Learning Parkinson's Disease Detection. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 25–30. https://doi.org/10.1109/ICDARW.2019.50111

Testa, R., Bennett, P., & Ponsford, J. (2012). Factor Analysis of Nineteen Executive Function Tests in a Healthy Adult Population. *Archives of Clinical Neuropsychology*, *27*(2), 213–224. https://doi.org/10.1093/arclin/acr112

Thaler, N. S., & Fujii, D. E. M. (2014). Cross-Cultural Considerations with Japanese American Clients: A Perspective on Psychological Assessment. In L. T. Benuto, N. S. Thaler, & B. D. Leany (Eds.), *Guide to Psychological Assessment with Asians*, 27–42. Springer New York. https://doi.org/10.1007/978-1-4939-0796-0_3

Troyer, A. K., & Wishart, H. A. (1997). A comparison of qualitative scoring systems for the Rey-Osterrieth Complex Figure Test. *The Clinical Neuropsychologist*, *11*(4), 381–390. https://doi.org/10.1080/13854049708400467

Tucha, L., Aschenbrenner, S., Koerts, J., & Lange, K. W. (2012). The Five-Point Test: Reliability, Validity and Normative Data for Children and Adults. *PLoS ONE*, *7*(9), e46080. https://doi.org/10.1371/journal.pone.0046080

Vogt, J., Kloosterman, H., Vermeent, S., Van Elswijk, G., Dotsch, R., & Schmand, B. (2019). Automated scoring of the Rey-Osterrieth Complex Figure Test using a deep-learning algorithm. *Archives of Clinical Neuropsychology*, *34*(6), 836–836. https://doi.org/10.1093/arclin/acz035.04

Wang, C., He, T., Zhou, H., Zhang, Z., & Lee, C. (2023). Artificial intelligence enhanced sensors—Enabling technologies to next-generation healthcare and biomedical platform. *Bioelectronic Medicine*, *9*(1), 17. https://doi.org/10.1186/s42234-023-00118-1

Webb, S. S., Moore, M. J., Yamshchikova, A., Kozik, V., Duta, M. D., Voiculescu, I., & Demeyere, N. (2021). Validation of an automated scoring program for a digital complex figure copy task within healthy aging and stroke. *Neuropsychology*, *35*(8), 847–862. https://doi.org/10.1037/neu0000748

Weintraub, S., Salmon, D., Mercaldo, N., Ferris, S., Graff-Radford, N. R., Chui, H., Cummings, J., DeCarli, C., Foster, N. L., Galasko, D., Peskind, E., Dietrich, W., Beekly, D. L., Kukull, W. A., & Morris, J. C. (2009). The Alzheimer's Disease Centers' Uniform Data Set (UDS): The Neuropsychologic Test Battery. *Alzheimer Disease & Associated Disorders*, *23*(2), 91–101. https://doi.org/10.1097/WAD.0b013e318191c7dd

Werry, A. E., Daniel, M., & Bergström, B. (2019). Group differences in normal neuropsychological test performance for older non-Hispanic White and Black/African American adults. *Neuropsychology*, *33*(8), 1089–1100. https://doi.org/10.1037/neu0000579

Wilson, L., Horton, L., Kunzmann, K., Sahakian, B. J., Newcombe, V. F., Stamatakis, E. A., Von Steinbuechel, N., Cunitz, K., Covic, A., Maas, A., Van Praag, D., & Menon, D. (2021). Understanding the relationship between cognitive performance and function in daily life after traumatic brain injury. *Journal of Neurology, Neurosurgery & Psychiatry*, *92*(4), 407–417. https://doi.org/10.1136/jnnp-2020-324492

Woods, D. L., Wyma, J. M., Herron, T. J., & Yund, E. W. (2016). Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PLOS ONE*, *11*(12), e0166439. https://doi.org/10.1371/journal.pone.0166439

Zhang, L., Ngo, A., Thomas, J. A., Burkhardt, H. A., Parsey, C. M., Au, R., & Hosseini Ghomi, R. (2021). Neuropsychological test validation of speech markers of cognitive impairment in the Framingham Cognitive Aging Cohort. *Exploration of Medicine*, *2*(3), 232–252. https://doi.org/10.37349/emed.2021.00044

Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., & Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods*, *50*(6), 2597–2605. https://doi.org/10.3758/s13428-018-1037-4

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

Chapter 2

An online platform and a dynamic database for neuropsychological assessment in Indonesia

Published as: Wahyuningrum, S. E., van Luijtelaar, G., & Sulastri, A. (2021). An online platform and a dynamic database for neuropsychological assessment in Indonesia. *Applied Neuropsychology: Adult*, *30*(3), 330–339. https://doi.org/10.1080/23279095.2021.1943397

Abstract

Proper use of neuropsychological tests in Indonesia is hindered by a lack of properly adapted neurocognitive tests as well as an absence of normative data. In 2016, we started adapting ten of these tests for use in Indonesia and collected data from healthy participants in Java. Here we introduce and propose a system that will facilitate the proper usage and interpretation of test scores: an online platform and a dynamic database. Newly collected data (492 healthy adults) of the Indonesian version of the Boston Naming Test (I-BNT) were used to illustrate the usefulness of the two functions. Analysis of variances, post-hoc tests, and a simulation study revealed the effects of age and education on the I-BNT, indicating that it is imperative to fine-tune the reference group based on these demographic factors. Putative inadequate sample size issues for obtaining reliable normative scores were overcome by employing regression analyses and the prediction of normative scores. It can be concluded that a flexible online platform is available for the calculation of normative scores either based on the whole population, on fine-tuned reference groups, or on predicted scores. The dynamic database's growth will allow to obtain even more fine-tuned and more reliable reference data as well as more accurate predictions. Fine-tuned reference data are badly needed for the heterogenous Indonesian population.

Keywords: I-BNT; I-ANDI; neuropsychological assessment; normative score; online platform.

Introduction

Normative scores are the key to indicate an individual's relative standing within a reference population. Traditionally, an individual's performance on one or series of neuropsychological tests is compared with a published reference group of normative data (Casaletto & Heaton, 2017; Zucchella et al., 2018). However, normative data may become out-of-date or drawn from another society, country, culture, ethnic group, or continent. An adequate interpretation of test scores demands that normative data be recently collected or recollected from a sample that mimics the respondent's demographic profile as much as possible.

The rapid development of technology in health practices and research led to the development of large online databases containing various types of clinical data, including normative scores e.g., Meyers (2013). An example of such a database is the one collected by Weintraub et al. (2009). They collected data using the UDS (Uniform Data Set), a neuropsychological test battery for patients with different levels of dementia. The project was continued by the creation of an online normative data calculator with demographic adjustments to facilitate the preclinical diagnosis of Alzheimer's Disease (Shirk et al., 2011). In addition to these examples, there is a newly developed neuropsychological online tool that compares a client's score with fine-tuned reference scores stored in a dynamic database. This system is currently used in the Netherlands. It is named Advanced Neuropsychological Diagnostics Infrastructure (ANDI), and was developed by researchers from the University of Amsterdam (de Vent et al., 2016). It consists of a dynamic database and an online platform that incorporates demographic data and neuropsychological test scores from a growing number of subjects. Its data are obtained from clinical neuropsychologists and researchers investigating neurocognition with a variety of tests in various healthy control groups in the Netherlands.

In the current study, we introduce and illustrate the Indonesian version of ANDI (named I-ANDI hereafter) with its two functions: a dynamic database and an online platform. The dynamic database is an infrastructure to store and incorporate various neuropsychological test scores from a large group of healthy subjects. The online platform allows neuropsychologists and clinicians in Indonesia to define a reference group based on demographic characteristics resembling those of their clients and thereby allow them to interpret their client's scores efficiently and adequately.

There are two differences between the Dutch-ANDI and I-ANDI. The Dutch-ANDI contains donated data from a variety of tests from various neuropsychologists and

practitioners in the Netherlands. We collected data ourselves by collaborating with university partners in various parts of Indonesia (consortium). Secondly, the data obtained with the ten adapted neuropsychological tests stored in I-ANDI's database were administered as a battery: that is, all subjects received all ten tests and this was not the case in the Dutch ANDI.

Indonesia is an archipelago with different ethnic groups that originate from different parts of the country. Therefore, it is necessary to spread the data collection over islands and different parts of islands in order to make sure that data from different ethnic groups will be stored in the database. These data can then be used for normative scores, for developing normative scores for separate ethnic groups, either alone or in combination with the common demographic factors education, age, and sex. Currently, I-ANDI provides facilities for establishing normative scores for the most common demographic factors: age, sex, and education. I-ANDI also provides the possibility of storing patient data. However, the patient data will remain separate from the healthy controls' data and will be used only for research purposes.

This paper illustrates the usefulness of I-ANDI for Indonesia's neuropsychological practices. Data of the Boston Naming Test, adapted for Indonesia (I-BNT) by Sulastri et al., (2018), will be used for that purpose. ANOVA's and regression analyses will be used to determine the influence of demographic factors on the I-BNT. If there is found to be an influence, then different groups will require different normative data. Finally, a simulation study will illustrate the flexibility of I-ANDI regarding the composition of the reference group for obtaining appropriate normative data.

Methods

The initial I-ANDI data collection

With supports from the European Union (Erasmus Grant outside Europe) and the Indonesian government for the project entitled "Development of I-ANDI," we established partnerships in 2016 with Radboud University, Nijmegen, the Netherlands and other Indonesian universities within the scheme of the consortium of researchers for our data collection. The consortium members were twelve researchers from Soegijapranata Catholic University in Semarang, Atma Jaya Catholic University in Jakarta, Widya Mandala Catholic University in Surabaya and Radboud University, the Netherlands. We started the data collection by choosing a battery consisting of ten clinically relevant neuropsychological tests covering the domains learning and memory, attention, executive function, and language and

then adapted the tests. The ten neuropsychological tests were the Auditory Verbal Learning, Figural Reproduction, Digit Span, Bourdon-Wiersma, Stroop Colour-word, Trail Making, I-Boston Naming Test, Token Test, Verbal Fluency, and Five Point Test. All ten tests were administered to all subjects and the data were stored in the I-ANDI database representing a sample of healthy subjects and further used to develop normative data. In early 2020 we embarked on new collaborations with mental health care providers and with hospitals. Going forward, we anticipate partnerships to grow with other universities on other islands of Indonesia. This will make it possible to increase the number of subjects in the data base and increase its ethnic diversity. The I-ANDI database is intended to contain two categories of subjects: healthy subjects and patients. However, because of Indonesia's Covid-19 outbreak in March of 2020, we temporarily postponed data collection from different patient categories. Now only the scores from the battery are stored, in the future, I-ANDI will allow data and datasets from single tests as well.

The number of subjects in a subgroup

At the time of the analyses, the database contained 492 subjects. Given that normative data of most cognitive tests are education-, age- and sex-dependent, we analyzed the effects of these factors, although, in our preliminary study reporting normative data on I-BNT (Sulastri et al., 2018), sex effects were not found. Given that education and age effects have been previously reported for the BNT (Neils et al., 1995), it can be expected that different normative scores for education-age categories will be necessary. Given the number and distribution of subjects in our dataset, we anticipate that not all education-age categories will contain the desired 50 subjects (Bridges & Holler, 2007). It is especially likely that the number of subjects with less than seven years of education and the number of older people with various degrees of education will be too low for obtaining reliable normative scores for these subgroups. Alternatively, normative scores could also be calculated with regression-based approaches (Tombaugh, 2004). The approaches have the advantage of requiring much smaller sample sizes (Oosterhuis et al., 2016) with improved diagnostic precision (Burggraaff et al., 2017).

Moreover, a regression-based normative formula that is simultaneously correct for all significant demographic variables might be more desirable for the clinician and more sensitive in detecting cognitive problems. In all, regression-based normative data is most desirable if there are fewer than 50 subjects to calculate the normative scores of a subgroup. Meanwhile, the need for normative scores from adapted neuropsychological tests is high, and Indonesia does not have a neuropsychological battery test database with normative data as yet.

Construction of the Indonesian Advanced Neuropsychological Diagnostics Infrastructure (I-ANDI)

Inspired by the Dutch ANDI normative database system developed by De Vent et al. (2016), we developed a dynamic database and online platform for Indonesia and named it I-ANDI. Figure 1 illustrates the stepwise construction and design of I-ANDI and how to use it. The I-ANDI user can submit a client's test score and retrieve his client's score as it relates to the scores of a reference group.

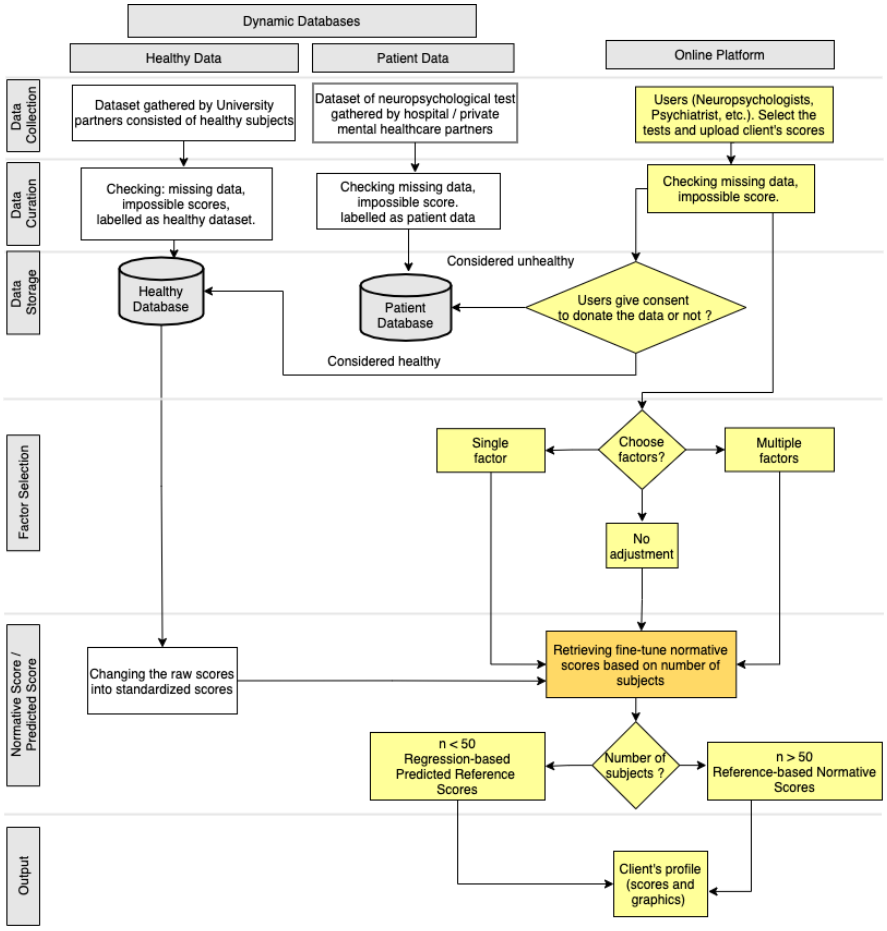


Figure 1. The stepwise construction and design of I-ANDI. The boxes under the dynamic databases (light grey) illustrate the data flow from data collection to data storage in two separate units. The boxes on the right side (yellow) represent the data flow in the online platform.

Data collection

I-ANDI provides two kinds of data entry to the dynamic database: (1) data collected by us, or by university, hospital and private mental healthcare providers that are collaborating under a Memorandum of Understanding; and (2) data provided via the online platform that is from other neuropsychologists, psychiatrists, and clinicians throughout the country.

Data gathered by us and by our university partners are from healthy subjects. These data will be used to develop normative scores and will be labeled and stored in the "healthy subjects" database. Data gathered by our university, hospital and mental healthcare partners under a similar collaborative arrangement are from various groups of diagnosed patients. These will be labeled and stored separately as patient data. To gain access to the I-ANDI database, a user will first register on the I-ANDI website. Then after having been approved by the Person in Charge (PIC) of the consortium responsible for the quality and security of the I ANDI database, the user will be able to select the test and upload their client data. I-ANDI will provide templates for the uploading of data in the form of spreadsheet and Comma Separated Values (CSV).

Data curation

This phase is designed to ensure the quality of the data entered. The submitted data are first checked for completeness. When there is incomplete or missing data, the system will alert the user to completes the data before resubmission and continuation of the process. The data are also checked for impossible scores, such as a score on any test that is higher than the test's maximum. In this manner, coding or typographic errors will be kept from contaminating the data file.

Data storage

I-ANDI has two separate databases: the healthy and the patient database. The online available data are stored anonymously. Before entering the data, a clinician or user first determines whether the data originate from a patient or not. In case of doubts, a second opinion expert will be consulted. Next, the user indicates whether one wishes to donate the data to the consortium.

Factor selection

By selecting from among several different demographic factors, users can select the normative score that best mimics the characteristics of his client(s). In other words, users can fine-tune the normative reference score. There are several different demographic choices available: multiple (e.g., age, education, sex), single (e.g., age), or none; the latter implies that the entire dataset will be used to calculate the normative scores.

In the future, when users want to select other or additional demographic factors in order to obtain a more specific or unique normative score, the system is capable of accommodating them. For example, factors such as the client's "daily spoken language" or the ethnicity of the client's parents can be added. Such information might be particularly helpful in deriving normative scores that are most suitable for language tests.

Normative score/predicted score

I-ANDI can generate normative scores and/or predictive normative scores from the healthy database. We transformed the raw scores of all tests into standardized scores (z-score). The system provides the fine-tuned normative scores based on the factor(s) chosen by the users. There are two possibilities for calculating normative scores that depend on the number of subjects. If the number of subjects is 50 or more, the system provides an output based on the normative reference-based scores, whereas when the number of subjects is less than 50, the system will recommend the use of regression-based predicted normative scores. The predicted score is obtained from the regression analysis's weighted score using existing data in the database.

Output

An individual's test score compared with the normative scores obtained from the chosen dataset is numerically and graphically displayed, next to the number of subjects on which the normative scores are calculated.

Participants

Data from the I-BNT were used to illustrate the newly developed system tools. They were collected during 2017–2019 in the island of Java (Jakarta, representing West Java, N = 193), Semarang (Central Java, N = 197), and Surabaya (East Java, N = 102). Java was chosen considering that it has the highest (57%) percentage of the total Indonesian population (267 million). Participants consisted of 295 females and 197 males, with an age range of 16 - 80 years old (M = 33.2; SD = 15.2). Participants were categorized into four age-by-decade groups (Palmer et al., 1998; Van Den Berg et al., 2009): (i) age 20–29 years, (ii) age 30–39 years, (iii) age 40–49 years, (iv) age 50–59 years and two other categories. Additional categories included (v) all persons older than 15 but younger than 20 and (vi) all persons over 60. The years of education varied between 0 and 22 years (M=13.9; SD=2.7) and were grouped into five lengths of education corresponding to natural divisions of the Indonesian education system: (i) educated for less than seven years, (ii) education between 7 and 9 years, (iii) education between 10 and 12 years, (iv) education between 13 and 16 years and (v) education over 17 years.

The tests were administered in "Bahasa Indonesia," the official language of Indonesia that is used nationwide in the educational system, in media, administration and business. Before the test was administered, researchers explained all study procedures and informed the participants that the data were to be used for scientific purposes. Then participants gave their consent to participate in the study. They received seventy five thousand rupiahs (equal to five US dollars) after finishing the series of tests. The I-BNT was one of the tests administered within the framework of a larger project in which nine other neuropsychological tests were adapted and data from healthy subjects were collected. All included participants with no reported history of psychiatric, neurological diseases, head trauma, drug abuse, or other illnesses that could influence test' performance. The current research was conducted in compliance with the Helsinki Declaration and the ethics committee of Soegijapranata University gave clearance for this research project (University Ethical Clearance number: 001B/B.7.5/FP.KEP/IV/2018). The design of the database and the transport and storage of private, sensitive information fulfills Indonesia's regulations as mentioned in ITE Law (Information and Electronic Transactions).

Table 1. I-BNT Participant demographics (N = 492).

Variables	**N**
Age (years)	
16-19	63
20-29	203
30-39	73
40-49	57
50-59	66
60 ++	30
Education (Years)	
0 – 6	15
7 -9	33
10-12	156
13-16	267
17 ++	21
Sex	
Female	295
Male	197

Measures

The BNT, introduced by Kaplan et al. (1983), consists of 60 line drawings graded on levels of difficulty (Alyahya & Druks, 2016; Kessels & Hendriks, 2016). Subjects have to report the name of the object in the drawing within a limited period of time. The I-BNT was administered using the original procedure of the BNT (Kaplan et al., 1983). The response time per item was restricted to 20s. Scores include the number of spontaneously produced correct responses, the number of correct responses given after semantic cueing, and the number of responses given after phonemic cueing. The total number of correct responses is the sum of the number of spontaneous correct responses and the number of correct responses after a stimulus cue is given (Strauss et al., 2006; Sulastri et al., 2018). In addition, the researcher or test administrator also noted the number of incorrect answers and measured the total time to complete the sixty-item test.

Statistical analysis

An analysis of variance (ANOVA) was used to find the effects of the demographic characteristics, of sex, age, and education on the total scores and total time of the I-BNT. Post-hoc analysis, according to Bonferroni, was used to further delineate main and interaction effects. Uni and multivariate linear regression analyses were used to obtain the factors and coefficients of the prediction model and formulae for calculating the predicted score; these will be used only when the number of subjects in a chosen reference group is fewer than 50.

Results

Table 1 provides information about the participants who completed the I-BNT and that were included in the database. The table presents the number of subjects in each age and education category, as well as the distribution of sexes in our current sample.

Table 2 shows the effects of age, education and sex on the I-BNT total score and total time. Total score and total time were significantly influenced by age and education as revealed by the outcomes of the three factor ANOVA (see also Figure 2). There was no main effect for sex; only first and second order interactions between sex and the other factors were found with small effect sizes (Cohen, 1988). Therefore, sex was no longer considered as a major factor affecting the performance of the I-BNT; this might be subject to change as more data are collected, or it may be found to differ for different ethnic groups.

The effects of demographic factors on the I-BNT total score and total time

The outcomes of the ANOVA showed that education is indeed significant for the total number correct (F (4.446) = 37.23, p < .001, $\acute{\eta}^2$ = .25; Table 2). The effect size was medium (Cohen, 1988), 25% of the total variance can be explained by education. Figure 2a shows the means and standard deviations for the different education categories. The highest number of correct items was achieved by people with 13–16 years of education, the lowest score by people in the group with less than 7 years of education. The post-hoc tests showed that all education groups differed from each other on the total number correct, the least educated group scoring significantly lower than other groups. The time to complete the I-BNT (total time) also showed a significant education effect (see Table 2). The fastest was the group that received 13–16 years of education (M = 262.7). The slowest were persons with the lowest education (M = 611.2).

Table 2. Analysis of variance for main effects and interactions between age, education, and sex on the total score and total time.

	Main Effect		**Interaction**	
Value BNT	**Variable**	***F***	**Variable**	***F***
Total Score	Education	37.23 ***	Education × age	1.89*
	Age	2.32 *	Education × sex	3.15*
	Sex	0.95	Age × sex	2.62*
			Education × age × sex	2.79**
Total Time	Education	12.45***	Education × age	0.83
	Age	2.88*	Education × sex	1.75
	Sex	0.81	Age × sex	1.47
			Education × age × sex	2.12*

Note: *p<0.05; **p<0.01; ***p<0.001

Figure 2b shows the mean (represented by a closed circle) and standard deviation (represented by a vertical line) of the total score based on different age groups; the highest score (56.70) was obtained in the 20–29 years group. The lowest score (51.70), comprised the ≥ 60 years age group. The main effect of age turned out to be rather small (Cohen, 1988) but significant (F (5,446) = 2.32, p < .05, $\acute{\eta}^2$ = .03, Table 2); only 3% of the variance in the total correct number can be ascribed to age. The post-hoc tests showed that all age groups scored better compared to the ≥ 60 years group. The post-hoc tests for age on the total time shows that the group ≥ 60 did less well in total time compared with others (M = 431.6). The fastest was the

40–49 years old group (M = 312.6). As mentioned earlier, the system allows the creation of a subset based on more than one demographic factor. The significant main effects, and to a lesser extent the first and second-order interaction effects, as given in Table 2, albeit with ή²'s smaller than 7%, already suggested that this might be useful. The second-order interaction's interpretation is that the effects of sex might be different in the different age and education groups. The interaction between age and education (ή² = .06) suggests that age effects on the total score of the I-BNT might be different for different education groups. The main and first-order interaction effects on both total number of correct items and on time to complete the test underline the importance of using fine-tuned normative scores for the I-BNT.

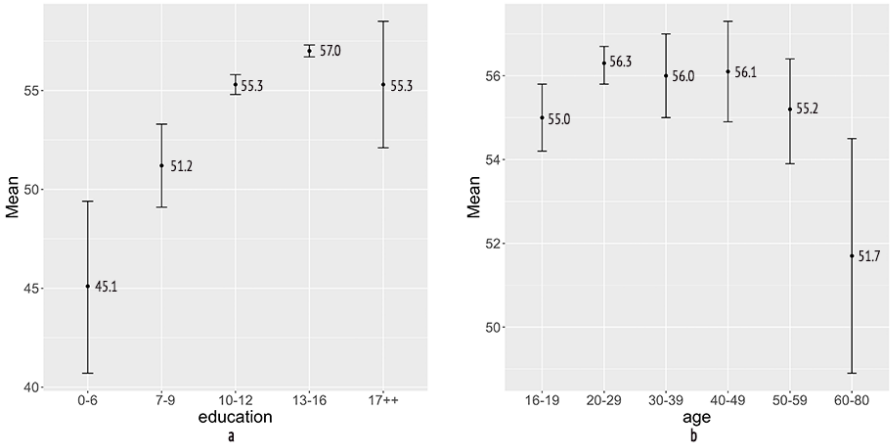


Figure 2. Mean and standard deviations of the I-BNT total score. The left side (2a) is education factor effect. Education on X-axis (elementary schools (0–6) to post graduate (17++)). The graph shows an education-dependent increase, except for the 17++ group. The larger standard deviation in the 0–6 and 17++ groups is indicative of a relatively small number of subjects in combination with large inter-individual differences. The right side (2b) is age factor effect. Age groups are illustrated on the X-axis. Note the small and non-significant age-related effects between the groups <60 years; only the oldest group (≥60) showed a significant decline.

The usefulness of different reference groups illustrated by simulation

As previously demonstrated, the demographic factors age and years of education significantly influence the number of correct I-BNT items. These factors have consequences for the comparisons of an individual's score with the reference group. The user of the online I-ANDI platform enters a client's demographic data and the raw I-BNT score into the system, and chooses the filters or factors in order to position his client's score in comparison with the optimally defined normative

group. The output of the system is in the form of a graph. The graph shows a person's raw and z-score score relative to the reference group score as well as the number of subjects (dataset) used for the calculation of the normative score.

The usefulness of I-ANDI, including the choice of different reference groups with different normative scores, is illustrated with the following simulation examples: a score of a single person aged = 68 years with two variants regarding the level of education, either only elementary school (6 years) or senior high (12 years), and I-BNT score of 48.

The results show, as can be seen in Figure 3, how the client's z-score relates to the by I-ANDI calculated scores of the clinician's chosen or preferred reference group. Noticeable differences between the client's score in relation to the different reference groups can be clearly seen.

Figure 3 illustrates that the relative standing of a client is highly dependent on the selected reference group. The vertical Y-axes in Figure 3 show the standardized performance scale (z-score) with zero as the mean of the intended reference groups. The numbers 1 and -1, 2 and -2, also 3 and -3 on the Y-axes indicate 1 to 3 standard deviations (SD) from the mean value. From left to right are the score of a client compared to (a) all data in the database (no filters), (b) 12 years of education, (c) 6 years of education (note the large effects of education), and (d) his age (>60) and 6 years of education.

Regression analyses

The data from all subjects were used for hierarchical multiple linear regression analyses in order to identify the contribution of the demographic factors of education, sex, and age, to both the total score and the total time of the I-BNT. Subsequently, a linear model was proposed to predict the normative scores for both variables. The beta weights and its significance scores are presented in Tables 3 and 4. Two models were analyzed. The first model considered only the education factor and was highly significant: it predicted total scores, $F(1,490) = 138.50$, $p < .001$ and total time, $F(1,490) = 98.17$, $p < .001$). The R^2 were 22% and 17% respectively. Both values show a medium effect size by Cohen's (1988) standards. This simplest model essentially shows that total score and total time can be predicted by identifying the client's education level. When the age and sex variables were added to the analysis model of total score and total time, there were no changes of R^2 (Tables 3 and 4).This agrees with the reported large $\acute{\eta}^2$ for education and the small $\acute{\eta}^2$ for both variables in the 3-factor ANOVA, as reported in Table 2.

From the model, the equation of the total score and total time model using education-category (1–5) are:

Predicted total score = 46.78 + (2.53 x education_category) (1)

Predicted total time = 655.80 + (-96.34 x education_category) (2)

The equations for the standardized z-score of the total score and total time are the same:

Predicted z-score = 1.8 + (-0.6 x education_category) (3)

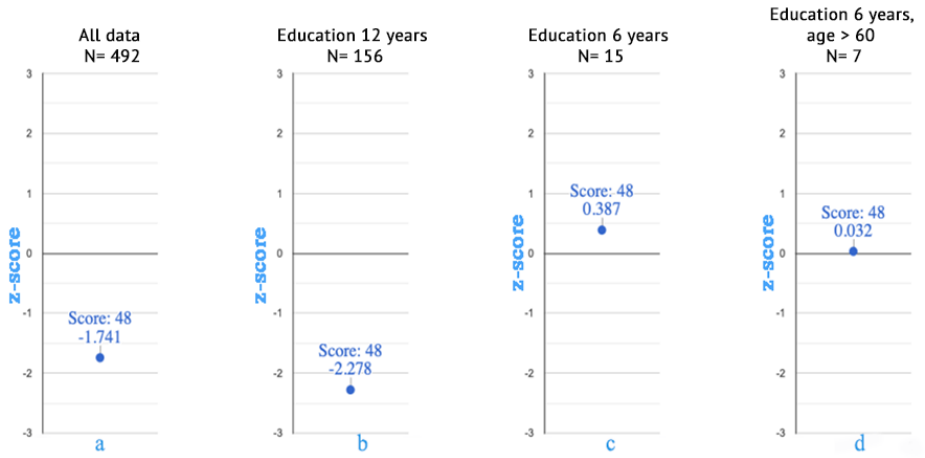


Figure 3. An illustration of the internet platform for obtaining client scores concerning different reference groups. The displayed plots are an individual's standardized z-score compared to (a) all data, (b) education in a 12-year subset, (c) education in a 6-year subset, (d) education in 6 years and ages above 60 years subset. In case the number of subjects is lower than 50, as in Figures 3(c) and 3(d), the user is advised to use the predicted normative score. Next, the number of subjects used for the calculation of the normative score is plotted on top of each graph.

Table 3. Hierarchical Multiple Linear Regression Analysis Predicting I-BNT Total Score for Age, Sex, and Education (*N*=492).

Variable	*B*	*SEB*	*β*	R^2	ΔR^2
Step 1				.22	.22
Education	2.53	.77	.47**		
Constant	46.78	.22			
Step 2			-.46**	.22	.00
Education	2.50	.22	-.03		
Age	-.08	.12	.01		
Sex	.13	.36			
Constant	46.97	34.93			

Notes. Unstandardized coefficient (B) and Standard Error of B (SEB), Beta weight (*β*).
**p < .001; Only education contributes significantly to the prediction of the scores.

Table 4. Hierarchical Multiple Linear Regression Analysis Predicting I-BNT Total Time for Age, Sex, and Education (*N*=492).

Variable	*B*	*SEB*	*β*	R^2	ΔR^2
Step 1				.17	.17
Education	-96.34	9.72	-.41**		
Constant	655.80	34.93			
Step 2				.17	.00
Education	-95.63	10.01	-.41**		
Age	.98	5.58	.01		
Sex	-13.47	16.06	-.04		
Constant	669.33	47.75			

Notes. Unstandardized coefficient (B) and Standard Error of B (SEB), Beta weight (β).
**p < .001; Only education contributes significantly to the prediction of the total time.

Discussion

The collection of normative scores for any psychological test requires substantial effort, and standardized procedures of data collection. Most psychological tests have been developed in WEIRD-specific (Western, Educated, Industrialized, Rich, and Democratic) cultures (Henrich et al., 2010) and monolingual societies. There exists a substantial likelihood that such cognitive tests are not culture free. Additionally, there may be changes in the norm-group demographics as well as changes in health across time that require recently collected data (De Vent et al., 2016). The effort to adapt tests to the Indonesian culture and to collect normative data based on the Indonesian population is a relatively recent development. Such efforts are hindered by publishers who claim copyrights, even when the normative scores are collected in another language, a different culture or time periods. Furthermore, middle and low-income countries often are unable to afford academic databases with neuropsychological norm scores. As a consequence, the most important stakeholders (in this case Indonesian neuropsychologists) are then unable to obtain recent and local normative scores by this means.

Another reason for a dynamic database and an internet platform that provides users the choice of fine-tuned reference data is Indonesia's tremendous cultural and linguistic diversity. This island nation has approximately 600 ethnic groups living on over 17,000 islands, speaking more than 700 native languages and no less than 1,100 dialects (Zein, 2020). To say that Indonesia is linguistically diverse is a huge understatement. Although Indonesia's official language, Bahasa Indonesia, is taught to all children in elementary school, it is not spoken daily by all Indonesians. This variety of languages and ethnic groups makes it even more important to find ways to customize normative scores in a manner that will take this large linguistic and ethnic diversity into account.

The current dataset is only a first step in this direction. We have been able to consider only the three demographic factors that are most often used in international studies: the factors of sex, age and educational level. A great deal remains to be done to further fine-tune the reference data set to more adequately reflect Indonesia's tremendous linguistic and cultural diversity.

Here we have demonstrated an online, continually evolving data base that provides for the determination and storage of normative scores, and the calculation of a client's score based on a customized, fined-tuned reference group. The advantages of our platform and dynamic database are that: (a) it has the possibility to obtain reference scores from a sub-dataset, (b) the normative data can be updated toward only recently collected data, (c) it has a procedure to obtain standardized Z -scores from all data or sub-datasets, and (d) it will allow the data base to grow. It will also offer possibilities for research into the influence that cultural, ethnic and language differences have on the cognitive domains. It will someday truly reflect the tremendous cultural, ethnic and language diversity of the Indonesian people.

However, the development of the normative database demands attention to certain issues. Besides a good and solid infrastructure and a good user-friendly interface, continuous control of the quality of the newly entered data is imperative. Currently the data comes only from the consortium with a quality check, including the outlier test on impossible scores and coding errors, as mentioned in the Methods section. In the future, when the system is fully operational and open for all registered (neuro)psychologists, it will include certain additional quality control procedures. Additional checks on the completeness of all submitted demographic data will be put in place and all neuropsychological tests will be based on appropriate and updated prediction models based on multivariate regression analyses. Only then will neuropsychologists be able to donate data from healthy and/or neuropsychological patients. Whenever there is uncertainty about the quality or the appropriateness of the data, the neuropsychologists of the consortium will decide whether or not to admit the new data to the existing dataset of healthy subjects. A final security measure that simply traces the login history of the system-users will also be added to the log-in monitoring system.

Furthermore, the database can provide a mix of neuropsychological test data. In addition to individual test scores, the database can be extended to provide combinations of test data. It can retrieve cognitive profiles rather than outcomes of just one single test. Finally, the dynamic database can also be used for participants with a clinical diagnosis; the data from these participants will be separated and

excluded from the normative sample. The data from various patient groups will be used for research purposes only.

The dynamic database and internet platform facilitate the availability of more reliable normative scores since the user can (through the online platform) define the reference or norm group by choosing different demographic filters (factors). The current system's version has four filters, and this number can be increased if necessary. The system can dynamically determine the normative score based on the user's requirements by directly comparing an individual's score results with normative scores from people who are in the same demographic categories. The same procedures can be applied to data from other neuropsychological tests. We collected normative data from some clinically relevant neuropsychological tests from different areas in Java and also from three other islands which will all be uploaded in the dynamic database as well. The more data that is uploaded, the more relevant the system will be. This includes the possibility to establish the contributing role of linguistic proficiency and type of ethnic group, for example. If these latter factors turn out to explain a statistically determined amount of variance, then the prediction models need to be adapted. It is worth noting that this dynamic database can also serve as an interesting research tool as it contains data from healthy controls' various cognitive domains since the same tests were administered to all subjects.

The scores that we obtained for the I-BNT agreed well with our recently reported data (Sulastri et al., 2018). We found significant and large effects of education on the total score and total time of the I-BNT: the lowest mean score (45) was generated by the group with the lowest level of education. Education's effects were rather large; other studies also conclude that educational background accounted for the greatest proportion of the variance in BNT data and more variance than age does (Henderson et al., 1998; Neils et al., 1995; Nicholas et al., 1989; Steinberg et al., 2005). In the current study, sex effects were not found, and this corresponds with data and conclusions from prior studies (Busch et al., 2006). The rather large effects of education and the small effects of age, as found in this study, contribute to the validity of the adapted version of the I-BNT, although a further clinical validation study is necessary with patients with mild forms of dementia or aphasia. Education's effects underline the importance of having separate normative scores for subjects with different levels of education. The age effects imply different normative scores only for people over 60 years. On the other hand, the first and second order interactions between the three demographic factors suggest that normative data for subgroups are to be preferred, perhaps not now, but in the future when more

data will be uploaded in the database. This is illustrated as well by our simulation: large differences in the standardized z-scores were obtained when the score of an individual was compared to the entire dataset or compared with groups with two different levels of education or an age-subgroup with lower levels of education. The z-scores varied from - 2.28 to .387 and mainly the level of education showed large effects in the simulation, similar to the relatively large effect size for level of education in the ANOVA. The simulation also shows that the system can be used in a flexible way in determining an individual's score in comparison with this reference group.

A note of caution is that the data presented in this paper should not be considered as normative scores. Normative data for the I-BNT were recently published (Sulastri et al., 2018), but it should be noted that in the published analysis only some of the cells were filled with a sufficient number of respondents (Bridges & Holler, 2007), and other cells were less or poorly filled. The same issue is present here as well, and this pertains to mostly young persons with less than seven years of education, to persons with the highest level of education, and elderly people in general. Moreover, our current sample may not be representative for the entire Indonesian population, perhaps only for the better-educated population in urbanized parts of Java island. It is imperative that clinicians should consider the characteristics of the normative data, including the representativeness of the sample and the sample size of individual cells in the published normative dataset when they interpret the client's score to avoid biases (Agelink Van Rentergem et al., 2017).

In our case most of the cells do not have enough subjects and therefore the option was created to have predicted normative scores based on a statistical prediction model. The use of predicted normative scores may have some benefits (Burggraaff et al., 2017) and can be recommended in case the scores of the test are not normally distributed, as is the case for total correct scores of the I-BNT, or when the sample is not fully representative of the whole population. Therefore, the present work should be considered only as an illustrative example regarding the creation and advantages of the dynamic database and the online platform with an extension to predict normative scores; not as a paper presenting normative scores for the Indonesian population at large. In the future, it is expected that more data from other islands and ethnic groups will be added to the database allowing for a better quality or more fine-tuned reference groups. More definitive normative scores for the I-BNT and other tests await a more complete dataset or more efficient norming procedure such as regression-based norming generated from a larger variety of demographic factors (Van Breukelen & Vlaeyen, 2005), including parental ethnicity.

Another issue, typical for verbal tests such as the BNT, is that linguistic proficiency in Bahasa, the daily spoken language, and the mother language might play a role in the test performance and consequently influence the normative data. This has not been explored as yet. If our database grows to include a larger variety of subjects, it will allow us to investigate the effects of daily spoken language as this demographic factor is included and stored in the database.

We would like to close with a general reminder against the usage of tests (including the I-BNT) imported from WEIRD and monolingual societies. The presence of normative data for a specific test, does not mean that this test is sufficiently valid in the Indonesian context. Therefore, research toward the validity of the I-BNT remains important. A general remark is that the use of tests from monolingual countries should be discouraged until their full psychometric properties have been explored in the Indonesian context including the role of ethnicity, mother and daily languages spoken. Currently, the test-retest reliability of the whole test battery is being explored, including those of the I-BNT. The preliminary outcome based on a sample of 50 healthy subjects varying in age between 21–64 and 12–22 years of education showed an excellent test-retest reliability for both the total score (.88) and total time (.84); data from other islands were recently collected in order to generate a better applicability to the Indonesian context and they await further analyses. Research to determine the validity of the different tests in the battery, which is considered as their clinical utility (clinical validation), is also foreseen. The data of the I-BNT has established rather large effects of education suggesting a sensitivity for this factor similar to that reported in other cultures and societies. However, until complete psychometrics for the different tests are available, caution is advised in the use and interpretation of these tests.

In all, I-ANDI was developed to assist neuropsychologists. It consists of a dynamic database and an online plat- form which makes it easy to find fined-tuned normative score based on the user's chosen demographic factor(s). The significant main and interaction effects, and the outcomes of the simulation, strongly support the use of fine-tuned reference data for the I-BNT. The system provides the capability of calculating the predicted score when the number of subjects in the selected reference dataset is insufficient. An individual's neuropsychological test performance is shown in the form of informative graphics. In the near future, data from other neuropsychological tests adapted for Indonesian sub- jects will be included in the database, as well as data from other parts (islands) of Indonesia to better reflect Indonesian's ethnic, and linguistic diversity. It should be noted that the values of the new instruments increase following their usage, and that internet-

enabled methods to compile normative data may help clinicians to efficiently characterize their patient's neuropsychological test scores.

Acknowledgement: The authors are indebted to Dr. Nathalie De Vent and her team from the University of Amsterdam for introducing Dutch ANDI and providing help with the code. We also thanked collaborating researchers from Radboud University, Nijmegen (Dr M.P.H. Hendriks) for advise throughout the initiating phase of the project, Soegijapranata Catholic University (Dr. Margaretha Sih Setija Utami, M.Kes., Drs. Haryo Guritno, M.Si., Lucia Trisni Widianingtanti, S.Psi., M.Si., CVR Abimanyu, S.Psi., M.Si), Atma Jaya Catholic University of Indonesia (Dr. Angela Oktavia Suryani M.Si., Justinus Budi Santoso, M.Psi., Psikolog, Dr. Magdalena Surjaningsih Halim, Psikolog, Dr. Yohana Ratrin Hestyanti, Psikolog) and Widya Mandala Catholic University of Surabaya (Florentina Yuni Apsari, S.Psi., M.Si., Psikolog), and the assistants who helped collecting the data. Dr. Frans van Haaren and Dr. John V. Keller helped with linguistic corrections.

Disclosure statement: No potential conflict of interest was reported by the author(s).

Funding: This work was supported by the Directorate of Higher Education General of Indonesia under grant number: 010/L6/AK/SP2H.1/ PENELITIAN/2019.

References

Agelink Van Rentergem, J. A., Murre, J. M. J., & Huizenga, H. M. (2017). Multivariate normative comparisons using an aggregated database. *PLOS ONE*, *12*(3), e0173218. https://doi.org/10.1371/journal.pone.0173218

Alyahya, R. S. W., & Druks, J. (2016). The adaptation of the Object and Action Naming Battery into Saudi Arabic. *Aphasiology*, *30*(4), 463–482. https://doi.org/10.1080/02687038.2015.1070947

Bridges, A. J., & Holler, K. A. (2007). How Many is Enough? Determining Optimal Sample Sizes for Normative Studies in Pediatric Neuropsychology. *Child Neuropsychology*, *13*(6), 528–538. https://doi.org/10.1080/09297040701233875

Burggraaff, J., Knol, D. L., & Uitdehaag, B. M. J. (2017). Regression-Based Norms for the Symbol Digit Modalities Test in the Dutch Population: Improving Detection of Cognitive Impairment in Multiple Sclerosis. *European Neurology*, *77*(5–6), 246–252. https://doi.org/10.1159/000464405

Busch, R. M., Chelune, G. J., & Suchy, Y. (2006). Using norms in neuropsychological assessment of the elderly. In D.K. Attix & K. A Welsh-Bohmer (Eds.), *Geriatric Neuropsychology: Assessment and Intervention Guilford Publications*, 133–157. Guilford Publications.

Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological Assessment: Past and Future. *Journal of the International Neuropsychological Society*, *23*(9–10), 778–790. https://doi.org/10.1017/S1355617717001060

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *Routledge Academic*.

De Vent, N. R., Agelink Van Rentergem, J. A., Schmand, B. A., Murre, J. M. J., Huizenga, H. M., & ANDI Consortium. (2016). Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A Normative Database Created from Control Datasets. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01601

Henderson, L. W., Frank, E. M., Pigatt, T., Abramson, Ruthak. K., & Houston, M. (1998). Race, gender, and educational level effects on Boston Naming Test scores. *Aphasiology*, *12*(10), 901–911. https://doi.org/10.1080/02687039808249458

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test*.

Kessels, R. P. C., & Hendriks, M. P. H. (2016). Neuropsychological Assessment. In *Encyclopedia of Mental Health* (pp. 197–201). Elsevier. https://doi.org/10.1016/B978-0-12-397045-9.00136-1

Meyers, J. E. (2013). *The Meyers neuropsychological system.* http://www.meyersneuropsychological.com

Neils, J., Baris, J. M., Carter, C., Dell'aira, A. L., Nordloh, S. J., Weiler, E., & Weisiger, B. (1995). Effects of Age, Education, and Living Environment on Boston Naming Test Performance. *Journal of Speech, Language, and Hearing Research*, *38*(5), 1143–1149. https://doi.org/10.1044/jshr.3805.1143

Nicholas, L. E., Brookshire, R. H., Maclennan, D. L., Schumacher, J. G., & Porrazzo, S. A. (1989). Revised administration and scoring procedures for the Boston Naming test and norms for non-brain-damaged adults. *Aphasiology*, *3*(6), 569–580. https://doi.org/10.1080/02687038908249023

Oosterhuis, H. E. M., Van Der Ark, L. A., & Sijtsma, K. (2016). Sample Size Requirements for Traditional and Regression-Based Norms. *Assessment*, *23*(2), 191–202. https://doi.org/10.1177/1073191115580638

Shirk, S. D., Mitchell, M. B., Shaughnessy, L. W., Sherman, J. C., Locascio, J. J., Weintraub, S., & Atri, A. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimer's Research & Therapy*, *3*(6), 32. https://doi.org/10.1186/alzrt94

Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., Langellotti, C., & Ivnik, R. J. (2005). Mayo's Older Americans Normative Studies: Age- and IQ-Adjusted Norms for the Boston Naming Test, the MAE Token Test, and the Judgment of Line Orientation Test. *The Clinical Neuropsychologist*, *19*(3–4), 280–328. https://doi.org/10.1080/13854040590945229

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary, Third Edition (3 rd). *Oxford University Press*.

Sulastri, A., Utami, M. S. S., Jongsma, M., Hendriks, M., & van Luijtelaar, G. (2018). The Indonesian Boston Naming Test: Normative Data among Healthy Adults and Effects of Age and Education on Naming Ability. *International Journal of Science and Research (IJSR)*, *8*(11), 134-139.

Tombaugh, T. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, *19*(2), 203–214. https://doi.org/10.1016/S0887-6177(03)00039-8

Van Breukelen, G. J. P., & Vlaeyen, J. W. S. (2005). Norming clinical questionnaires with multiple regression: The Pain Cognition List. *Psychological Assessment*, *17*(3), 336–344. https://doi.org/10.1037/1040-3590.17.3.336

Weintraub, S., Salmon, D., Mercaldo, N., Ferris, S., Graff-Radford, N. R., Chui, H., Cummings, J., DeCarli, C., Foster, N. L., Galasko, D., Peskind, E., Dietrich, W., Beekly, D. L., Kukull, W. A., & Morris, J. C. (2009). The Alzheimer's Disease Centers' Uniform Data Set (UDS): The Neuropsychologic Test Battery. *Alzheimer Disease & Associated Disorders*, *23*(2), 91–101. https://doi.org/10.1097/WAD.0b013e318191c7dd

Zein, S. (2020). Language policy in superdiverse Indonesia (1st ed). *Routledge*.

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

Chapter 3a

Indonesia Neuropsychological Test Battery: Normative Score, Reliability, Age and Education Effects

Published as: Wahyuningrum, Shinta Estri and Hendriks, M.P.H and Luijtelaar, Gilles van and Sulastri, Augustina *Indonesia Neuropsychological Test Battery: Normative Score, Reliability, Age and Education Effects.* Proceeding of Biopsychosocial Issue. pp. 264-273. ISSN 2985-8429

Abstract

Neuropsychological tests are sensitive tools for measuring cognitive abilities, and they refer to the Assessment of healthy persons and various categories of patients. Some cognitive abilities are influenced by age and education, and some are not. Recently, ten neuropsychology tests were adapted for Indonesia, forming the Indonesian Neuropsychological Test Battery (INTB). It is the first neuropsychological test battery administered in Bahasa, Indonesia. This study presents preliminary normative scores, test-retest reliability, and the effects of age and education on each test. Data from four hundred and ninety healthy participants from Java (Jakarta, Semarang, and Surabaya), stored in a dynamic database (Indonesian-ANDI), were used. Preliminary normative scores of INTB are presented. All tests showed a moderate to good test-retest correlation coefficient ranging from 0.51 to 0.84, and an exception was the short and long-term recall scores of the RAVLT. Analysis of variance revealed that eighteen subtests were significantly age-dependent, and the scores tended to decline with the ageing process. Only the time to complete the Bourdon and RAVLT learning over trials did not decrease. In contrast, cognitive performance was increased along with a higher education level. The only exceptions were the time to complete the Bourdon and the RAVLT's learning over trials and delayed recall. These different effects of age and education on the tests of the INTB demonstrate the necessity to correct the normative score of the tests in a tailored way for these factors.

Keywords: INTB, age effect, education effects, neuropsychology, normative score

Introduction

Neuropsychological tests assess and evaluate human cognitive abilities (Zillmer et al., 2008). It is widely recognized that demographic factors may influence a person's cognitive ability. The most influential factors are age and education (Jansen et al., 2021; Murman, 2015). These influences can be positive or negative (Guerra-Carrillo et al., 2017; Jansen et al., 2021; Weber & Skirbekk, 2014). Most cognitive abilities decline along with normal ageing, whereas some do not, or there might be an increase followed by a decrease (Glisky, 2007). Moreover, the peak and decline in cognitive performance vary widely for individuals or populations (Hartshorne & Germine, 2015). In the majority of the studies, education has a positive impact on cognitive abilities: the performance on most of the tests will increase with higher education levels. However, there is also the condition when the ageing process and education do not change cognitive performance, or the change is seemingly insignificant.

Neuropsychological tests have been widely used worldwide, showing adequate evidence to measure individual cognitive abilities. However, in Indonesia, this is rarely done. The interpretation of the test scores requires normative data, and recently collected normative scores are mostly lacking for the Indonesian population. Therefore, recently a group of researchers adapted a series of neuropsychological tests for Indonesia, the Indonesian Neuropsychological Test Battery (INTB). The INTB consists of ten neuro- psychological tests covering three significant domains: learning and memory, language and executive function, including attention. The INTB was administered as a series of tests using "Bahasa Indonesia, or called Bahasa hereafter," in six different mainly urbanized parts of West, Middle and East Java, Bali, South Sulawesi, and East Kalimantan. Data gathered before the covid19 pandemic were collected from 890 healthy subjects. All data were stored in a "dynamic" database and an online platform called the Indonesian Neuropsychological Dynamic Infrastructure (I-ANDI) (Wahyuningrum et al., 2021). Like many other psychological tests to be used by practitioners, the INTB should have normative scores to interpret the tests' results correctly. A normative score is the expected test score from a sample mimicking the clients' demographic factors as much as possible (Zillmer et al., 2008). We used this data set to propose normative scores of all these ten tests for Indonesia, which are now standard practices worldwide (Lezak et al., 2012; Lovden et al., 2020). We wanted to know whether the newly adapted tests for Indonesia are sensitive to demographic effects. If that is the case, the normative scores must be corrected for these demographic factors. In case age and education effects are found, and they mimic what is

internationally reported, this contributes to the validity of the tests. Two earlier studies on a smaller sample, one on the Indonesian version of the Boston Naming Test (Sulastri et al., 2018) and one on the Trail Making Test (Widhianingtanti et al., 2022) already suggested that this is indeed the case, now this will be investigated for all tests in the battery and a much larger sample.

Here we report the effects of education and age on the whole battery of the Javanese sample. More specifically, we investigate whether there are also age-dependent increases in cognitive performances, the age of a putative peak and a decline in the performances. Next, we also present preliminary normative scores and outcomes of the reliability analyses of all ten tests.

Methods

INTB was conducted using the paper and pencil method. The tests were: The Token Test (TT), Boston Naming Test (BNT), and phonemic Verbal Fluency Test (VFT), which were intended to assess language domain. The Stroop Test (ST), Bourdon Test (BWT), Trail Making Test (TMT), and Five Point test (FPT) were administered and aimed to assess the attention and executive function domain. The Digit Span (DS), Figural Reproduction (FR), and Rey Auditory Verbal Learning (RAVLT) represented the learning and memory domain. All participants completed the test battery, which took approximately two and a half hours.

Here we report the data from the Javanese population; four hundred and ninety healthy participants from three big cities in Java Island, Indonesia (Jakarta, Semarang, and Surabaya) were involved. The age range of the participants was from 16-80 years, with education levels ranging from elementary school (6 years) to doctoral programs (over 17 years). The age was categorized into six with decade intervals except for the youngest and oldest groups. The categories were: 16-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, and 60-80 years. Education was divided into four categories: 0-9 years, 10-12 years, 13-16 years, and above 17 years. Tests were conducted by a research assistant, a second-grade student from the psychology faculty who trained to collect data. Data collection was conducted in compliance with the Helsinki Declaration, and the ethics committee of Soegijapranata University gave clearance for this research project (University Ethical Clearance number: 001B/B.7.5/ FP.KEP/IV/2018).

Another fifty participants were involved in determining the test-retest reliability of the INTB. We used *Pearson Correlation Coefficient* with a different dataset. The test-retest interval was three weeks. The score of each test was converted into a z-score to facilitate comparisons between scores of the various tests. The inverse scores of all variables that measured time and number of errors were used.

Analyses of Variance (*ANOVA)* were used to determine the age and education effects of the ten test scores. Basic rules of thumb for *partial eta squared* (η^2) as a measure of the effect size are that η^2= .01 indicates a small effect; η^2= .06 indicates a medium effect; η^2= .14 indicates a significant effect.

Results

The minimum and maximum score, mean and standard deviation of 490 participants for the ten tests are shown in Table 1. Means and standard deviations were calculated using data from all participants (N= 490). Normative scores for different combinations of age and education could not be given as yet because the number of participants in certain groups is still less than 50 (Bridges & Holler, 2007). For example, in the low education group, with a government program of nine years of compulsory education, it becomes challenging to find participants in this group. In addition, minimum and maximum scores explain the range of performance for all participants.

Table 1. Mean, standard deviation, minimum, and maximum of ten tests (n = 490).

Variables	Min; Max	Mean	Standard Deviation
BNT score	34; 60	51.37	4.95
BNT Time	60; 848	308.81	161.27
FR score	4; 15	11.87	2.54
DS forward	1; 14	7.48	2.24
DS backward	0; 15	6.28	2.40
DS sequence	0; 16	7.74	2.89
BWT score	118; 327	193.14	13.82
BWT error	0; 58	10.07	9.57
ST score card 3	81; 100	97.76	3.01
ST score card 3-2	-19; 18	-1.16	3.37
ST total error	0; 22	3.29	3.56
RAVLT Mean A1-A5	4; 14.60	10.05	2.00
RAVLT LOT	-5; 46	16.87	7.80
RAVLT STPR	37.5; 150	90.01	16.68

Table 1. Continued

Variables	Min; Max	Mean	Standard Deviation
RAVLT LTPR	36.36; 137.50	88.62	17.17
VFT total score	6; 90	40.59	13.36
TT score	9; 163	146.86	22.85
FPT unique number	3; 58	25.97	9.52
TMT Time A	9; 134	44.99	18.79
TMT Time B	17; 426	87.52	50.67

Pearson correlation coefficient revealed the test- retest reliability of the ten tests. The data are presented in Table 2. Good reliability (> .80) was introduced by TT, BNT, and TMT time B scores. Moderate coefficients were revealed by DS forward, DS backwards, DS sequence, FR score, FPT unique number, TMT time A, ST card 3, BWT score, RAVLT mean A1-A5, and VFT score with a range of .52 to .78. At the same time, a low correlation coefficient was found for RAVLT LOT, RAVLT STPR, and RAVLT LTPR, ST card 3-2, ST error with coefficient values less than .49.

Table 2. Pearson correlation coefficient of ten tests (n=50).

Variables	r
BNT score	.88
BNT Time	.84
FR score	.73
DS forward	.78
DS backward	.76
DS sequence	.60
BWT score	.54
BWT error	.86
ST score card 3	.52
ST score card 3-2	.35
ST total error	.49
RAVLT Mean A1-A5	.78
RAVLT LOT	-.15
RAVLT STPR	.19
RAVLT LTPR	.10
VFT total score	.72
TT score	.84
FPT unique number	.67
TMT Time A	.66
TMT Time B	.81

Table 3 shows the results of the ANOVA on the age effects for ten tests. Age had a large impact as expressed by a measure of effect size on RAVLT mean A1-A5 (η^2= .210), FPT unique number (η^2 = .141), TMT time A (η^2 = .228), and TMT time B (η^2 = .203), the other variables showed a medium to large effect size. There was no statistically significant age effect for the BWT score and RAVLT LOT (p=.207 and p =.308, respectively).

Table 3. Age effects of ten tests.

Variables	F	sig	Partial eta squared (η^2)
BNT score	8.98	.000	.085
BNT Time	3.92	.002	.039
FR score	8.47	.000	.108
DS forward	11.77	.000	.108
DS backward	8.71	.000	.083
DS sequence	15.84	.000	.083
BT score	1.44	.207	.015
BT error	11.417	.000	.106
ST score card 3	5.52	.000	.054
ST score card 3 - 2	2.49	.031	.025
ST total error	5.84	.000	.057
RAVLT mean A1-A5	25.68	.000	.210
RAVLT LOT	1.20	.308	.012
RAVLT STPR	6.07	.000	.059
RAVLT LTPR	3.86	.002	.038
VFT total score	6.56	.000	.063
TT score	4.45	.001	.044
FPT unique number	15.85	.000	.141
TMT Time A	28.66	.000	.228
TMT Time B	24.73	.000	.203

Note: BT score and RAVLT LOT have no significant age-effects

Figure 1, left below, illustrates the domain attention and executive function age trends. The overall trend for this domain tends to decline except for performance on the BWT time score. The time to complete the test was not age dependent; in contrast, the number of errors increased with ageing. The lines with colours light blue (TMT time A), green (TMT time B), and black (FPT unique) show the age-dependent effects of these tests. The three tests show the same onset decline at 30 years old and a significant decrease in the most senior age group (60+). The three variables of the Stroop Test show a substantial reduction in the age above 60, although, for the Stroop card 3, the decline started a bit earlier, at the age of 50 and continued with a slight decrease at age 60 and above.

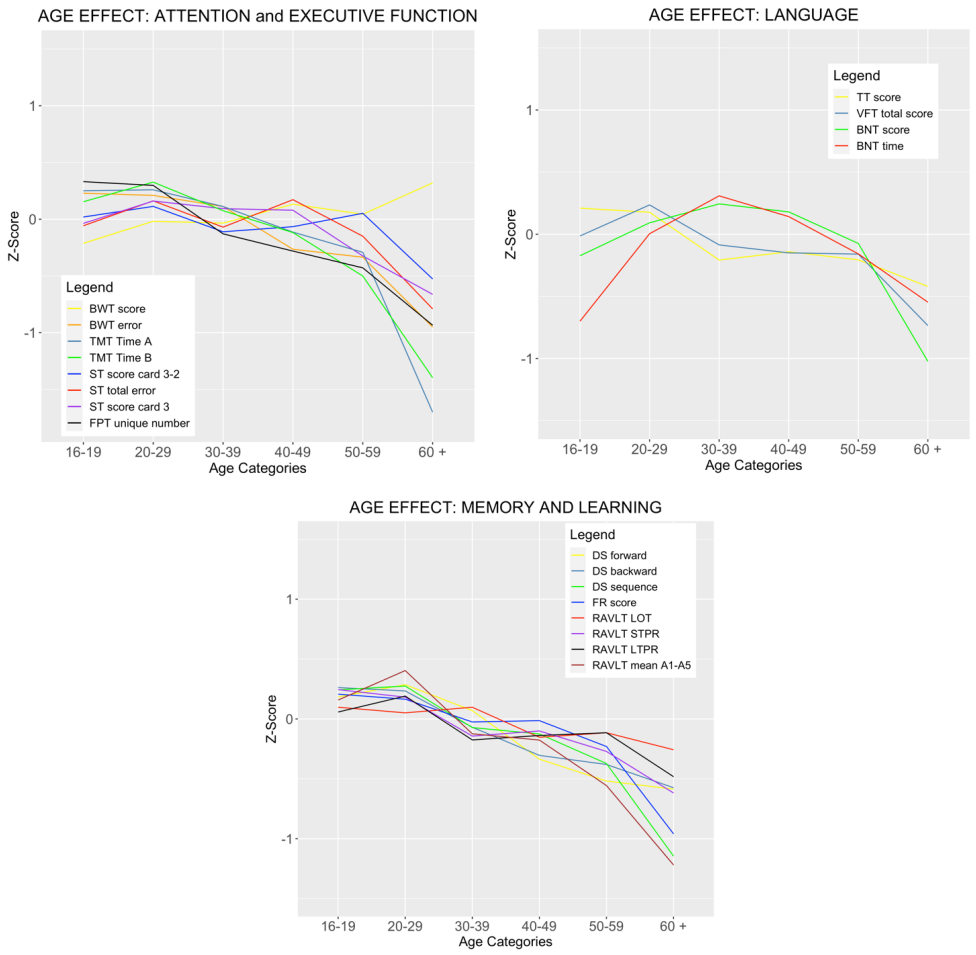


Figure 1. Age-effects for attention and executive learning domain (left side), language domain (middle side), and memory and learning (right side).

The following graph on the middle illustrates the language domain's age trends all subtests on this domain show medium to significant effects. The oldest category showed the most significant decline. Interestingly, we found the peak in the performance on the BNT, not for the younger category but the 30- 39 years category. In contrast, VFT and Token tests showed a decline at 30, but both tests' performances remained stable in the following age category.

The last graph illustrated the memory and learning domain. All subtests show a declining trend except for RAVLT LOT, the line with the light red colour. As

mentioned in Table 2, LOT was not significantly found to be affected by the ageing process; the figure shows a straight line. This means that this verbal learning test's "learning over trials" was relatively stable for all age categories. In contrast, RAVLT represents A1-A5 or the mean of the recall of the five first recall trials, which showed a significant age effect. This trend, demonstrated by the line-colour dark red, shows that the peak performance occurred at age 20 and continues with two declines, first at age 30 and later at 50. Other variables show a similar trend, starting at age 30 and significantly declining for the oldest categories at 60+.

Table 4. Education effects of ten tests.

Variables	F	sig	Partial eta squared (η^2)
BNT score	43.04	.000	.210
BNT time	30.42	.000	.158
FR score	18.25	.000	.101
DS forward	22.52	.000	.122
DS backward	18.73	.000	.104
DS sequence	22.88	.000	.124
BT score	.92	.429	.006
BT Error	10.60	.000	.061
ST score card 3	12.67	.000	.073
ST score card 3 - 2	10.54	.000	.061
ST total error	17.61	.000	.098
RAVLT mean A1-A5	14.27	.000	.081
RAVLT LOT	.66	.576	.004
RAVLT STPR	5.08	.002	.030
RAVLT LTPR	.043	.988	.000
VFT total score	41.19	.000	.203
TT score	22.43	.000	.122
FPT unique number	17.38	.000	.097
TMT time A	27.13	.000	.143
TMT time B	64.34	.000	.284

Note: Bourdon score, RAVLT LTPR and RAVLT LOT have no significant education-effects.

As presented in Table 4, the education effect revealed only three subtests were insignificant. In contrast, the rest of the subtests were significant, with a p-value below .01. The subtests were BWT score (p=.429), RAVLT LOT (p=.576), and RAVLT

LTPR (p=.988). Interestingly, on the subtest, TMT time A, TMT time B, VFT score, BNT score, and BNT time found significant education effects with an effect size above .14.

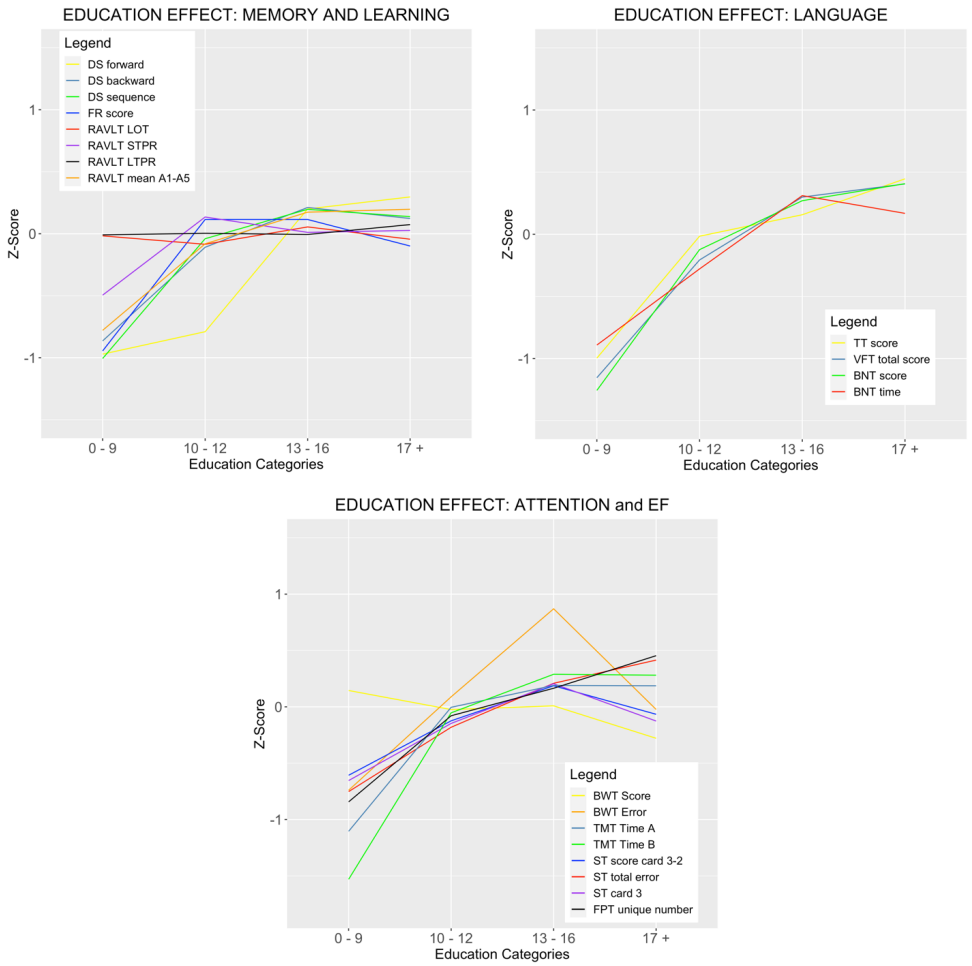


Figure 2. Education-effects for attention and executive learning domain (left side), language domain (middle side), and memory and learning (right side).

The effect of education on the ten tests is illustrated in Figure 2. Overall, cognitive trends in education categories increased along with educational attainment. The worst performance was the lowest education categories, except for BWT score and RAVLT LOT and LTPR, which are not significantly affected by education.

The left side graph demonstrates performance in the attention and executive function domain. On TMT performance, both subtests show a significant increase

in the second education category followed by a slight increase in the following categories and tend to be stable on the differences for the last two categories. The consistent increasing performance found on the FPT and Stroop total error, illustrated by black and red lines, shows a positive linear correlation. While on the three subtests (Stroop Card 3-2, Stroop score card 3, and Bourdon-Wiersma error), there is an increase in the first three education categories and continues with a slight decline for the highest category.

The language domain tests correlate positively with education attainment, which is presented on the middle side graph. Performance on language rose in the second and third categories but did not significantly increase for the highest category. The last domain, learning and memory, is shown in the right side graph in the middle. Some variables in this domain significantly increase from the lowest to the next education category, followed by a flatline. A different trend is DS forward, which shows a significant increase a bit later, at 13 years of education. In contrast, a trend is shown by the black and red flatline describing the performance on RAVLT LOT and LTPR, which have no significant changes in those performances along the education categories.

Discussions

We analyzed the effects of the demographic factors using 490 subjects to see whether the tests were sensitive to commonly reported age and education effects. The implication would be that the normative scores for these tests needed to be adapted and corrected for these demographic factors. We propose only preliminary scores of the whole Javanese sample because age and education effects were found, demanding different normative scores for different age and education groups. Since, for most of the combination of age and education categories, the number of subjects was less than 50, and this number is necessary to obtain a reliable and representative score per subcategory, we present only preliminary normative scores for ten tests (Bridges & Holler, 2007). Next, we offer the reliability of the tests, next to age and education effects.

In general, the reliability of the ten tests was also good enough, proving that the tests were reliable for the Indonesian population. However, we found a notable trend in the result of the test-retest reliability of the RAVLT on three of four investigation scores (LOT, STPR, and LTPR). Apart from the low reliability of STPR (0.19) and LTPR (0.10), we found negative results for LOT. The latter indicates that the mean score of the retest was lower than the test. The South African adolescent study also revealed a

lower mean score on the second test (Blumenau & Broom, 2011). The low test-retest reliability in our analysis of the RAVLT was most probably due to the that we used the same word list in both sessions, and this had a consequence that the fifteen items were still in the participants' memory as expressed by higher memory scores and a lower increase over trials. The reliability score for the recall of the first five trials was close to .8, suggesting that this aspect was less bothered by using the same wordlist.

In general, the trend on the age effect shows that almost all tests showed the anticipated decline (Cohen et al., 2019; Elkana et al., 2015). In nearly all tests or variables, younger people outperform older, except in the language domain, where the middle age categories perform better than others. From all subtests, we conclude that the most senior category performed poorer in all tests, and an accelerated decline occurred around sixty. As expected, not all variables or tests start to decline at the same age, most start at age 30, and some come later (Glisky, 2007). A very general conclusion is that this study confirms that the Indonesian population's cognitive performance is influenced by ageing but that the ageing process is different for the different tests.

Regarding the education effect, we found that participants with low education performed the worst, in agreement with what is generally reported (Guerra-Carrillo et al., 2017). The significantly better performance was shown by those with at least ten years or more of education. For the domain language and executive function, including attention, better performance gradually increased until the category with the most years of schooling. Different from both domains, there is no significant improvement for memory and learning domains on the educational level after 13 years of education.

From the ten tests, we emphasize that the Bourdon-Wiersma score time and RAVLT LOT have no significant effects on either age or educational level. Bourdon-Wiersma score was used to measure the ability to maintain accuracy and concentration while looking for a group of 4 dots. Being attentive and scanning the lines fast is an ability that remains intact in older people and seems already present even after elementary school. The number of errors of the Bourdon-Wiersma test was age-dependent and did not deviate much from the cognitive decline found for

other executive functions. Interestingly, although there is no significant age effect, the Bourdon scores show a slight increase in the last age categories. We expect our older participants to be more patient and careful when completing this test. RAVLT LOT is used to measure the ability to learn over the trial. Our population shows this ability remains stable even for the oldest and less educated groups.

We found that three subtests have a significant impact on the ageing process. The first was the RAVLT mean of trial A1 to trial A5. This variable represented the average immediate verbal recall. A sudden decline happened from the early young older category (40 years old) and continued to the most senior category. The second was FPT, a score of unique number design associated with creativity and mental flexibility, the decline starts earlier, at age 30, and the decline continues until the oldest category. The third was TMT time A and time B; both subtests were about speed to visually detect and recognize the sequence of numbers or letters. The knowledge that speed processing is also age-dependent decline (Salthouse, 2010) might be helpful for the next researcher on the neuropsychological test, which measures speed or visual or auditory ability with participants who are elderly, to have a pre-test for the perception abilities for fairness of the results.

Furthermore, education level largely influences the performance of VFT total score and BNT time and total correct score, and these tests belong to the language domain. Some previous studies conclude that performance in the language domain remained stable and increased along with educational attainment (Murman, 2015). In BNT performance, we expect that the vocabulary and knowledge were raised during the years of education and the better jobs that the better educated have. Other subtests which gave significant effects were TMT time A and B. In addition to measuring speed, this test also measures working memory. Concerning the VFT total score, a participant was asked to produce words beginning with a specific letter. Both test scores increase at higher levels of education (Troyer, 2000).

Conclusions

The scores on ten cognitive tests for the Javanese population highly depended on age and education. The performance of the elderly was the lowest compared to all age categories. Participants with the lowest education level performed poorly on almost all cognitive tests. Tests in the language domain were the most sensitive for education, while attention, executive function, memory, and learning became more stable after ten years of schooling. The age and education effects of the tests of the INTB imply that representative and valid normative data need to be developed for these different categories of Javanese people, adapted for age and education.

Acknowledgement: This work was supported by the Directorate of Higher Education General of Indonesia with Dikti grant number 0317/AK.04/2022. We also thanked the Indonesian Neuropsychological Consortium, research assistants, and all participants who contributed to the current project.

References

Blumenau, J., & Broom, Y. (2011). Performance of South African Adolescents on Two Versions of the Rey Auditory Verbal Learning Test. *South African Journal of Psychology*, *41*(2), 228–238. https://doi.org/10.1177/008124631104100211

Bridges, A. J., & Holler, K. A. (2007). How Many is Enough? Determining Optimal Sample Sizes for Normative Studies in Pediatric Neuropsychology. *Child Neuropsychology*, *13*(6), 528–538. https://doi.org/10.1080/09297040701233875

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

Elkana, O., Eisikovits, O. R., Oren, N., Betzale, V., Giladi, N., & Ash, E. L. (2015). Sensitivity of Neuropsychological Tests to Identify Cognitive Decline in Highly Educated Elderly Individuals: 12 Months Follow up. *Journal of Alzheimer's Disease*, *49*(3), 607–616. https://doi.org/10.3233/JAD-150562

Glisky, E. (2007). Changes in Cognitive Function in Human Aging. In D. Riddle (Ed.), *Brain Aging* (Vol. 20072731, pp. 3–20). CRC Press. https://doi.org/10.1201/9781420005523.sec1

Guerra-Carrillo, B., Katovich, K., & Bunge, S. A. (2017). Does higher education hone cognitive functioning and learning efficacy? Findings from a large and diverse sample. *PLOS ONE*, *12*(8), e0182276. https://doi.org/10.1371/journal.pone.0182276

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

Jansen, M. G., Geerligs, L., Claassen, J. A. H. R., Overdorp, E. J., Brazil, I. A., Kessels, R. P. C., & Oosterman, J. M. (2021). Positive Effects of Education on Cognitive Functioning Depend on Clinical Status and Neuropathological Severity. *Frontiers in Human Neuroscience*, *15*, 723728. https://doi.org/10.3389/fnhum.2021.723728

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (Fifth edition). Oxford University Press.

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

Murman, D. (2015). The Impact of Age on Cognition. *Seminars in Hearing*, *36*(03), 111–121. https://doi.org/10.1055/s-0035-1555115

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society*, *16*(5), 754–760. https://doi.org/10.1017/S1355617710000706

Sulastri, A., Utami, M. S. S., Jongsma, M., Hendriks, M., & van Luijtelaar, G. (2018). The Indonesian Boston Naming Test: Normative Data among Healthy Adults and Effects of Age and Education on Naming Ability. *International Journal of Science and Research (IJSR)*, *8*(11), 134-139.

Troyer, A. K. (2000). Normative Data for Clustering and Switching on Verbal Fluency Tasks. *Journal of Clinical and Experimental Neuropsychology*, *22*(3), 370–378. https://doi.org/10.1076/1380-3395(200006)22:3;1-V;FT370

Wahyuningrum, S. E., van Luijtelaar, G., & Sulastri, A. (2021). An online platform and a dynamic database for neuropsychological assessment in Indonesia. *Applied Neuropsychology: Adult*, 30(3), 330–339. https://doi.org/10.1080/23279095.2021.1943397

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

Widhianingtanti, L. T., Luijtelaar, G. V., Suryani, A. O., Hestyanti, Y. R., & Sulastri, A. (2022). Indonesian Trail Making Test: Analysis of Psychometric Properties, Effects of Demographic Variables, and Norms for Javanese Adults. *Jurnal Psikologi*, *49*(2), 104. https://doi.org/10.22146/jpsi.68953

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

Chapter 3b

The Indonesian Neuropsychological Test Battery (INTB): Psychometric Properties, Preliminary Normative Scores, The Underlying Cognitive Constructs, And The Effects Of Age And Education

Published as: Wahyuningrum, S. E., Sulastri, A., Hendriks, M. P. H., & van Luijtelaar, G. (2022). The Indonesian Neuropsychological Test Battery (INTB): Psychometric properties, preliminary normative scores, the underlying cognitive constructs, and the effects of age and education. *Acta Neuropsychologica, 20*(4), 445–470. https://doi.org/10.5604/01.3001.0016.1339

Abstract

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

Material and Methods: Four hundred ninety healthy adults living on Java Island participated in this study. All subjects completed all tests. The test-retest reliability was determined in a parallel study with fifty participants.

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

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

Keywords: Cognitive construct, Indonesian Neuropsychology Test Battery, Psychometric, age-effect, education-effect.

Introduction

A neuropsychological test is a tool that assesses a person's specific cognitive abilities (Elkana et al., 2015; Kessels & Hendriks, 2023; Lezak et al., 2012). With rapid modern technologies in brain imaging, neuropsychological tests (NPTs) are imperative to establish disturbances in any specific cognitive function domains. NPTs are ideally suited to establish or complement cognitive diagnostics (Zillmer et al., 2008). Assessment using NPTs also can be used to illustrate the strength and weaknesses of an individual's cognitive patterns by comparing his/her score with scores from healthy subjects with similar demographic characteristics, called normative scores (Zucchella et al., 2018).

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

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

Occasionally, subtests from one NPT can reflect more than one cognitive domain (Bialystok et al., 2014; Mengual-Macenlle et al., 2015; Nielsen et al., 2018; Santos et al., 2015; Tucha et al., 2012). Previous studies grouped NPTs variables to reveal

a specific domain (Chapman et al., 2011; Siedlecki et al., 2008). For illustration, the construct validity of the Montreal Cognitive Assessment battery, some variables of different tests measure the same cognitive construct (Vogel et al., 2015). Many methods can be used to explore the underlying cognitive constructs of a series of tests. One of them is Principal Component Analysis (PCA). PCA is data-driven and based solely on analysing the data itself, without having to make any assumptions about the underlying mechanisms. It may reduce high-dimensional data sets to a small number of modes or constructs. The constructs depend on the dataset (Santos et al., 2015). Therefore, they will be specifically meant for interpretation and mainly restricted to the population being studied (Chapman et al., 2011; Fong et al., 2019).

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

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

Method

Participants

Four hundred and ninety participants were recruited from three different cities in Java (West Indonesia) to collect normative data and investigate the underlying cognitive factors in the data set. The data were collected in collaboration with two universities as consortium members (Wahyuningrum et al., 2022). Exclusion criteria were age less than 16 years and self-reported history of any developmental or acquired brain injury or brain-related diseases. All participants had been informed of their privacy rights and knew that their data would be used for scientific purposes. The ethical issue compliance was in line with the WMA Declaration of Helsinki and approved by the local ethics committee of Soegijapranata Catholic University. The participants received seventy-five thousand rupiahs (equal to five US dollars) after finishing series of tests.

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

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

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

Measurement and Assessment

The INTB consists of ten paper and pencil-based tests conducted using the official Indonesian language, "Bahasa Indonesia". All participants can speak and read Bahasa Indonesia at least at a primary school level. Participants were allotted approximately two hours to complete the series of the ten tests. The tests are presented in the same order and highly standardized for all participants. Trained

research assistants were students in the final year of their psychology study, called the tester of the INTB. All ten tests were administered and scored according to standard instructions. These tests were presented in the following order:

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

Rey Auditory-Verbal Learning Test was adapted from RAVLT. The test measures verbal episodic memory, learning over trials, and short- and long-term recall (after 20 minutes delay) (Strauss & Spreen, 1991). There were seven trials in two parts: part one was the learning and memorizing condition of 15 words nouns (list A) presented five times (A1-A5), followed by a distraction in the form of the presentation of 15 different words (list B). In part two, subjects were asked to recall the 15 words from list A (A6), and after a 20-minute delay, they were asked once more to recall list A (A7). Four variables were used: first, the mean score of trial 1 to trial 5. Second, a learning score over the five trials (LOT). This was calculated by dividing the difference in the scores between trial five and trial one by five. The third variable (A6) is short-term retention (STPR), expressed as the percentage of recalled items of A5. The last variable is long-term retention (LTPR); this delayed recall score was expressed as a percentage of A5 as well. The test was adapted for Indonesia by the replacement of the majority of the words with more familiar ones to Indonesian people. We translated and adapted the words of the RAVLT created by Geffen et al. (1994) using direct translation from English, and some words were chosen by their closest meaning rather than the number of syllables/pronunciations/phonemes of the words and also by the familiarity of the terms for Indonesians (Utami et al., 2024).

The long version of the *Boston Naming Test (BNT)* originated from Kaplan et al. (2016) and was adapted for Indonesia by Sulastri et al. (2018). This test, measuring verbal naming ability, consists of 60 black and white drawings of objects. Participants are asked to name the objects within 20 seconds. If they fail, a phonemic and semantic cue is given, respectively. Both the total number of correct items and the total time to complete the test are the most commonly used dependent variables.

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

by the participant, as well as the number of perseveration errors (the number of repeated designs).

The Trail Making Test (TMT) (Reitan & Wolfson, 1995) is used to measure executive functioning and divided attention. TMT is divided into two parts, part A and part B, and both measure the time spent completing the test. In part A, the participant draws a line connecting circles within numbers 1-25. Furthermore, in part B, the participant must connect two sets of stimuli (number and letter) in an alternating sequence. The dependent variables were both time consumption of part A and part B.

The *Verbal Fluency test (VF)* evaluates the spontaneous production of words under restricted search conditions (Strauss & Spreen, 1991), and the phonemic version that has been used in this study is also considered as an executive function test. In this test, the participants have to produce words of three phonemic categories. For each category, the subject is given 60 seconds. Participants are asked to produce as many words that begin with the letter K, second with the letter S, and third with the letter T. The constraints are not to mention a name of a person or place. The number of the correct words for each phonemic category was used as the dependent variable, as well as the sum of the three categories. The test was adapted for Indonesia, and the choice of the three letters was based on (Hendrawan & Hatta, 2010).

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

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

The *Bourdon Wiersma test (BW)* measures concentration and sustained attention. The test consists of 50 lines for each with 25 groups of dots with a varying (three

to five) number. Participants were instructed to mark the group with four dots. The dependent variables were the (mean) time to finish the rows and the number of errors (both misses and false positives).

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

Statistical Analysis

The reliability of the tests was determined in three different ways. First, by using *a test-retest method* (n = 50), with an interval of seven to fourteen days. The first and second test administration were correlated (*Pearson's correlation coefficient*) and compared using Student's *t-test* for dependent groups. Next, the *Intraclass Correlation Coefficient* (ICC) and *Standard Error of Measurement* (SEM) were used as reliability coefficients. ICC was used to measure the internal consistency reliability or discrepancy between subjects, and SEM was defined as the determination of the amount of variation or spread in the measurement errors for a test and is subsequently an indicator of the reliability of a test (Geerinck et al., 2019). Low levels of SEM (close to 0) indicate high levels of score accuracy, and high level of SEM (close to the SD of the observed score) indicates low levels of score accuracy. SEM for test and retest was calculated by the difference of standard deviation between the scores of the two assessments divided by the square root of 2 (Geerinck et al., 2019; Palta et al., 2011). ICC values were interpreted as 'excellent' when > 0.90, as 'good' between 0.75-0.90, as "moderate" between 0.50 and 0.75 and 'poor' when < 0.50 (Koo & Li, 2016).

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

Preliminary normative scores are presented using *mean*, *standard deviation*, *median*, and *percentile scores* of twenty-four variables of the ten NPTs. All INTB variables were standardized to *z-scores* (with a mean of zero and an SD of 1) as our dataset

for factor analysis. Furthermore, eleven variables representing error scores or time to complete the test were reversed by multiplying the z-score value by minus one.

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

Finally, *a two-factor Multivariate Analysis of Variance* (MANOVA) followed one factor ANOVA's and *Bonferroni post-hoc tests* were employed to investigate the effect of age and education as between-subject factors on the underlying cognitive constructs. Considering the age groups' education differences as shown in Table 1 and the age differences between the education groups (data not presented) we used ANCOVA's to control for these factors in the *post-hoc* one factor ANOVA's. *The post-hoc tests* were aimed to establish both global differences between age groups (global decline) as well as significant declines between two neighbouring age categories, namely a fast decline. The latter is indicative of rather large declines between adjacent age groups. The *mean* and the *standard error* of each age and education category for each cognitive construct or factor are presented in Figure 1. In addition, the *orthogonal polynomial contrasts* were used to describe the age-dependent changes for each of the cognitive constructs, more precisely, whether a significant amount of variance could be explained by linear, quadratic, or cubic trends. As a rule of thumb, to determine the effect size, we used *eta squared* (η^2), a value of < 0.06 is a small effect, between 0.06 and 0.14 is a medium-size effect, and > 0.14 is a large effect (J. Cohen, 1988).

Results

Preliminary Normative Score

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

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

Assessment	Variable	Mean	SD	Min; max	Median	Percentile (5 and 95)
Digit Span (DS)	Forward	7.48	2.24	1; 14	7.00	(4.0; 11.0)
	Backward	6.28	2.40	0; 15	6.00	(3.0; 11.0)
	Sequence	7.74	2.89	0; 16	8.00	(3.5; 14.0)
RAVLT	Learning over trials	16.87	7.80	-5; 46	17.0	(3.0; 29.0)
	Mean A1-A5	10.05	2.01	4; 14.6	10.3	(6.4; 13.0)
	Short term recall	90.01	16.68	37.50; 15	91.7	(60.0; 116.7)
	Long term recall	88.61	17.17	36.36; 137.5	90.9	(57.14; 115.4)
Boston Naming Test (BNT)	Total correct	51.37	4.95	34; 60	52.0	(42.0; 58.0)
	Total time	309	161	60; 848	260	(125;643)
Ruff's Five Point Test (FP)	Unique number	25.97	9.52	3; 58	25.5	(10.0; 41.45)
	Perseverance errors	3.35	6.19	0; 51	1.0	(0.0; 13.45)
Trail Making Test (TMT)	Time A	44.99	18.79	9; 134	41.0	(23.0; 81.0)
	Time B	87.52	50.67	17; 426	75.0	(41.0;191.25)
	Time B - time A	42.53	41.75	-51; 340	32.0	(6.0; 114.45)
	Errors A	0.50	1.82	0; 21	0	(0.; 2.)
	Errors B	1.48	3.64	0; 24	0	(0; 10.0)
Verbal Fluency Test (VF)	Letter K	14.44	5.02	3; 41	14.5	(7.0; 23.0)
	Letter S	13.72	5.41	1; 44	14.0	(6.0; 23.0)
	Letter T	12.43	4.74	1; 33	12.0	(5.0; 20.0)
	Total score	40.59	13.36	6; 90	41.0	(6.0; 90.0)
Stroop Coulor Word Test (S)	Card 1 time	49.73	17.62	10; 240	45.5	(35.0; 75.0)
	Card 2 time	58.55	17.79	31; 150	54.0	(41.0; 95.5)
	Card 3 time	86.93	2.28	18; 255	80.0	(55.6;143.4)
	Time card 3 – card 2	28.38	19.57	-40; 149	26.0	(3.44; 62.0)
	Card 3 correct	97.76	3.01	81; 100	99.0	(92.0; 100.0)
	Total Errors	3.29	3.56	0; 22	2.0	(0.0; 11.0)
Figural Reproduction (FR)	Total correct	11.87	2.54	4; 15	12.0	(7.0; 15.0)
	Total time	61.21	28.54	17; 277.9	54.55	(29.0; 115.9)
Bourdon Wiersma Test (B)	Mean time	11.1	2.7	6.07; 25.8	10.58	(7.9; 15.9)
	Number of errors	10.1	9.6	0; 58	7.00	(1.0; 30.4)
	SD time	2.1	1.5	0.78; 20.6	1.75	(0.99; 3.95)
Token Test (T)	Total correct	146.86	22.85	9;163	156	(98.6; 163.0)
	Number of errors	14.13	17.77	0; 84	7.00	(0.0; 55.0)

Reliability measures

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

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

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

Nine tests showed moderate to excellent reliability on the ICC index (0.60 - 0.91). However, it was poor for three of the four RAVLT variables learning over trials, short-term percent retention, and long-term percent retention (under 0.25). Only the

mean performance A1-A5 (0.87) showed excellent reliability, which confirms the outcomes of a previous study (De Sousa Magalhães et al., 2012). We used the same version of the RAVLT to determine the test-retest reliability, others used a parallel version. Therefore, learning over trials and both recall scores were affected by the performance one or two weeks earlier. Table 3 also shows the SEM indices for all variables. All the SEM values indicate lower values than the standard deviation of the experimental test. The variables that measured "time" seemed slightly higher, but all of them were below their *standard deviation*.

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

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

Factor Loadings and Their Interpretations

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

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

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

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

All variables from the DS test loaded exclusively on a single factor with factor loadings ranging between 0.65 to 0.82, also TMT time B (.333), errors of Token test (.395), and RAVLT mean A1-A5 (.482) loaded on this construct. The score of DS forward is generally interpreted as a short-term auditory memory buffer or the phonological loop as part of Baddeley's working memory model and relies on attention. The backward and sequence scores are considered and interpreted as the result of the phonological and the central executive parts of Baddeley's working memory model to manipulate the digits from back to forth. The errors of the Token test, TMT time B, and RAVLT mean A1-A5 might also contain attention and working memory component. Therefore, cognitive factor two revealed the cognitive ability of **attention, auditory short-term, and working memory**.

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

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

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

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

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

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

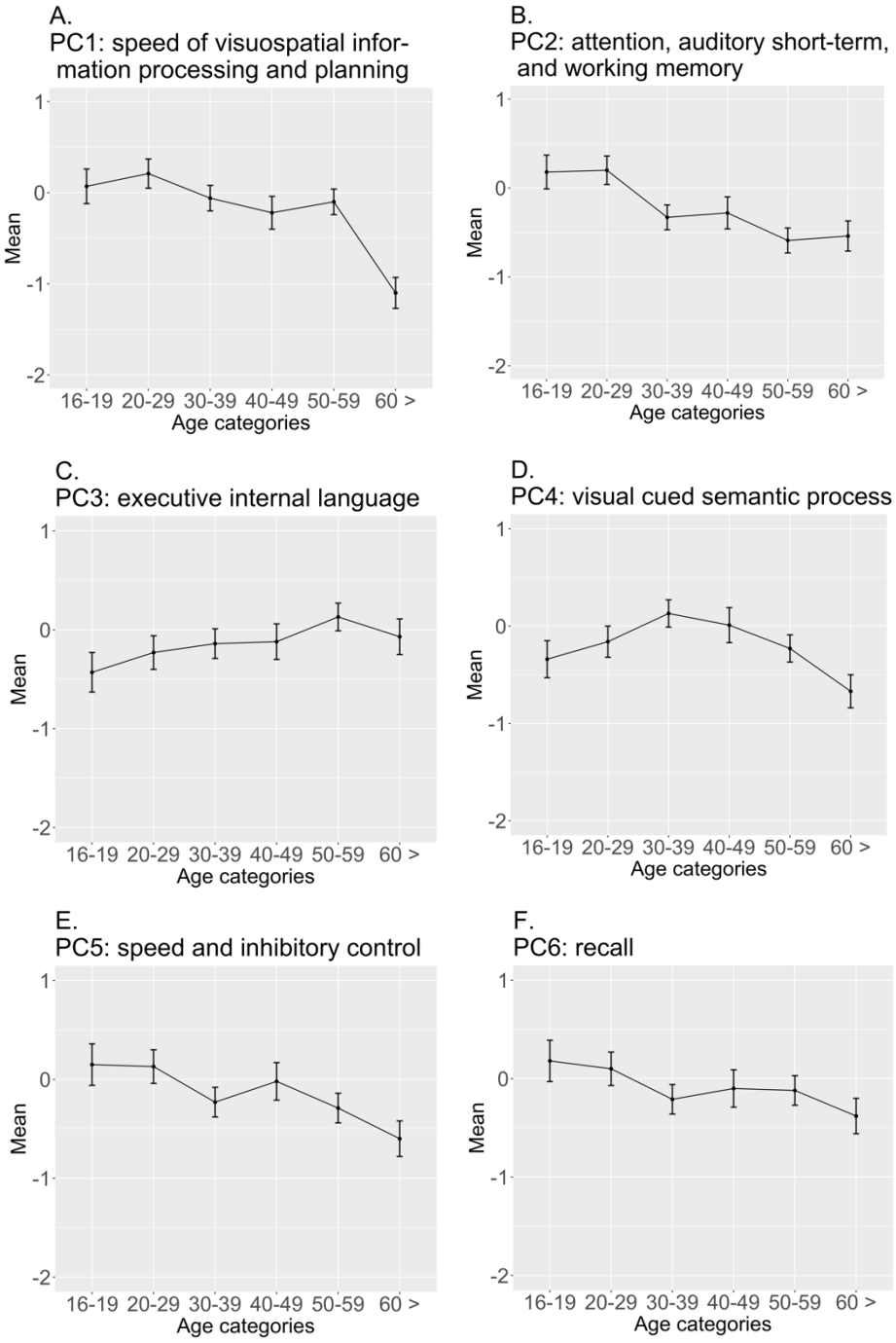
Table 5. The outcome of ANCOVA for age and education and effect sizes for seven PC's as well as an outcome on Bonferroni post-hoc test and presence of orthogonal trends.

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

MANOVA

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

Age Effects


A.
PC1: speed of visuospatial information processing and planning
B.
PC2: attention, auditory short-term, and working memory
C.
PC3: executive internal language
D.
PC4: visual cued semantic process
E.
PC5: speed and inhibitory control
F.
PC6: recall
Mean
1
0
-1
-2
16-19 20-29 30-39 40-49 50-59 60 >
Age categories

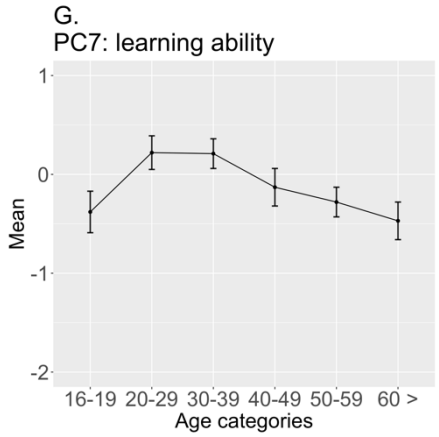


Figure. 1. The mean of the Z-score and standard error for age effects on seven PCs. Four PCs (A, B, E, F) have a significant age effect. Linear trend declines were found for PC 2 (B) and PC 5 (E), and a significant quadratic trend was found for PC 4 (D) and PC 7 (G). Both linear and quadratic trends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

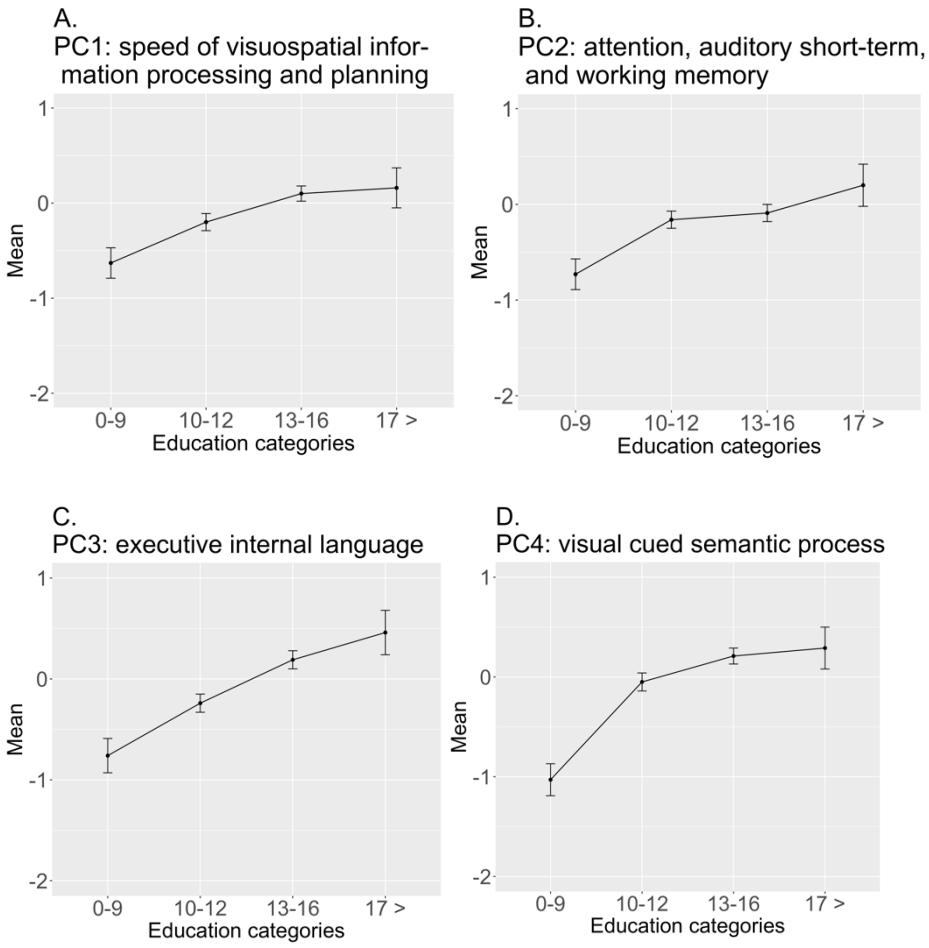
AGE EFFECTS

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

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

PC4 and 20-39 for PC7, followed by an age-dependent decrease with the lowest scores for the oldest group. PC4 represents **visual cued semantic processes,** and PC7 represents **learning ability.** Orthogonal trend analyses confirmed the lack of significant age effects for PC3 (**executive internal language).** PC6 **(recall ability)** showed a sudden early decline at age 30 and a further decline at 50+.

Education Effects



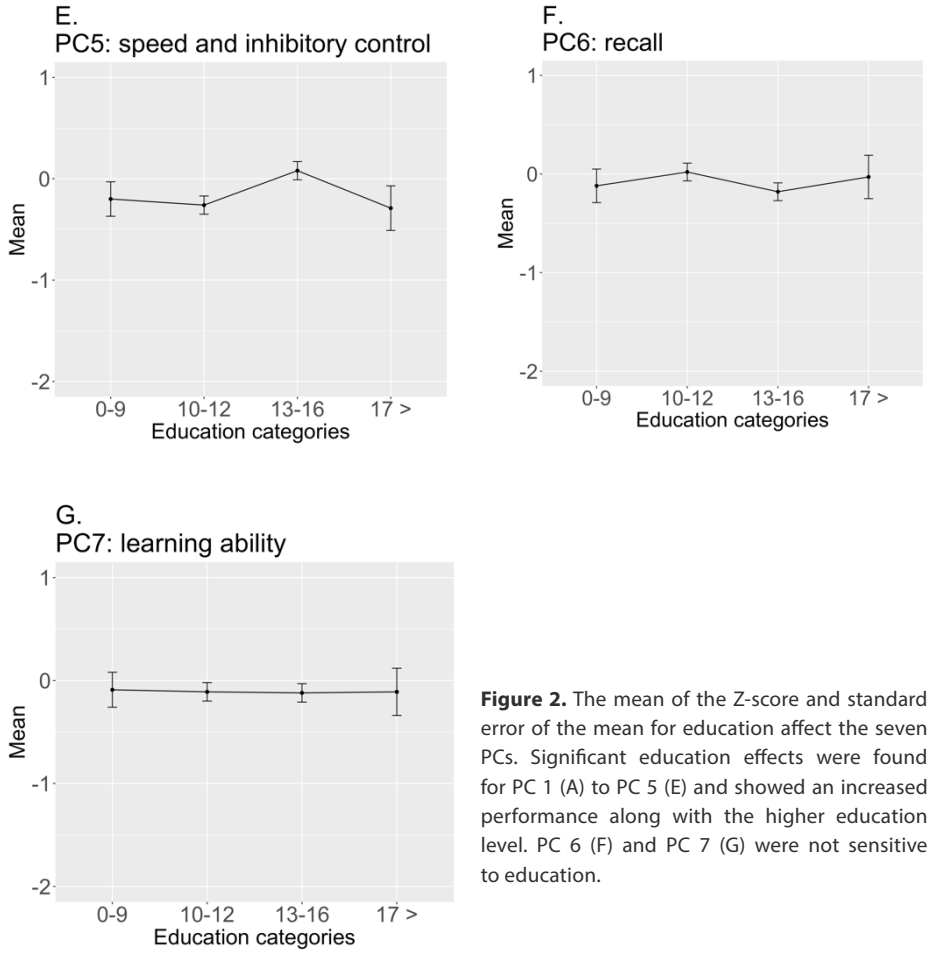


Figure 2. The mean of the Z-score and standard error of the mean for education affect the seven PCs. Significant education effects were found for PC 1 (A) to PC 5 (E) and showed an increased performance along with the higher education level. PC 6 (F) and PC 7 (G) were not sensitive to education.

EDUCATION EFFECTS

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

Regarding PC 1 (**Speed of visuo-spatial information processing and planning**) and PC 2 **(attention**, **auditory short-term, and working memory)**, all education groups outperformed the group with the least years of education (0-9, junior high)

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

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

Discussion

Preliminary normative scores

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

We chose to collect only data in Java because in Java Island lives about 56% of the total population of Indonesia and the increase in urbanization in Java has reached 80% (Sub-directorate of Statistical Demographic, 2013). We assume that the urbanized population in Java Island can be used as a pilot for the development prior to more complete normative scores representing a more comprehensive

geographical range of the Indonesian population. Currently, data are collected on three other islands to increase the representation Indonesia's norms and to analyse possible local differences. Normative scores in the form of mean, median, and SD are the parameters for the success of the adaptation of the various test tools.

The normative data from the tests can be compared with data obtained elsewhere: First, DS, both forward and backward, with a similar wide age range were close to previous studies reported in South Africa, Brazilian, and seven European countries (Ostrosky-Solís & Lozano, 2006; Zimmermann et al., 2015). Our verbal memory scores of the RAVLT were lower for the variable STPR, LTPR, and LOT compared to a study by De Sousa Magalhães et al. (2012). However, their participants were younger (age ranging between 17-40 years) than ours, and considering the well-known age-dependent decrease in memory, this can be expected. Our average score for trials one to five came close to what was reported by De Sousa Magalhães et al. (2012).

The Five Point test was similar in the number of participants and age range as studied by Cattelani et al. (2011) were associated with producing the unique design. For three other tests, the Token test, Boston Naming Test, and Stroop Color Word Test, the means showed comparable and close to other countries' reports (Ktaiche et al., 2021; Troyer et al., 2006). The performances of our sample on the TMT were a bit slower on trial A but almost the same for trial B compared to the results from a Scandinavian sample of 170 participants of 41-84 years (Espenes et al., 2020). We used words starting with the letters K, S, and T for the adapted phonemic word fluency test based on Hendrawan & Hatta (2010). Although the number of correct words is somewhat dependent on the number of words starting with these letters in Bahasa, there were no large differences between the three subtests and the international scores. Also, the FR reproduction was adapted, and this is the first report of its normative score. Our normative scores were rather similar to an earlier study (de Brito-Marques et al., 2012). The test was conducted in Brazil and reported the correlation between the score of visual reproduction for normal older adults and their education level. The mean row time in seconds of the Bourdon Wiersma test was also comparable with normative scores in a Dutch population with the Indonesian sample was about 2 seconds per row faster. In contrast, Indonesians made more errors than the Dutch (https://andi.nl/tests/aandacht-en-werkgeheugen/bw/) and this might reflect a speed-accuracy trade-off. Finally, our test score from our sample comes close to what is internationally reported. However, it should be kept in mind that a detailed comparison between our outcomes and what is internationally reported is less meaningful considering differences in the assessment's circumstances and demographic factors. Moreover,

language performance in Bahasa Indonesia (for many Indonesians, Bahasa Indonesia is their second language) and cultural and socioeconomic factors might also contribute to differences between scores of other countries and our sample.

Reliability of the test battery

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

Cognitive constructs and age and education effects

Principal Component Analysis was used as an exploratory factor analysis capable of inferring fundamental cognitive construct functions, while the result does not depend on certain assumptions. We identified seven constructs, some contained a mixture of different variables from different tests, and others were determined by the outcomes of a single test. We identified, among others, speed of visual-spatial information processing, visual cued semantic process, verbal recall, verbal learning, attention/working memory, executive internal language, and inhibitory control, most of which are commonly acknowledged cognitive constructs. The outcomes of confirmatory factor analysis demonstrated the goodness of fit of the seven-factor model. These outcomes contribute to the construct validity of the INTB as a group of tests measuring different aspects of cognition. Earlier, two studies of the Montreal Cognitive Assessment (MoCA) test, covering eight different cognitive domains, explored the underlying factor from a different subset of MoCA items. The earlier study yielded five factors (memory, attention/processing speed, visuospatial, language, and executive function), and the later study yielded four factors (visuospatial/executive function, memory, attention, and language) (Moafmashhadi & Koski, 2013; Vogel et al., 2015). Some of our results resemble those reported by these authors with one obvious difference regarded the language tests. In INTB,

we used three language tests that measure different aspects. The PCA showed that the three language tests loaded on two different constructs: VF loaded on the executive internal language function. In contrast, the BNT and Token loaded on a construct named visual cued semantic process together with FR. Interestingly, also the TMT-time B loaded in this factor. In all these four tests, the presentation of visual stimuli is followed by a motor act (drawing, speaking, and pointing). The distinction between executive internal language and the visual cued semantic process is more often found in the psycho-linguistic literature; differences between semantic and executive aspects and deficits of language function have been found in different patient categories (Dick et al., 2001).

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

Our constructs also fitted partly in the Cattel-Horn-Carrol (CHC) model: a model trying to encompass the theoretical structure of intelligence as broad domains of cognition (van Rentergem et al., 2020). The current CHC taxonomy distinguishes four conceptual groupings (i.e., motor abilities, perceptual processing, controlled attention, and acquired knowledge) on two levels (i.e., speed and level). Our PC1 belongs to the theoretical CHC constructs "perceptual processing and motor abilities". Our PC2 and PC5 to CHC construct "controlled attention" with PC2 narrower to working memory. Three of our PCs, PC3, PC6, and PC7, might have fit in the broad cognitive abilities associated with "acquired knowledge" and seem also close to long-term storage and retrieval fluency, while PC7 may reflect also "learning efficiently". PC4 is associated with the three broad abilities "acquired knowledge, perceptual processing, and controlled attention". However, more quantitative analyses are necessary to see how well the data from the I-NTB fit into the CHC model.

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

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

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

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

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

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

In addition to age and education as main factors, an interaction of age x education effects on the PC measuring “attention, auditory short term, and working memory” was found. As revealed by post-hoc tests, it was found that those with a senior high education (also for those with only junior high) showed a sharp decrease in

attention, auditory short-term, and working memory from 30 years onward, while this was not the case for the undergraduates; their decline is non-significant, and if it happened it started later. This emphasizes the relevance of education in the prevention of cognitive decline.

Limitations

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

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

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

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

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

Conclusions

We conclude that all test scores were in good agreement with what is internationally reported. The psychometric analyses showed objectives concerning reliability and validity of the tests in the INTB are considered promising. Interestingly, as

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

Acknowledgment: This work was supported by the Directorate of Higher Education General of Indonesia with number 0317/AK.04/2022. We thank Prof. Dr. R.P.C. Kessels, Dr. Vitoria Piai, Dr. Loes van Aken, and Dr. J.M. Oosterman for the valuable feedback and discussion.

References

Adioetomo, S. M., & Ghazy, M. (2014). Indonesia on the Threshold of Population Ageing. *UNFPA Indonesia Monograph Series*: No.1.

Agelink Van Rentergem, J. A., De Vent, N. R., Schmand, B. A., Murre, J. M. J., Staaks, J. P. C., Huizenga, H. M., & ANDI Consortium. (2020). The Factor Structure of Cognitive Functioning in Cognitively Healthy Participants: A Meta-Analysis and Meta-Analysis of Individual Participant Data. *Neuropsychology Review*, *30*(1), 51–96. https://doi.org/10.1007/s11065-019-09423-6

Ananta, A., Arifin, E. N., Hasbullah, M. S., Budi Handayani, N., Pramono, A., & Handayani, N. B. (2015). Demography of Indonesia's ethnicity. ISEAS, *Institute of Southeast Asian Studies*.

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

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

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

Bridges, A. J., & Holler, K. A. (2007). How Many is Enough? Determining Optimal Sample Sizes for Normative Studies in Pediatric Neuropsychology. *Child Neuropsychology*, *13*(6), 528–538. https://doi.org/10.1080/09297040701233875

Brito-Marques, P. R. D., Cabral-Filho, J. E., & Miranda, R. M. (2012). Visual reproduction test in normal elderly: Influence of schooling and visual task complexity. *Dementia & Neuropsychologia*, *6*(2), 91–96. https://doi.org/10.1590/S1980-57642012DN06020005

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

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *Routledge Academic*.

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

De Renzi, E., & Vignolo, L. A. (1962). The Token Test: a Sensitive Test to Detect Receptive Disturbance in Aphasics. *Brain*, *85*(4), 665–678. https://doi.org/10.1093/brain/85.4.665

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

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

Elkana, O., Eisikovits, O. R., Oren, N., Betzale, V., Giladi, N., & Ash, E. L. (2015). Sensitivity of Neuropsychological Tests to Identify Cognitive Decline in Highly Educated Elderly Individuals: 12 Months Follow up. *Journal of Alzheimer's Disease*, *49*(3), 607–616. https://doi.org/10.3233/JAD-150562

Espenes, J., Hessen, E., Eliassen, I. V., Waterloo, K., Eckerström, M., Sando, S. B., Timón, S., Wallin, A., Fladby, T., & Kirsebom, B.-E. (2020). Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years. *The Clinical Neuropsychologist*, *34*(sup1), 110–126. https://doi.org/10.1080/13854046.2020.1829068

Fong, M. W. M., Van Patten, R., & Fucetola, R. P. (2019). The Factor Structure of the Boston Diagnostic Aphasia Examination, Third Edition. *Journal of the International Neuropsychological Society*, *25*(7), 772–776. https://doi.org/10.1017/S1355617719000237

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Geerinck, A., Alekna, V., Beaudart, C., Bautmans, I., Cooper, C., De Souza Orlandi, F., Konstantynowicz, J., Montero-Errasquín, B., Topinková, E., Tsekoura, M., Reginster, J.-Y., & Bruyère, O. (2019). Standard error of measurement and smallest detectable change of the Sarcopenia Quality of Life (SarQoL) questionnaire: An analysis of subjects from 9 validation studies. *PLOS ONE*, *14*(4), e0216065. https://doi.org/10.1371/journal.pone.0216065

Geffen, G. M., Butterworth, & Geffen, L. B. (1994). Test-retest reliability of a new form of the auditory verbal learning test (AVLT). *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, *9*(4), 303–316.

Glisky, E. (2007). Changes in Cognitive Function in Human Aging. In D. Riddle (Ed.), *Brain Aging* (Vol. 20072731, pp. 3–20). CRC Press. https://doi.org/10.1201/9781420005523.sec1

Guerra-Carrillo, B., Katovich, K., & Bunge, S. A. (2017). Does higher education hone cognitive functioning and learning efficacy? Findings from a large and diverse sample. *PLOS ONE*, *12*(8), e0182276. https://doi.org/10.1371/journal.pone.0182276

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

Hendrawan, D., & Hatta, T. (2010). Evaluation of Stimuli for Development of the Indonesian Version of Verbal Fluency Task Using Ranking Method. *Psychologia*, *53*(1), 14–26. https://doi.org/10.2117/psysoc.2010.14

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

Jansen, M. G., Geerligs, L., Claassen, J. A. H. R., Overdorp, E. J., Brazil, I. A., Kessels, R. P. C., & Oosterman, J. M. (2021). Positive Effects of Education on Cognitive Functioning Depend on Clinical Status and Neuropathological Severity. *Frontiers in Human Neuroscience*, *15*, 723728. https://doi.org/10.3389/fnhum.2021.723728

Kaplan, E., Goodglass, H., & Weintraub, S. (2016). *Boston Naming Test* [Dataset]. https://doi.org/10.1037/t27208-000

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

Kessels, R. P. C., & Hendriks, M. P. H. (2023). Neuropsychological assessment. In *Encyclopedia of Mental Health* (pp. 622–628). Elsevier. https://doi.org/10.1016/B978-0-323-91497-0.00017-5

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

Ktaiche, M., Fares, Y., & Abou-Abbas, L. (2021). Stroop color and word test (SCWT): Normative data for the Lebanese adult population. *Applied Neuropsychology: Adult*, *29*(6), 1578–1586. https://doi.org/10.1080/23279095.2021.1901101

Lamar, M., Swenson, R., Penney, D., Hospital, L., & Libon, D. (2018). Encyclopedia of Clinical Neuropsychology. *Encyclopedia of Clinical Neuropsychology*.

Lezak, M. D. (Ed.). (2004). Neuropsychological assessment (4th ed). *Oxford University Press*.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). Neuropsychological assessment (Fifth edition). *Oxford University Press*.

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

Mengual-Macenlle, N., Marcos, P. J., Golpe, R., & González-Rivas, D. (2015). Multivariate analysis in thoracic research. *Journal of Thoracic Disease*, *7*(3), E2–E6.

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

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

Murman, D. (2015). The Impact of Age on Cognition. *Seminars in Hearing*, *36*(03), 111–121. https://doi.org/10.1055/s-0035-1555115

Nielsen, T. R., Segers, K., Vanderaspoilden, V., Bekkhus-Wetterberg, P., Minthon, L., Pissiota, A., Bjørkløf, G. H., Beinhoff, U., Tsolaki, M., Gkioka, M., & Waldemar, G. (2018). Performance of middle-aged and elderly European minority and majority populations on a Cross-Cultural Neuropsychological Test Battery (CNTB). *The Clinical Neuropsychologist*, *32*(8), 1411–1430. https://doi.org/10.1080/13854046.2018.1430256

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

Palta, M., Chen, H.-Y., Kaplan, R. M., Feeny, D., Cherepanov, D., & Fryback, D. G. (2011). Standard Error of Measurement of 5 Health Utility Indexes across the Range of Health for Use in Estimating Reliability and Responsiveness. *Medical Decision Making*, *31*(2), 260–269. https://doi.org/10.1177/0272989X10380925

Pena-Casanova, J., Blesa, R., Aguilar, M., Gramunt-Fombuena, N., Gomez-Anson, B., Oliva, R., Molinuevo, J. L., Robles, A., Barquero, M. S., Antunez, C., Martinez-Parra, C., Frank-Garcia, A., Fernandez, M., Alfonso, V., Sol, J. M., & for the NEURONORMA Study Team. (2009). Spanish Multicenter Normative Studies (NEURONORMA Project): Methods and Sample Characteristics. *Archives of Clinical Neuropsychology*, *24*(4), 307–319. https://doi.org/10.1093/arclin/acp027

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

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

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

Santos, N. C., Costa, P. S., Amorim, L., Moreira, P. S., Cunha, P., Cotter, J., & Sousa, N. (2015). Exploring the Factor Structure of Neurocognitive Measures in Older Individuals. *PLOS ONE*, *10*(4), e0124229. https://doi.org/10.1371/journal.pone.0124229

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

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

Stroop, J. R. (1935). Studies of Interference in Serial Verbal Reactions. Journal of Experimental Psychology, 18, 643-662. https://doi.org/10.1037/h0054651

https://www.scribd.com/document/49683284/Stroop-Stroop-1935

Sulastri, A., Utami, M. S. S., Jongsma, M., Hendriks, M., & van Luijtelaar, G. (2018). The Indonesian Boston Naming Test: Normative Data among Healthy Adults and Effects of Age and Education on Naming Ability. *International Journal of Science and Research (IJSR)*, *8*(11), 134-139.

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*, *05*(14), 1611–1619. https://doi.org/10.4236/psych.2014.514171

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

Tucha, L., Aschenbrenner, S., Koerts, J., & Lange, K. W. (2012). The Five-Point Test: Reliability, Validity and Normative Data for Children and Adults. *PLOS ONE*, *7*(9), e46080. https://doi.org/10.1371/journal.pone.0046080

Tucker-Drob, E. M., Brandmaier, A. M., & Lindenberger, U. (2019). Coupled cognitive changes in adulthood: A meta-analysis. *Psychological Bulletin*, *145*(3), 273–301. https://doi.org/10.1037/bul0000179

Utami, M. S. S., Sulastri, A., Santoso, J., Suryani, A., Goeritno, H., Widhianingtanti, L., & Van Luijtelaar, G. (2024). Adaptation of Rey Auditory Verbal Learning Test for Indonesia: Its Validity and Reliability. *Acta Neuropsychologica*, *22*(1), 15–34. https://doi.org/10.5604/01.3001.0053.9735

Vogel, S. J., Banks, S. J., Cummings, J. L., & Miller, J. B. (2015). Concordance of the Montreal cognitive assessment with standard neuropsychological measures. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(3), 289–294. https://doi.org/10.1016/j.dadm.2015.05.002

Wahyuningrum, S. E., Sulastri, A., Hendriks, M. P. H., & van Luijtelaar, G. (2022). The Indonesian Neuropsychological Test Battery (INTB): Psychometric Properties, Preliminary Normative Scores, The Underlying Cognitive Constructs, and the Effects of Age and Education. *Acta Neuropsychologica*, *20*(4), 445–470. https://doi.org/10.5604/01.3001.0016.1339

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

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

Zimmermann, N., Cardoso, C. D. O., Trentini, C. M., Grassi-Oliveira, R., & Fonseca, R. P. (2015). Brazilian preliminary norms and investigation of age and education effects on the Modified Wisconsin Card Sorting Test, Stroop Color and Word test and Digit Span test in adults. *Dementia & Neuropsychologia*, *9*(2), 120–127. https://doi.org/10.1590/1980-57642015DN92000006

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

Chapter 4

A Computer Vision System for an Automated Scoring of a Hand-drawn Geometric Figure

Published as Wahyuningrum, S. E., van Luijtelaar, G., Sulastri, A., Hendriks, M. P. H., Sanjaya, R., & Heskes, T. (2024). A Computer Vision System for an Automated Scoring of a Hand-drawn Geometric Figure. Sage Open, 14(4). https://doi.org/10.1177/21582440241294142

Abstract

Visual Reproduction is a condition to measure Visual Spatial Memory as one of the cognitive domains commonly used to measure visuo-spatial memory. Geometric figures serve as stimulus material, and probands have to reproduce the figures from memory through a hand drawing. The scoring of the drawing has subjective elements. This study aims to evaluate the scoring criteria for the Figural Reproduction Test (FRT), part of the Indonesian Neuropsychological Test Battery, and to develop and evaluate an automated scoring system based on computer vision technology (FRT-CVAS). Scoring evaluation conducted by Cohen Kappa analysis, accuracy, sensitivity, and specificity. The analyses of the three criteria of the manual confirmed a subjective element in the scoring of the shape of triangles by a moderate (.74) inter-rater agreement; this agreement could be improved to .84 by a slight modification of its criteria. FRT-CVAS, based on computer vision's identification of the different elements of the hand drawing, was developed and trained using 290 drawings. The system was additionally tested by comparing its scoring with the scoring of two independent raters on 120 drawings from a second data set. FRT-CVAS recognized all elements, and its comparison between human raters showed a high accuracy and sensitivity (minimally .91), while the specificity was .80 for one of the three criteria. FRT-CVAS offers a highly standardized, consistent, precise, and objective output from the first card in the FRT. This approach is advantageous to data-hungry alternatives such as deep learning when applied to the automated scoring of hand drawings with relatively little data available for training.

Keywords: angle size; computer vision; FRT-CVAS; geometrical figures; triangle shape.

Introduction

Visually presented stimuli are indispensable to any cognitive test battery. Moetesum et al. (2022) reviewed and highlighted the two most commonly used techniques in visual neuropsychological assessment: visual analysis and procedural analysis. The visual analysis assesses visual functions such as object recognition, visual memory, and visual attention. At the same time, procedural analysis evaluates an individual's ability to perform tasks that require the integration of visual perception and motor coordination.

In the visual-spatial test, the participants were tasked with producing a copy of a figure stimulus by hand drawing. The hand drawing scoring method involves quantifying the precision of each component and its spatial alignment reflecting the degree to which the drawn image matches the original design presented (Zhang et al., 2021). According to Awad et al. (2004), the scoring system of hand drawing tends to contribute to the subjective nature of the rater. They suggest that employing a simple, objective, and explicit scoring system can mitigate this issue. They found that explicitly defining accuracy and placement separately for each element in Rey-Osterrieth Complex Figure's drawing (ROCF; Rey, 1941; Osterrieth, 1944) reduces the flexibility for the scorer in determining what constitutes accurate reproduction. Moreover, adjustment of scoring criteria is necessary when transitioning manual tests to computerized formats (Kim et al., 2020).

Technology trends in recent decades have significantly impacted the automated scoring of drawings and hand-writings by more precise and efficient assessment techniques (Pereira et al., 2015). Examples of computerized assessments on hand-drawing and hand-writing are the Clock Drawing Test, ROCF, and the Bender-Gestalt Test (Canham et al., 2000, 2005; Chen et al., 2020; Pereira et al., 2015; Vogt et al., 2019), handwriting segmentation (Vessio, 2019), handwriting classification (Taleb et al., 2019), and spiral drawing. The key to an automated scoring system was algorithms used in system development. Those algorithms were based on geometric template matching, fuzzy logic, statistics (Histogram, Local Binary Pattern), Convolution Neural Networks (CNN), Support Vector Machines (SVM), and the Decision Tree with accuracies ranging from 63% to 99%.

An automated scoring system standardizes scoring, removes all rater variability (Canham et al., 2005; Diaz-Orueta et al., 2022), and it does not fatigue. It also reduces processing time and relieves psychologists and expert physicians of tedious tasks. Vogt et al. (2019) developed a method of scoring hand drawings that used a deep

learning approach, and they then compared this system's judgments with those of six expert raters. They obtained a high (.88) correlation between automatic and manual scoring methods. They suggested adding segment detection to the automated system to achieve an accuracy equal to that of human raters.

The Indonesia Neuropsychology Consortium adapted a visual memory test (visual reproduction test) from the Wechsler Memory Scale (WMS) and named it the Figural Reproduction Test (FRT). It is part of the Indonesian Neuropsychological Test Battery (INTB), which includes nine other cognitive tests (Sulastri et al., 2018; Wahyuningrum et al., 2022). The FRT consists of three cards with different geometric designs. Figure 1 presents the sample of hand-drawing responses from the first of the three stimulus cards. As can be seen, the large variation in the drawings by different subjects leaves room for subjective scorer judgment. Langer et al., (2022) mentioned that the diagnostic effectiveness of this test is limited by a lack of clear scoring criteria, vagueness as to the extent of deviation from the standard that is allowable, and differences among scorers in their interpretation of complex criteria. Diaz-Orueta et al. (2022) reviewed the visual reproduction test (as a subset of the WMS/Wechsler Memory Scale) and noted that the WMS had been revised several times with regard to the scoring criteria (WMS-R), the stimulus material (WMS-III), and the scoring procedure (WMS-IV) thus demonstrating a perceived need for improved precision in the scoring of this test.

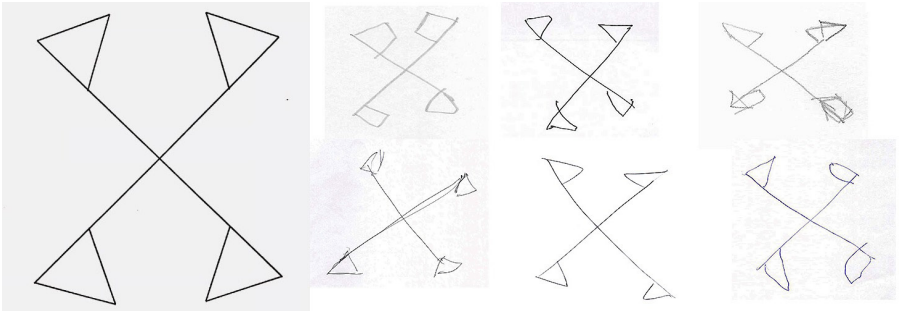


Figure 1. The left side is the original first card. The right side illustrates some scanned hand-drawn responses.

The present study aims to evaluate the scoring criteria provided in the scorer's manual of the Figural Reproduction Test (Wahyuningrum et al., 2022). Next, we seek to develop an automated scoring system for the first figure of the FRT. This new system we named the FRT-CVAS (Figural Reproduction Test Computer Vision Automated Scoring). It relies heavily upon state-of-the-art computer vision

techniques to identify various features of the drawings, particularly the triangle's shape (the flags), the orientation of the four triangles, the position of the various elements, the crossing lines, and their angles. Finally, we evaluate the performance of FRT-CVAS in terms of its specificity, sensitivity, and accuracy, and we compare its performance with that of human raters, including the computer's performance when two raters disagreed.

Methods

The Figural Reproduction Test: Procedure

The FRT data was collected as part of a larger study in which other cognitive tests were also administered under the INTB data collection protocol. We used the drawings of Card 1 (Figure 1) of the FRT as performed by 410 healthy Indonesian participants called image responses. The age-range of the participants was 16 - 80 years old and their years of education varied from 6 to 22. The drawings were scanned manually to obtain the digital images. All scanned images were randomly separated into two parts: a training dataset (n = 290) and an evaluation dataset (n = 120). The first dataset was used to develop FRT-CVAS for an initial evaluation, the second dataset was used to compare the performance of FRT-CVAS with that of two independent raters.

The scores of the FRT were determined by investigating the object's presence, orientation, size, and the shapes of the different elements in the hand drawing. A previous study reported the normative scores obtained with the Indonesian version of FRT, and the results agreed with what is internationally reported (Wahyuningrum et al., 2022). The FRT consists of different geometric stimuli presented on three cards. The test requires the participant to reproduce the figures one by one on a separate blank sheet of paper after the tester has shown each card for 10 seconds. After that, the tester manually scores the participant responses according to the test manual's instructions.

Manual Scoring Rules

We used intra-rater and inter-rater reliability for the evaluation of the scoring criteria. Two senior clinical psychologists with three to four years of experience in the practice of neuropsychological assessment independently graded one hundred and twenty hand drawings. Raters were first asked to grade the responses using the scoring criteria provided in the manual. Three weeks later, they again rated the drawings according to a slight change: the three compound criteria were now

subdivided into six single criteria. It was hypothesized that the scoring accuracy would improve if the raters used six single criteria (Serra, 1986) instead of three compound criteria.

We hypothesize that there is a lack of manual scoring instruction when combining two detail criteria in one parameter. We propose a modified scoring to make the result more accurate. We provide the different within the original and modified scoring criteria in Table 1.

Table 1. Two ways of scoring: the left column contains the compound criteria, and the right column the single scoring criteria.

Criteria	First round (Compound criteria scoring)	Second round (Single criteria scoring)	
1.	Two crossing lines and four flags	Two crossing lines	b. Four flags
2.	Two flags facing each other above and below	Two flags facing each other above	Two flags facing each other below
3.	Angles of crossing line (60° to 120°) and the four flags should have the shape of a triangle.	Angles of crossing line (60° to 120°)	Four flags should be in triangle shape

The Figural Reproduction Test: Computer Vision Automated Scoring

System development: FRT-CVAS was developed using Python 3 and OpenCV, two programming languages widely used to develop artificial intelligence applications. Specifically, OpenCV is a cross-platform library primarily focusing on image processing. Python was employed to extract image elements and automate the grading process (Hyun et al., 2018). Both programs were used to identify and judge the correctness of the image elements - the two crossing lines, their angles, the presence and direction of the four pointing flags as well as their proper shape, and to calculate the total score. Figure 2 presents the steps taken by FRT-CVAS.

Pre-Processing: Images of hand drawings vary in many aspects (such as line thickness, brightness, and size), and they may sometimes contain misplaced scores written by a research assistant. To standardize the varying line thickness present in the different hand drawings, algorithms developed from mathematical morphology were used: an erosion procedure when the lines were too thick and a dilation procedure when the lines were too thin (Guoquan et al., 2008). Furthermore, the Gaussian Filter Algorithm (Gong et al., 2018) was used to remove elements present on the drawing that were not part of the respondent's original drawing, such as a score or comment written by the research assistant.

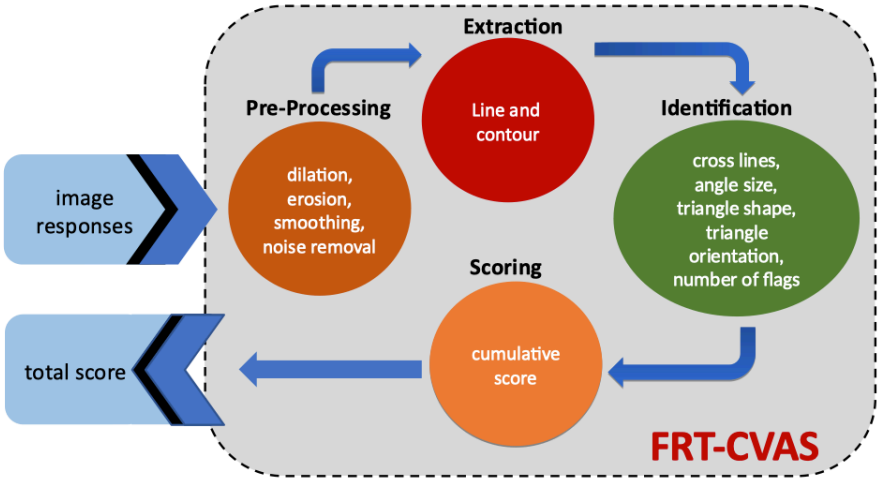


Figure 2. Flow diagram of the automated scoring process (FRT-CVAS).

Line extraction: This extraction was used to identify the two crossing lines. The Hough Transform algorithm determined straight lines by selecting the candidate lines (Sim & Wright, 2005). A median formula was used to determine the desired line. Then, the automated system identified two straight lines that cross. Next, the intersection point was assessed. This intersection point was used to determine the angles of the intersecting lines. In the scoring criteria, the angle had to be equal to or greater than 60 degrees and less than or equal to 120 degrees.

Contour Extraction: Flag extraction was conducted by dividing the image into four parts and extracting each part using a contour detection algorithm. This algorithm detected the borders of every object of interest as a representative feature (Awad et al., 2004). The individual flags had to be located at the end of the crossing lines, at either the top left, top right, bottom left, or bottom right.

Shape Identification and Orientation: The extracted contour is a set of polygon areas throughout the image. We needed to extract the shapes of the four triangle flags by calculating the number of sides from the polygon using the approxPolyDP library (www.opencv.org). Finally, the orientation of the flags was determined by identifying the moment of the polygon relative to the flag's line. Figure 3 illustrated the result of the unit processing.

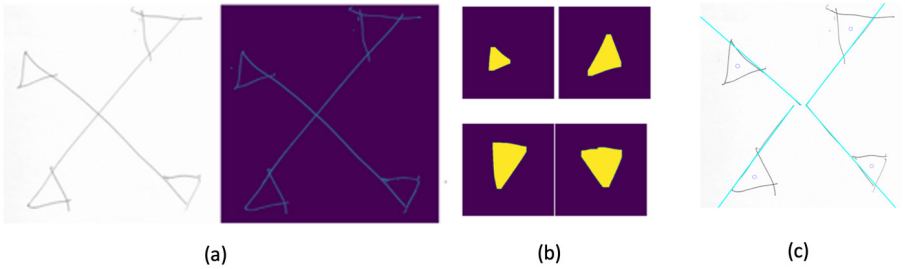


Figure 3. Sample of the FRT-CVAS process. (a) remove noise and line extraction, (b) Contour detection, and (c) shape position and orientation

Statistical analysis

Evaluation was conducted in three different parts. In the first evaluation training data was to revealed the performance of automated scoring system. Second the FRT score from two raters was evaluated. In the third evaluation the two different scoring (manual and automatic scoring) were compared.

The computer performance was evaluated in several ways. The training dataset was first evaluated by determining, for each element, its sensitivity (True Positives / (True Positives + False Negatives)), specificity (True Negatives / (True Negatives + False Positives)), and accuracy ((True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)). Those calculation was also used for the evaluation dataset to assess FRT-CVAS on an independent dataset that had not been previously used. True Positive means that an element was identified correctly by a single rater (training dataset) and by two different raters' (test dataset) and the computer. Conversely, True Negative means that both rater(s) and the computer agreed and identified an element as false. False Negatives occurred when the computer recognized an element as incorrect, but the raters considered it correct. In contrast, False Positives occurred when the computer identified an element correctly, but the rater scored it incorrectly.

The scoring of the different elements of the drawing yielded binary data: zero for a missing or incorrect item and one for a correct item. Therefore, Cohen's Kappa was used to investigate the intra-rater reliability for the two scoring methods (compound vs. single criteria). Next, the inter-rater reliability for both ways of scoring were assessed, as well as the agreement between FRT-CVAS and human raters. The benchmarks of Cohen's Kappa used in this study were: < 0.20 = no agreement; 0.21 - 0.39 = minimal; 0.40 - 0.59 = weak; 0.60-0.79 = moderate, 0.80-0.90 = strong, and >.90 = almost perfect agreement.

The scores of the items agreed upon by the two raters agreed on were used to score the FRT-CVAS. For the occasions in which raters disagreed, a new rater was recruited to score these instances of disagreement, and his or her judgment was considered decisive.

Results

The results section is divided into three parts: the first part contains the analysis of how the neuropsychologist scored manually, using intra- and inter-rater reliability indices. In the second part, FRT-CVAS is described and its preliminary evaluation is provided using the training dataset (N=290). In the third part, FRT-CVAS is evaluated against the judgment of two independent raters and on a different dataset (N=120).

Manual Scoring Evaluation

Two neuropsychologist scored the 120 hand drawings twice, with an intervals of three weeks. The intra-rater reliability, expressed by Cohen's Kappa, determined the degree of agreement for each rater between the first and second rounds, and the inter-rater reliability coefficients measured the degree of agreement between the two raters separately for rounds 1 and 2.

Intra-rater reliability: Cohen's Kappa for the total score for rater 1 was = .75, $p < .01$, with an inconsistency of 11.65%. For rater 2, it was =.86, $p < .01$, with an inconsistency of 7.5%. The analyses of each of the criteria showed excellent agreement (>.90) for both criterion 1 (with the combined score of 1a and 1b) and criterion 2 (with the combined score of 2a and 2b, as shown in Table 1). Only moderate agreement with K=.66, $p < .01$ for criterion 3 was found for the combined score of 3a and 3b.

Inter-rater reliability: Table 2 presents the result of Cohen's Kappa analysis showing the degree to which the two raters agreed in their scoring. In round one, when compound criteria were used, the raters were in excellent agreement (K above .90 for criteria 1 and 2). In contrast, the agreement for criterion 3 was weak (K = .46). Consequently, the agreement of the total scores were only moderate agreement of total scores (K = .74, moderate). In the second round, using single criteria, the K value showed almost perfect agreement for criterion 2a (K = .93) as well as for criterion 2b (K = .89). In contrast, very low agreement was found for criterion 3a (K = -.01). This negative value indicates that agreement between two raters was even less than expected by chance (Sim & Wright, 2005). The low agreement score is indicating a far greater subjective element in the scoring of this criterion than

of any other. On the other hand for criterion 3b, there was a moderate agreement (K = .65). The agreement on the total score was even stronger (K = .84).

Table 2. Inter-rater correlation coefficients (ICC) between raters on compound (Round 1) and single (Round 2) criterion scoring, p-value, F, and Confidence Interval 95% between raters for criterion two, criterion three, and total score.

Between raters	ICC	P value	F ($_{119,119}$)	Lower bound	Upper bound
			Round 1		
Compound 2	.921	<.001	24.40	.889	.944
Compound 3	.467	<.001	2.75	.314	.596
			Round 2		
Single 2a	.928	<.001	26.60	.898	.949
Single 2b	.892	<.001	17.53	.849	.924
Single 3a	-.011	.549	.98	-.189	.168
Single 3b	.657	<.001	4.83	.542	.748
Total score Card 1	.754	<.001	7.15	.666	.822

Note: the scoring of drawings using compound 1 and single criteria 1a and 1b has no variance since both raters fully agreed.

FRT-CVAS Scoring Evaluation

The automatic scoring system was developed, optimized, and given a preliminary evaluated using 290 hand drawings. It was once more independently tested using 120 hand drawings that had not been previously used.

Despite high variability in the drawings of the training dataset (see Figure 1), FRT-CVAS resolved this problem well and achieved high accuracy. Further, Hough Transform and median methods were used to determine the correct crossing lines. Next, specific elements of figures (the four flags or triangles) were identified by choosing the four biggest areas and ignoring the smaller ones, if present. The area had to have a sufficient surface to be regarded as a flag. After the flag areas were defined, the system calculated the number of sides from the defined areas.

The results of the evaluation of FRT-CVAS using the training dataset are presented in Table 3. The outcomes of the FRT-CVAS were first compared with those of a single rater, and the data of the training set were used. The scoring by CVAS-FRT was done separately for each of the different elements characterizing the figure. The crossing lines and the presence of the four flags were always correctly identified, with only

true positives, no true negatives and no false detections, indicating a sensitivity and accuracy score of 1, with unidentified specificity.

The next four elements concerned the orientation of the four facing flags. The participants made quite a number of errors regarding this part of the task, and this was indicated by a high number of true negatives for these elements. FRT-CVAS identified correctly almost all incorrect flag orientations. This resulted in sensitivity, specificity, and accuracy scores that were all above .95. The orientation of three of the four flags showed a small number of false negatives, and the top left flag showed somewhat more (13 out of 290 drawings). The angle detection by FRT-CVAS was also excellent, yielding three excellent grades for sensitivity, specificity, and accuracy. FRT-CVAS also revealed good sensitivity and accuracy: both were above .96 on the triangle identification.

Table 3. Element-wise analyses of the training data set. The accuracy and sensitivity of the recognition of all elements are high (> or equal to .95), while the specificity of two elements was undefined, it was relatively low for three of the four triangles shape elements and high for the other elements.

Element	True positive	True negative	False positive	False negative	Sensitivity	Specificity	Accuracy
Crossline	290	0	0	0	1	undefined	1
Four flags	290	0	0	0	1	undefined	1
Top left flag	225	51	1	13	.95	.96	.95
Bottom left flag	225	58	3	4	.98	.95	.98
Top right flag	233	51	3	3	.99	.94	.98
Bottom right flag	229	58	1	2	.99	.98	.99
Angle	283	7	0	1	.99	1	.99
Triangle top left	276	4	5	6	.98	.44	.96
Triangle bottom left	274	4	5	8	.97	.44	.96
Triangle top right	281	6	1	3	.99	.86	.99
Triangle bottom right	272	8	6	5	.98	.57	.96

The specificity on the identification of the triangle was low (between .44 and .56), except for the top right triangle (specificity = .86). Figure 4 provides some examples of the hand-drawing that were misidentified by the CVAS-FRT and which contributed to the low specificity. Figure 4a illustrates the difficulty the system had in identifying the form of a triangle on the bottom side when the triangles were quite close to one another. Figure 4b shows that the computer misidentified the triangle on the left bottom as the correct triangle, and Figures 4c and 4d reveal that the computer misidentified the triangle because of the wide gap between the

lines. In other words, the lines of the triangle did not connect sufficiently. Figures 4e and 4f show that the system may identify a triangle as correct when it was judged to be incorrect by human raters. The system found a closed line which it then misidentified as a triangle.

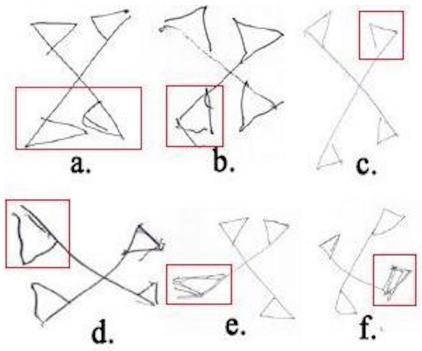


Figure 4(a-f). (a-f) Hand-drawing examples that were misidentified by system

The accuracy of the computerized versus manual scoring

The accuracy of the computerized assessment was also evaluated against the manually assigned scoring of the evaluation data set. Interestingly, most of the participants' errors regarded the proper recall of the orientation of the flags: 22% of the subjects made such errors. We included in this first comparison only element scores in which both raters agreed this concerned criterion 1 of all 120 figures of both scoring rounds.

Table 4 shows the result of the analysis of all image elements. Overall, automated scoring exhibited excellent sensitivity and accuracy. For all image elements, it was between 91 and 100%. Regarding compound criterion one, all responses were correctly scored by FRT-CVAS. Therefore, the sensitivity and accuracy were 1.00, and the value for specificity for compound criterion 1 remained undefined. Similarly, the scoring according to the single criteria 1a and 1b produced no errors and yielded perfect sensitivity and accuracy with undefined specificity. FRT-CVAS was also excellent at determining the true negative responses made by the participants: the score was 99% for all criteria. FRT-CVAS made a few false negatives: this regarded a small number for compound criteria 2 (5) and 3 (9) and single criteria 2a (2), 2b (3), and 3b (7).

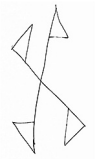
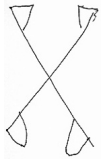
Table 4. Comparison of FRT-CVAS and two ways of human scoring. N is the number of human-rater agreements on scoring. The results revealed high sensitivity, accuracy, and specificity.

Scoring criteria	True positive	True negative	False positive	False negative	Sensitivity	Specificity	Accuracy
Round 1							
Compound 1 (N = 120)	120	0	0	0	1	undefined	1
Compound 2 (N = 116)	77	34	0	5	.92	1	.94
Compound 3 (N = 110)	96	4	1	9	.91	.80	.91
Round 2							
Single 1a (N = 120)	120	0	0	0	1	undefined	1
Single 1b (N = 120)	120	0	0	0	1	undefined	1
Single 2a (N = 117)	90	24	1	2	.98	.96	.98
Single 2b (N = 115)	85	27	0	3	.96	1	.97
Single 3a (N = 117)	117	0	0	0	1	undefined	1
Single 3b (N = 115)	103	4	1	7	.94	.80	.93

Note: N is the number of responses where both raters agreed

The subjectivity of human raters is also illustrated in Table 5. The two raters judged some hand-drawings differently. We found that the raters disagreed on five of the 120 drawings of the evaluation data set regarding the direction of the facing flags (criterion 2) and on seven drawings regarding the correct shape of the flags (criterion 3). More specifically, this regarded three drawings based on criteria 2a and 2b, two on 2b, and one drawing on 3a and 3b, four on only 3b, and two for only 3b. To get an objective judgment of what should be considered as correct, we introduced a third independent judge to rate the figures where there was disagreement. Next, we compared the scoring from CVAS-FRT against the majority score. The automated scoring agreed with this rater in four of the five drawings for criterion 2. Furthermore, regarding criteria three, CVAS-FRT shows benefits in determining the correct angle size of the crossing lines. In all cases its score matched that of the majority of raters. While on the triangle shape identification, the CVAS-FRT score matches the majority of the raters on four out of five drawing responses.

Table 5. FRT-CVAS agreement with three human raters. The automated scoring is consistent with the majority of the raters.

Figures	Rater 1		Rater 2		Rater 3		FRT-CVAS		Number of raters complying with FRT-CVAS
	Angle size	Triangle form	Angle size	Triangle form	Angle size	Triangle form	Angle size	Triangle form	
	0	-	1	-	0	-	0	-	Agree with 2 raters on angle size
	1	1	0	0	1	1	1	1	Agree with 2 raters on angle size and triangle shape
	-	1	-	0	-	1	-	1	Agree with 2 raters on triangle shape
	-	1	-	0	-	1	-	0	Agree with 1 rater on triangle shape
	-	1	-	0	-	1	-	1	Agree with 2 raters on triangle shape
	0	-	1	-	0	-	0	-	Agree with 2 raters on angle size
	-	1	-	0	-	1	-	1	Agree with 2 raters on triangle shape

Discussion

This study evaluated the scoring of a visual-spatial reproduction/memory test with either human or computer scoring. This was done by comparing the scores of the FRT-CVAS with those of a single rater on the training data set and with the shared opinion of two independent raters on the test data set.

Based on two ways of applying scoring criteria, we found that the consistency of the two raters was excellent regarding the presence of elements (i.e. the crossed lines and the flags). Conversely, there was disagreements about whether the two flags at the top and bottom were facing each other, the angular size of the intersecting lines, and the judgement of the shape of the flags. Using an arc to measure the angle of the intersecting lines resulted more likely in inconsistent judgments of line angle. However, this seems rather time-consuming and inefficient (Awad et al., 2004). The identification of the shape of the flags also led to a more general disagreement between raters because the shape of the drawn flags varied not only between drawings but also within a single drawing. This difficulty in judging whether the shape is correct or not was found in a previous study regarding a complex figure test (Langer et al., 2022). The difficulty in its scoring might need to be addressed by having the scorers receive additional practice in scoring geometric figures. Alternatively, as demonstrated, complex drawings like these can be accurately and objectively scored by employing an automated system (Guoquan et al., 2008).

In round two of the study, the utilization of six single criteria, as opposed to three compound criteria, resulted in a discernible enhancement in inter-rater agreement. This observation parallels findings reported by Awad et al. (2004), who conducted a comparative analysis of two scoring methodologies for Taylor's complex figure. The use of six single criteria contributes to reducing human subjectivity. This finding aligns with previous studies in which using single criteria increases the sensitivity and specificity of scoring (Jamus et al., 2023; Troyer & Wishart, 1997; Zhang et al., 2021). In addition, our intra-rater analysis (see Table 3) also shows that using single criteria helps raters score more accurately. Using single criteria in the scoring may also maintain the consistency of the raters when scoring a large number of hand drawings. For the future, we suggest that, as long as the FRT is still administered manually, the raters should consider using single criteria rather than compound criteria to improve the accuracy of their judgements.

Li et al. (2013) defined that identifying units is the pivotal challenge in automating the analysis and scoring process of the ROCF. Apart from contending with

ambiguous drawing features, the necessity to discern individual units and their constituent elements such as line segments, circles, and points adds complexity. Moreover, variations in the number and geometric relations of these components across units further complicate the scoring procedure. The methodologies by which computer vision decomposes complex images into their constituent parts were reported in previous studies (El-gayar et al., 2013; Fleuret et al., 2011). In this study, we extracted the individual elements of the drawing following the suggestion of Vogt et al. (2019), who predicted it would increase the scoring accuracy. Indeed, the accuracy of our system was high, and all individual elements were at least 91% and often close to 95% in accuracy.

The inception of automated scoring methodologies for visual-spatial memory tasks, notably exemplified by the ROCF, dates back over two decades (Canham et al., 2000). Initial endeavours primarily relied on rudimentary feature extraction techniques to discern select components of hand-drawn responses. The computer vision approach has also achieved significant improvement in the automated scoring of the Complex Figure Test (Canham et al., 2005; Gao et al., 2018; Li et al., 2013; Webb et al., 2021) and other figural tests such as the Trail Making Test (Dahmen et al., 2017) and Ruff's Figural Fluency Test (Elderson et al., 2016).

There are two different ways of collecting the hand drawing response, firstly using a digital device such as a tablet and drawing pad, and secondly using a scanner to digitalize the paper hand drawing. In prior investigations conducted by Webb et al. (2021) and (Li et al. (2013), which leveraged digital devices for data collection, participants in those studies were instructed to directly produce drawings using the provided digital interfaces. The adoption of digital devices in data collection by the desire to mitigate noise artifacts and the occurrence of overlapping strokes that challenge handwriting image processing. In practice, many practitioners still employ the paper and pencil method. In this study, we adopt a distinct methodological approach by utilizing scanned representations of hand-drawn responses, using paper and pencil. To address inherent noise artifacts within the scanned images, the Gaussian Filter technique is employed, serving to attenuate extraneous signal fluctuations. Moreover, to disentangle overlapping stroke lines and facilitate accurate angle calculations, a line extraction method is applied, thereby enhancing the fidelity of image analysis outcomes.

Recent advances in computer image recognition techniques have significantly improved medical disease classification and diagnosis. A computer-vision approach to handwriting analysis has enabled the early diagnosis of Parkinson’s disease

(Pereira et al., 2015; Souza et al., 2018). One of the advantages of any automated scoring system is accuracy: in our case, determining the angle of the intersecting lines (Webb et al., 2021). Subsequent advantage, particularly in the field of artificial intelligence (AI), facilitated the emergence of deep learning algorithms tailored for automated scoring applications. Recent studies by Langer et al. (2022) and Park et al. (2023) underscore the utility of deep learning methodologies in assessing memory deficits through automated scoring protocols, extending beyond the ROCF.

The efficacy of deep learning algorithms is contingent upon access to training data, typically comprising thousands of images. This prompts an inquiry into the optimal quantity of data requisite for the engagement of the deep learning algorithms. In their investigation, Langer et al. (2022) posit that once the dataset surpasses thousands of images, deep learning techniques can be effectively utilized for system development. They note that with around 3000 images, there is a decrease in mean MAEs (Mean Absolute Errors), and beyond approximately 10000 images, additional data inclusion does not lead to significant improvements. Another study conducted by Li et al. (2013) emphasized the necessity for caution when employing deep learning methodologies, particularly when confronted with limited sample sizes, and underscored the importance of ensuring congruence between the training data and the target images.

To evaluate the subjectivity among raters, we looked at the drawings in which two raters scored differently than FRT-CVAS. The points of disagreement were all about the evaluation of the correct shape of the triangles. The raters had a different tolerance than the automated system for the number of triangle sides. For instance, a triangle was sometimes accepted as correct when a small fourth side or a curved line had been added to the drawing. The variable shape of the triangles was also responsible for the false negative results provided by the FRT-CVAS: 11.7% and 10%, respectively from the compound and single criteria, respectively. The difference in the scores given by the raters is due to more generous interpretation of the factors required for "triangle-shaped flag" concept. In daily practice, we recommend that if the computer interprets an item as incorrect it should be re-examined by a neuropsychologist, who will then makes a final judgment. This re-examination can be done quickly if the digitized drawing and the scores can be displayed simultaneously. Our computer-rater comparisons showed that this automated system, with its standardized procedures and high agreement with human raters, can help neuropsychologists score more objectively. Kenda et al. (2022) also reported that the problem of inter-rater variability can be eliminated by using a standardized automated method.

The subjective nature of scoring visual-spatial tests is a perceptual organization that is commonly happens in human vision (Nevatia, 2000). Some studies note that human raters, even well trained clinicians, may not be consistent in giving scoring to hand-drawing (Canham et al., 2000; Langer et al., 2022). In this study two raters were initially involved for scoring the hand drawing. Later, one rater was added to provide more votes for determining the final decision. Adding more human raters would help to better estimate the threshold or deviation parameter in determining the shape and then could also improve the computer's accuracy in identifying a triangle. However, the main problem vagueness in the definition of what should be considered as a triangle, is not solved by adding more raters. Therefore, to reduce the difference between raters, we advise that the manual of the test will be adapted to facilitate more detailed and unambiguous criteria.

We expect that the automatic scoring system can be made more precise concerning the scoring of the triangles by involving more human raters in the determination of what is a correct triangle and by comparing the outcome of FRT-CVAS with the shared opinion of these raters. These judging results can then be used as a new training dataset for further development of the automated scoring system, thereby improving its agreement with human judgment.

Limitations of this study are that the hand-drawing responses were produced by self-described "healthy" participants, while the test is designed for application to clinical populations. Therefore, as a next step, we plan to clinically validate the test and its automated scoring in various groups of patients. We did find some drawings that seemed diagnostic of mild or moderate motor problems, but these hand drawings did not seem to have a negative effect on the performance score of the automated system. Another limitation is that we developed an

automated scoring for the first card of the FRT only. However, the principles that we applied from computer vision and the techniques of feature extraction can be extended to the other figures of the FRT and the automated analyses of other hand drawings of geometrical figures.

Conclusion and Outlooks

Human judgments of hand drawn geometrical figure involve subjective elements. In the Figure Reproduction Test (FRT) these subjective factors are particularly noticeable in the judgment of the correct shape of triangles. The accuracy of the scoring can be improved somewhat by using single criteria rather than of the compound criteria suggested in the scorer's manual. FRT-CVAS, which is a computer

vision approach, further removes this subjectivity: by extracting and evaluating all of the hand-drawn elements in detail it achieves a high-level of accuracy, sensitivity, and good specificity. FRT-CVAS produces a more standardized, consistent, precise, and objective result. Given the leniency of the test's scoring instructions with regarding acceptable triangle shapes, the FRT-CVAS cannot be expected to produce a results expert scorers will always agree with. In the meantime, the best course of action is for the individual clinician to check the shape of the triangles for themselves if in serious doubts.

It should be noted that the improvement in scoring which was achieved by using the six single criteria instead of three compound criteria, was based not only one of the three figures of the FRT. If the scoring of the other two even more complex figures of the test had been divided in this way, the difference between the two scoring methods would probably have been greater. This simple innovation would therefore have a beneficial effect on the accurate diagnosis of pathology.

Automated scoring systems offer professionals the advantage of accessing to objective measurements, thereby enhancing the reliability and standardization of assessments. Employing computer vision techniques, even datasets comprising fewer than a thousand training samples of hand drawing can be effectively processed. This methodology proves instrumental in the development of automated scoring mechanisms tailored to newly adapted visual-spatial assessments. Furthermore, the integration of the FRT-CVAS into online platforms facilitates widespread accessibility, enabling neuropsychologists in across different geographical regions to use tools such as the hand-drawing response of the Figural Reproduction Test with greater ease and efficiency.

Concerning future research plans our focus is on developing automated scoring capabilities specifically designed for the remaining cards in the Figural Reproduction Test (FRT). The approach we use in FRT-CVAS is a more of a comprehensive process that includes image segmentation, feature extraction, classification, and grouping of its constituent small image units. This methodology shows promise as a pre-processing tool conducive to subsequent integration with deep learning algorithms. Our future research trajectory covers a wide range, starting with the collecting of hand-drawing results from visual tests such as the Five Point Test, Trail Making Test and FRT not only in healthy groups but also in diverse patient groups. Next, data from various patient categories will be used as training data to improve the effectiveness of our system for early detection and identification of latent neurological pathology using a deep learning paradigm.

Funding: This research was funded by DIKTI (Directorate of Higher Education General of Indonesia, number: 076/E5/PG.02.00.PL/2023)

Acknowledgements: The authors are indebted to Dr. John V. Keller, who helped us with linguistic corrections.

References

Awad, N., Tsiakas, M., Gagnon, M., Mertens, V. B., Hill, E., & Messier, C. (2004). Explicit and Objective Scoring Criteria for the Taylor Complex Figure Test. *Journal of Clinical and Experimental Neuropsychology*, *26*(3), 405–415. https://doi.org/10.1080/13803390490510112

Canham, R. O., Smith, S. L., & Tyrrell, A. M. (2000). Recognition and Grading of Severely Distorted Geometric Shapes from within a Complex Figure. *Pattern Analysis & Applications*, *3*(4), 335–347. https://doi.org/10.1007/s100440070005

Canham, R. O., Smith, S. L., & Tyrrell, A. M. (2005). Location of structural sections from within a highly distorted complex line drawing. *IEE Proceedings - Vision, Image, and Signal Processing*, *152*(6), 741. https://doi.org/10.1049/ip-vis:20045166

Chen, S., Stromer, D., Alabdalrahim, H. A., Schwab, S., Weih, M., & Maier, A. (2020). Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, *10*(1), 20854. https://doi.org/10.1038/s41598-020-74710-9

Dahmen, J., Cook, D., Fellows, R., & Maureen, S.-E. (2017). An analysis of a digital variant of the Trail Making Test using machine learning techniques. *Technology and Health Care*, *25*, 251–264. https://doi.org/10.3233/THC-161274.

Diaz-Orueta, U., Rogers, B. M., Blanco-Campal, A., & Burke, T. (2022). The challenge of neuropsychological assessment of visual/visuo-spatial memory: A critical, historical review, and lessons for the present and future. *Frontiers in Psychology*, *13*, 962025. https://doi.org/10.3389/fpsyg.2022.962025

Elderson, M. F., Pham, S., Van Eersel, M. E. A., LifeLines Cohort Study, Wolffenbuttel, B. H. R., Kok, J., Gansevoort, R. T., Tucha, O., Van Der Klauw, M. M., Slaets, J. P. J., & Izaks, G. J. (2016). Agreement between Computerized and Human Assessment of Performance on the Ruff Figural Fluency Test. *PLOS ONE*, *11*(9), e0163286. https://doi.org/10.1371/journal.pone.0163286

El-gayar, M. M., Soliman, H., & Meky, N. (2013). A comparative study of image low level feature extraction algorithms. *Egyptian Informatics Journal*, *14*(2), 175–181. https://doi.org/10.1016/j.eij.2013.06.003

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, *108*(43), 17621–17625. https://doi.org/10.1073/pnas.1109168108

Gao, J., Yang, Y., Lin, P., & Park, D. S. (2018). Computer Vision in Healthcare Applications. *Journal of Healthcare Engineering*, *2018*, 1–4. https://doi.org/10.1155/2018/5157020

Gong, X.-Y., Su, H., Xu, D., Zhang, Z.-T., Shen, F., & Yang, H.-B. (2018). An Overview of Contour Detection Approaches. *International Journal of Automation and Computing*, *15*(6), 656–672. https://doi.org/10.1007/s11633-018-1117-z

Guoquan Jiang, Xing Ke, Shangfeng Du, & Jiao Chen. (2008). A straight line detection based on randomized method. *2008 9th International Conference on Signal Processing*, 1149–1152. https://doi.org/10.1109/ICOSP.2008.4697333

Hyun, G. J., Park, J. W., Kim, J. H., Min, K. J., Lee, Y. S., Kim, S. M., & Han, D. H. (2018). Visuospatial working memory assessment using a digital tablet in adolescents with attention deficit hyperactivity disorder. *Computer Methods and Programs in Biomedicine*, *157*, 137–143. https://doi.org/10.1016/j.cmpb.2018.01.022

Jamus, D. R., Mäder-Joaquim, M. J., De Paula Souza, L., De Paola, L., Claro-Höpker, C. D., Terra, V. C., & Soares Silvado, C. E. (2023). Rey–Osterrieth complex figure test: Comparison of traditional and

qualitative scoring systems after unilateral temporal lobectomy. *The Clinical Neuropsychologist*, *37*(2), 416–431. https://doi.org/10.1080/13854046.2022.2047790

Jiang, G., Ke, X., Du, S., & Chen, J. (2008, October 26–29). A straight line detection based on randomized method [Confer- ence session]. 9th International Conference on Signal Pro- cessing, Beijing (pp. 1149–1152). *IEEE*. https://doi.org/10. 1109/ICOSP.2008.4697333

Kenda, M., Cheng, Z., Guettler, C., Storm, C., Ploner, C. J., Leithner, C., & Scheel, M. (2022). Inter-rater agreement between humans and computer in quantitative assessment of computed tomography after cardiac arrest. *Frontiers in Neurology*, *13*, 990208. https://doi.org/10.3389/fneur.2022.990208

Kim, K. W., Lee, S. Y., Choi, J., Chin, J., Lee, B. H., Na, D. L., & Choi, J. H. (2020). A Comprehensive Evaluation of the Process of Copying a Complex Figure in Early- and Late-Onset Alzheimer Disease: A Quantitative Analysis of Digital Pen Data. *Journal of Medical Internet Research*, *22*(8), e18136. https://doi.org/10.2196/18136

Langer, N., Weber, M., Vieira, B. H., Strzelczyk, D., Wolf, L., Pedroni, A., Heitz, J., Müller, S., Schultheiss, C., Tröndle, M., Lasprilla, C. A., Rivera, D., Scarpina, F., Zhao, Q., Leuthold, R., Jenni, O. G., Brugger, P., Zaehle, T., Lorenz, R., & Zhang, C. (2022). The AI Neuropsychologist: Automatic scoring of memory deficits with deep learning. *Https://Www.Biorxiv.Org/Content/10.1101/2022.06.15.496291v4*. https://doi.org/10.1101/2022.06.15.496291

Li, Y., Clamann, M., & Kaber, D. B. (2013). Validation of a Haptic-Based Simulation to Test Complex Figure Reproduction Capability. *IEEE Transactions on Human-Machine Systems*, *43*(6), 547–557. https://doi.org/10.1109/TSMC.2013.2287341

Moetesum, M., Diaz, M., Masroor, U., Siddiqi, I., & Vessio, G. (2022). A survey of visual and procedural handwriting analysis for neuropsychological assessment. *Neural Computing and Applications*, *34*(12), 9561–9578. https://doi.org/10.1007/s00521-022-07185-6

Nevatia, R. (2000). Perceptual Organization for Generic Object Descriptions. In K. L. Boyer & S. Sarkar (Eds.), *Perceptual Organization for Artificial Vision Systems* (Vol. 546, pp. 173–189). Springer US. https://doi.org/10.1007/978-1-4615-4413-5_10

Osterrieth, P. A. (1944). Test of copying a complex figure; contribution to the study of perception and memory. *Archives of Psychology*, 30, 206–356.

Park, J. Y., Seo, E. H., Yoon, H.-J., Won, S., & Lee, K. H. (2023). Automating Rey Complex Figure Test scoring using a deep learning-based approach: A potential large-scale screening tool for cognitive decline. *Alzheimer's Research & Therapy*, *15*(1), 145. https://doi.org/10.1186/s13195-023-01283-w

Pereira, C. R., Pereira, D. R., Silva, F. A. D., Hook, C., Weber, S. A. T., Pereira, L. A. M., & Papa, J. P. (2015). A Step Towards the Automated Diagnosis of Parkinson's Disease: Analyzing Handwriting Movements. *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, 171–176. https://doi.org/10.1109/CBMS.2015.34

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Arch. Psychol*, *28*, 286–340.

Serra, J. (1986). Introduction to Mathematical Morphology. Computer Vision, Graphics, and Image Processing, *35*(3), 283–305. https://doi.org/10.1016/0734- 189X(86)90002-2.

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Souza, J. W. M. D., Alves, S. S. A., Rebouças, E. D. S., Almeida, J. S., & Rebouças Filho, P. P. (2018). A New Approach to Diagnose Parkinson's Disease Using a Structural Cooccurrence Matrix for a Similarity Analysis. *Computational Intelligence and Neuroscience*, *2018*, 1–8. https://doi.org/10.1155/2018/7613282

Sulastri, A., Utami, M. S. S., Jongsma, M., Hendriks, M., & van Luijtelaar, G. (2018). The Indonesian Boston Naming Test: Normative Data among Healthy Adults and Effects of Age and Education on Naming Ability. *International Journal of Science and Research (IJSR)*, *8*(11), 134-139.

Taleb, C., Khachab, M., Mokbel, C., & Likforman-Sulem, L. (2019). Visual Representation of Online Handwriting Time Series for Deep Learning Parkinson's Disease Detection. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 25–30. https://doi.org/10.1109/ICDARW.2019.50111

Troyer, A. K., & Wishart, H. A. (1997). A comparison of qualitative scoring systems for the Rey-Osterrieth Complex Figure Test. *The Clinical Neuropsychologist*, *11*(4), 381–390. https://doi.org/10.1080/13854049708400467

Vessio, G. (2019). Dynamic Handwriting Analysis for Neurodegenerative Disease Assessment: A Literary Review. *Applied Sciences*, *9*(21), 4666. https://doi.org/10.3390/app9214666

Vogt, J., Kloosterman, H., Vermeent, S., Van Elswijk, G., Dotsch, R., & Schmand, B. (2019). Automated scoring of the Rey-Osterrieth Complex Figure Test using a deep-learning algorithm. *Archives of Clinical Neuropsychology*, *34*(6), 836–836. https://doi.org/10.1093/arclin/acz035.04

Wahyuningrum, S. E., Sulastri, A., Hendriks, M. P. H., & van Luijtelaar, G. (2022). The Indonesian Neuropsychological Test Battery (INTB): Psychometric Properties, Preliminary Normative Scores, The Underlying Cognitive Constructs, and the Effects of Age and Education. *Acta Neuropsychologica*, *20*(4), 445–470. https://doi.org/10.5604/01.3001.0016.1339

Webb, S. S., Moore, M. J., Yamshchikova, A., Kozik, V., Duta, M. D., Voiculescu, I., & Demeyere, N. (2021). Validation of an automated scoring program for a digital complex figure copy task within healthy aging and stroke. *Neuropsychology*, *35*(8), 847–862. https://doi.org/10.1037/neu0000748

Zhang, X., Lv, L., Min, G., Wang, Q., Zhao, Y., & Li, Y. (2021). Overview of the Complex Figure Test and Its Clinical Application in Neuropsychiatric Disorders, Including Copying and Recall. *Frontiers in Neurology*, *12*, 680474. https://doi.org/10.3389/fneur.2021.680474

CHAPTER 5

Automated Speech Recognition in Bahasa Indonesia for Verbal Neuropsychological Tests

Submitted by Shinta Estri Wahyuningrum, Gilles van Luijtelaar, Augustina Sulastri, Marc P.H. Hendriks, Ridwan Sanjaya, Tom Heskes, & David van Leeuwen.

Abstract

Neuropsychological assessment in the verbal domain typically relies on traditional paper-based approaches. However, the advent of the Automatic Speech Recognition (ASR) system is enabling a transformative change in human-computer interaction. Nevertheless, implementing ASR presents difficulties in areas where languages such as Bahasa Indonesia are not widely spoken. The challenges arise from the limited availability of transcribed audio sample data and the complex and costly implementation. In this study, we aim to evaluate a transcription model using Wav2vec2 and grammar constriction techniques in the Indonesian language for verbal (neuro)psychological tests using the audio sample from the Indonesian-Boston Naming Test (I-BNT), a 60-word naming assessment. This model has the potential to be used by neuropsychologists in the transcription of verbal response-based test results. We collected verbal responses from a diverse cohort of 100 participants, varying in sex, accent, and recording quality. We compared two different ASR decoding methods: Wav2Vec2 letter and Wav2Vec2 plus Viterbi decoding using a bi-gram Language Model (LM). The performance of the models was assessed using Word Error Rate (WER). Our results indicate that both the inclusion of bi-gram LM and the quality of the dataset significantly ($p < 0.001$) influenced speech to text performance. The ASR model utilizing a bi-gram LM demonstrated a high accuracy (~95%) in transcribing the 60-word I-BNT. The model demonstrated a good-to-excellent performance irrespective of sex and accent of the speakers. While some words remain challenging to transcribe accurately, we propose several opportunities to address this issue. The successful application of automatic ASR using Wav2Vec2 for Bahasa Indonesia holds promise for integration into other verbal (neuro)psychological tests tailored to specific assessment needs.

Keywords: Automated speech-to-text; Bahasa Indonesia; Bigram; Indonesian Boston Naming Test; Verbal domain test; Wav2Vec2.

Introduction

Automated Speech Recognition (ASR), also known as Automatic Speech-to-Text or Automatic Speech Transcription, is a technology that enables programs to convert human speech into a written text. Contemporary ASR systems use neural networks to transform spoken language (audio signal) into written text (Alharbi et al., 2021). ASR is the backbone of virtual assistants such as SIRI, ALEXA, and Google Assistant enabling those applications to understand and respond to voice commands (Pahwa et al., 2020). ASR is a versatile technology with applications across various aspects of daily life. However, its development requires a targeted approach, tailoring the system to the specific needs of each application (Filippidou & Moussiades, 2020; Kitzing et al., 2009).

The ASR system has various applications and functionalities, such as being able to dictate messages (Pragt et al., 2022), smart home device control, transcribe meetings, and subtitling films. ASR is also revolutionizing the way cognitive functions can be assessed. It offers valuable insights for healthcare professionals by automatically converting audio to text that can then be further processed. An application was conducted for early detection of dementia and Alzheimer's disease by identifying speech patterns, such as fluency, pronunciation, and silence intervals. Subtle changes in speech patterns that might be unnoticed in a traditional interview could potentially be used to identify issues in word pronunciation, such as prolonged pauses between words and difficulty word finding (Chien et al., 2019; De La Fuente Garcia et al., 2020; Iglesias et al., 2022; Toth et al., 2018). Researchers have compared speech patterns of healthy individuals with those of patients diagnosed with cognitive decline (König et al., 2015, 2018; Roshanzamir et al., 2021; Zhang et al., 2021). This approach showed potential for the early detection of cognitive decline. Additionally, it highlights the need for language adaptation in its application.

ASR also can be used to help therapists conduct remote therapy sessions using a standardized online test. This application makes crucial interventions accessible to a wider range of patients regardless of geographical limitations (Abad et al., 2013). Additionally, ASR is being explored for automated scoring in specific tests (van den Noort et al., 2008). For example, in verbal fluency tests where patients name as many words as possible within a time limit, ASR can transcribe their responses and analyze them to generate a score (König et al., 2018). ASR eliminates the need for manual scoring, thereby saving time, potentially increasing objectivity (Giauque et al., 2023; Tran et al., 2022), and helping the clinician to analyze mistakes that are not observed directly such as changes in speech, speech fluency, and articulation.

This transcription approach can also be conducted particularly in assessments that require verbal responses such as the Boston Naming Test, the Verbal Fluency Test, Digit Span, and the Rey-Auditory Verbal Learning Test (Strauss et al., 2006). These advances demonstrate the versatility of adapting the ASR to various needs within the field of neuropsychology.

Traditionally, Hidden Markov Models (HMMs) formed the cornerstone of speech recognition systems. About a decade ago, the introduction of neural networks was a turning point by enhancing the accuracy of ASR (Kim et al., 2020). Both approaches to ASR require a substantial quantity of labelled data for training. The further development of deep learning approaches, large open data sets, and faster computing led to ASR models that directly map the acoustic input to text without the need for an intermediate phonetic representation. These methods require labeled audio data and typically have increased accuracy as more data is fed into the neural networks (Sharma et al., 2020).

Labelled audio data is expensive and time-consuming to produce, especially for the languages that are less commonly spoken worldwide. Wav2vec2 is a model that uses self-supervised learning (Baevski et al., 2020; Schneider et al., 2019) which allows it to learn from unlabelled audio data. The usability of Wav2vec2 is generally considered to be developer-friendly, with pre-trained models readily available for use and fine-tuning to new languages (Maxwell-Smith & Foley, 2023) or even other speech tasks (Vaessen & van Leeuwen, 2022). The Wav2Vec2 model allows the ASR to process the raw audio data directly, and further fine-tuning existing pre-trained models offers more effective training by requiring less labelled audio data compared to previous approaches. The main advantage of this model is the flexibility to adapt to different languages and accents making this model versatile for various ASR.

Although ASR offers a wide range of benefits such as real-time transcription, language learning, and assistance for various devices there are still significant challenges. The main issues in developing the ASR system include language variation, speaker accent, and dialect coverage (Silber Varod et al., 2021; van den Noort et al., 2008). With over 7,000 languages and countless accents and dialects worldwide, it is difficult for an ASR system to accommodate such diversity. Even with extensive training data, some systems may fail to accurately recognize speech from speakers with less common languages and accents. Other issues that can affect the performance include background noise, the sex of the speaker (Alsharhan & Ramsay, 2020; Garnerin et al., 2019), health conditions (Ngueajio & Washington, 2022), and the quality of the audio sample. ASR struggles to accurately transcribe

speech in noisy environments. Background noise such as traffic, music, and cross-talk typically have a negative effect on the transcription accuracy. Furthermore, a previous study found that racial differences also influence the performance of ASR (Koenecke et al., 2020).

Indonesia, a vast archipelago nation, is characterized by a variety of languages spoken by different ethnic groups with numerous dialects and accents. However, the lingua franca, Bahasa Indonesia, only introduced in 1928 and further promoted and standardized after Indonesia gained independence 1945, is the official language, spoken in public, at schools, and in the mass media. For most Indonesians, Bahasa Indonesia is not their first language and only acquired at elementary school.

ASR technology for Bahasa Indonesia continues to evolve, with recent developments demonstrating notable improvements in accuracy. A recent study assesses the performance of the XLS-R 300m model across datasets containing Indonesia, Javanese, and Sundanese showing that incorporationg a 5-gram KenLM significantly reduces Word Rate (WER) (Ariasaputra et al., 2024). Furthermore, the application of XLSR-53 pretrained models for cross-lingual transfer learning proves effective in enhancing WER performance when compared to Wav2Vec2, even with limited data (Arisaputra & Zahra, 2022). Given that Bahasa Indonesia agglutinative language, speech recognition models may benefit from syllable-based recognition approaches (Hidayat, 2016). Indonesia still requires a well-structured methodology to process verbal responses, optimize ASR performance in a low-resource language setting, and overcome linguistic challenges unique to Bahasa Indonesia.

This study investigates the implementation of an ASR system for transcribing spoken Bahasa Indonesia. An existing Wav2vec2 model fine-tuned for Bahasa Indonesian is leveraged to address the challenge of limited resources specific to the Indonesian language. Additionally, an analysis is conducted to evaluate the effect of factors such as dataset quality, sex of the speaker, and speaker accents. This study aim to gain a deeper understanding of the ASR system's strengths and limitations within the context of the I-BNT. The findings will contributes to both speech processing research for specialized task and clinical neuropsychology, bridging a crucial gap in automated cognitive assessments. This article is structured as follows: Section 2 corresponds to methodology of data collection explaining how the data are collected and the system design. This section explaining the method of speech transcription system and how to evaluate the system as well. Section 3 presents the results obtained from the model comparation, factors that influence the ASR performance and the word level analysis. Section 4 presents conclusions.

Methodology of data collection and system design

The Indonesian Boston Naming Test, or I-BNT, is an adaptation of the BNT (Kaplan et al., 2016) test designed to assess confrontational naming abilities within the Indonesian context. Confrontational naming simply refers to the ability to name objects presented visually. During the test, the examiner presents the examinee with drawings of 60 objects, one at a time. The examinee is then asked to name each object aloud and in case no or a false response is given, cues are given by the examiner and again a verbal response can be given. The I-BNT test traditionally relies on examiners to manually score responses as correct or incorrect. This process can be time-consuming and prone to errors. We believe the use of ASR systems can offer a solution. ASR can automatically transcribe the examinee's answers, potentially streamlining the administration process. At the end, the examiner sums all the individual item scores to determine a total score for the I-BNT.

Dataset collection

The dataset that we used was collected from around 100 participants residing in major cities across three Indonesian islands. Participant ages ranged from 20 to 55 years old, with the majority being undergraduate students. The audio data were collected using digital devices such as web online and smartphone (Pesau & Luijtelaar, 2021; van Leeuwen et al., 2016). Audio segmentation was carried out during the audio recording process, and this followed by an automated word transcription steps. The dataset was divided based on the recording procedures into two groups with the following characteristics:

I-BNT I: The initial dataset was collected during the live administration of the I-BNT. Audio recordings were captured using an application embedded within the examiner's smartphone. The smartphone was placed on the table and recorded all sounds that might occur during the test administration. Each recording was made immediately after the participant responded to a stimulus image, ensuring that the audio was already segmented by word. However, since the examiner’s voice occasionally appeared in the recordings during test administration, additional manual segmentation was required to isolate only the examinee’s speech. After recording, the audio data were converted into WAV format. The final dataset consisted of 1,042 audio files. Not all recordings contained the full set of 60 words, as some responses were omitted or removed due to excessive overlapping speech (cross-talk). However, we retained files that included background noise, minor cross-talk, and repeated words within a single audio file.

Table 1. The distribution of the number of audio recordings for each word from I-BNT I and I-BNT II datasets.

Indonesian Word	English Word	N	Indonesian Word	English Word	N
Anak Panah	arrow	63	Helikopter	Helicopter	69
Badak	Rhinoceros	74	Jamur	Mushroom	69
Bel	Bell	71	Kacang Tanah	Peanut	66
Bola Dunia	Globe	66	Kaktus	Cactus	69
Bunga	Flower	84	Karangan Bunga	Bouquet	66
Burung	Bird	76	Kartu	Card	69
Busur	Bow	73	Kompas	Compass	69
Cangkul	Hoe	70	Kuda Laut	Seahorse	68
Corong	Funnel	62	Kursi	Chair	70
Cumi-cumi	Squid	67	Kursi Roda	Wheelchair	70
Enggrang	Stilts	63	Monas	National Monument	70
Engsel Pintu	Door Hinge	63	Moncong	Snout	65
Eskalator	Escalator	65	Onta	Camel	70
Gantungan Baju	Clothes Hanger	68	Palet	Palette	58
Gendang	Drum	70	Pelana	Saddle	66
Gergaji	Saw	73	Peluit	Whistle	69
Gitar	Guitar	70	Pensil	Pencil	65
Gulungan Kertas	Paper Roll	66	Perahu	Boat	68
Gunting	Scissors	68	Piramida	Pyramid	70
Gunung Berapi	Volcano	67	Pohon	Tree	70
Raket	Racket	70	Roti Tawar	Load Bread	69
Rumah	House	66	Rumah Gadang	Traditional House	54
Sapu	Broom	66	Sempoa	Abacus	68
Sikat gigi	Toothbrush	73	Simpul	Knot	68
Siput	Snail	68	Sisir	Comb	73
Stetoskop	Stethoscope	66	Suling	Flute	67
Tempat Tidur	Bed	71	Tenda	Tent	75
Teralis	Window Grille	60	Teropong	Telescop	66
Tikus	Mouse	70	Topeng	Mask	70
Wayang	Puppet	69	Wortel	Carrot	92

I-BNT II: The subsequent dataset was acquired through an online assessment using the audio recording feature. Since participants administered the test independently, the recording conditions varied, even though they were instructed to find a quiet place to minimize background noise. Audio was recorded after participants viewed

the visual stimuli, which were presented via a web browser interface. They were then prompted to verbally articulate their word associations by activating the recording function through the designated button. As a result, the system automatically segmented the recorded audio. This approach allowed participants to use either their smartphones or laptops and adjust their speaking distance as needed. Although designed to reduce noise, some audio samples still contained residual background noise. The total number of audio recordings collected was 3,074.

The I-BNT I dataset serves as a sample of the real-world conditions under which the I-BNT is typically administered, providing insights into the recording environment and challenges encountered. The I-BNT II dataset was compiled with the objective of reducing the effects of cross-talk and background noise, and it had the added advantage of being recorded directly in .wav format. Audio recordings from both datasets were initially captured at a sampling rate of 44.1 kHz and subsequently down-sampled to 16 kHz to align with the acoustic model requirements for the Wav2vec2 model. The distribution of audio recordings for each word is presented in Table 1; it comprises the total set of audio recordings from both I-BNT datasets.

Manual Transcription for Ground Truth Labels

A ground truth transcription is used in ASR development as a reference for measuring the accuracy of our ASR system. The manual transcription process was conducted by a professional transcriber who meticulously listened to each utterance within the audio files. The transcription rules required the transcriber to document every word present in the audio, along with noting additional information. Such information includes speech elongation, cross-talk, corrupted audio files resulting from conversion errors or truncations, and additional vocalizations (e.g., "eee," "hmm," "oh," etc.).

ASR Performance Accuracy

The performance of ASR was measured using the Word Error Rate (WER), which is used as a standard metric for evaluating the system's performance. The WER is determined by aligning the transcription generated by the ASR system with the ground truth transcription and calculating the number of errors. Firstly, substitutions are words that are incorrectly transcribed by the ASR system. Secondly, insertions are words that are present in the ASR system but not in the ground truth transcription. Finally, deletions are words that are omitted in the ASR system but not present in the ground truth transcription (Filippidou & Moussiades, 2020). The WER is computed as:

$$WER = (S + I + D) / N \quad (1)$$

Where: S is the number of substitutions, I is the number of insertions, D is the number of deletions and N is the number of words in the ground truth transcription.

A lower WER in the speech-to-text system indicates a better accuracy for the transcription model. To illustrate the calculation of WER, consider the following example: the human-transcribed data serves as the reference, while the system-transcribed data is termed the hypothesis. In this instance, the reference is "karangan bunga" and the hypothesis is "kuda". The error analysis reveals a substitution of 1 (where "karangan" is substituted with "kuda"), a deletion of 1 (where "bunga" is deleted), no insertions (as no additional words are added), and a total of 2 words in the reference text. Consequently, the WER is calculated as (1 substitution + 0 insertions + 1 deletion) divided by 2 (the number of words in the reference), resulting in a WER of 1.

ASR System Decoding Methods

In this study, we employed the Wav2Vec2-Large-XLSR-Indonesian model, a fine-tuned facebook/wav2vec2-large-xlsr-53 (Wirawan, 2021), renowned for its adaptability to multiple languages (Baevski et al., 2020), in conjunction with the bi-gram language model (Pakoci & Popović, 2021). The Wav2Vec2 model leverages self-supervised learning on large unlabeled datasets, making it particularly effective for multilingual speech recognition tasks. We used two different configurations for the models.

Our first speech recognition model (Model 1) used wav2vec2-xlsr (Cross-Lingual Speech Representation), which was pretrained on multiple languages. The process begins with wav2vec2 converting spoken audio into character-level logits, which serve as the foundation for generating potential transcriptions. A beam search decoder then processed these logits, generating multiple possible word sequences as interpretations of the input audio.

Our second model (Model 2) improved transcription accuracy by combining wav2vec2 with bigram Language Model (LM). To further refine the output from wav2vec2, we incorporated KenLM, a language model that ranked the generated sequences based on their linguistic probability. This rescoring process prioritized the most likely word sequences, ensuring better alignment with natural language patterns. Finally, the system selected the most probable sentence as the final output, enhancing overall recognition accuracy and reliability.

Figure 1 shows how the ASR system works. The system takes raw audio recordings in WAV format with a sample rate of 16 kHz as input and converts them into word transcriptions in Bahasa Indonesia. The Wav2vec2 processes audio data, and represents the entire utterance as a sequence of letters. To enhance accuracy, Viterbi decoding is applied in conjunction with an n-gram language model. This method incorporates a list of all possible responses for I-BNT words to guide the decoding process and improve the selection of the most likely word sequence.

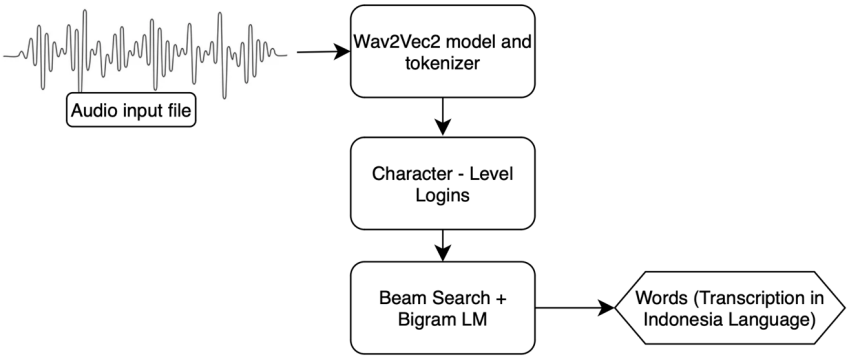


Figure 1. ASR system for I-BNT speech to text using wav2vec2 model and bigram language model.

The evaluation was conducted on three aspects: model performance, the effect of dataset quality and demographic factors, and word-level analysis.

Investigating the Performance of the Two Models

This examined how effectively the models (Model 1 and Model 2) transcribed the audio sample files. We assessed the accuracy and efficiency of the model by calculating the WER of each model for all 4,116 labelled audio samples from the two datasets (I-BNT I and I-BNT II). A t-test statistic was used to investigate the difference in accuracy between the two models in capturing the spoken content from the I-BNT recordings.

Investigating the Effects of Dataset Quality and Demographic Factors

This evaluation explored the impact of three key areas: dataset quality, speaker sex, and speaker accent. Dataset quality was assessed by comparing two datasets (Dataset I-BNT I and Dataset I-BNT II) with the following audio file characteristics. Characteristics of I-BNT I: crosstalk occurrence, audio file formats require conversion before they can be processed by ASR, some audio recordings need additional segmentation to minimize crosstalk caused by the tester's interaction,

controlled environment, the tester has the flexibility to adjust the positioning of both participants and recording devices to optimize audio quality. Characteristics of I-BNT II: the quality of the internet connection varies, which may affect the audio transmission; no need for audio file format conversion, as the data is already in a compatible format; some participants may not be mindful of their microphone distance, which can impact audio clarity. Additionally, we investigated the influence of speaker sex by categorizing speakers as male or female. Finally, we examined the impact of speaker accent by analyzing regional pronunciation variations within the Indonesian language, focusing on speakers from Java, Bali, and Sulawesi.

The WER was calculated for each subgroup, providing a quantitative measure of system performance across these different speakers and recording characteristics. Finally, an Analysis of Variance (ANOVA) was conducted to identify statistically significant differences in WER between the groups. This analysis revealed the impact of dataset quality, speaker sex, and speaker accent on ASR system accuracy.

Investigating WER on the 60 Words

Previous evaluations of the ASR system employed holistic metrics, analyzing performance at the entire audio sample level. This approach provided valuable insights into overall system accuracy. However, for a more detailed understanding of the system's strengths and weaknesses, the current investigation focuses on WER at the word level. By segmenting each audio file into individual words and calculating WER for each word, this analysis allows for the identification of specific word types or pronunciations that pose a challenge for the ASR system. This fine-grained analysis provides crucial insights for targeted improvements aimed at enhancing the system's robustness and accuracy. Taking this a step further, we conducted a closer examination of the words with the highest error rates. This aimed to understand the reasons behind these difficulties and to propose potential improvements in future ASR system iterations.

Results

Comparative Analysis of Two ASR models

Figure 2 depicts the mean WER values for both models. Based on the initial observations of the mean WER results, Model 2 appears to outperform Model 1. However, to confirm this statistically (independent groups), an Unpaired Two-Sample T-Test was conducted on the WER data from 4116 audio samples.

The t-test results provide strong evidence of a statistically significant difference between the means of WER (Model 1) and WER (Model 2). The analysis reveals a mean WER for Model 1 of 0.488 with a standard error of 0.003, while the mean WER of Model 2 is 0.050 with a standard error of 0.003. Additionally, the T-statistic $t(126) = 16.60$ indicates a strong difference between the means, and the highly significant p-value ($p\text{-value} < .001$) confirms that this difference is unlikely due to chance. Finally, the 95% confidence interval [0.385, 0.490] for the difference in means excluded zero, further reinforces the conclusion that Model 2 has a statistically lower mean WER. We conclude that a WER of 0.05 for Model 2 indicates that, on average, 5% of the words in the transcribed audio differed from the reference transcript. This signifies a considerably lower error rate compared to Model 1's WER of 0.488, where nearly half (48.8%) of the words were incorrectly recognized.

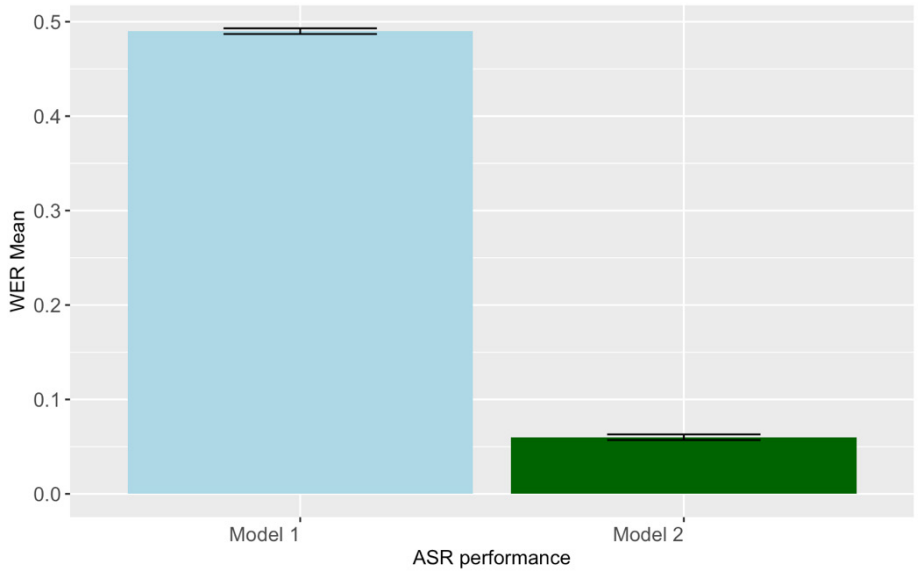


Figure 2. Word Error Rate (WER) of two speech recognition models. Model 2 achieves a significantly lower WER compared to Model 1. This indicates that Model 2 produces transcripts with fewer errors, demonstrating its superior accuracy in ASR compared to Model 1.

Table 2 shows the transcription from the two models. The results showed that model 1 does not align with the context of the I-BNT word list. Noteworthy errors include letter substitutions within words (e.g., "simpul" - "cimpul", "simpul" - "fimpul", and "siput" - "sibet"), the addition of extraneous letters (e.g., "simpul" - "seimpul", "sikat gigi" - "seikat gigi", "wayang" - "kuayang"), and letter reductions (e.g., "anak

panah" - "anakpana", "kacang tanah" - "kacang tana", "siput" - "sipu"). Additionally, Model 1 demonstrates sensitivity to short pauses within words, resulting in the transcription output being erroneously split into two words (e.g., "simpul" - "sim pul", "wayang" - "ua yang"). Furthermore, this model also captured the elongations and vocalizationsd in the transcription, as were identified from the original utterance (for example, "simpul" - "siiiimpul", "simpul" - "na simpul").

In contrast, Model 2 yields a discernible improvement in transcription outcomes. Notably, various factors, such as letter substitutions, additions, and deletions, demonstrate a substantial reduction. Furthermore, instances of elongation, short pauses, and vocalizations are conspicuously absent in the transcription generated by Model 2. The imposition of vocabulary constraints, guided by the I-BNT word list, ensures that unsuccessful transcription attempts yield words derived exclusively from the I-BNT lexicon. For example, "kacang tanah" being transcribed as "kacang panah", importantly, these errors occur far less frequently with the bi-gram model.

Table 2. Illustrative Examples of Transcription Outputs Using Model 1 and Model 2. The examples show the outcomes of each model transcribing the same spoken input.

Reference	Transcription		Word Error Rate	
	Model 1	Model 2	model 1	model 2
siput	sebut	siput	1	0
siput	tyciput	siput	1	0
siput	sibut	siput	1	0
siput	sipu	siput	1	0
siput	sibet	siput	1	0
simpul	simpel	simpul	1	0
simpul	fimpul	simpul	1	0
simpul	sim pul	simpul	1	0
simpul	cimpul	simpul	1	0
simpul	simpun	simpul	1	0
na simpul	a simpul	simpul	1	0
simpul	seimpul	simpul	1	0
simpul	stimpul	simpul	1	0
sikat gigi	esikat gigi	sikat gigi	0,5	0
sikat gigi	sikat digi	sikat gigi	0,5	0
sikat gigi	sikat kiki	sikat gigi	0,5	0
sikat gigi	sika kici	sikat gigi	0,5	0
sikat gigi	tikat gigi	sikat gigi	0,5	0

Table 2. Continued

Reference	Transcription		Word Error Rate	
	Model 1	Model 2	model 1	model 2
wayang	ua yang	wayang	1	0
wayang	kwaya	wayang	1	0
wayang	kuayang	wayang	1	0
wayang	kuayang	kacang	1	1
wayang	uayir	api	1	1
wayang	ayang	kacang	1	1
wayang	kanmaya	api	1	1
wayang	wayang	anak	1	1
wayang	kuaya	kuda	1	1
roti tawar	erati talar	berapi tawar	1	0,5
roti tawar	tosi tawar	kursi tawar	0,5	0,5

The Impact of Dataset Quality (dataset), Accent, and Sex on ASR Performance

In light of the finding that Model 2 exhibited superior performance, all subsequent evaluations consistently utilized the transcription results generated by Model 2. The dataset used for system evaluation included both I-BNT I and I-BNT II. Table 3 provides a detailed breakdown of the number of audio files, mean of WER, and standard deviation (SD) by dataset, sex and accent.

Table 3. Distribution of two datasets (total 4116 audio samples) by factor groups from the Indonesian Boston Naming Test (I-BNT) dataset.

Factor	Group	N	Word Error Rate	
			mean	Standard deviation
Dataset quality	I-BNT I	1042	.111	.32
	I-BNT II	3074	.040	.19
Sex	Female	2588	.059	.23
	Male	1528	.055	.23
Accent	Java	2184	.044	.20
	Bali	1040	.111	.32
	Sulawesi	892	.032	.17

The ANOVA analysis conducted on sex, accent, and dataset quality identified a significant difference solely in the factor of dataset quality between the two datasets. This difference was evidenced by a calculated F-value of $F(1,4109) = 76.276$, p-value $< .001$, and $\eta^2 = 0.02$. Conversely, the factor of sex exhibited no statistically significant difference in mean WER between female and male groups, as denoted by an F-value of $F(1, 4109) = 0.357$, p-value = 0.55, and $\eta^2 < 0.01$. Similarly, no significant disparities were observed among the three accent groups: Java, Bali, and Sulawesi, with an F-value of $F(1, 4103) = 0.395$, p-value = .53, and $\eta^2 < 0.01$. Furthermore, no significant interactions were found between factors across groups concerning ASR performance.

In assessing dataset quality (dataset), we observed several patterns in both audio utterances and transcriptions. Notably, the I-BNT I dataset exhibited frequent word repetition, and there were instances where two-word sequences appeared in reverse order (e.g., "gadang rumah gadang" instead of "rumah gadang"). The factors impacting transcription quality often emerged within this dataset, including background noise, cross talk, and audio corruption due to trimming processes. These factors may have occurred individually or combined into dual-factor scenarios within a single file.

In contrast, the I-BNT II dataset maintained a lower level of background noise. However, elongation in pronunciation was more prevalent, and we encountered several files that appeared to be corrupted, possibly during the transfer from online to local storage. In the case elongation, a neuropsychologist provides an initial recommendation on whether elongation can be considered acceptable and classified as a correct answer.

Word-Level Transcription Performance

The word-level performance analysis was conducted by measuring the performance of each model on the transcription accuracy (WER) of each word. We aimed to evaluate how well the model performs across different word types, including variations in phonetic complexity and pronunciation challenges. Figure 3 illustrates the distribution of word accuracy for both models. The data suggests that Model 2 achieves higher transcription accuracy for all words compared to Model 1. This can be observed by the overall distribution of points, likely concentrated towards the lower WER range on the Y-axis for Model 2. The figure also illustrates that some words are spread out towards the right side, indicating that these words pose a significant challenge for both models especially for model 1.

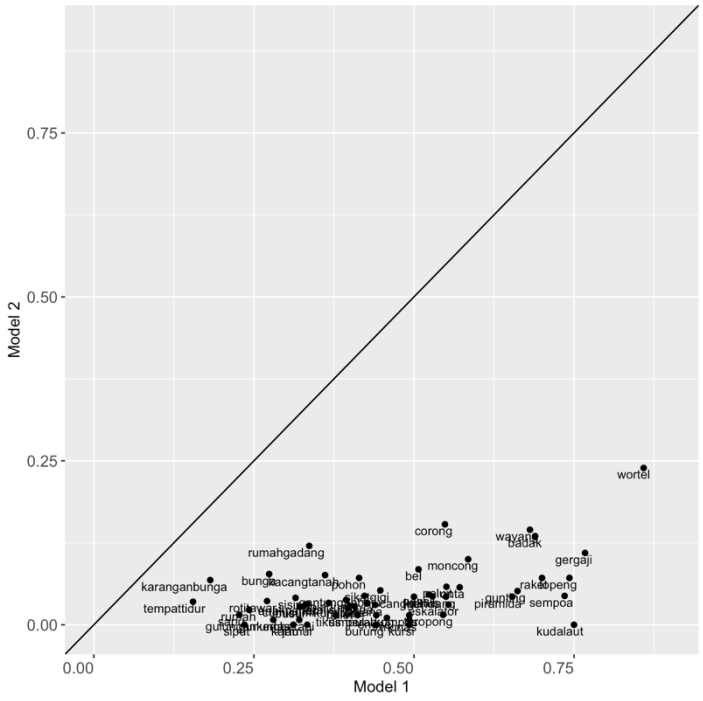


Figure 3. Word-Level transcription performance on model 1 vs. model 2. The scatter plot compares the performance of Model 1 and Model 2 on word-level transcription. Each point represents a word, with its WER on the X-axis for model 1 and the Y-axis for model 2. The diagonal line represents the relationship between the two models for each word. Areas on the right side indicate words that are challenging for both models.

Our word-level evaluation revealed that Model 2 consistently outperformed Model 1. Subsequently, we delved into the distribution of WER (Word Error Rate) for Model 2 across both datasets. As depicted in Figure 4, Dataset I-BNT II demonstrated superior performance compared to Dataset I-BNT I. Despite the overall better performance of Dataset I-BNT II, there remains a subset of words with lower accuracy levels in both datasets.

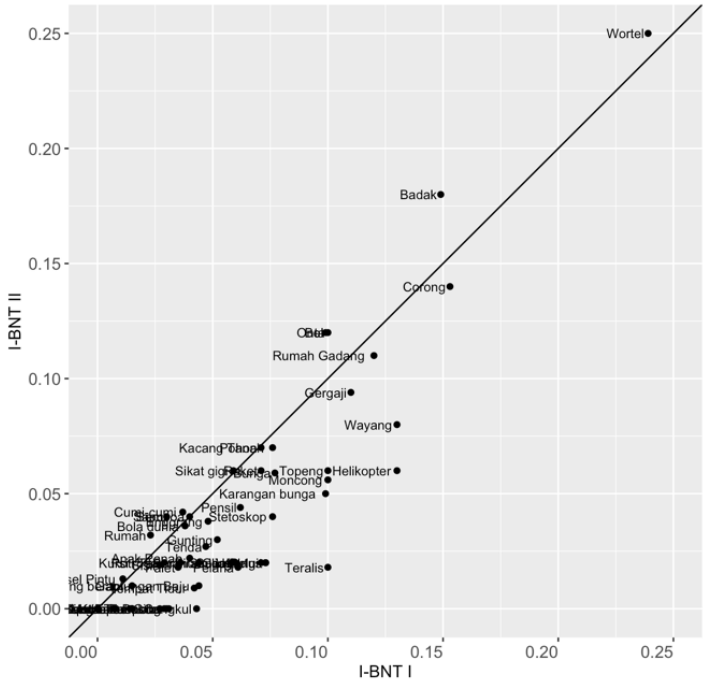


Figure 4. Word-Level transcription performance on two datasets. The scatter plot compares the performance of transcription accuracy on both databases. The X-axis for I-BNT I and the Y-axis for I-BNT II. The diagonal line represents the relationship between the two models for each word.

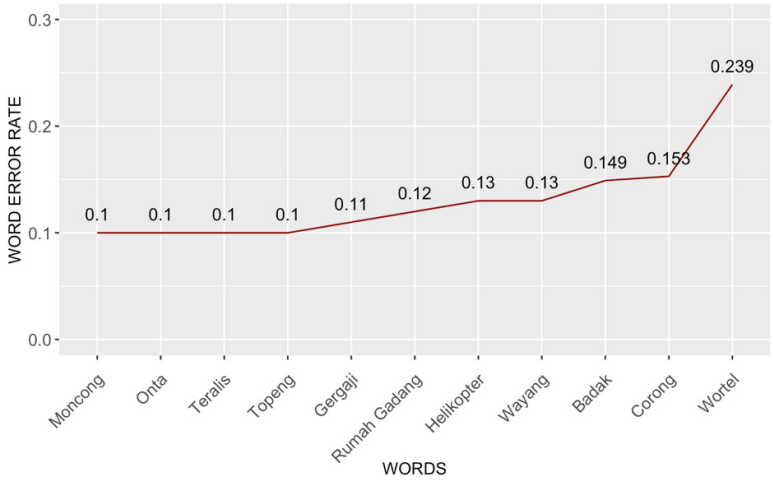


Figure 5. Distribution of WER (Word Error Rate) Across Word Groups as found with Model 2.The focus here is on word groups with an average WER exceeding 10%, indicating words that posed a great challenge for accurate transcription.

Figure 5 presents the eleven highest WER values from the I-BNT II dataset using model 2. The evaluation identified a specific set of words within the I-BNT list that exhibited consistently high WER values. These words included: "wortel" (carrot), "corong" (funnel), "badak" (rhinoceros), "rumah gadang" (large traditional house), "gergaji" (saw), , "onta" (camel), and “bel” (doorbell).

Discussion

ASR systems convert spoken language into text. Their accuracy depends on the language they are processing and the underlying technology. Studies show that ASR currently performs best with English (Silber Varod et al., 2021; van den Noort et al., 2008) because English is the most data available of all languages, it is the longest and best studied for ASR. Researchers are looking for ways to improve ASR for other languages. One promising approach is the Wav2vec2 model, which can be adapted to new languages with less training data (Kozhirbayev, 2023; Sriram et al., 2022). However, existing studies using Wav2vec2 for Indonesian and Sundanese languages still have high error rates (above 20%) (Cryssiover & Zahra, 2024; Maxwell-Smith & Foley, 2023). In this study, we evaluated the accuracy and fairness regarding sex and accent of an Indonesian ASR system specifically designed for verbal neuropsychological tests.

Model 1 offers a simple approach to text generation, its lack of consideration for context severely limits the quality of the output. It can generate nonsensical or repetitive text, lacking the coherence and structure achieved by methods that leverage language models. People naturally pronounce words differently, including stretching sounds (elongation) or slightly changing them (substitution, deletion, addition). These variations posed a challenge to Model 1’s ability to accurately match speech to the reference text. Therefore, this initial model lacks the requisite accuracy for naming tasks such as I-BNT.

Despite its unsuitability for scoring the verbal test, this model exhibits promising potential for specific clinical evaluations. The transcription results generated by model 1 can be utilized to analyze error patterns in patients. These patterns, such as extended pauses time within phrases (“com--puter”), and frequent use of placeholders ("uh," "um") or general words ("that," "like this") (Novogrodsky & Kreiser, 2015), can be extracted in speech recordings. By analyzing these patterns, researchers can then identify characteristics associated with various neuropsychological disorders (König et al., 2015; Zhang et al., 2021). Additionally,

the model holds promise in identifying emotional states and differentiating dialects. These capabilities highlight the model's potential for specialized applications within clinical evaluation (Brown, 2016; Favaro et al., 2023; Tits et al., 2018).

Model 2 achieved a remarkably low average error rate of around 5% across the entire dataset, significantly outperforming Model 1 on the I-BNT. This improvement highlights the effectiveness of integrating a bigram grammar constraint technique (Creutz et al., 2007) into ASR systems for specific domains, as supported by studies demonstrating accuracy gains (Ao & Ko, 2021; Habeeb et al., 2021). This technique refines the model's output by focusing on a predefined vocabulary relevant to the task (Sennrich et al., 2016; Ye-Yi Wang et al., 2003). This study demonstrates that the use of wav2vec2 combined with a bigram language model (LM) effectively reduces the Word Error Rate (WER) compared to previous studies that employed other ASR methods (Muljono et al., 2017; Prakoso et al., 2016). Unlike traditional ASR approaches that rely on manually designed components (acoustic models, lexicon, and language models), Wav2Vec2 is an end-to-end model that learns directly from raw audio using self-supervised learning. This allows it to perform better in low-resource settings (Arisaputra et al., 2024) and noisy environments, making it particularly suitable for verbal neuropsychological tests.

In the I-BNT context, bigrams effectively address contextual errors arising from letter-level misrecognition, such as single-letter insertions, deletions, and substitutions. The cornerstone of Model 2's remarkable performance lies in its manually defined vocabulary list, carefully compiled by experienced clinicians and neuropsychologists. This vocabulary list is a list of all possible correct responses that could appear and be considered to be valid answers.

However, this approach offers a trade-off. While beneficial for domain-specific applications, it might limit the model's ability to learn complex relationships between words, potentially restricting the fluency and variety of generated text. Therefore, this model is best suited for verbal tests with a well-defined, contextual vocabulary.

Additionally, accurate ASR output is significantly affected by background noise (Alapetite, 2008; Rajnoha & Pollák, 2011; Rodrigues et al., 2019; Yadav et al., 2018). Our findings confirmed this, as the I-BNT II dataset which has lower noise levels, achieved higher transcription accuracy. Similarly, previous studies on Indonesian ASR have shown that noise can lead to word substitutions, such as *"saya"* (I) being transcribed as "kaya" (rich) or "daya" (power) (Ferdiansyah & Purwarianti, 2012).

Sex bias in ASR performance has been documented. Garnerin et al. (2019) reported a substantial 24% increase in WER for female speakers, while Alsharhan & Ramsay (2020) observed a smaller sex gap of approximately 5%. In contrast, our study, conducted within a controlled context and vocabulary, revealed a negligible sex difference in ASR performance, with a WER disparity of less than 1%.

Our investigation into accent's impact on the I-BNT transcription yields intriguing insights. Although statistically significant variations across all accent groups were not observed, the study did identify specific mispronunciations demonstrably linked to distinct regional speech patterns. For instance, speakers with a Balinese accent often added the letter "h" to words ending in the vowel "a," transforming "sempoa" (abacus) into "sempoah" and "bunga" (flower) into "bungah." Our second model version effectively addressed these instances. However, a crucial limitation of this study is the limited variation of sample size; we only analyzed data from three regional accents. Considering the linguistic landscape of Indonesia, with an estimated 700 accents, capturing the full spectrum of speech patterns will require incorporating a broader range of samples in the future. Expanding the dataset will enhance the system's generalizability, making it applicable to a broader population, which aligns with findings from previous studies (Fergadiotis et al., 2019; Hula et al., 2015).

We identified several factors contributing to the high WER for 11 words in the I-BNT list. The issue is elongation speech and substitution-selection-addition letters, which also inflate WER. When a speaker stretches out a sound, the ASR system may struggle to capture all the intended sounds, leading to omissions or misinterpretations. Similarly, even minor human transcription errors, such as hearing "terali" instead of "teralis," are counted as full errors in WER calculations. This highlights the impact of slight variations in human perception.

Loan-word discrepancy is another factor. Loanwords, words borrowed from other languages and used in Indonesian without translation, leading to discrepancies between their pronunciation and written form. Addressing loanwords is crucial as they can impact ASR accuracy due to orthographic differences (Zhang & Osth, 2024). In this case, the relationship between the orthography (spelling) and pronunciation of the loan-word differs from the relationship the Wav2vec2 model learned during fine-tuning for Bahasa. For instance, "wortel" (Dutch for "carrot"), potentially causing the ASR system to misinterpret the sounds and assign an incorrect word. Similarly, regional terms like "corong" (a Javanese word for "funnel"), which are loanwords from local languages, may exhibit pronunciation variations for people from other islands in Indonesia. To mitigate loadwords challenges, we recommend

adding alternative pronunciations in the pronunciation dictionary for these words, or incorporating additional audio data samples for loanwords and retraining the ASR system to account for their diverse variations (Brown, 2016; Prasad & Jyothi, 2020; Radzikowski et al., 2021).

Prior to integrating an ASR system into an automated scoring system for verbal tests, several considerations are crucial for optimal performance. One critical step is addressing variations in word correctness, encompassing all possible correct responses. In our study, we initially included 77 words in the bigram list. However, we encountered instances where different word arrangements conveyed the same meaning, leading to ASR misrecognition. For example, "gulungan kertas" (paper roll) and "kertas gulung" (rolled paper) represent the same object but it has a different word order.

To address this challenge, we recommend incorporating these variations into the system. Additionally, expanding the I-BNT word list to include words highlighting these problematic issues is advisable. This can be achieved by continually adding more possible I-BNT vocabulary that neuropsychologists across the archipelago accept as correct responses from subjects. Next, careful preparation of audio data collection is essential. Minimizing noise exposure and standardizing microphone distance during test administration in controlled environments can significantly improve ASR performance (Abad et al., 2013; Jamal et al., 2017; Rodrigues et al., 2019).

The potential of ASR system extends beyond a fully automated I-BNT assessment. It holds particular promise for various neuropsychological assessments in the verbal language domain in Indonesian. Examples include the Digit Span and Rey Auditory Verbal Learning Test (RAVLT) and the Stroop test. The Digit Span Test has a very limited vocabulary (numbers zero to nine), limited verbal responses are also shared by the two other mentioned tests of the INTB. The RAVLT memory test regards the learning of 15 repeatedly presented words and during the verbal recall omission and repetition errors can be detected by ASR. The number of words in the STROOP tests is also limited (the four color names) and also here correct answers and interference errors can be automatically detected, making it another suitable application for ASR.

Context-aware ASR systems, like our Model 2, excel in handling diverse Indonesian speech, adapting to vocabulary variations across assessment types. As highlighted in a previous study (Ziman et al., 2018), speech-to-text systems offer significant advantages like cost-effectiveness, reliability, and speed in automatically transcribing

data from psychological experiments (Hula et al., 2015). Successful integration of ASR into these assessments can significantly enhance research efficiency and potentially streamline the clinical evaluation process in the future (Fergadiotis et al., 2019; König et al., 2018).

A distinct challenge emerges when dealing with verbal tests that require semantic meaning analysis, such as the semantic Verbal Fluency Test, for example animal names. This test necessitates an additional layer of processing – labeling and root word determination for each semantically meaningful word. This critical step requires collaboration between neuropsychologists and linguistic experts (König et al., 2018; Woods et al., 2016), while computer scientists can help on the classification of the word using supervised machine learning. By accurately capturing these semantic nuances in verbal responses, an automated system could significantly enhance, or even transform, the efficiency and objectivity of such neuropsychological evaluations.

As previously emphasized, a cornerstone of effective ASR system development is a diverse audio sample database encompassing the rich linguistic Indonesian context. Fortunately, Indonesia possesses a valuable resource in the I-ANDI database for neuropsychological test results (Wahyuningrum et al., 2021). This comprehensive database can accommodate speech data from a wide range of speakers across the archipelago. It presents a significant advantage for constructing a robust Indonesian-language ASR system specifically designed for automated verbal testing.

By capitalizing on this comprehensive dataset encompassing both existing data and data from clinical populations, we aim to develop an ASR system tailored for Indonesian clinical applications. This system has the potential to revolutionize various aspects of clinical neuropsychology, such as early detection of cognitive decline in the elderly (König et al., 2015; Gosztolya et al., 2019), accurate diagnosis of conditions like aphasia (Jamal et al., 2017; Le & Provost, 2016) and effective management of dementia (Xue et al., 2021).

Conclusion

Our case study of the Indonesian-BNT (I-BNT) demonstrates the efficacy of bi-grams in reducing transcription error rates. The provision of a comprehensive list of all possible responses for each test item is a crucial step in successfully implementing ASR within an automated scoring system for I-BNT. The n-gram implementation proves particularly well-suited for ASR systems dealing with a limited and contextual vocabulary. It effectively handles letter-level omissions, additions, and substitutions, enhancing overall transcription accuracy. It is crucial to emphasize that the ASR system should be carefully tailored to meet the specific requirements of the verbal test.

This study indicates also the potential of utilizing solely Wav2Vec2 models for pre-processing speech patterns to identify potential speech impairments. This approach offers a valuable tool for researchers, as it allows for the detection of subtle changes in speech production that might be missed during a traditional interview. Our current dataset primarily focuses on data from healthy individuals. However, the inherent flexibility of the database allows for its adaptation to incorporate audio samples from clinical populations as well. This opens doors for exploring the use of Wav2Vec2-based ASR in identifying and monitoring speech-related disorders within the field of neuropsychology.

The success of ASR in the I-BNT assessment transcends the mere achievement of high accuracy. This study underscores the potential for integrating ASR into the fully automated assessment of a broader range of verbal (neuro)psychological tests. Promising candidates for such integration include the Digit Span, RAVLT, STROOP, Wechsler vocabulary and comprehension subtests, Thematic Apperception Test (TAT), and Verbal Fluency Tests.

Acknowledgements: this research was funded by DIKTI (Directorate of Higher Education General of Indonesia, number: 076/E5/PG.02.00.PL/2023)

Data availability: Data will be made available on request.

References

Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., Farrajota, L., & Martins, I. P. (2013). Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, *27*(6), 1235–1248. https://doi.org/10.1016/j.csl.2012.10.003

Alapetite, A. (2008). Impact of noise and other factors on speech recognition in anaesthesia. *International Journal of Medical Informatics*, *77*(1), 68–77. https://doi.org/10.1016/j.ijmedinf.2006.11.007

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, *9*, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Alsharhan, E., & Ramsay, A. (2020). Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition. *Language Resources and Evaluation*, *54*(4), 975–998. https://doi.org/10.1007/s10579-020-09505-5

Ao, J., & Ko, T. (2021). Improving Attention-based End-to-end ASR by Incorporating an N-gram Neural Network. *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5. https://doi.org/10.1109/ISCSLP49672.2021.9362055

Arisaputra, P., Handoyo, A. T., & Zahra, A. (2024). XLS-R Deep Learning Model for Multilingual ASR on Low- Resource Languages: Indonesian, Javanese, and Sundanese (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2401.06832.

Arisaputra, P., & Zahra, A. (2022). Indonesian Automatic Speech Recognition with XLSR-53. Ingénierie Des Systèmes d Information, 27(6), 973–982. https://doi.org/10.18280/isi.270614.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (arXiv:2006.11477). arXiv. http://arxiv.org/abs/2006.11477

Brown, G. (2016). Automatic Accent Recognition Systems and the Effects of Data on Performance. *The Speaker and Language Recognition Workshop (Odyssey 2016)*, 94–100. https://doi.org/10.21437/Odyssey.2016-14

Chien, Y.-W., Hong, S.-Y., Cheah, W.-T., Yao, L.-H., Chang, Y.-L., & Fu, L.-C. (2019). An Automatic Assessment System for Alzheimer's Disease Based on Speech Using Feature Sequence Generator and Recurrent Neural Network. *Scientific Reports*, *9*(1), 19597. https://doi.org/10.1038/s41598-019-56020-x

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., & Stolcke, A. (2007). Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing*, *5*(1), 1–29. https://doi.org/10.1145/1322391.1322394

Cryssiover, A., & Zahra, A. (2024). Speech recognition model design for Sundanese language using WAV2VEC 2.0. *International Journal of Speech Technology*, *27*(1), 171–177. https://doi.org/10.1007/s10772-023-10066-5

De La Fuente Garcia, S., Ritchie, C. W., & Luz, S. (2020). Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *Journal of Alzheimer's Disease*, *78*(4), 1547–1574. https://doi.org/10.3233/JAD-200888

Favaro, A., Tsai, Y.-T., Butala, A., Thebaud, T., Villalba, J., Dehak, N., & Moro-Velázquez, L. (2023). Interpretable speech features vs. DNN embeddings: What to use in the automatic assessment of Parkinson's disease in multi-lingual scenarios. *Computers in Biology and Medicine*, *166*, 107559. https://doi.org/10.1016/j.compbiomed.2023.107559

Ferdiansyah, V., & Purwarianti, A. (2012). Indonesian Automatic Speech Recognition System Using English-Based Acoustic Model. American Journal of Signal Processing, 2(4), 60–63. https://doi.org/10.5923/j.ajsp.20120204.01.

Fergadiotis, G., Hula, W. D., Swiderski, A. M., Lei, C.-M., & Kellough, S. (2019). Enhancing the Efficiency of Confrontation Naming Assessment for Aphasia Using Computer Adaptive Testing. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1724–1738. https://doi.org/10.1044/2018_JSLHR-L-18-0344

Filippidou, F., & Moussiades, L. (2020). A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (Vol. 583, pp. 73–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-49161-1_7

Garnerin, M., Rossato, S., & Besacier, L. (2019). *Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance* (arXiv:1908.08717). arXiv. http://arxiv.org/abs/1908.08717

Giauque, T. A., Shaughnessy, L., Lin, G., Smith, M., & Press, D. Z. (2023). Developing a brief, accurate and adaptive version of the Boston Naming Test based on high-fidelity estimates of item properties. *Alzheimer's & Dementia*, *19*(S18), e079827. https://doi.org/10.1002/alz.079827

Habeeb, I. Q., Abdulkhudhur, H. N., & Al-Zaydi, Z. Q. (2021). Three N-grams Based Language Model for Auto-correction of Speech Recognition Errors. In A. M. Al-Bakry, S. O. Al-Mamory, M. A. Sahib, L. E. George, J. A. Aldhaibani, H. S. Hasan, & G. S. Oreku (Eds.), New Trends in Information and Communications Technology Applications, Vol. 1511, 131–143. *Springer International Publishing*. https://doi.org/10.1007/978-3-030-93417-0_9

Hidayat, S. (2016). SPEECH RECOGNITION OF KV-PATTERNED INDONESIAN SYLLABLE USING MFCC, WAVELET AND HMM. Kursor, 8(2), 67. https://doi.org/10.28961/kursor.v8i2.63.

Hula, W. D., Kellough, S., & Fergadiotis, G. (2015). Development and Simulation Testing of a Computerized Adaptive Version of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research*, *58*(3), 878–890. https://doi.org/10.1044/2015_JSLHR-L-14-0297

Iglesias, M., Favaro, A., Motley, C., Oh, E. S., Stevens, R. D., Butala, A., Moro-Velazquez, L., & Dehak, N. (2022). Cognitive and Acoustic Speech and Language Patterns Occurring in Different Neurodegenerative Disorders while Performing Neuropsychological Tests. *2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1–6. https://doi.org/10.1109/SPMB55497.2022.10014965

Jamal, N., Shanta, S., Mahmud, F., & Sha'abani, M. (2017). *Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review*. 020028. https://doi.org/10.1063/1.5002046

Kaplan, E., Goodglass, H., & Weintraub, S. (2016). *Boston Naming Test* [Dataset]. https://doi.org/10.1037/t27208-000

Kim, K. W., Lee, S. Y., Choi, J., Chin, J., Lee, B. H., Na, D. L., & Choi, J. H. (2020). A Comprehensive Evaluation of the Process of Copying a Complex Figure in Early- and Late-Onset Alzheimer Disease: A Quantitative Analysis of Digital Pen Data. *Journal of Medical Internet Research*, *22*(8), e18136. https://doi.org/10.2196/18136

Kitzing, P., Maier, A., & Åhlander, V. L. (2009). Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phoniatrics Vocology*, *34*(2), 91–96. https://doi.org/10.1080/14015430802657216

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., & Robert, P. (2018). Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task. *Dementia and Geriatric Cognitive Disorders*, *45*(3–4), 198–209. https://doi.org/10.1159/000487852

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., & David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(1), 112–124. https://doi.org/10.1016/j.dadm.2014.11.012

Kozhirbayev, Z. (2023). Kazakh Speech Recognition: Wav2vec2.0 vs. Whisper. *Journal of Advances in Information Technology*, *14*(6), 1382–1389. https://doi.org/10.12720/jait.14.6.1382-1389

Le, D., & Provost, E. M. (2016). Improving Automatic Recognition of Aphasic Speech with AphasiaBank. *Interspeech 2016*, 2681–2685. https://doi.org/10.21437/Interspeech.2016-213

Maxwell-Smith, Z., & Foley, B. (2023). Automated speech recognition of Indonesian-English language lessons on YouTube using transfer learning. *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 1-16.

Muljono, Syadida, A. Q., Setiadi, D. R. I. M., & Setyono, A. (2017). Sphinx4 for Indonesian continuous speech recognition system. 2017 International Seminar on Application for Technology of Information and Communication (iSemantic), 264–267. https://doi.org/10.1109/ISEMANTIC.2017.8251881.

Ngueajio, M. K., & Washington, G. (2022). Hey ASR System! Why Aren't You More Inclusive?: Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review. In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence* (Vol. 13518, pp. 421–440). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-21707-4_30

Novogrodsky, R., & Kreiser, V. (2015). What can errors tell us about specific language impairment deficits? Semantic and morphological cuing in a sentence completion task. *Clinical Linguistics & Phonetics*, *29*(11), 812–825. https://doi.org/10.3109/02699206.2015.1051239

Pahwa, R., Tanwar, H., & Sharma, D. S. (2020). Speech Recognition System: A review. *International Journal of Future Generation Communication and Networking*, *13*(3).

Pakoci, E., & Popović, B. (2021). Methods for Using Class Based N-gram Language Models in the Kaldi Toolkit. In A. Karpov & R. Potapova (Eds.), *Speech and Computer* (Vol. 12997, pp. 492–503). Springer International Publishing. https://doi.org/10.1007/978-3-030-87802-3_45

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

Pragt, L., van Hengel, P., Grob, D., & Wasmann, J.-W. A. (2022). Preliminary Evaluation of Automated Speech Recognition Apps for the Hearing Impaired and Deaf. *Frontiers in Digital Health*, *4*, 806076. https://doi.org/10.3389/fdgth.2022.806076

Prakoso, H., Ferdiana, R., & Hartanto, R. (2016). Indonesian Automatic Speech Recognition system using CMUSphinx toolkit and limited dataset. 2016 International Symposium on Electronics and Smart Devices (ISESD), 283–286. https://doi.org/10.1109/ISESD.2016.7886734.

Prasad, A., & Jyothi, P. (2020). How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3739–3753. https://doi.org/10.18653/v1/2020.acl-main.345

Radzikowski, K., Wang, L., Yoshie, O., & Nowak, R. (2021). Accent modification for speech recognition of non-native speakers using neural style transfer. *EURASIP Journal on Audio, Speech, and Music Processing*, *2021*(1), 11. https://doi.org/10.1186/s13636-021-00199-3

Rajnoha, J., & Pollák, P. (2011). *ASR systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness*. *Radioengineering*, *20*(1).

Rodrigues, A., Santos, R., Abreu, J., Beça, P., Almeida, P., & Fernandes, S. (2019). Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender. *Proceedings of the XX International Conference on Human Computer Interaction*, 1–8. https://doi.org/10.1145/3335595.3335635

Roshanzamir, A., Aghajan, H., & Soleymani Baghshah, M. (2021). Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, *21*(1), 92. https://doi.org/10.1186/s12911-021-01456-3

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). *wav2vec: Unsupervised Pre-training for Speech Recognition* (arXiv:1904.05862). arXiv. http://arxiv.org/abs/1904.05862

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, *158*, 107020. https://doi.org/10.1016/j.apacoust.2019.107020

Silber Varod, V., Siegert, I., Jokisch, O., Sinha, Y., & Geri, N. (2021). A cross-language study of speech recognition systems for English, German, and Hebrew. *Online Journal of Applied Knowledge Management*, *9*(1), 1–15. https://doi.org/10.36965/OJAKM.2021.9(1)1-15

Sriram, A., Auli, M., & Baevski, A. (2022). Wav2Vec-Aug: Improved self-supervised training with limited data (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2206.13654

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary, Third Edition (3 rd). *Oxford University Press.*

Tits, N., Haddad, K. E., & Dutoit, T. (2018). ASR-based Features for Emotion Recognition: A Transfer Learning Approach (arXiv:1805.09197). *arXiv*. http://arxiv.org/abs/1805.09197

Toth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., Pakaski, M., & Kalman, J. (2018). A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. *Current Alzheimer Research*, *15*(2), 130–138. https://doi.org/10.2174/1567205014666171121114930

Tran, B. D., Mangu, R., Tai-Seale, M., Lafata, J. E., & Zheng, K. (2022). Automatic speech recognition performance for digital scribes: A performance comparison between general-purpose and specialized models tuned for patient-clinician conversations. *Annual Symposium Proceedings. AMIA Symposium*, *2022*, 1072–1080.

Vaessen, N., & van Leeuwen, D. A. (2022). Fine-Tuning Wav2Vec2 for Speaker Recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7967–7971. https://doi.org/10.1109/ICASSP43922.2022.9746952

van den Noort, M., Bosch, P., Haverkort, M., & Hugdahl, K. (2008). A Standard Computerized Version of the Reading Span Test in Different Languages. *European Journal of Psychological Assessment*, *24*(1), 35–42. https://doi.org/10.1027/1015-5759.24.1.35

van Leeuwen, D. A., Hinskens, F., Martinovic, B., van Hessen, A., Grondelaers, S., & Orr, R. (2016). Sprekend Nederland: A heterogeneous speech data collection. *Computational Linguistics in the Netherlands Journal*, *6*, 21–38.

Wahyuningrum, S. E., van Luijtelaar, G., & Sulastri, A. (2021). An online platform and a dynamic database for neuropsychological assessment in Indonesia. *Applied Neuropsychology: Adult*, 30(3), 330–339. https://doi.org/10.1080/23279095.2021.1943397

Wirawan, C. (2021). Wav2Vec2-Large-XLSR-Indonesian retrieved from https://huggingface.co/indonesian-nlp/wav2vec2-large-xlsr-indonesia

Woods, D. L., Wyma, J. M., Herron, T. J., & Yund, E. W. (2016). Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PLOS ONE*, *11*(12), e0166439. https://doi.org/10.1371/journal.pone.0166439

Xue, C., Karjadi, C., Paschalidis, I. Ch., Au, R., & Kolachalama, V. B. (2021). Detection of dementia on voice recordings using deep learning: A Framingham Heart Study. *Alzheimer's Research & Therapy*, *13*(1), 146. https://doi.org/10.1186/s13195-021-00888-3

Yadav, I. C., Shahnawazuddin, S., Govind, D., & Pradhan, G. (2018). Spectral Smoothing by Variational Mode Decomposition and its Effect on Noise and Pitch Robustness of ASR System. *ICASSP IEEE Xplore*.

Zhang, L., & Osth, Adam. F. (2024). Modelling orthographic similarity effects in recognition memory reveals support for open bigram representations of letter coding. *Cognitive Psychology*, *148*, 101619. https://doi.org/10.1016/j.cogpsych.2023.101619

Zhang, X., Lv, L., Min, G., Wang, Q., Zhao, Y., & Li, Y. (2021). Overview of the Complex Figure Test and Its Clinical Application in Neuropsychiatric Disorders, Including Copying and Recall. *Frontiers in Neurology*, *12*, 680474. https://doi.org/10.3389/fneur.2021.680474

Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., & Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods*, *50*(6), 2597–2605. https://doi.org/10.3758/s13428-018-1037-4

CHAPTER 6

Summary and General Discussion

The development of neuropsychological testing from face-to-face paper and pencil to various types of computerized approaches has opened up a new frontier of collaboration between computer scientists and neuropsychologists (Karyotaki et al., 2021; Karyotaki & Drigas, 2015; Miller & Barr, 2017). This convergence has resulted in significant advancements in the field, offering benefits for both research and clinical practice (Carotenuto et al., 2021; Mancioppi et al., 2019).

Computer science is revolutionizing the field of neuropsychology by introducing innovations that improve the efficiency, accuracy, and accessibility of Neuropsychological tests (Bilder & Reise, 2019; Miller, 2019). In clinical practice, computerization facilitates data visualization and representation and facilitates the analysis and interpretation of the results (Langer et al., 2022; Vogt et al., 2019). Automated scoring systems further increase efficiency while reducing the risk of human error (Cavedoni et al., 2020; Chen et al., 2020; Kang et al., 2019; Zhang et al., 2021). In addition, advancements in artificial intelligence have the potential to improve the early detection of cognitive decline.

On the technical side, the availability of large digital datasets enables streamlined data collection from both healthy individuals and clinical populations. This digitization simplifies data management and allows for the automated generation of normative scores (Aoki et al., 2023; De Vent et al., 2016; Kiselica et al., 2020; Ruan et al., 2020; Shirk et al., 2011). Moreover, tele-neuropsychology extends access to consultancy, assessment, and care for individuals in remote areas (Carotenuto et al., 2021; Takenga et al., 2020).

This study explored the advantages of using computer science techniques in developing normative scores for neuropsychological battery tests specifically tailored to the Indonesian population. Additionally, the study investigated the potential of automated scoring systems to improve the objectivity and consistency of the evaluation of neuropsychological tests.

Database and Online Platform I-ANDI

Building upon the original Dutch Advanced Neuropsychological Diagnostic Infrastructures ANDI (De Vent et al., 2016), a specialized database (I-ANDI) was designed and developed for Indonesian neuropsychological data. Due to the sensitivity of demographic factors in neuropsychological test results and the concern that standards from Western countries might not accurately represent the current Indonesian population, the I-ANDI system helps in generating culturally appropriate normative scores. This will improve the validity and reliability of

neuropsychological assessments in Indonesia (Fernández & Abe, 2018). Chapter 2 summarizes the functions of I-ANDI, which include an online platform, a dynamic database, and generated adjusted normative scores. Normative scores for neuropsychological tests are commonly adjusted for demographic data, level of formal education, age, and sometimes sex.

The I-ANDI online platform offers a promising solution for advancing neuropsychological testing in Indonesia, a country characterized by its archipelagic geography (Fox-Fuller et al., 2022; Jacobsen et al., 2003; Sumpter et al., 2023). This web-based platform offers convenient accessibility from any device, such as a computer, tablet, or smartphone (Bertini et al., 2023; Demeyere et al., 2021). Researchers and neuropsychologists can leverage this flexibility to access the platform anytime, anywhere. Additionally, the platform utilizes a single data input method, ensuring standardized storage of neuropsychological data. This uniformity fosters data consistency and facilitates analysis (Kiselica et al., 2020). Currently, there are ten neuropsychological tests accommodated in the dynamic database. The system is designed to be flexible and adaptable to other tests commonly used in Indonesia, such as the Indonesian version of the Wechsler Adult Intelligence Scale - Fourth Edition (Suwartono et al., 2014) and Performance Validity Tests (PVT) (Adhiatma et al., 2023).

This platform offers significant advantages in the interpretation of psychological tests in Indonesia. By providing adjusted normative scores, neuropsychologists can more accurately tailor assessments and interpretation of test results to the demographic factors of their clients (Aoki et al., 2023; De Vent et al., 2016; Kiselica et al., 2020; Ruan et al., 2020; Shirk et al., 2011). Unlike traditional normative score tables, the platform includes an option to generate predictive normative scores for populations with relatively fewer data points per cell, by combining variables by a more flexible selection of demographic factors (Bridges & Holler, 2007; Oosterhuis et al., 2016; Tombaugh, 2004). Additionally, the platform can serve as an alternative for generating updated normative scores based on the cumulative data in the system (D'Alfonso, 2020; Dash et al., 2019). The dynamic database provides further advantages, including the flexibility to incorporate data from both current neuropsychological tests and other (neuro)psychological tests, patient-related medical data, easy data searchability, scalability for handling larger datasets, and the ability to generate sub-datasets for focused research (Le & Provost, 2016; Miller, 2019; Parsons & Duffield, 2020; Smith et al., 2011; Utami et al., 2024; Abimanyu et al., submitted).

The current system has several limitations in data analysis and client profiles. Comparisons between client performance and normative data are currently

limited to the subtest level, and the system is not yet equipped with a module to build a comprehensive cognitive test profile (Burggraaff et al., 2017; Kiselica et al., 2020). Furthermore, the data used in this study is currently limited to the Javanese population, thus failing to represent the Indonesian population. In the future, with the addition of data from various islands, it is expected to represent the cognitive performance of a larger part of the Indonesian population. The absence of patient data also hinders the evaluation of diagnosis, treatment, and prediction of cognitive deficits (Picanco Diniz et al., 2014; Slot et al., 2019). Supporting features such as automated data cleaning and the generation of predictive scores through statistical analysis in I-ANDI are still under development to ensure quality data assurance and quality data control (McLaughlin et al., 2021).

Indonesian Neuropsychological Test Battery (INTB)

Neuropsychologists rely on normative scores to evaluate an individual's cognitive strengths and weaknesses. These scores are established by comparing an individual's test results to those of healthy people with similar demographics, such as age and education (Mitrushina et al., 2005; Strauss et al., 2006). These demographic factors significantly influence cognitive function, making them crucial for accurate interpretation (Kiselica et al., 2024). Chapter three delves into the psychometric properties, normative data, and demographic factors that influence the cognitive performance of INTB within the Indonesian population. Preliminary normative scores for ten neuropsychological tests covering four distinct cognitive domains were established using a sample of 490 healthy Javanese participants. These norms provide a baseline for developing future normative scores representative of the entire Indonesian population. Our study reported in Chapter 3 demonstrated satisfactory test-retest reliability scores, confirming the consistency of these measures over time.

Our findings align with international reports demonstrating that demographic factors such as age and education significantly impact cognitive performance (Mitrushina et al., 2005; Strauss et al., 2006). A general decline in cognitive abilities accompanies aging, while higher education levels correlate with enhanced performance (Falch & Sandgren Massih, 2011; Le Carret et al., 2003; Lovden et al., 2020). We also highlight the variability of these factors across cognitive domains. For instance, language and executive functions show improvements with increased education, whereas memory and learning domains exhibit less pronounced relationships (Pagliarin et al., 2014; Park et al., 2015; Rohde et al., 2022). These findings underscore the necessity of considering demographic variables when interpreting neuropsychological assessment results for the Indonesian population.

Chapter 3b delves into the influence of age and education on the INTB's underlying cognitive domains in the Indonesian population. The study employed Principal Component Analysis (PCA) on data collected from 490 participants to identify underlying cognitive patterns within the INTB revealing seven distinct cognitive constructs, presumably representing different cognitive abilities. Our findings revealed distinct age-dependent trends for different cognitive constructs. While constructs related to visuospatial information, planning, speed, inhibitory control, attention, auditory short-term, and working memory exhibited a linear decline (Cohen et al., 2019; Glisky, 2007), visually cued semantic processes and learning ability demonstrated a quadratic trend (Lezak et al., 2012; Tucker-Drob et al., 2019). In addition, each construct exhibited a unique pattern of age dependency. . Education effects were generally more pronounced than age effects (Picanco Diniz et al., 2014; Ruan et al., 2020). In particular, the interaction between education and age was evident in the construct measuring attention, auditory short-term, and working memory. Individuals with a senior high or junior high education experienced a sharp decline in these cognitive functions after the age of 30, unlike undergraduates, for whom the decline was not significant, and if it occurred, it started later. This emphasizes the relevance of education in mitigating cognitive decline (Park et al., 2023). The interaction between education and age underlines the relevance of education in preventing early aging.

This study provides insights into the differential effects of aging and education on seven cognitive abilities. By leveraging this knowledge, it might be possible to develop tailored strategies to promote and maintain cognitive health throughout life. Our results may also suggest that educational programs designed to promote cognitive development could be beneficial across the lifespan (Picanco Diniz et al., 2014; Ruan et al., 2020). Notably, higher education levels potentially protect against age-related decline in attention and working memory, implying that lifelong learning might be a viable strategy for maintaining cognitive health (Lovden et al., 2020; Oosterman et al., 2021).

Although the current study offers valuable findings, its restricted focus on a single Indonesian island limits the extent to which its conclusions can be applied to the broader Indonesian population. Currently, a consortium of neuropsychologists in Indonesia is actively collecting data from various regions across the archipelago. To comprehensively understand cognitive functions in Indonesia, future research should address several key areas. First, comparative studies between different cultural groups are essential to identify and mitigate cultural biases in neurocognitive tests (Al-Jawahiri & Nielsen, 2021; Fujii, 2018).

Second, while some studies have found no significant differences between groups that use one or two spoken languages in daily life (Immanuel et al., 2024; Pesau et al., 2023), these findings do not guarantee applicability to the entire Indonesian population (Franzen et al., 2023). Therefore, it is important to investigate whether differences in cognitive performance are related to ethnic groups or influenced by bilingualism and multilingualism across Indonesia's diverse linguistic landscape (Pagliarin et al., 2014). To build local research capacity and ensure the sustainability of neuropsychological assessment efforts, online training programs for clinicians conducting neuropsychological research and test development should be prioritized. Additionally, establishing normative data for cognitive abilities tailored to the Indonesian population is necessary to accurately evaluate cognitive growth and decline.

To enhance accessibility and efficiency in neuropsychological assessment, especially in resource-limited regions, technology integration is imperative (Carotenuto et al., 2021; Takenga et al., 2020). A centralized database facilitates collaboration among neuropsychologists in Indonesia. By enabling data and resource sharing, it empowers researchers to analyze regional trends in cognitive function. Existing sub-datasets, such as those used by Pesau et al. (2023) on language domain tests, by Immanuel et al. (2024) on non-language cognitive domains, Utami et al. (2024) on Rey Auditory Verbal Learning Test data, Widhianingtanti et al. (2022) on the Trail Making Test data, and (Wulanyani et al., 2024) on executive function tests exemplify the platform's potential for advancing research.

Computer Vision for Visuo-Spatial Tests

A persistent challenge in neuropsychological test scoring lies in the subjectivity inherent to evaluating hand-drawn responses from the client's reproduction of visually presented test materials. This subjectivity can hinder the standardization of scoring across different examiners. Several approaches can be explored to address this limitation. One possibility is the development of an automated scoring system, such as those utilizing computer vision with feature extraction techniques. Computer vision necessitates human expertise to determine the correctness of relevant features. Computer vision, particularly with feature extraction techniques, presents a promising approach (El-gayar et al., 2013; Fleuret et al., 2011).

Since computer vision excels at identifying sub-objects within images, it aligns well with the demands of scoring visual tests like the Figural Reproduction Test (FRT). For instance, in the FRT's three cards, neuropsychologists assess the constituent parts of the objects. Feature extraction can be employed to identify relevant features

such as orientation, shape, and lines, offering an objective and standardized scoring method (Jamus et al., 2023; Troyer & Wishart, 1997; Zhang et al., 2021).

Chapter 4 validates the effectiveness of computer vision in analyzing the first card of the FRT, demonstrating that the FRT-CVAS accurately identifies objects relevant to scoring criteria with over 91% accuracy using fewer than 1,000 training images. The system excels at recognizing object size, angle, and shape direction. While improvements are needed in determining triangle thresholds, the primary challenge lies in the inter-rater variability in determining the threshold tolerance, which hinders the widespread adoption of computer vision in FRT scoring.

A limitation of this study is the scope of the newly developed system, which is currently limited to identifying only one stimulus card. Future research should focus on expanding the system to include the remaining two cards (consisting of three objects). Additionally, our findings highlight the need for a more detailed, transparent, and unbiased scoring system. As recommended by Moetesum et al. (2022), a collaborative effort between system developers and neuropsychologists is essential to achieve this. The meticulous approach adopted by Di Febbo et al. (2023) in defining expanded scoring rules as in the manual of the FRT, provides a valuable blueprint for future developments in this area.

Enhancing neuropsychological assessment in Indonesia requires technological integration. Computerized visual tests, and traditional paper-and-pencil tests like the TMT, the Five Point Test, and Rey's Complex Figure Test are other tests that can be modernized by using tablets. This digital shift enhances accuracy and efficiency in the assessment process, however equivalence between the traditional way and digital way of assessment needs to be established as had been done in another test (Brown et al., 2018; Elderson et al., 2016). Additionally, digitizing the assessment and scoring helps the scoring procedure to be more objective and reduces time and burden of the tester on the whole assessment process.

Automated Speech to Text for Verbal Tests

Chapter 5 investigated the potential for adapting Automated Speech Recognition (ASR) models specifically to the Indonesian language. These models offer significant advantages for clinicians by automatically transcribing spoken responses from test subjects into text format. This demonstrably reduced the burden and time required for test administration, freeing valuable clinician time (König et al., 2015, 2018). The study successfully employed the Wav2vec2 decoding method, eliminating the need for a large corpus of speech data for training (Baevski et al., 2020; Schneider et al.,

2019). In addition, this method simplified the adaptation process for Indonesians by avoiding the need to build training data from scratch.

Furthermore, this study identified two distinct approaches to ASR models that could be beneficial for clinicians when using verbal tests as used in the INTB. The first approach, utilizing only the Wav2vec2 model, has proven to be adept at capturing nuanced aspects of speech, such as pauses, intonation, and speech rate. This feature proved invaluable in clinical settings, allowing the analysis of speech disorders in both children and adults (Iglesias et al., 2022; Kitzing et al., 2009). The second approach combined Wav2vec2 with a bi-gram model, which is better suited to tests measuring phonemic speech (individual word sounds) (Ao & Ko, 2021; Habeeb et al., 2021; Pakoci & Popović, 2021). This combination should work well for specific, limited-context tests like the Digit Span (10-digit vocabulary), RAVLT (15-word list), and Stroop test (color names). The advantages of a speech-to-text system include the ability to automatically score responses, ensuring consistent and objective evaluation, as well as providing accurate and precise time tracking for each verbal task. Additionally, the system can collect audio recordings of the test-takers speech, which can be stored and analyzed for future research, offering valuable data for studying speech patterns, cognitive performance, or linguistic abilities over time.

This chapter explained how an ASR model is capable of recognizing various Indonesian accents. The research results indicate that the ASR model that we developed and evaluated can recognize three Indonesian accents speaking Bahasa Indonesia with a good accuracy level. Moreover, the scores of male and female voices were equally correct, demonstrating the robustness of the ASR model. However, this model still has several limitations, particularly in terms of data representation from all regions of Indonesia. To address this, further research should focus on collecting more comprehensive data and developing a more advanced automated scoring system.

Future Research

The development of the Indonesia Neuropsychological Database offers a significant leap forward for neuropsychological research in the country. One of its key strengths lies in its ability to serve as a centralized repository for user data, mirroring the success of similar databases for specific conditions like aphasia (Casaletto & Heaton, 2017; Zucchella et al., 2018), dementia (https://dementia.talkbank.org/), and Alzheimer's (https://naccdata.org/). This centralized approach unlocks several valuable opportunities for future research.

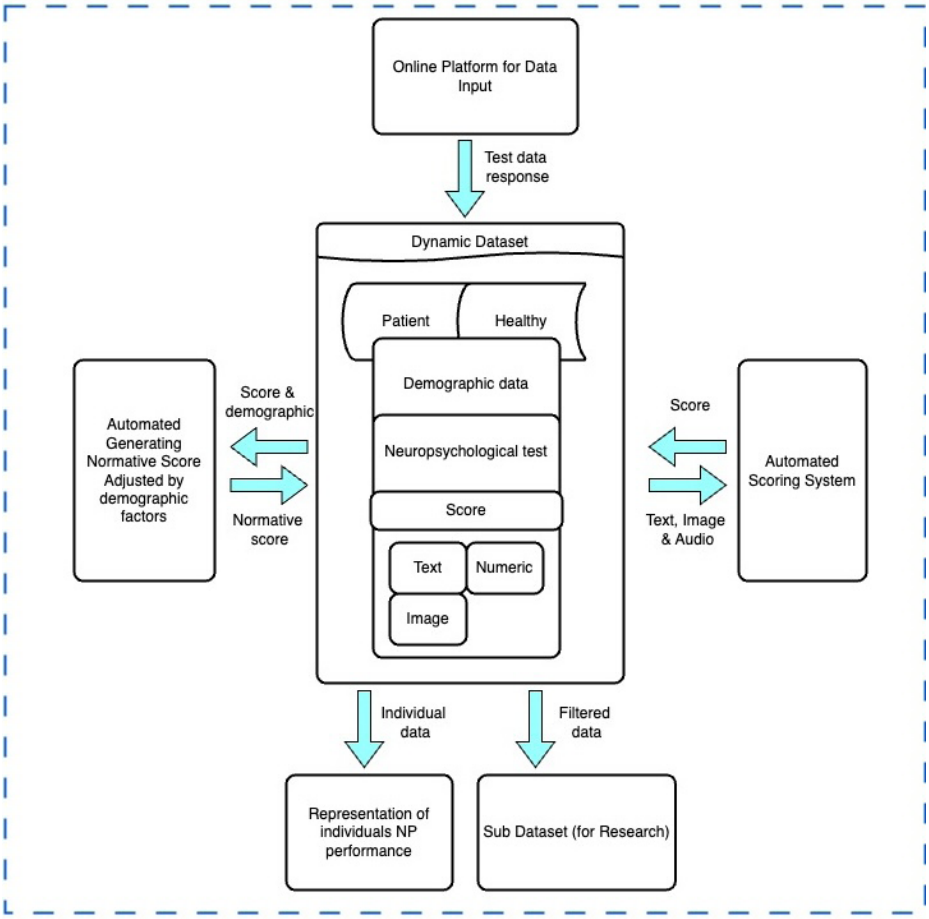


Figure 1 showcases how the existing system serves as a foundation to accommodate the future needs of neuropsychologists, other clinicians, researchers, and patients.

I-ANDI is a platform designed to store neuropsychological test data, along with existing data from various regions in Indonesia, including cities from Java (Jakarta, Bandung, Semarang, Surabaya), Bali (Denpasar, Bandung), Kalimantan (Samarinda), and Sulawesi (Makassar). The platform is used to generate normative scores based on the data in the I-ANDI database, focusing initially on data representation from Java. This platform can be found in the www.norm.indonesian-andi.id.

The Indonesian-ANDI platform has shown significant potential in advancing the field of neuropsychology in Indonesia through the digitalization of data. Moving forward, future research will concentrate on several issues:

1. Geographical Representation: Expanding data collection to cover all regions and islands in Indonesia, including rural and remote areas. This will ensure a more representative understanding of the entire Indonesian population.
2. Social Diversity: Collecting data from various groups based on age, education level, socio-economic status, ethnicity, language, occupation, and minority groups. This aims to provide a comprehensive view of neuropsychological conditions across different segments of the population.
3. Demographic Variables: Incorporating additional demographic factors such as language, ethnicity, and migration status to gain deeper insights into potential neurocognitive differences influenced by these variables.
4. Clinical Data Integration: Including data from various cognitive impairments will enhance predictive models for more accurate diagnosis. This data will consist of medical histories, cognitive test results, and intervention responses, combined with population data to better analyze patterns in neurocognitive disorders.
5. Machine Learning Algorithms: Leveraging machine learning to analyze increasingly complex neuropsychological data, identifying patterns and insights that may not yet be visible through current research methods. Furthermore, developing predictive models to identify individuals at risk of neurocognitive disorders early, while also tracking the progression of these conditions over time.
6. Cognitive Profile Evaluations: Assessing cognitive profiles both individually and regionally, will provide valuable data for policy decisions in different regions of Indonesia.
7. Development of digital assessment:
    - o Extending the number of hand-drawing stimuli: Addressing current limitations, including using a single stimulus image in testing, by focusing future research on a broader range of stimulus types, among others the 5-Point Test.
    - o Automated verbal scoring test: Implementing fully automated scoring for verbal tests, such as the Stroop Test and Digit Span, which use a limited vocabulary, making the process more efficient and standardized.

Through these initiatives, I-ANDI aims to create a more robust, representative, and insightful understanding of neuropsychological health in Indonesia.

Conclusion

The I-ANDI system was designed to assist neuropsychologists by providing a dynamic database and an online platform that simplifies the process of generating fine-tuned normative scores based on specific demographic factors. Its advanced capabilities allow for the calculation of predicted scores in instances where the available reference dataset lacks a sufficient number of subjects. The platform, accessible at http://norm.indonesian-andi.id, currently houses normative data for the Javanese population, based on an initial dataset of 490 individuals. This data collection is ongoing, with the current dataset exceeding 1,000.

Cognitive test scores for the Javanese population are strongly influenced by both age and education. Elderly participants consistently performed the lowest across all age groups, while individuals with less education scored poorly on most tests. Language-based tests were the most sensitive to educational background, while attention, executive function, memory, and learning abilities stabilized after approximately ten years of schooling. These findings emphasize the need for normative data tailored to different age and educational levels within the Javanese population. However, establishing definitive normative scores across Indonesia presents challenges due to the country's ethnic and linguistic diversity. Seven distinct cognitive constructs displayed varying patterns of age-related decline, with education having a greater impact than age. The interaction between age and education highlights the significant role of education in delaying early cognitive decline.

In the assessment of visual-spatial abilities, human judgments of hand-drawn geometrical figures, particularly in the Figure Reproduction Test (FRT), tend to be subjective, especially when evaluating shapes like triangles. The introduction of FRT-CVAS, a computer vision-based approach, reduces this subjectivity by providing detailed, standardized, and objective analyses of hand-drawn elements. Scoring accuracy improves when using single criteria rather than the more complex compound criteria suggested in the manual, underscoring the importance of clear and detailed scoring guidelines for achieving consistent results.

Lastly, the study on the verbal test, the Indonesian-BNT (I-BNT), highlights the effectiveness of using bi-grams to reduce transcription errors, particularly when integrated with Automatic Speech Recognition (ASR) systems. ASR technology, which converts spoken language into text, often struggles with accuracy in limited vocabulary scenarios or when encountering regional accents. However,

the incorporation of bi-grams and N-gram techniques has proven effective in addressing common issues such as omissions, additions, and substitutions, significantly improving transcription accuracy. Additionally, the Wav2Vec2 models used in this research show potential for pre-processing speech and detecting subtle impairments in otherwise healthy individuals. This opens up exciting possibilities for ASR in broader clinical applications, particularly for identifying early signs of speech-related disorders. Beyond I-BNT, the integration of ASR holds promise for automating and improving the accuracy of other verbal neuropsychological tests, such as the Digit Span, RAVLT, STROOP, and Verbal Fluency Tests, making assessments more efficient and reliable.

References

Abimanyu, C.V.R., Wahyuningrum, S.E., Goeritno, H., Sulastri, A., van Luijtelaar, G., (2024). Submitted.

Adhiatma, W., Hendriks, M. P. H., Halim, M. S., & Kessels, R. P. C. (2023). The Development of a Performance Validity Test (PVT) for Indonesia. *Buletin Psikologi*, *31*(1), 94. https://doi.org/10.22146/buletinpsikologi.73390

Al-Jawahiri, F., & Nielsen, T. R. (2021). Effects of Acculturation on the Cross-Cultural Neuropsychological Test Battery (CNTB) in a Culturally and Linguistically Diverse Population in Denmark. *Archives of Clinical Neuropsychology*, *36*(3), 381–393. https://doi.org/10.1093/arclin/acz083

Ao, J., & Ko, T. (2021). Improving Attention-based End-to-end ASR by Incorporating an N-gram Neural Network. *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5. https://doi.org/10.1109/ISCSLP49672.2021.9362055

Aoki, S., Nagatani, F., Kagitani-Shimono, K., Ohno, Y., Taniike, M., & Mohri, I. (2023). Examining normative values using the Cambridge neuropsychological test automated battery and developmental traits of executive functions among elementary school-aged children in Japan. *Frontiers in Psychology*, *14*, 1141628. https://doi.org/10.3389/fpsyg.2023.1141628

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (arXiv:2006.11477). *arXiv*. http://arxiv.org/abs/2006.11477

Bertini, F., Beltrami, D., Barakati, P., Calzà, L., Ghidoni, E., & Montesi, D. (2023). A Web-Based Application for Screening Alzheimer's Disease in the Preclinical Phase. *2023 IEEE Symposium on Computers and Communications (ISCC)*, 1–6. https://doi.org/10.1109/ISCC58397.2023.10218229

Bilder, R. M., & Reise, S. P. (2019). Neuropsychological tests of the future: How do we get there from here?. *The Clinical Neuropsychologist*, *33*(2), 220–245. https://doi.org/10.1080/13854046.2018.1521993

Bridges, A. J., & Holler, K. A. (2007). How Many is Enough? Determining Optimal Sample Sizes for Normative Studies in Pediatric Neuropsychology. *Child Neuropsychology*, *13*(6), 528–538. https://doi.org/10.1080/09297040701233875

Brown, F. C., O'Connor, B. P., Vitelli, K. M., Heinly, M., Rommel, G. C., & Davis, R. N. (2018). Comparison of the Computer and Hand Administered Versions of the Brown Location Test (BLT). *Archives of Clinical Neuropsychology*, *33*(1), 47–56. https://doi.org/10.1093/arclin/acx049

Burggraaff, J., Knol, D. L., & Uitdehaag, B. M. J. (2017). Regression-Based Norms for the Symbol Digit Modalities Test in the Dutch Population: Improving Detection of Cognitive Impairment in Multiple Sclerosis. *European Neurology*, *77*(5–6), 246–252. https://doi.org/10.1159/000464405

Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological Assessment: Past and Future. *Journal of the International Neuropsychological Society*, *23*(9–10), 778–790. https://doi.org/10.1017/S1355617717001060

Carotenuto, A., Traini, E., Fasanaro, A. M., Battineni, G., & Amenta, F. (2021). Tele-Neuropsychological Assessment of Alzheimer's Disease. *Journal of Personalized Medicine*, *11*(8), 688. https://doi.org/10.3390/jpm11080688

Cavedoni, S., Chirico, A., Pedroli, E., Cipresso, P., & Riva, G. (2020). Digital Biomarkers for the Early Detection of Mild Cognitive Impairment: Artificial Intelligence Meets Virtual Reality. *Frontiers in Human Neuroscience*, *14*, 245. https://doi.org/10.3389/fnhum.2020.00245

Chen, S., Stromer, D., Alabdalrahim, H. A., Schwab, S., Weih, M., & Maier, A. (2020). Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, *10*(1), 20854. https://doi.org/10.1038/s41598-020-74710-9

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

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology*, *36*, 112–117. https://doi.org/10.1016/j.copsyc.2020.04.005

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data*, *6*(1), 54. https://doi.org/10.1186/s40537-019-0217-0

De Vent, N. R., Agelink Van Rentergem, J. A., Schmand, B. A., Murre, J. M. J., Huizenga, H. M., & ANDI Consortium. (2016). Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A Normative Database Created from Control Datasets. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01601

Demeyere, N., Haupt, M., Webb, S. S., Strobel, L., Milosevich, E. T., Moore, M. J., Wright, H., Finke, K., & Duta, M. D. (2021). Introducing the tablet-based Oxford Cognitive Screen-Plus (OCS-Plus) as an assessment tool for subtle cognitive impairments. *Scientific Reports*, *11*(1), 8000. https://doi.org/10.1038/s41598-021-87287-8

Di Febbo, D., Ferrante, S., Baratta, M., Luperto, M., Abbate, C., Trimarchi, P. D., Giunco, F., & Matteucci, M. (2023). A decision support system for Rey–Osterrieth complex figure evaluation. *Expert Systems with Applications*, *213*, 119226. https://doi.org/10.1016/j.eswa.2022.119226

Elderson, M. F., Pham, S., Van Eersel, M. E. A., LifeLines Cohort Study, Wolffenbuttel, B. H. R., Kok, J., Gansevoort, R. T., Tucha, O., Van Der Klauw, M. M., Slaets, J. P. J., & Izaks, G. J. (2016). Agreement between Computerized and Human Assessment of Performance on the Ruff Figural Fluency Test. *PLOS ONE*, *11*(9), e0163286. https://doi.org/10.1371/journal.pone.0163286

El-gayar, M. M., Soliman, H., & Meky, N. (2013). A comparative study of image low level feature extraction algorithms. *Egyptian Informatics Journal*, *14*(2), 175–181. https://doi.org/10.1016/j.eij.2013.06.003

Falch, T., & Sandgren Massih, S. (2011). The Effect of Education on Cognitive Ability. *Economic Inquiry*, *49*(3), 838–856. https://doi.org/10.1111/j.1465-7295.2010.00312.x

Fernández, A. L., & Abe, J. (2018). Bias in cross-cultural neuropsychological testing: Problems and possible solutions. *Culture and Brain*, *6*(1), 1–35. https://doi.org/10.1007/s40167-017-0050-2

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, *108*(43), 17621–17625. https://doi.org/10.1073/pnas.1109168108

Fox-Fuller, J. T., Rizer, S., Andersen, S. L., & Sunderaraman, P. (2022). Survey Findings About the Experiences, Challenges, and Practical Advice/Solutions Regarding Teleneuropsychological Assessment in Adults. *Archives of Clinical Neuropsychology*, *37*(2), 274–291. https://doi.org/10.1093/arclin/acab076

Franzen, S., Van Den Berg, E., Bossenbroek, W., Kranenburg, J., Scheffers, E. A., Van Hout, M., Van De Wiel, L., Goudsmit, M., Van Bruchem-Visser, R. L., Van Hemmen, J., Jiskoot, L. C., & Papma, J. M. (2023). Neuropsychological assessment in the multicultural memory clinic: Development and feasibility of the TULIPA battery. *The Clinical Neuropsychologist*, *37*(1), 60–80. https://doi.org/10.1080/13854046.2022.2043447

Fujii, D. E. M. (2018). Developing a cultural context for conducting a neuropsychological evaluation with a culturally diverse client: The ECLECTIC framework. *The Clinical Neuropsychologist*, *32*(8), 1356–1392. https://doi.org/10.1080/13854046.2018.1435826

Glisky, E. (2007). Changes in Cognitive Function in Human Aging. In D. Riddle (Ed.), *Brain Aging* (Vol. 20072731, pp. 3–20). CRC Press. https://doi.org/10.1201/9781420005523.sec1

Habeeb, I. Q., Abdulkhudhur, H. N., & Al-Zaydi, Z. Q. (2021). Three N-grams Based Language Model for Auto-correction of Speech Recognition Errors. In A. M. Al-Bakry, S. O. Al-Mamory, M. A.

Sahib, L. E. George, J. A. Aldhaibani, H. S. Hasan, & G. S. Oreku (Eds.), *New Trends in Information and Communications Technology Applications* (Vol. 1511, pp. 131–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-93417-0_9

Iglesias, M., Favaro, A., Motley, C., Oh, E. S., Stevens, R. D., Butala, A., Moro-Velazquez, L., & Dehak, N. (2022). Cognitive and Acoustic Speech and Language Patterns Occurring in Different Neurodegenerative Disorders while Performing Neuropsychological Tests. *2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1–6. https://doi.org/10.1109/SPMB55497.2022.10014965

Immanuel, A. S., Pesau, H. G., Wulanyani, N. M. S., Sulastri, A., & Van Luijtelaar, G. (2024). The Role of Spoken Language on Performance of Cognitive Tests: The Indonesian Experience. *Journal of Cognition and Culture*, *24*(3–4), 207–240. https://doi.org/10.1163/15685373-12340187

Jacobsen, S. E., Sprenger, T., Andersson, S., & Krogstad, J.-M. (2003). Neuropsychological assessment and telemedicine: A preliminary study examining the reliability of neuropsychology services performed via telecommunication. *Journal of the International Neuropsychological Society*, *9*(3), 472–478. https://doi.org/10.1017/S1355617703930128

Jamus, D. R., Mäder-Joaquim, M. J., De Paula Souza, L., De Paola, L., Claro-Höpker, C. D., Terra, V. C., & Soares Silvado, C. E. (2023). Rey–Osterrieth complex figure test: Comparison of traditional and qualitative scoring systems after unilateral temporal lobectomy. *The Clinical Neuropsychologist*, *37*(2), 416–431. https://doi.org/10.1080/13854046.2022.2047790

Kang, M. J., Kim, S. Y., Na, D. L., Kim, B. C., Yang, D. W., Kim, E.-J., Na, H. R., Han, H. J., Lee, J.-H., Kim, J. H., Park, K. H., Park, K. W., Han, S.-H., Kim, S. Y., Yoon, S. J., Yoon, B., Seo, S. W., Moon, S. Y., Yang, Y., … Youn, Y. C. (2019). Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data. *BMC Medical Informatics and Decision Making*, *19*(1), 231. https://doi.org/10.1186/s12911-019-0974-x

Karyotaki, E., Efthimiou, O., Miguel, C., Maas Genannt Bermpohl, F., Furukawa, T. A., Cuijpers, P., Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, Riper, H., Patel, V., Mira, A., Gemmil, A. W., Yeung, A. S., Lange, A., Williams, A. D., Mackinnon, A., Geraedts, A., Van Straten, A., Meyer, B., Björkelund, C., et al. (2021). Internet-Based Cognitive Behavioral Therapy for Depression: A Systematic Review and Individual Patient Data Network Meta-analysis. *JAMA Psychiatry*, *78*(4), 361. https://doi.org/10.1001/jamapsychiatry.2020.4364

Karyotaki, M., & Drigas, A. (2015). Online and other ICT Applications for Cognitive Training and Assessment. *International Journal of Online and Biomedical Engineering (iJOE)*, *11*(2), 36. https://doi.org/10.3991/ijoe.v11i2.4360

Kiselica, A. M., Karr, J. E., Mikula, C. M., Ranum, R. M., Benge, J. F., Medina, L. D., & Woods, S. P. (2024). Recent Advances in Neuropsychological Test Interpretation for Clinical Practice. *Neuropsychology Review*, *34*(2), 637–667. https://doi.org/10.1007/s11065-023-09596-1

Kiselica, A. M., Webber, T. A., & Benge, J. F. (2020). The Uniform Dataset 3.0 Neuropsychological Battery: Factor Structure, Invariance Testing, and Demographically Adjusted Factor Score Calculation. *Journal of the International Neuropsychological Society*, *26*(6), 576–586. https://doi.org/10.1017/S135561772000003X

Kitzing, P., Maier, A., & Åhlander, V. L. (2009). Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phoniatrics Vocology*, *34*(2), 91–96. https://doi.org/10.1080/14015430802657216

König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., & Robert, P. (2018). Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task. *Dementia and Geriatric Cognitive Disorders*, *45*(3–4), 198–209. https://doi.org/10.1159/000487852

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., & David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(1), 112–124. https://doi.org/10.1016/j.dadm.2014.11.012

Langer, N., Weber, M., Vieira, B. H., Strzelczyk, D., Wolf, L., Pedroni, A., Heitz, J., Müller, S., Schultheiss, C., Tröndle, M., Lasprilla, C. A., Rivera, D., Scarpina, F., Zhao, Q., Leuthold, R., Jenni, O. G., Brugger, P., Zaehle, T., Lorenz, R., & Zhang, C. (2022). The AI Neuropsychologist: Automatic scoring of memory deficits with deep learning. *Https://Www.Biorxiv.Org/Content/10.1101/2022.06.15.496291v4*. https://doi.org/10.1101/2022.06.15.496291

Le Carret, N., Lafont, S., Letenneur, L., Dartigues, J.-F., Mayo, W., & Fabrigoule, C. (2003). The Effect of Education on Cognitive Performances and Its Implication for the Constitution of the Cognitive Reserve. *Developmental Neuropsychology*, *23*(3), 317–337. https://doi.org/10.1207/S15326942DN2303_1

Le, D., & Provost, E. M. (2016). Improving Automatic Recognition of Aphasic Speech with AphasiaBank. *Interspeech 2016*, 2681–2685. https://doi.org/10.21437/Interspeech.2016-213

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). Neuropsychological assessment (Fifth edition). *Oxford University Press*.

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

Mancioppi, G., Fiorini, L., Timpano Sportiello, M., & Cavallo, F. (2019). Novel Technological Solutions for Assessment, Treatment, and Assistance in Mild Cognitive Impairment. *Frontiers in Neuroinformatics*, *13*, 58. https://doi.org/10.3389/fninf.2019.00058

McLaughlin, P. M., Sunderland, K. M., Beaton, D., Binns, M. A., Kwan, D., Levine, B., Orange, J. B., Peltsch, A. J., Roberts, A. C., Strother, S. C., & Troyer, A. K. (2021). The Quality Assurance and Quality Control Protocol for Neuropsychological Data Collection and Curation in the Ontario Neurodegenerative Disease Research Initiative (ONDRI) Study. *Assessment*, *28*(5), 1267–1286. https://doi.org/10.1177/1073191120913933

Miller, J. B. (2019). Big data and biomedical informatics: Preparing for the modernization of clinical neuropsychology. *The Clinical Neuropsychologist*, *33*(2), 287–304. https://doi.org/10.1080/13854046.2018.1523466

Miller, J. B., & Barr, W. B. (2017). The Technology Crisis in Neuropsychology. *Archives of Clinical Neuropsychology*, *32*(5), 541–554. https://doi.org/10.1093/arclin/acx050

Mitrushina, M., Boone, K. B., Razani, J., & D'elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed). Oxford University Press.

Moetesum, M., Diaz, M., Masroor, U., Siddiqi, I., & Vessio, G. (2022). A survey of visual and procedural handwriting analysis for neuropsychological assessment. *Neural Computing and Applications*, *34*(12), 9561–9578. https://doi.org/10.1007/s00521-022-07185-6

Oosterhuis, H. E. M., Van Der Ark, L. A., & Sijtsma, K. (2016). Sample Size Requirements for Traditional and Regression-Based Norms. *Assessment*, *23*(2), 191–202. https://doi.org/10.1177/1073191115580638

Oosterman, J. M., Jansen, M. G., Scherder, E. J. A., & Kessels, R. P. C. (2021). Cognitive reserve relates to executive functioning in the old–old. *Aging Clinical and Experimental Research*, *33*(9), 2587–2592. https://doi.org/10.1007/s40520-020-01758-y

Pagliarin, K. C., Ortiz, K. Z., De Mattos Pimenta Parente, M. A., Arteche, A., Joanette, Y., Nespoulous, J.-L., & Fonseca, R. P. (2014). Montreal-Toulouse Language Assessment Battery for aphasia: Validity and reliability evidence. *NeuroRehabilitation*, *34*(3), 463–471. https://doi.org/10.3233/NRE-141057

Pakoci, E., & Popović, B. (2021). Methods for Using Class Based N-gram Language Models in the Kaldi Toolkit. In A. Karpov & R. Potapova (Eds.). *Speech and Computer* (Vol. 12997, pp. 492–503). Springer International Publishing. https://doi.org/10.1007/978-3-030-87802-3_45

Park, J. Y., Seo, E. H., Yoon, H.-J., Won, S., & Lee, K. H. (2023). Automating Rey Complex Figure Test scoring using a deep learning-based approach: A potential large-scale screening tool for cognitive decline. *Alzheimer's Research & Therapy*, *15*(1), 145. https://doi.org/10.1186/s13195-023-01283-w

Park, S.-C., Jang, E. Y., Lee, K. U., Lee, K., Lee, H.-Y., & Choi, J. (2015). Reliability and validity of the Korean version of the Scale for the Assessment of Thought, Language, and Communication. *Comprehensive Psychiatry*, *61*, 122–130. https://doi.org/10.1016/j.comppsych.2015.04.002

Parsons, T., & Duffield, T. (2020). Paradigm Shift Toward Digital Neuropsychology and High-Dimensional Neuropsychological Assessments: Review. *Journal of Medical Internet Research*, *22*(12), e23777. https://doi.org/10.2196/23777

Pesau, H. G., Immanuel, A. S., Sulastri, A., & Van Luijtelaar, G. (2023). The role of daily spoken language on the performance of language tests: The Indonesian experience. *Bilingualism: Language and Cognition*, *26*(3), 538–549. https://doi.org/10.1017/S136672892200075X

Picanco Diniz, C., Cabral Soares, F., Galdino De Oliveira, T. C., Dias E Dias Macedo, L., Wanderley Picanco Diniz, D. L., Valim Oliver Bento-Torres, N., Bento-Torres, J., & Tomás, A. (2014). CANTAB object recognition and language tests to detect aging cognitive decline: An exploratory comparative study. *Clinical Interventions in Aging*, 37. https://doi.org/10.2147/CIA.S68186

Rohde, A., McCracken, M., Worrall, L., Farrell, A., O'Halloran, R., Godecke, E., David, M., & Doi, S. A. (2022). Inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane Evidence-Based Language Test. *Disability and Rehabilitation*, *44*(4), 637–645. https://doi.org/10.1080/09638288.2020.1776774

Ruan, Q., Xiao, F., Gong, K., Zhang, W., Zhang, M., Ruan, J., Zhang, X., Chen, Q., & Yu, Z. (2020). Demographically Corrected Normative *Z* Scores on the Neuropsychological Test Battery in Cognitively Normal Older Chinese Adults. *Dementia and Geriatric Cognitive Disorders*, *49*(4), 375–383. https://doi.org/10.1159/000505618

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition (arXiv:1904.05862). *arXiv*. http://arxiv.org/abs/1904.05862

Shirk, S. D., Mitchell, M. B., Shaughnessy, L. W., Sherman, J. C., Locascio, J. J., Weintraub, S., & Atri, A. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimer's Research & Therapy*, *3*(6), 32. https://doi.org/10.1186/alzrt94

Slot, R. E. R., Sikkes, S. A. M., Berkhof, J., Brodaty, H., Buckley, R., Cavedo, E., Dardiotis, E., Guillo-Benarous, F., Hampel, H., Kochan, N. A., Lista, S., Luck, T., Maruff, P., Molinuevo, J. L., Kornhuber, J., Reisberg, B., Riedel-Heller, S. G., Risacher, S. L., Roehr, S., et al. (2019). Subjective cognitive decline and rates of incident Alzheimer's disease and non–Alzheimer's disease dementia. *Alzheimer's & Dementia*, *15*(3), 465–476. https://doi.org/10.1016/j.jalz.2018.10.003

Smith, A. K., Ayanian, J. Z., Covinsky, K. E., Landon, B. E., McCarthy, E. P., Wee, C. C., & Steinman, M. A. (2011). Conducting High-Value Secondary Dataset Analysis: An Introductory Guide and Resources. *Journal of General Internal Medicine*, *26*(8), 920–929. https://doi.org/10.1007/s11606-010-1621-5

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary, Third Edition (3 rd). *Oxford University Press*.

Sumpter, R., Camsey, E., Meldrum, S., Alford, M., Campbell, I., Bois, C., O'Connell, S., & Flood, J. (2023). Remote neuropsychological assessment: Acceptability and feasibility of direct-to-home teleneuropsychology methodology during the COVID-19 pandemic. *The Clinical Neuropsychologist*, *37*(2), 432–447. https://doi.org/10.1080/13854046.2022.2056922

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*, *05*(14), 1611–1619. https://doi.org/10.4236/psych.2014.514171

Takenga, C., Preik, P., Berndt, R.-D., Arnold, A., Juenger, M., & Vikandy, M. (2020). Design of a telehealth system adapted for health care delivery in rural areas: Tele-consults between general practitioners and specialists. *International Journal of Innovation and Scientific. 48*(2), 38-46.

Tombaugh, T. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, *19*(2), 203–214. https://doi.org/10.1016/S0887-6177(03)00039-8

Troyer, A. K., & Wishart, H. A. (1997). A comparison of qualitative scoring systems for the Rey-Osterrieth Complex Figure Test. *The Clinical Neuropsychologist*, *11*(4), 381–390. https://doi.org/10.1080/13854049708400467

Tucker-Drob, E. M., Brandmaier, A. M., & Lindenberger, U. (2019). Coupled cognitive changes in adulthood: A meta-analysis. *Psychological Bulletin*, *145*(3), 273–301. https://doi.org/10.1037/bul0000179

Utami, M. S. S., Sulastri, A., Santoso, J., Suryani, A., Goeritno, H., Widhianingtanti, L., & Van Luijtelaar, G. (2024). Adaptation of Rey Auditory Verbal Learning Test fir Indonesia: Its Validity and Reliability. *Acta Neuropsychologica*, *22*(1), 15–34. https://doi.org/10.5604/01.3001.0053.9735

Vogt, J., Kloosterman, H., Vermeent, S., Van Elswijk, G., Dotsch, R., & Schmand, B. (2019). Automated scoring of the Rey-Osterrieth Complex Figure Test using a deep-learning algorithm. *Archives of Clinical Neuropsychology*, *34*(6), 836–836. https://doi.org/10.1093/arclin/acz035.04

Widhianingtanti, L. T., Luijtelaar, G. V., Suryani, A. O., Hestyanti, Y. R., & Sulastri, A. (2022). Indonesian Trail Making Test: Analysis of Psychometric Properties, Effects of Demographic Variables, and Norms for Javanese Adults. *Jurnal Psikologi*, *49*(2), 104. https://doi.org/10.22146/jpsi.68953

Wulanyani, N. M. S., Widhianingtanti, L. T., Immanuel, A. S., Aisyah, A. R. K., Hendriks, M. P. H., Hestyanti, Y. R., & van Luijtelaar, G. (2024). Psychometric properties of the Five-executive Function Tests in Indonesian samples. *Psikohumaniora: Jurnal Penelitian Psikologi*, *9*(1).

Zhang, X., Lv, L., Min, G., Wang, Q., Zhao, Y., & Li, Y. (2021). Overview of the Complex Figure Test and Its Clinical Application in Neuropsychiatric Disorders, Including Copying and Recall. *Frontiers in Neurology*, *12*, 680474. https://doi.org/10.3389/fneur.2021.680474

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

APPENDIX

English Summary

The importance of supporting neuropsychologists in Indonesia is growing due to increasing awareness of mental health, the ageing population, and the increased prevalence of neurological conditions such as stroke, dementia, and traumatic brain injury. Stroke, one of the leading causes of disability in Indonesia, has a high incidence rate. As the population ages, Indonesia will face a higher incidence of neurodegenerative diseases such as dementia and Parkinson's. In clinical settings, neuropsychologists play a critical role in early diagnosis and anticipating cognitive decline, providing essential support not only to patients but also to their families throughout the care processes. In educational settings, neuropsychological assessment and intervention can also assist in the development of individualised educational plans that can subsequently help children reach their full potential.

However, Indonesia faces challenges in fully integrating neuropsychological services. There is a lack of trained neuropsychologists and specialized programs throughout the country. In addition, neuropsychological testing requires standardized tools and expertise that are often not available, especially in rural areas. Many hospitals and clinics outside urban centres lack the knowledge and resources to conduct appropriate, specialized cognitive assessments. Additionally, the collection and use of normative data is essential to ensure that neuropsychological assessments are accurate and culturally relevant. Normative data serve as benchmarks for clinicians, helping them to interpret cognitive, emotional, and behavioural test results by comparing an individual's performance to that of others with similar demographic characteristics (e.g., age, education, and cultural background). In a diverse country like Indonesia, where languages spoken , years and quality of education, and socio-economic status vary significantly between ethnic groups and regions, the use of locally collected normative data helps to avoid preventing misinterpretation and ensure diagnostic accuracy.

Many neuropsychological tests developed in Western countries may not fully reflect the cognitive functioning of Indonesians who speak Bahasa Indonesia or predominantly local languages. The latter is particularly true in rural and remote areas of the archipelago. Additionally, differences in access to education and health care between richer and poorer provinces and between urban and rural parts of the country further highlight the need for locally relevant data to ensure meaningful and accurate cognitive assessments across different regions. To support Indonesian neuropsychologists working across different regions and islands, we propose the use of technology to improve their workflow and effectiveness. It is suggested

that an Indonesian platform could facilitate remote connections between neuropsychologists across the country. In Chapter 2, we introduced such an online platform, initially designed to provide access to normative scores for the Indonesian Boston Naming Test (I-BNT), thus facilitating a more accurate and fair interpretation of test results. We also examined the influence of demographic factors on I-BNT performance. The results showed that age and education significantly impacted test scores. Recognizing the importance of demographic factors in cognitive assessment, the proposed platform includes a feature that allows users to fine-tune reference groups based on these factors. In addition, this study highlighted the need for a growing database with more subjects including those living outside Java island. To address this, we developed a dynamic database with an architecture that supports future expansion and enables the generation of normative scores that reflect Indonesia's diverse population. A notable advantage of this design is its ability to adapt to changing data needs, ensuring a broader and more accurate representation of the population. However, the current study relies primarily on neuropsychological data from populations in Java, highlighting a limitation: the need for data from more diverse regions, including rural and urban areas across Indonesia. Despite this limitation, the database has significant advantages. Its integration with the online platform provides neuropsychologists across Indonesia with streamlined access to normative scores, including those adjusted for demographic differences within the Indonesian dataset. This innovation not only simplifies research processes but also encourages collaboration and improves access to cognitive assessment tools across the country.

In Chapter 3a, we introduced a neuropsychological test battery that has been adapted by a consortium of psychological researchers for use in Indonesia. The battery consists of ten neuropsychological tests covering domains such as executive function, attention, memory and learning, and language. This research examined the influence of three demographic factors on the scores of these adapted tests, and the results emphasized the importance of normative scores adjusted for demographic variables. Preliminary normative scores of the ten tests of the Indonesian Neuropsychological Test Battery (INTB) were presented. In this chapter it was shown that all tests showed moderate to good test-retest reliability except for the short and long-term recall scores of the RAVLT. The eighteen subtests were significantly influenced by age, with scores generally decreasing with age. Notably, the time to complete the Bourdon test and the RAVLT learning over trials did not decrease with age. In contrast, cognitive performance improved with levels of education, except for the time to complete the Bourdon test and the RAVLT's learning over trials and delayed recall. These differential effects of age and

education on the INTB tests highlight the need to adjust normative scores for these demographic factors in a tailored manner.

We analysed the psychometric properties of the cognitive constructs, such as reliability, presented preliminary normative data, and the effects of age and education as indicators of validity, all of which are presented in Chapter 3b. To assess the underlying cognitive construct, we used a data-driven method, Principal Component Analysis (PCA), to identify seven main components for the Indonesian population that showed good model fit. PCA successfully revealed these components. ANOVA results indicated significant age effects on six constructs: visuospatial processing speed, auditory short-term and working memory, processing speed, inhibitory control, and verbal learning ability. However, no age effects were found for executive internal language. Education significantly influenced all constructs except recall and verbal learning ability. These findings improve our understanding of cognitive functioning across different demographic groups in Indonesia. They also provide a basis for developing more targeted interventions and support systems to meet individual needs. Future research could extend the data to better represent the entire Indonesian population.

Chapter 4 presented a computer vision approach for developing an automated scoring system for an adapted visual-spatial test, using a dataset of fewer than 1,000 samples. This approach enables practitioners to develop a standardized, objective scoring system for the Figural Reproduction Test-Computer Vision Automated Scoring System (FRT-CVAS), a widely used and nationally adapted spatial test. The approach is particularly suitable for small datasets, as demonstrated in the initial adaptation phase. In this study, we collected 420 hand-drawn samples from the FRT as part of the INTB. We compared the automated scoring system with traditional manual scoring, based on subjective judgment. The result showed a high accuracy and sensitivity (minimum .91), while the specificity was .80 for one of the three criteria. By systematically extracting and analysing each hand-drawn component, FRT-CVAS significantly reduces subjectivity, achieving high accuracy, sensitivity, and reasonable specificity. This ensures a more standardized, consistent, and objective scoring result. The study is Currently focused on a single FRT card. In order to develop a fully automated scoring system for the full visual-spatial test, additional sample data from other FRT stimulus cards is required.

In Chapter 5, we introduced a transcription model using Wav2vec2 and grammar constriction techniques in the Indonesian language for verbal neuropsychological testing using the audio sample of the Indonesian-Boston Naming Test (I-BNT), a

60-word naming assessment. We collected verbal responses from a diverse cohort of 100 participants, varying in sex, accent, and recording quality. We compared two different ASR decoding methods: Wav2Vec2 letter and Wav2Vec2 plus Viterbi decoding using a bi-gram Language Model (LM). The performance of the models was evaluated using the Word Error Rate (WER). Our results showed that both the inclusion of the bi-gram LM and the quality of the dataset significantly influenced the speech-to-text performance. The ASR model using a bi-gram LM demonstrated a high accuracy in transcribing the 60-word I-BNT. While some words remain challenging to transcribe accurately, we proposed several ways to address this issue.

The successful application of automatic ASR using Wav2Vec2 for Bahasa Indonesia holds promise for integration into other verbal neuropsychological tests tailored to specific assessment needs. This AI-based speech recognition approach is designed to assist neuropsychologists in creating a more objective automated scoring system for language tests by transcribing spoken responses. The algorithm used, Wav2vec2, is adaptable to different languages, including underrepresented ones. This automated language test scoring system for language test provides consistent, rapid, and reliable results in neuropsychological practice.

We conclude that the integration of advanced neuropsychological services and technologies in Indonesia is essential to meet the growing need for mental health support, particularly in light of an ageing population and increasing rates of neurological conditions. The present study emphasizes the importance of establishing a centralized neuropsychological database and collecting of culturally relevant normative data to improve the accuracy and applicability of cognitive assessments across Indonesia's diverse demographic landscape. Furthermore, innovative methods, such as automated computer vision and AI-based speech recognition for scoring cognitive assessments, offer significant improvements in objectivity, consistency, and efficiency. These advances, particularly when adapted for local languages and small datasets, pave the way for more accessible and reliable neuropsychological services across urban and rural settings. By modernizing assessment methods and building supportive infrastructure, Indonesia can improve the quality and reach of neuropsychological care, ultimately improving outcomes for patients and their families.

Nederlandse Samenvatting

Het belang van het ondersteunen van neuropsychologen in Indonesië groeit door de toenemende bewustwording over mentale gezondheid, de vergrijzing van de bevolking en de toegenomen prevalentie van neurologische aandoeningen zoals beroertes, dementie en traumatisch hersenletsel. Beroertes, een van de belangrijkste oorzaken van invaliditeit in Indonesië, hebben een toenemende incidentie. Naarmate de bevolking veroudert, zullen ook in Indonesië meer gevallen van neurodegeneratieve ziekten zoals dementie en de ziekte van Parkinson voorkomen. In een klinische setting spelen neuropsychologen een cruciale rol bij de vroege diagnose en het anticiperen op cognitieve achteruitgang, waarbij zij essentiële steun bieden aan patiënten, maar ook aan hun families gedurende het zorgproces. In een onderwijssetting kunnen neuropsychologische beoordelingen en interventies ook helpen bij het ontwikkelen van geïndividualiseerde onderwijsplannen, die kinderen vervolgens kunnen helpen hun volledige potentieel te bereiken.

Indonesië kampt echter met uitdagingen bij het volledig integreren van neuropsychologische diensten. Er is een tekort aan opgeleide neuropsychologen en gespecialiseerde programma's neuropsychologie in het hele land. Bovendien vereist neuropsychologische diagnostiek specialistische kennis, die vaak niet beschikbaar is, en vooral niet in plattelandsgebieden. Veel ziekenhuizen en klinieken buiten de stedelijke centra hebben tot op zekere hoogte nog steeds niet de noodzakelijke middelen voor gespecialiseerde cognitieve beoordelingen. Hiervoor is een betere integratie van neuropsychologische diensten in het bredere zorgsysteem cruciaal. Bovendien zijn het verzamelen en gebruiken van normatieve gegevens essentieel om ervoor te zorgen dat neuropsychologische beoordelingen nauwkeurig en cultureel relevant zijn. Normatieve gegevens dienen als benchmarks voor clinici en helpen hen bij het interpreteren van cognitieve, emotionele en gedragsmatige testresultaten door de prestaties van een individu te vergelijken met die van anderen met vergelijkbare demografische kenmerken (bijv. leeftijd, opleiding en culturele achtergrond). In een divers land als Indonesië, waar taal, opleiding en sociaaleconomische status aanzienlijk verschillen tussen regio's en bevolkingsgroepen, helpt het gebruik van lokaal verzamelde normatieve gegevens om verkeerde interpretaties te voorkomen en diagnostische nauwkeurigheid te garanderen.

Veel neuropsychologische tests die in westerse landen zijn ontwikkeld, weerspiegelen mogelijk niet volledig het cognitieve functioneren in Indonesië. Al is Bahasa Indonesia de officiële taal in Indonesië, de meeste Indonesiërs spreken eerst en vooral een van de 700 lokale talen. Bovendien benadrukken verschillen in onderwijs en toegang

tot gezondheidszorg tussen stedelijke en plattelandsbevolking de noodzaak van lokaal relevante demografische gegevens om zinvolle en nauwkeurige cognitieve beoordelingen in verschillende regio's en voor verschillende bevolkingsgroepen te garanderen.

Om Indonesische neuropsychologen te ondersteunen die in deze regio's en op deze eilanden werken, stellen we voor om technologie te implementeren om hun workflow en effectiviteit te verbeteren. We stellen voor dat Indonesië een platform nodig heeft om externe verbindingen tussen neuropsychologen te vergemakkelijken. In Hoofdstuk 2 introduceerden we een online platform dat is ontworpen om toegang te bieden aan psychologen en illustreren de toepassing van dit platform aan de hand van de normatieve scores van de aan Indonesië aangepaste Boston Naming Test (I-BNT). Dit maakt een nauwkeurige interpretatie van de testresultaten mogelijk. Deze studie omvat verschillende analyses, waaronder Analysis of Variance (ANOVA), post-hoc tests en systeem simulaties, om de invloed van demografische factoren op I-BNT-prestaties te onderzoeken. De bevindingen laten zien dat leeftijd en opleiding een aanzienlijke invloed hebben op de testresultaten. Omdat het platform het belang van demografische factoren bij cognitieve beoordelingen erkent, bevat het een functie waarmee gebruikers referentiegroepen kunnen verfijnen op basis van deze variabelen. De simulaties tonen de toepassing van verschillende referentiegroepen aan. Bovendien benadrukte deze studie de noodzaak van een groeiende database. Hiervoor hebben we een dynamische database architectuur ontwikkeld die een toekomstige uitbreiding ondersteunt en het genereren van normatieve scores mogelijk maakt die de diverse bevolking van Indonesië weerspiegelt. Een voordeel van dit ontwerp is dat de capaciteit aangepast kan worden aan veranderende data behoeften, wat zorgt voor een bredere en nauwkeurigere representatie van de bevolking. De huidige studie is echter voornamelijk gebaseerd op neuropsychologische gegevens van bevolkingsgroepen in Java. Deze beperking benadrukt de behoefte aan gegevens uit meer diverse regio's, waaronder plattelands- en stedelijke gebieden in heel Indonesië. Ondanks deze beperking biedt de database aanzienlijke voordelen. De integratie met het online platform biedt neuropsychologen in heel Indonesië toegang tot normatieve scores, waaronder scores die zijn aangepast voor demografische verschillen binnen de Indonesische dataset. Deze innovatie vereenvoudigt niet alleen onderzoeksprocessen, maar bevordert ook samenwerking en verbetert de toegankelijkheid tot cognitieve beoordeling middelen, gevalideerde en betrouwbare neuropsychologische tests, in het hele land.

In Hoofdstuk 3a introduceerden we een neuropsychologische testbatterij die toegepast is in een consortium van neuropsychologie. De Indonesian Neuro-

psychological Test Battery (INTB) testbatterij bestaat uit een tiental neuropsychologische tests die domeinen onderzoekenals de executieve functies, verschillende vormen van aandacht, leren en geheugen, en taal. Onderzocht werd de invloed van drie demografische factoren op de uitkomsten van deze tests: de bevindingen benadrukken het belang van normatieve scores die zijn aangepast voor demografische variabelen. Voorlopige normatieve scores van de tests van de INTB werden gepresenteerd. Let wel, voorlopig omdat alleen deelnemers uit verschillende delen van Java geïncludeerd zijn. In dit hoofdstuk werd duidelijk dat alle tests een matige tot goede test-hertestbetrouwbaarheid vertoonden, met uitzondering van de scores van de RAVLT op korte- en langetermijngeheugen . Achttien subtests werden significant beïnvloed door leeftijd, waarbij de scores over het algemeen afnamen naarmate deelnemers ouder werden. Opvallend is dat de tijd om de Bourdon-test en RAVLT variabelen leren-over-trials te voltooien niet afnam met de leeftijd. Daarentegen verbeterden de cognitieve prestaties met een hoger opleidingsniveau, behalve de tijd om de Bourdon-test en het RAVLT-leren over trials en vertraagde herinnering te voltooien. Deze uiteenlopende effecten van leeftijd en opleiding op de INTB-testen benadrukken de noodzaak om de normatieve scores voor deze demografische factoren op maat aan te passen.

In Hoofdstuk 3b zijn de psychometrische eigenschappen van cognitieve constructen zoals, de voorlopige normatieve gegevens, betrouwbaarheid en de effecten van leeftijd en opleiding, meer specifiek verder gerapporteerd. Om de onderliggende cognitieve constructen van de INTB te kunnen vaststellen, gebruikten we een datagestuurde, Principal Component Analysis (PCA). Zeven cognitieve hoofdcomponenten of constructen werden geïdentificeerd voor de INTB voor de Indonesische steekproef. De geïdentificeerde constructen gaven significante leeftijdseffecten aan op zes constructen: snelheid van visuospatiële informatieverwerking, auditief korte termijn geheugen en werkgeheugen, verwerkingssnelheid, cognitieve inhibitie en verbaal leervermogen. Er werden geen leeftijdseffecten gevonden voor taal. Opleiding had een significante invloed op alle constructen, behalve herinnering en verbaal leervermogen. Deze bevindingen verbeteren het begrip van cognitief functioneren in verschillende demografische groepen in Indonesië. Ze vormen ook een basis voor het ontwikkelen van meer gerichte interventies en ondersteuningssystemen om aan individuele behoeften te voldoen. Toekomstig onderzoek zou de gegevens kunnen uitbreiden om de Indonesische bevolkingsgroepen beter te vertegenwoordigen.

Hoofdstuk 4 presenteert een “computer vision” benadering voor het ontwikkelen van een geautomatiseerd scoresysteem voor een aangepaste visueel-ruimtelijke

test, met behulp van een dataset van minder dan 1.000 deelnemers. Deze benadering stelt professionals in staat om een gestandaardiseerd, objectief scoresysteem te ontwikkelen voor de Figural Reproduction Test—Computer Vision Automated Scoring System (FRT-CVAS), een veelgebruikte ruimtelijke test die is aangepast in verschillende landen. De benadering is met name geschikt voor kleine datasets, zoals aangetoond in de initiële aanpassingsfase. In deze studie verzamelden we FRT data bestaande uit steekproef van 420 willekeurig getrokken proefpersonen. De FRT is een onderdeel van de INTB en meet o.a. ruimtelijk geheugen. We vergeleken het geautomatiseerde scoresysteem met traditioneel handmatig scoren, wat vaak afhankelijk is van subjectieve oordelen. Het resultaat toonde een hoge nauwkeurigheid en gevoeligheid (minimaal .91), terwijl de specificiteit .80 was voor een van de drie criteria. Door systematisch elk met de hand getrokken component te extraheren en te analyseren, vermindert FRT-CVAS de subjectiviteit aanzienlijk, wat een hoge nauwkeurigheid, gevoeligheid en redelijke specificiteit oplevert. Dit zorgt voor een meer gestandaardiseerd, consistent en objectief resultaat. De studie richtte zich op één FRT-kaart, maar de ontwikkelde analysetechnieken geven mogelijkheden voor een automatisering van de volledige test. Om een volledig geautomatiseerd scoresysteem voor de gehele visueel-ruimtelijke test te ontwikkelen, zijn aanvullende voorbeeldgegevens van de andere stimulus kaarten in de FRT vereist.

In hoofdstuk 5 introduceerden we een transcriptie model met behulp van Wav2vec2 en grammaticale vernauwing technieken in de Indonesische taal (Bahasa Indonesia) voor verbale neuropsychologische tests met behulp van het audio steekproef van de I-BNT. We verzamelden verbale reacties van een divers cohort van 100 deelnemers, variërend in geslacht, en accent onder twee opname-omstandigheden (met en zonder veel omgevingsruis). We vergeleken twee verschillende automatic speech recognition (ASR) de coderingsmethoden: Wav2Vec2-letter en Wav2Vec2 plus Viterbi-decodering met behulp van een bigram Language Model (LM). De prestaties van de modellen werden beoordeeld met behulp van Word Error Rate (WER). Onze resultaten geven aan dat zowel de opname van bigram LM als de kwaliteit van de dataset de spraak-naar-tekst prestaties aanzienlijk beïnvloeden. Het ASR-model met hulp van een bigram LM toonde een hoge nauwkeurigheid bij het transcriberen van de 60-woorden I-BNT. Hoewel sommige woorden nog steeds moeilijk nauwkeurig te transcriberen waren, hebben we verschillende mogelijkheden voorgesteld om dit probleem aan te pakken. De succesvolle toepassing van ASR met Wav2Vec2 voor Bahasa Indonesia biedt perspectief voor integratie in andere verbale neuropsychologische tests die zijn afgestemd op specifieke beoordeling behoeften. Deze op AI gebaseerde spraakherkenning

benadering is ontworpen om neuropsychologen te helpen bij het creëren van een objectiever geautomatiseerd scoresysteem voor taaltests door transcriptie van gesproken antwoorden. Het gebruikte algoritme Wav2vec2 is aan te passen aan verschillende talen, inclusief ondervertegenwoordigde talen, waaraan Indonesia zo rijk is. Dit geautomatiseerde scoresysteem voor taaltests biedt consistente, snelle en betrouwbare resultaten in de neuropsychologische praktijk.

We concluderen dat de integratie van geavanceerde neuropsychologische diensten en technologieën in Indonesië essentieel is om de toenemende behoefte aan ondersteuning van de gezondheidszorg aan te pakken, met name in het licht van een vergrijzende bevolking en toenemende percentages neurologische aandoeningen. De huidige studie benadrukt het belang van het opzetten van een gecentraliseerde neuropsychologische database en het verzamelen van cultureel relevante normatieve gegevens om de nauwkeurigheid en toepasbaarheid van cognitieve beoordelingen in het kader van de demografische verschillen in Indonesië te verbeteren. Bovendien bieden innovatieve methoden, zoals geautomatiseerde computer vision en AI-gebaseerde spraakherkenning voor het scoren van cognitieve beoordelingen, aanzienlijke verbeteringen in objectiviteit, consistentie en efficiëntie. Deze ontwikkelingen, met name wanneer ze worden aangepast voor lokale talen en kleine datasets, banen de weg voor toegankelijkere en betrouwbaardere neuropsychologische diensten in stedelijke en landelijke omgevingen. Door beoordelingsmethoden te moderniseren en ondersteunende infrastructuur op te zetten, kan Indonesië de kwaliteit en het bereik van neuropsychologische zorg verbeteren, wat uiteindelijk de zorg voor patiënten en hun families zal verbeteren.

Ringkasan dalam Bahasa Indonesia

Seiring dengan meningkatnya kesadaran masyarakat akan pentingnya kesehatan fungsi otak, peran neuropsikolog di Indonesia semakin krusial. Hal ini juga diperkuat dengan peningkatan populasi usia lanjut, dan peningkatan prevalensi kondisi neurologis seperti stroke, demensia, dan cedera otak secara fisik maupun yang disebabkan oleh trauma. Sampai saat ini, stroke masih menjadi salah satu penyebab utama disabilitas di Indonesia, dengan angka kejadian yang tinggi. Kemudian terkait dengan bertambahnya populasi lansia, Indonesia akan menghadapi peningkatan kejadian penyakit *neurodegeneratif* seperti demensia dan Parkinson. Di beberapa negara maju, untuk kegiatan preventif pemerintah tidak hanya melibatkan neurolog, neuropsikolog juga memiliki peran penting dalam diagnosis dini dan antisipasi penurunan kognitif serta memberikan dukungan penuh tidak hanya kepada pasien tetapi juga kepada keluarga mereka sepanjang proses perawatan. Kondisi kognitif seseorang perlu untuk diukur karena memiliki peran dalam dunia pendidikan, terutama dalam membantu mengembangkan rencana pendidikan individual yang selanjutnya dapat membantu anak-anak mencapai potensi penuh mereka.

Dalam rangka integrasi layanan neuropsikologi di Indonesia masih terkendala oleh beberapa faktor. Faktor yang mempengaruhi antara lain kekurangan tenaga ahli neuropsikolog, keterbatasan alat dan peralatan standar untuk tes neuropsikologis, dan assessment neuropsikologi, baik di perkotaan maupun di pedesaan terutama di propinsi yang masih tertinggal. Oleh karena itu perlu adanya upaya peningkatan untuk mengintegrasikan layanan neuropsikologi di Indonesia. Hal utama yang penting untuk dilakukan adalah pengumpulan dan penggunaan data normatif di Indonesia untuk memastikan bahwa penilaian neuropsikologis akurat dan relevan secara budaya. Data normatif tes neuropsikologis akan berfungsi sebagai tolak ukur bagi klinisi. Terutama membantu neuropsikolog dalam menginterpretasikan hasil tes kognitif, emosional, dan perilaku dengan membandingkan kinerja individu dengan kinerja orang lain dengan karakteristik demografis serupa (misalnya, usia, pendidikan, dan latar belakang budaya). Di negara yang beragam seperti Indonesia, di mana bahasa, pendidikan, dan status sosial ekonomi sangat bervariasi, menggunakan data normatif yang dikumpulkan berdasarkan populasi Indonesia membantu mencegah salah tafsir dan memastikan akurasi diagnostik.

Banyak tes neuropsikologi yang dikembangkan di negara-negara Barat mungkin tidak sepenuhnya mencerminkan fungsi kognitif di Indonesia, terutama karena perbedaan kultur Indonesia dengan negara Barat. Selain itu di Indonesia,

perbedaan tingkat pendidikan dan akses kesehatan antara populasi perkotaan dan pedesaan juga menjadi pertimbangan penting untuk mengukur penilaian kognitif di masing-masing wilayah di Indonesia. Hal ini menjadi alasan untuk perlunya mengembangkan tes neuropsikologi yang sudah diadaptasi ke dalam bahasa Indonesia dan mengumpulkan data skor tes untuk kemudian dijadikan parameter acuan sendiri untuk populasi Indonesia.

Dalam rangka mendukung neuropsikolog Indonesia yang bekerja di berbagai wilayah perkotaan dan pedesaan baik dalam satu pulau maupun antar pulau, kami mengusulkan penggunaan teknologi untuk meningkatkan alur kerja dan efektivitas. Neuropsikolog di Indonesia membutuhkan *platform* untuk memfasilitasi koneksi jarak jauh diantara mereka. Pada Bab 2, kami memperkenalkan *platform online* yang dirancang untuk menyediakan akses ke skor normatif untuk Boston Naming Test versi bahasa Indonesia (I-BNT). Platform ini memfasilitasi interpretasi hasil I-BNT yang lebih akurat. Studi ini menggabungkan berbagai analisis seperti Analisis Varian (ANOVA), tes post-hoc, dan simulasi sistem, untuk melihat pengaruh faktor demografis terhadap kinerja I-BNT. Hasil analisis menunjukkan bahwa usia dan pendidikan secara signifikan mempengaruhi skor tes. Kelebihan dalam sistem ini adalah neuropsikolog bisa membandingkan kemampuan *client* nya dengan *normative* data yang sesuai dengan kategori usia dan pendidikan dengan memilih kelompok referensi yang sesuai dengan karakteristik demografis *client*. Dengan demikian, neuropsikolog dapat memperoleh pemahaman yang lebih baik tentang kinerja kognitif individu. Selain itu, dalam studi 2 juga memperkenalkan sebuah database dinamis yang menfasilitasi adanya penambahan data skor neuropsikologis secara berkelanjutan. Database ini juga mendukung memperoleh skor normatif yang mencerminkan populasi Indonesia setelah database lengkap berisi representative data dari seluruh penjuru Indonesia. Pada saat studi ini dilaksanakan data neuropsikologi yang tersedia baru dari populasi Pulau Jawa. Sehingga perlu adanya penambahan data dari wilayah yang lebih beragam, termasuk daerah pedesaan dan perkotaan di seluruh Indonesia. Meskipun ada keterbatasan ini, database menawarkan keuntungan signifikan. Integrasinya dengan platform online menyediakan fasilitas bagi neuropsikolog di seluruh Indonesia untuk melakukan akses yang lebih mudah ke skor normatif, termasuk yang disesuaikan dengan perbedaan demografis dalam dataset Indonesia. Inovasi ini tidak hanya memudahkan proses penelitian tetapi juga mendorong kolaborasi dan meningkatkan aksesibilitas ke alat penilaian kognitif di seluruh negeri.

Dalam bab 3a kami memperkenalkan baterai tes neuropsikologis yang telah diadaptasi oleh konsorsium peneliti psikologi untuk digunakan di Indonesia. Baterai

ini terdiri dari 10 tes neuropsikologis yang mencakup domain fungsi e*xecutive*, *attention*, *memory* dan *learning*, serta *language*. Dalam bab ini juga dilaporkan hasil evaluasi terhadap pengaruh tiga faktor demografis pada 10 tes yang diadaptasi. Hasil dari evaluasi menunjukkan pentingnya skor normatif yang disesuaikan dengan variabel demografis. Selain itu dalam bab ini juga ditemukan bahwa semua tes menunjukkan reliabilitas *test-retest* yang *moderate* hingga *good*, kecuali pada skor *recall* jangka pendek dan jangka panjang dari RAVLT. Analisis *varians* menunjukkan bahwa delapan belas subtes secara signifikan dipengaruhi oleh usia, dengan pola skor yang menurun seiring bertambahnya usia. Kami menemukan hal menarik pada subtes yang mengukur waktu untuk menyelesaikan tes Bourdon dan proses pembelajaran pada RAVLT, kedua variabel ini tidak menunjukkan pola menurun seiring bertambahnya usia. Dari sisi pendidikan, kami menemukan korelasi positif antara level pendidikan dan kemampuan kognitif yaitu kinerja kognitif meningkat pada individu dengan tingkat pendidikan yang lebih tinggi, kecuali untuk waktu menyelesaikan tes Bourdon dan pembelajaran serta *delay recall* pada RAVLT. Efek yang bervariasi dari usia dan pendidikan pada tes INTB menggaris bawahi perlunya menyesuaikan skor normatif untuk faktor-faktor demografis ini secara khusus.

Pada bab 3b kami memaparkan hasil psikometri, konstruksi kognitif, normatif score awal, reliabilitas, dan pengaruh usia dan pendidikan sebagai indikator validitas. Kognitif konstruk ditentukan dengan menggunakan metode berbasis data yang disebut *Principal Component Analysis* (PCA), dari empat ratus sembilan puluh data dihasilkan tujuh konstruk pada dataset pulau Jawa Indonesia. Hasil ANOVA pada tujuh konstruk terhadap faktor demografis menunjukkan efek signifikan berdasarkan kelompok usia yang pada enam konstruk yaitu: kecepatan pemrosesan informasi visual-spasial, memori jangka pendek dan kerja auditori, kecepatan pemrosesan, kontrol penghambatan, dan kemampuan belajar verbal. Namun, tidak ditemukan efek usia pada bahasa internal eksekutif. Faktor pendidikan secara signifikan mempengaruhi semua konstruk kecuali recall dan kemampuan belajar verbal. Temuan-temuan ini meningkatkan pemahaman kita tentang fungsi kognitif di berbagai kelompok demografis di Indonesia. Informasi ini juga menjadi dasar untuk mengembangkan intervensi dan sistem dukungan yang lebih akurat untuk memenuhi kebutuhan individu.

Pada bab 4, kami membahas tentang pendekatan komputer vision untuk mengembangkan sistem penilaian otomatis pada tes visual-spasial yang diadaptasi, menggunakan kumpulan data gambar tangan dengan jumlah data kurang dari 1.000 sampel. Pendekatan ini memungkinkan para praktisi untuk mengembangkan sistem penilaian objektif yang disebut *Figural Reproduction Tes-Computer Vision*

Automated System (FRT-CVAS), sebuah tes spasial (*Figural reproduction*) yang banyak digunakan dan diadaptasi di berbagai negara. Pendekatan ini tepat digunakan pada penelitian dengan dataset yang masih terbatas. Dalam penelitian ini, kami mengumpulkan 420 gambar tangan dari FRT yang merupakan bagian dari Baterai Tes Neuropsikologi yang diadaptasi untuk Indonesia. Pada studi 4, kami membandingkan sistem penilaian otomatis dengan penilaian manual tradisional, yang seringkali dipengaruhi oleh penilaian subjektif. Hasilnya menunjukkan *accuracy* dan *sensitivity* tinggi (minimal .91), sementara *specificity* menunjukkan angka .80 untuk salah satu dari tiga kriteria. HAsil ini menunjukkan bahwa sistem ini bisa diimplementasikan. Pendekatan ini dilakukan secara sistematis dengan mengekstrak dan menganalisis setiap komponen gambar tangan. FRT-CVAS mendukung tujuan awal untuk mengembangkan sistem penilaian yang lebih standar, konsisten, dan objektif. Saat ini, penelitian ini berfokus pada satu kartu FRT. Untuk mengembangkan sistem penilaian otomatis penuh untuk seluruh tes visual-spasial, diperlukan data sampel tambahan dari kartu stimulus lainnya dalam FRT.

Pada bab 5 kami memperkenalkan model transkripsi bahasa dengan menggunakan dua metode yaitu Wav2Vec2 dan teknik pembatasan gramatikal dalam bahasa Indonesia untuk tes neuropsikologis verbal menggunakan sampel audio dari *Boston Naming Test* versi Bahasa (I-BNT), sebuah tes penamaan 60 gambar. Kami mengumpulkan respons verbal dari 100 peserta yang beragam dalam jenis kelamin, aksen, dan kualitas rekaman. Kami membuat model dengan membandingkan dua metode dekode *Automated Speech Recognition* (ASR): Wav2Vec2 dan Wav2Vec2 ditambah dekode *Viterbi* menggunakan *Language Model* (LM) *bi-gram*. Kinerja model dinilai menggunakan *Word Error Rate* (WER). Hasil dari uji coba menunjukkan bahwa dengan model dengan dan tanpa bi-gram dan kualitas dataset secara signifikan mempengaruhi kinerja transkripsi suara ke teks. Model ASR yang menggunakan LM *bi-gram* menunjukkan akurasi lebih tinggi dalam mentranskripsikan 60 kata I-BNT. Meskipun beberapa kata masih sulit untuk ditranskripsikan secara akurat, dalam studi ini kami mengusulkan beberapa peluang untuk mengatasi masalah ini. Penerapan ASR otomatis yang sukses menggunakan Wav2Vec2 untuk Bahasa Indonesia menjanjikan integrasi ke dalam tes neuropsikologis verbal lainnya yang disesuaikan dengan kebutuhan penilaian spesifik. Pendekatan pengenalan ucapan berbasis AI ini dirancang untuk membantu neuropsikolog dalam menciptakan sistem penilaian otomatis yang lebih objektif untuk tes bahasa melalui transkripsi respons lisan. Sistem penilaian tes bahasa otomatis ini menawarkan hasil yang konsisten, cepat, dan andal dalam praktik neuropsikologi.

Berdasarkan studi yang telah kami lakukan, kami menyimpulkan bahwa integrasi layanan dan teknologi neuropsikologi yang canggih di Indonesia sangat penting dalam mengatasi peningkatan kebutuhan dukungan kesehatan mental, terutama mengingat peningkatan populasi lansia dan meningkatnya angka kondisi neurologis. Studi ini menekankan pentingnya membangun database neuropsikologi terpusat dan pengumpulan data normatif yang relevan secara budaya untuk meningkatkan akurasi dan penerapan penilaian kognitif di seluruh lanskap demografis Indonesia yang beragam. Selain itu, metode inovatif, seperti *automated scoring* dengan pendekatan *computer vision* dan pengenalan ucapan berbasis AI untuk penilaian kognitif, menawarkan peningkatan substansial dalam objektivitas, konsistensi, dan efisiensi. Tawaran integrasi teknologi ini terutama yang mengakomodasi adaptasi pada bahasa lokal dan jumlah dataset yang terbatas memberi peluang bagi layanan neuropsikologi agar lebih mudah diakses dan handal untuk digunakan di seluruh wilayah Indonesia. Dengan otomatisasi metode penilaian dan membangun infrastruktur pendukung, Indonesia dapat meningkatkan kualitas dan jangkauan layanan neuropsikologis, pada akhirnya meningkatkan hasil bagi pasien dan keluarga mereka.

Research Data Management

Data Collection and Storage

The data for this study were provided by the Indonesian Neuropsychology Consortium, collected through a research project funded by the Directorate of Higher Education (DIKTI) under grant number 010/L6/AK/SP2H.1/PENELITIAN/2019.

Data for Chapters 2, 3a, and 3b were obtained through the Indonesian Neuropsychology Consortium. Chapter 2 includes scores and demographic information (age, education, and gender) from one neuropsychological test. Chapters 3a and 3b provide scores and demographic information (age, education, gender) from 10 neuropsychological tests.

Data for Chapter 4 were collected by manually scanning participants' hand-drawn responses, while data for Chapter 5 were gathered through audio-recorded interviews. These data for Chapters 4 and 5 were collected as part of another research project funded by DIKTI under grant number 076/E5/PG.02.00.PL/2023.

All pseudonymized data from these studies are securely stored on the server (norm.indonesian-andi.id). Technical and organizational measures were implemented to ensure data availability, integrity, and confidentiality. Physical (paper-based) data are kept in locked cabinets at the Psychology Faculty of Soegijapranata Catholic University, Semarang, Indonesia.

Availability of Data and Code

All personal data have been anonymized. Data from Chapters 2-5 are available upon request to the corresponding author. Participant scores and demographic data are stored on a secure, password-protected online platform (www.norm.indonesian-andi.id), with demographic data stored separately from research data to ensure privacy. All data, including scores, audio recordings, and hand-drawn images, have been anonymized.

Ethical Approval and Informed Consent

The Research Ethics Committee of Soegijapranata Catholic University, Indonesia, approved the data collection, storage, and use for scientific purposes in the studies presented in Chapters 3-5 (approval number: 001B/B.7.5/FP.KEP/IV/2018). This approval adheres to the principles of the Declaration of Helsinki and local legislation.

Curriculum vitae

Shinta Estri Wahyuningrum, S.Si., M.Cs., was born on September 27, 1982. She is currently a faculty member at the Informatics Engineering Department, Faculty of Computer Science, Soegijapranata Catholic University, Indonesia. She can be reached via email at shinta@unika.ac.id. She pursued her higher education in the field of Computer Science. She obtained her Master's degree from the Faculty of Mathematics and Natural Sciences at Gadjah Mada University, Indonesia, between 2009 and 2011. Her thesis, titled "Dual Method Steganography in Audio Signal," focused on security in audio data processing. Prior to that, she completed her Bachelor's degree in Computer Science at Sanata Dharma University, Indonesia, from 2000 to 2004. Her undergraduate thesis, "Javanese Letter Recognition Using Fuzzy," explored pattern recognition and artificial intelligence techniques. Throughout her career, Shinta has held various managerial positions at Soegijapranata Catholic University. She currently serves as the Secretary of the Institute for Research and Community Service (2024-2025). Previously, she was the Secretary of the Informatics Engineering Department from 2016 to 2019 and the Head of the same department from 2013 to 2015. She has also contributed as the Vice Dean for Academic Affairs in 2012 and the Vice Dean for Finance and Administration from 2007 to 2009.

Over the years, she has been actively involved in securing research grants. Some of her notable projects include the development of digital and automated neuropsychological visual and audio tests funded by RISTEKDIKTI (2022-2023) and the Erasmus+ Neuropsychological Assessment at Radboud University in 2018. She also contributed to the Development of Advanced Neuropsychological Diagnostics Infrastructures (ANDI) in Indonesia (2018-2020), the creation of an electronic consulting system for pediatric congenital heart disease diagnosis (2017), and the development of an electronic medical record data access security model using mobile technology (2015). Beyond academia, Shinta is engaged in professional organizations. She is an active staff member of the Indonesian Neuropsychology Association (ANI) for the 2024-2028 period and was part of the curriculum team of the Association of Computer Higher Education (APTIKOM) in Central Java from 2016 to 2020. Additionally, she has contributed as an IT staff member for the INTB development in www.norm.indonesian-andi.id

Portfolio

Training activities Courses	hours
DGS - Graduate School Day 2 (2024)	7
RU - Writing Scientific Articles (2022)	84
RU - Language Development for Academic Writing (2021	42
RU - Statistics for PhD's by using SPSS (2021)	56
EU-ASEAN High-Performance Computing (HPC) School (2021)	40
DGS - Graduate School Day (2021)	7
RU - Effective Writing Strategies (2021	84
DGS - Scientific Integrity Course (2021)	7
DGS - Graduate School Day (2020)	7
RU - Writing a Conference Abstract (2020)	14

List of Publication

Wahyuningrum S.E., van Luijtelaar G., Sulastri A., Hendriks M.P.H., Sanjaya R., Heskes T. 2024. "A Computer Vision System for an Automated Scoring of a Hand-drawn Geometric Figure". *Sage Open* 14(4). DOI:10.1177/21582440241294142

Wahyuningrum S.E., Hendriks M.P.H., Sulastri A. van Luijtelaar G. "Online Platform untuk Interpretasi Hasil Tes Neuropsychologi". Presented at Kongres Pelajar Indonesia di Belanda "Membangun Visibilitas Riset Peneliti Muda Indonesia. 28 Oktober 2023 – KBRI Denhaag.

Wahyuningrum S.E., van Luijtelaar G., Hendriks M.P.H., Sulastri A., Sanjaya R. 2023. "Protocol of Generating Sub-Datasets from a Neuropsychological Database". Poster presentation at the *International Neuropsychological Society* is holding INS 2023 Taiwan Meeting.

Wahyuningrum S.E., Hendriks M.P.H., van Luijtelaar G., Sulastri A. 2023. "Indonesia Neuropsychological Test Battery: Normative Score, Reliability, Age and Education Effects." *The 2nd International Conference on Biopsychosocial Issues*. June 2023.

Wahyuningrum S.E., **Sulastri A.**, Hendriks M.PH, Consortium Indonesian NP, van Luijtelaar G. 2022. "The Indonesian Neuropsychological Test Battery (INTB): Psychometric Properties, Preliminary Normative Scores, The Underlying Cognitive Constructs, and the Effects of Age and Education". *Acta Neuropsychologica* 20 (4): 445-470. DOI: 10.5604/01.3001.0016.1339

Wahyuningrum S.E., Sulastri A., van Luijtelaar G. 2021. "An online platform and a dynamic database for neuropsychological assessment in Indonesia". *Applied Neuropsychology: Adult* 30(1): 1-10. DOI: 10.1080/23279095.2021.1943397

Setianto Y.B.D., **Wahyuningrum S.E.** 2021. "Multi-Tier Model with JSON-RPC in Telemedicine Devices Authentication and Authorization Protocol". Conference: 2021 7th *International Conference on Engineering, Applied Sciences and Technology (ICEAST)*. DOI:10.1109/ICEAST52143.2021.9426308

Hartanu D.A.S., **Wahyuningrum S.E.** 2021. "Short FCC and CCCto Detect Eyes and Mouth in Different Image Size". *Proxies Jurnal* 3(1):52-62. DOI: 10.24167/proxies.v3i1.3627

Setianto Y.B.D., **Wahyuningrum S.E.** 2019. "Medical Device Authentication and Authorization Protocol in Indonesian Telemedicine Systems". *Conference: 2019 4th International Conference on Information Technology (InCIT)*. DOI: 10.1109/INCIT.2019.8912058

Wahyuningrum, S.E., Sulastri A., Sanjaya R. 2019. "Information System databases for Neuropsychology Tests: case study in Boston Naming Test". *Sisforma* 6(1): 28. DOI: 10.24167/sisforma.v6i1.2274.

Santoso Y.O, Sukiatmodjo A., Setiady D.A., **Wahyuningrum S.E.**, YB. Dwi Setianto. 2018. "Buku Saku Berbasis Mobile bagi Orang Tua dengan Anak PJB". *Seminar Nasional Aplikasi Sainsa Dan Teknologi (SNAST)*. 15 September 2018.

Cahyono D.S., **Wahyuningrum S.E.** 2018. "Perbandingan Identifikasi Gambar Huruf yang telah dimanipulasi Menggunakan Algoritma ALBP and Chain Code" (Comparison Letter Recognition Using ALBP and Chain Code). *Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA)*. ISSN : 2089-9815.

Setianto Y.B.D., **Wahyuningrum S.E.** 2016. "Centralized Availability Assurance for Distributed Electronic Medical Record Data". *Journal of Theoretical and Applied Information Technology* 83(3), 423.

Wahyuningrum S.E. 2016. "e-posyandu untuk anak-anak PJB" (E-Posyandu for children with Congenital Heart Disease). *Proceeding Energizing Innovative Culture*. Catholic Soegijapranata University.

Acknowledgment

I want to sincerely thank to Prof. Gilles van Luijtelaar Thank you for believing in me and allowing me to learn much about research from you. Thank you for your patience, guidance, and support. I am deeply grateful to Prof. Tom Heskes and Prof. Marc Hendrik for their unwavering guidance, mentorship, and invaluable support throughout this research endeavor. Their expertise and scholarly input have significantly enriched the content and rigor of my study. My deepest thanks to Dr. Augustina Sulastri for the golden opportunity, trust, and mentorship during the study process. Appreciation to Prof. Ridwan and Prof. David, their expertise and constructive feedback have been instrumental in shaping the direction and quality of my work.

To the manuscript committee: Prof. dr. J.M. Oosterman, Prof. dr. E.O. Postma (Tilburg University, Netherlands), Prof. dr. ing. R. Pulungan (Gadjah Mada University Indonesia) I am grateful for your valuable input.

I extend my gratitude to okedech family, loving husband Anto, Rara and Bima for their continuous encouragement, understanding, and motivation during the challenging phases of my doctoral journey. Appreciate to my parent Bapak Ibu Anton for unwavering support and belief in me have been truly inspiring. I am thankful to Bapak Ibu Sunardi, My lovely sister Nisi and Vita, Kenzi, Una and Om Ari, Mba Rini, who are always supporting me and my family.

My heartfelt gratitude goes to the core team of Neuro Soegijapranata: bu Lastri as the leader, Bu Cicih, and Pak Haryo, who were always delightful and generous in sharing their expertise in statistics; Bu Trisni, my roommate during the initial two months in Nijmegen; and Pak Abi and Pak Danis, whose cheerful and enthusiastic presence always lifted spirits. I also extend my appreciation to the extended family of the Indonesian Neuropsychology Consortium: Bu Angela, Bu Heni Gerda, Bu Wulan, Pak Arya, Bu Yohana, and all others who cannot be mentioned individually.

I am equally grateful to the extended family of the Informatics Engineering Department: Mrs. Rosita, who consistently provides support and assistance, as well as Mr. Aji, Mr. Marlon, Mr. Yonathan, and Mr. Yuli. My profound thanks also go to colleagues who became like family during my time in Nijmegen Widhi, for the insightful discussions and feedback; Tiara, for making weekends enjoyable; Evelyne, for her unwavering presence from start to finish; Kayan, a delightful housemate; Mba. Nuri, Mas Pras dan Umar, who graciously welcomed me into their home; Mas. Fajar, Mba Zaimah, Mas Afnan, Mba Ainul, Mba Ira for sharing their knowledge;

as well as Mas. Wisnu and Derry. Special thanks to Bu Arwin for her expertise in designing my beautiful book cover.

I wish to express my sincere gratitude to the DCC International Office, particularly Dr. Robin Kayser, Ms. Miranda, and DCC secretary Ms. Karin Potter, for their invaluable support and assistance during every visit. Last but not least, for those people who have directly and indirectly supported me, whom I cannot mention one by one.

Donders Graduate School

For a successful research Institute, it is vital to train the next generation of scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School in 2009. The mission of the Donders Graduate School is to guide our graduates to become skilled academics who are equipped for a wide range of professions. To achieve this, we do our utmost to ensure that our PhD candidates receive support and supervision of the highest quality.

Since 2009, the Donders Graduate School has grown into a vibrant community of highly talented national and international PhD candidates, with over 500 PhD candidates enrolled. Their backgrounds cover a wide range of disciplines, from physics to psychology, medicine to psycholinguistics, and biology to artificial intelligence. Similarly, their interdisciplinary research covers genetic, molecular, and cellular processes at one end and computational, system-level neuroscience with cognitive and behavioural analysis at the other end. We ask all PhD candidates within the Donders Graduate School to publish their PhD thesis in de Donders Thesis Series. This series currently includes over 600 PhD theses from our PhD graduates and thereby provides a comprehensive overview of the diverse types of research performed at the Donders Institute. A complete overview of the Donders Thesis Series can be found on our website: https://www.ru.nl/donders/donders-series

The Donders Graduate School tracks the careers of our PhD graduates carefully. In general, the PhD graduates end up at high-quality positions in different sectors, for a complete overview see https://www.ru.nl/donders/destination-our-former-phd. A large proportion of our PhD alumni continue in academia (>50%). Most of them first work as a postdoc before growing into more senior research positions. They work at top institutes worldwide, such as University of Oxford, University of Cambridge, Stanford University, Princeton University, UCL London, MPI Leipzig, Karolinska Institute, UC Berkeley, EPFL Lausanne, and many others. In addition, a large group of PhD graduates continue in clinical positions, sometimes combining it with academic research. Clinical positions can be divided into medical doctors, for instance, in genetics, geriatrics, psychiatry, or neurology, and in psychologists, for instance as healthcare psychologist, clinical neuropsychologist, or clinical psychologist. Furthermore, there are PhD graduates who continue to work as researchers outside academia, for instance at non-profit or government organizations, or in pharmaceutical companies. There are also PhD graduates who work in education, such as teachers in high school, or as lecturers in higher education. Others continue in a wide range of positions, such as policy advisors,

project managers, consultants, data scientists, web- or software developers, business owners, regulatory affairs specialists, engineers, managers, or IT architects. As such, the career paths of Donders PhD graduates span a broad range of sectors and professions, but the common factor is that they almost all have become successful professionals.

For more information on the Donders Graduate School, as well as past and upcoming defences please visit: http://www.ru.nl/donders/graduate-school/phd/

9 789465 150956

DONDERS INSTITUTE

Radboud University

Radboudumc