Assessment of Cognitive Ageing: Applying Generalisability Theory, Rasch and Network Analyses to Establish Reliability and Validity of Measures

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Psychology

at

The University of Waikato

by

QUOC CUONG TRUONG

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Year of submission

2023

Abstract

Accurate assessment of cognitive ageing is critical to improve healthcare for older adults with cognitive decline. While bio-physiological assessments, such as laboratory, physical, and neurological examinations, offer high precision and interval-level scaling, they are limited by not reflecting perceptual experiences of individuals. In other words, it is impossible to find out what an individual is experiencing without asking questions, which requires psychometric measures, often using ordinal level scales. Psychometric measures are widely used to assess, diagnose, and research cognitive ageing but their reliability and validity may bias the results. Firstly, differentiating between enduring and dynamic patterns captured by a measure helps to distinguish an enduring condition while monitoring changes over time. This type of differentiation is also important to determine the reliability and validity of ordinal scales focused on enduring or dynamic patterns. Secondly, all measures should be ideally at least interval-level to meet the basic assumptions of parametric statistics for comparative analyses with interval level variables such as neuroimaging data. Thirdly, the cognitive components assessed using psychometric measures of cognitive ageing interact with each other, and this may impact on the validity of the outcomes measured at the same time, as well as the ability to predict the future outcomes of cognitive ageing. Together, these issues challenge the reliability and validity of the current psychometric measures of cognitive ageing. The present thesis addresses these reliability and validity issues by using three modern theories of measurement, namely, Generalisability theory (G-Theory), Rasch analysis, and Network analysis.

The first part of this thesis focuses on the application of G-Theory to examine dynamic and enduring aspects of cognitive ageing as measured by the Memory Assessment Clinic Questionnaire (MAC-Q), the 16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16), the Mini Mental State Examination (MMSE), and the Modified Telephone Interview for Cognitive Status (TICS-M). The findings of these studies highlighted the importance of differentiating between dynamic and enduring aspects of cognition and revealed the specific benefits of each measure in assessing cognitive ageing. The second part of the thesis included three studies that applied Rasch methodology to the IQCODE-16 and the TICS-M. These studies generated algorithms that allow the transformation of ordinal scales scores into interval-level data, which improved the precision of the measures. The final two studies used network analysis to evaluate the validity of the cognitive domains of TICS-M and MMSE in relation to neuropsychological test performance, advancing the understanding of global network mechanisms involved in assessments of cognitive ageing and relations between cognitive functions. Overall, this thesis involved analyses of three longitudinal datasets, totalling 2705 participants. The reported findings enhanced reliability and validity of widely used assessment tools for older adults and contributed to better understanding of cognitive ageing processes.

Keywords: Cognitive Ageing, Assessment, Psychometrics, Diagnostics, Rasch Analysis, Generalisability Theory, Network Analysis, Reliability, Validity

Acknowledgements

The completion of this PhD thesis has been a long and challenging journey, and I could not have achieved it without the support and encouragement of numerous individuals. I am grateful for their support, inspiration, and guidance throughout this process.

First and foremost, I express my deep gratitude to my chief supervisor, Dr. Oleg Medvedev, for his invaluable guidance, mentorship, and support. His expertise, feedback, and encouragement have been instrumental in shaping this research project. I am grateful for his time, effort, and patience in helping me to develop my research skills and to navigate through the challenges of the PhD program. Thank you, Dr. Oleg Medvedev, for your unwavering support and dedication to my success.

I also wish to acknowledge my co-supervisors, Dr. Carol Choo and Dr. Matti Cervin, for their significant contributions to this work. Their expertise, feedback, and constructive criticism have helped to refine and improve the quality of this research. I appreciate their guidance and support, and I am grateful for their time and effort in reading and providing feedback on my articles and thesis. I am also grateful for the support and hard work of the research team members and the collaboration of the Center of CheBa. I acknowledge their invaluable contributions, including their technical expertise, research assistance, and intellectual insights. Thank you for your dedication and commitment to this research project.

I would also like to express my gratitude to the participants who took part in this research project. Without their participation and contribution, this research would not have been possible. Their willingness to share their experiences, insights, and knowledge has been invaluable in shaping this study. I am thankful to the University of Waikato with awarding me the University of Waikato Doctoral Scholarship that allowed me to undertake this research project.

I express my profound gratitude to my wife, Loan, for her constant love, support, and encouragement during this journey. Her patience, understanding, and unwavering support have been an incredible source of motivation for me. Thank you for being my constant source of inspiration and for always being there for me through thick and thin. My two sons, Bach and Khoa (Ken), have been a significant source of motivation and inspiration for me. Thank you for your patience and understanding during the long hours of research and thesis writing. I hope that this thesis inspires you to follow your passions and to pursue your dreams.

I also acknowledge my parents and parents in law, for their love, support, and encouragement throughout my academic journey. They have been a constant source of inspiration and guidance, and I cannot thank them enough for their sacrifices and dedication to my success. Thank you for instilling in me the values of hard work, determination, and perseverance. All my siblings and siblings in law have been a source of strength, inspiration, and support. Thank you for your unwavering support, encouragement, and belief in my abilities.

Finally, I acknowledge my friends, who have been a constant source of positivity, inspiration, and support throughout my academic journey. Thank you for your encouragement, motivation, and for always being there for me.

Table of Contents

List of Tables

List of Figures

Commonly Used Abbreviations

AD	Alzheimer's disease
BGGM	Bayesian Gaussian Graphical Models
CI	credible interval
CTT	Classical Test Theory
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
DSM-V	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
GDS	Geriatric Depression Scale
G-Theory	Generalisability Theory
ICD-10	International Statistical Classification of Diseases and Related Health Problems- 10^{th}
ICD-11	International Statistical Classification of Diseases and Related Health Problems- 11^{th}
IQCODE-16	Informant Questionnaire on Cognitive Decline in the Elderly
MAS	Memory and Ageing Sydney Study
MCI	Mild Cognitive Impairment
MMSE	Mini Mental State Examination
PP	Posterior Probability
SCC	Subjective Cognitive Complains
SCD	Subjective Cognitive Decline
TICS-M	Modified Telephone Interview for Cognitive Status

Chapter 1 General Introduction

Healthy and Abnormal Cognitive Ageing

Cognitive ageing refers to the process in which cognitive functioning gradually deteriorates when we get older (Harada et al., 2013). The decline in reasoning, memory, learning, thinking, planning and processing speed related to age can arise during adulthood and then progress into the later years (Blazer et al., 2015). Minor loss of cognitive capacities can be viewed as a normal part of ageing if it does not significantly impact an individual's daily functioning. In other words, if the loss of cognitive capacities related to the ageing process is relatively trivial (i.e., does not influence significantly an individual's life), this is considered healthy or normal cognitive ageing.

Some older adults could be considered as having abnormal cognitive ageing or cognitive decline (as compared with age-related peers) if they experience significant amount of decline in cognitive performance, which adversely affects their daily function and leads to clinical problems. According to Lipnicki et al. (2013), cognitive decline is defined as a deterioration of cognitive performance on variables measuring cognitive processes, whereby diagnosis of incident mild cognitive impairment (MCI) or dementia could be made based on the diagnostic criteria from the Diagnostic and Statistical Manual of Mental Disorders, Fourth or Fifth Edition (DSM-IV, DSM-V; American Psychiatry Association, 1994, 2013) or International Statistical Classification of Diseases and Related Health Problems- 10th or 11th Edition (ICD-10, ICD-11; World Health Organization, 2003, 2018).

Dementia and MCI are the two most common types of cognitive decline in ageing population, and these conditions significantly impact both individuals and communities. MCI is a prevalent cognitive condition among older adults and affects 19-42% of individuals over the age of 65. People with MCI are six to twelve times more likely to develop dementia, in comparison to their cognitively healthy peers (Roth et al., 2018). The continuous challenge

which clinicians face is to accurately assess cognitive decline that is considered as clinically significant (Hugo & Ganguli, 2014).

Clinical assessment of cognitive ageing typically involves history taking, bio-physiological assessments, and psychometric and/or neuropsychological assessments (Chin et al., 2019; Sachdev et al., 2010). Specifically, clinicians conduct a thorough history taking and utilize tools to assess cognitive functioning. They then formulate a clinical conclusion, with reference to standard criteria, such as those outlined in the relevant version of the DSM or ICD. Neuroimaging and biological markers can offer valuable insights and further evidence for the diagnosis and the identification of the condition subtype, or to support differential diagnoses, and contribute to the overall assessment process. While bio-physiological assessments, such as laboratory, physical, and neurological examinations, offer high precision and interval-level scaling, they are inherently limited as they do not reflect perceptual and subjective experiences of individuals. This underscores the need for psychometric measures that employ ordinal level scales to inquire about individual experiences. Moreover, most psychometric measures are constructed such that they include a number of statements or questions to which the respondent answers by selecting a response option on an ordinal scale, for example, a 5 Likert-scale ranging from "1 = Not at all" to "5 = Always". In addition, accurate assessment of cognitive ageing is essential, as it plays a critical role in precisely identifying cognitive aging conditions, and it ensures that assessment results are free from bias. Therefore, the use of accurate, reliable, and valid psychometric measures for cognitive aging can help identify cognitive decline at earlier stages and enhance the outcomes of early interventions.

Dementia and Mild Cognitive Impairments

The burden of dementia has been rising in New Zealand (NZ) and around the world (Roth et al., 2018; Wortmann, 2011). According to Roth et al. (2018), there were 40,000 people living with dementia in 2012 in NZ, and this number will grow to 75,000 by the year 2026.

This means that the economic burden will be approximately $19 billion in 2026 (Roth et al., 2018). There were approximately 35.6 million individuals worldwide living with dementia in 2010, but the number is expected to reach 115.4 million in 2050 (Prince et al., 2013). Evidence suggests that if appropriate preventive interventions can be initiated earlier, the onset of dementia may be delayed, and by eliminating potential risk factors which are modifiable, this can decrease the incidence of dementia by 35% to 50% (Rait et al., 2005). As such, early identification is key to reduce the individual and societal burden of dementia.

Dementia has a profound impact on older people and their families as well as the healthcare system. In the United States (US), Alzheimer's disease (AD) is not only one of the leading causes of death in older people but also contributes significantly to hospital admissions and the need for highly skilled nursing facilities in both hospital and home care settings (Kramarow & Tejada-Vera, 2019; Langa et al., 2017; Xu et al., 2010). Studies have revealed that the costs of health services including both formal and informal caregiving for individuals with dementia are substantial and continue to rise (Roth et al., 2018). Besides, family members who care for people with dementia also experience higher levels of distress (i.e., stress, anxiety, depression), and other health conditions, in comparison to typical families (Alzheimer's Association et al., 2013).

Mild cognitive impairment (MCI) refers to the transitional state between normal cognitive ageing and dementia, whereby functional abilities are preserved (Petersen, 2004). Detecting MCI can be challenging as individuals with MCI are often able to function independently or may only need minor support in their daily lives, despite objective evidence of cognitive decline (Albert et al., 2011). It is estimated that approximately 5% to 20% of adults aged over 65 years have MCI in the United Kingdom (UK) (Alzheimer's Society, 2020). The inconsistencies in the reporting on the prevalence of MCI can be explained by the lack of consensus on its definitions and subtypes (Ward et al., 2012).

As people with MCI are more likely to go on to develop dementia, research suggests that the link between MCI and dementia is significant, and early diagnosis and treatment of MCI can thus reduce the probability of its progress into severe cognitive decline, such as dementia or Alzheimer's disease (AD) (Petersen et al., 2005). The description of AD by Centers for Disease Control and Prevention (CDC; Centers for Disease Control and Prevention, 2020) will be used for this thesis, where AD is defined as a common subtype of dementia, characterized by impaired abilities in memory, thinking, and/or decision-making that interferes with doing everyday activities. It is important to note that although dementia mostly affects older adults, it is not considered a part of normal ageing (CDC, 2020).

Predictors of Cognitive Ageing

Generally, predictors of cognitive ageing can be classified into protective and risk factors. Risk factors are associated with the onset of the disorders, increase the incidence rate, or contribute to the development of illness (World Health Organization, 2022). Conversely, protective factors are associated with the delay of the onset of the disorders, contributing to early prevention, or a decrease in incidence rates (Brasso et al., 2021). Accurate identification and evaluation of risk and protective factors directly depend upon accuracy of measurement and the type and appropriateness of the statistical analysis that is conducted. Growing evidence has demonstrated the crucial contributions of both risk and protective factors in the field of cognitive decline in older people (Mehta & Yeo, 2017; Patnode et al., 2020).

Demographic Factors. Demographic risk factors for cognitively healthy aging include age, sex, educational background, and ethnicity (Garcia et al., 2018; 2021). In research on cognitive ageing, increasing age was consistently considered the strongest risk factor for the development of dementia, particularly for those who are above 80 years old (Langa et al., 2017). Sex was another demographic risk factor, which suggested that the prevalence of dementia is higher in women compared to men (Beam et al., 2018). For example, approximately 70,000 of

New Zealand population were living with dementia, with around 30% more women with dementia as compared to men (Alzheimers New Zealand, 2020). However, this difference might be due to longer life-expectancy in women (Gao et al., 1998; Jorm & Jolley, 1998). Low educational level and ethnicity were also considered vital risk factors for the prevalence of cognitive decline (Skoog et al., 2017). Studies in the US reported elevated prevalence in African American and Latino populations; but some authors have noted that these findings may be explained by lower education levels (Crimmins et al., 2018; Farina et al., 2020). In New Zealand, Māori and Pacific people with dementia were younger compared to New Zealand Europeans, and Pacifika seemed to have more serious dementia (Cullum et al., 2018). However, Cullum et al. (2018) did not report the education levels of the sample making it difficult to support such a speculation.

Meng and D'arcy (2012) conducted a meta-analysis to investigate the association between education and dementia. The authors reported that although education can be reflective of socioeconomic status (e.g., nutrition) or quality of environmental factors (e.g., medical system), the level of education is negatively associated with the prevalence of dementia. Bilingualism has been reported as a factor that can delay the onset of dementia (Alladi et al., 2013), and some studies suggest that bilingualism may specifically protect individual from a decline in attention and executive functioning (Craik et al., 2010; Mortimer et al., 2014).

Genetic factors. Deterministic autosomal dominant genes have been found to be a cause of some forms of dementia, while several genes were known to increase the likelihood of having dementia-like conditions such as Alzheimer's disease (Hugo & Ganguli, 2014). The most well-known is the Apolipoprotein E (APOE) polymorphism on Chromosome 19. Specifically, the APOE*4 allele was found to be linked with elevated risks of dementia due to hypercholesterolemia and heart disease, which were comorbid with Alzheimer's and

Parkinson's diseases (Chuang et al., 2010; Roses, 1996; Srinivasan et al., 2006; Tsuang et al., 2013; Yin et al., 2012). Twin studies indicated that homozygous individuals for APOE*4 were at higher risk of dementia than those who were heterozygous (Hugo & Ganguli, 2014). However, literature reviews showed that APOE*4 is considered a risk but not an indicator for diagnosing ageing-related cognitive dysfunctions as its effect on risk decreases when individuals were getting to 80 years of age or older. For example, older adults in their eighties with normal cognitive ageing who were APOE*4 positive did not show an increased vulnerability to developing dementia compared to those of the same age who were APOE*4 negative (Hugo & Ganguli, 2014).

Medical Factors. Cardiovascular and heart diseases have been identified as significant risk factors for dementia (Justin et al., 2013). Individuals with cardiovascular conditions in midlife, such as diabetes mellitus, high cholesterol, hypertension, and high body mass index have increased risk of dementia later in life (Beeri et al., 2004; Kivipelto et al., 2005). It has been observed that experiencing cardiovascular problems early in life may increase the risk of dementia (Justin et al., 2013). Moreover, atrial fibrillation and heart failure are risk factors for cognitive dysfunctions in older people (Qiu et al., 2006; Santangeli et al., 2012; Tilvis et al., 2004). Heart diseases can significantly affect cerebral hypoperfusion, which produces abnormal cellular energy leading to the release of toxic proteins (de La Torre, 2012). Studies have recently suggested that there is an association between elevated blood pressure and alterations in biomarkers, which link to the onset of Alzheimer's disease in cognitively healthy ageing individuals (Nation et al., 2013).

Furthermore, early inflammation and blood inflammatory markers (e.g., C-reactive protein, cytokines, and interleukins) have been reported as a risk factor of dementias (Ravaglia et al., 2007; Schmidt et al., 2002). Stroke and obstructive sleep apnea were also found to be associative with the onset of dementia (Budhiraja et al., 2010; Leys, 2009; Pendlebury &

Rothwell, 2009). Obstructive sleep apnea is also a known risk for cognitive decline in some older individuals because it increases the risk of hypertension, cardiovascular disease and stroke (Kim et al., 2013). Growing evidence indicates that white matter abnormalities are also associated with an increased risk of dementia (Kim et al., 2013). Head injury is another risk factor for a variety of cognitive dysfunctions in older adults including dementia, MCI as well as AD (Fleminger et al., 2003; Jellinger, 2004; Mayeux et al., 1993; Plassman et al., 2000). Furthermore, the severity of the injury is directly associated with the risk in older age (Barnes et al., 2014; Guo et al., 2000; Sundström et al., 2007).

Some studies have indicated a protective effect against dementia when using non-steroidal anti-inflammatory drugs (NSAIDs) (Launer et al., 1998; Szekely et al., 2008; Yip et al., 2005). However, a meta-analysis revealed that many positive results might contain recall bias, prescription bias, or publication bias (De Craen et al., 2005). Additionally, the lipid-lowering Hydroxymethylglutaryl-CoA (HMG-CoA) reductase inhibitors, also called “statins”, were reported to play a role as a protective factor against dementia in meta-analyses (Swiger et al., 2013; Wong et al., 2013).

The use of oestrogen therapy have been assumed to play a protective role against dementia, but research has shown contradictory evidence regarding its protective effects. The Women's Health Initiative Memory Study (WHIMS) trial found that there was no protective effect and perhaps an elevated risk to dementia if older women used combination hormone therapy (Shumaker et al., 2004; Shumaker et al., 2003). A meta-analysis reported similar findings (O'Brien et al., 2014). However, some research with repeated observations over a longer period found that the timing of hormone therapy could be a critical factor in protecting against cognitive decline, and particularly at menopause (Shao et al., 2012; Whitmer et al., 2011; Zandi et al., 2002).

Psychological Factors. Depression has been reported to be a complex and bi-directional risk factor associated with dementia. Young adults who have recurrent episodes of major depressive disorders are more likely to develop dementia or MCI in later life in comparison to those without mental health conditions (Dotson et al., 2010). Late-life depression is believed to be associated with vascular or degenerative diseases, contributing to abnormal cognitive ageing associated with dementia (Panza et al., 2010; Richard et al., 2013). Studies have also reported that anxiety in older age is associated with cognitive dysfunctions (Beaudreau & O'Hara, 2008; Mantella et al., 2007). Similarly, post-traumatic stress disorder has been reported to be a risk factor for dementia in both sexes, and the risk is increased in individuals who also suffer from depression (Flatt et al., 2018; Yaffe et al., 2010).

Moreover, cognitive activity plays an important role, and people who engage in lifelong occupations that do not require high skilled training or education are at increased risk of dementia (Bickel & Kurz, 2009; Bonaiuto et al., 1995). Engaging in popular leisure activities or hobbies (e.g., reading books, gardening, or participating in social games) has been associated with a reduced risk of dementia (Akbaraly et al., 2009; Karp et al., 2006; Verghese et al., 2006; Verghese et al., 2003). Engaging in cognitively stimulating activities, such as playing card games, puzzles, computer games or using the internet, has also been shown to be protective against cognitive decline and to improve cognitive processes in individuals of all ages (Hopkins et al., 2023; Hughes, 2010).

Lifestyle and Environmental Factors. Various environmental and occupational exposures have been found to have varying associations with cognitive decline in older adults (Brown et al., 2005). Exposure to pesticides is linked to increased risk of cognitive decline, and its mechanism has been established at the molecular level (Franco et al., 2010). Smoking has been identified as elevating the risk of dementia (Quik et al., 2007; Quik et al., 2012), even

though some researchers have found that there may be a selection bias in older people (Hernán et al., 2008; Quik et al., 2012).

Excessive consumption of alcohol has been reported to increase the probability of developing dementia (Anttila et al., 2004; Mukamal et al., 2003). However, other studies have found that alcohol consumption at mild and moderate levels is associated with a reduced risk of cognitive impairments and dementia (Anttila et al., 2004; Ganguli et al., 2005; Mukamal et al., 2003; Ruitenberg et al., 2002). Following a Mediterranean diet with mild to moderate alcohol consumption has been associated with improved cognitive functioning and a decreased risk of cognitive decline (Lourida et al., 2013).

Studies have also indicated that people who participated in physical activities were more likely to have a reduced risk of cognitive decline conditions (Hamer & Chida, 2009). Other characteristics, such as regularly engaging in mental, social or creative activities, were also protective factors against the risk of dementia (Kim et al., 2016; Wang et al., 2002). In addition, having a large social interaction network was associated with a lower incidence of dementia, particularly in older women (Crooks et al., 2008).

Conclusion

The predictors of cognitive ageing can be broadly classified into protective and risk factors, both of which appear to play important roles in cognitive decline in older individuals. Accurate identification of these risk and protective factors is essential for understanding cognitive ageing and in developing appropriate prevention and intervention strategies (Lim et al., 2019). However, the identification of these factors and their unique contributions is only possible when reliable and valid assessment tools are used in research and clinical practice. It is also important to consider multiple factors and their interactions when assessing an individual's risk of cognitive decline. Additionally, interventions targeting modifiable risk

factors, such as promoting healthy lifestyles and providing educational opportunities, may help delay the onset and reduce the incidence of cognitive decline in older adults. By understanding and addressing these predictors, we can strive to improve the cognitive health and quality of life for ageing populations. In other words, accurate assessment of cognitive ageing can help to identify cognitive decline at earlier stages and improve the outcomes of early interventions.

Assessments of Cognitive Ageing

Accurate assessment of cognitive aging and its relevant predictors can help facilitate strategies to improve healthcare for older individuals experiencing abnormal cognitive aging or cognitive decline. A clinical assessment typically involves history-taking, the collection of bio-physiological health indices, and objective cognitive testing. These assessments are often supplemented by reports from family or caregivers, which are obtained using psychometric measures (Chin et al., 2019). However, the reliability and validity of psychometric measures used to assess cognitive aging may be subject to scrutiny. Specifically, there has been limited evidence to comprehensively differentiate between enduring and dynamic patterns in psychometric measures employed for assessing cognitive aging. Moreover, while bio-physiological assessments obtained through appropriate laboratory, physical, and neurological examinations offer high precision and interval-level scaling, psychometric measures came with limitations and constraints, including the limited precision of ordinal scores or the use of ordinal scaling that may not be suitable for parametric statistical testing. Furthermore, predictors of cognitive ageing or their components interact with each other, and such interactions may then impact on the outcomes measured and thus challenging the validity of scales used to assess cognitive ageing.

Limited Evidence Exists to Comprehensively Differentiate Between Enduring and Dynamic Patterns in the Psychometric Measures Used to Assess Cognitive Ageing

Psychometric measures or scales are undeniably the most widely used tools in health sciences, particularly in psychology. However, there is limited robust evidence available to identify sources of measurement error and their unique contributions to assessments of cognitive aging. This estimation helps establish true reliability and distinguish between state

(dynamic) and trait (enduring) patterns in psychometric scales used in the field of cognitive aging.

In psychology, a trait or enduring pattern refers to a relatively stable characteristic of a person, while a state or dynamic pattern is described as a characteristic that a person adapts to a given situation and environment (Hamaker et al., 2007). For example, personality is considered a trait because it endures over time, while anxiety can manifest as both state and trait patterns. To gain a clearer understanding of the dynamic and enduring patterns of a phenomenon, all relevant aspects (i.e., state, trait, and their interactions) must be considered (Buss, 1989; Epstein, 1984).

A reliable measure of enduring patterns would reflect stable changes over time (e.g., alterations in cognition) and remain unaffected by an individual's transient changes (e.g., mood or current stress level). In opposition, a measure of dynamic patterns would be sensitive to transitory changes over time, meaning that the assessment score of a state measure is influenced by individual state changes. The clear differentiation of enduring and dynamic patterns of a measure is important because it helps to monitor both dynamic and enduring changes over time and distinguish clearly between them, which can improve reliability and validity of scales (Paterson et al., 2018).

For example, if a study can accurately separate enduring patterns from the dynamic patterns of a measure of cognitive ageing, the measure can then be used as the main indicator to evaluate the risks of a cognitive ageing condition (e.g., dementia). In contrary, if a measure of state aspects or dynamic patterns can be reliably identified, this assessment tool can be the main target of preventive interventions. Generalisability theory (G-Theory; Cronbach et al., 1963) has been advocated as the most appropriate method to examine sources of measurement error, as well as to investigate dynamic and enduring patterns in psychometric measures, and

examine reliability and generalisability of assessment scores (Medvedev et al., 2017; Truong et al., 2020).

Limitations in Using Ordinal Scales for Assessment, Diagnosis and Research

Stevens (1946) proposed the theory of scales of measurement, which stated that all measurements in science were conducted utilising four different types of scales: nominal, ordinal, interval, and ratio. The nominal scales differentiate between subjects based merely on their classifications, such as sex or ethnicity. The ordinal scales contain items, which are supposed to place objects of assessment in order, such as a four-point Likert scale ranging from one to four (e.g., "*strongly agree*" to "*strongly disagree*"). The interval scales measure a variable based on specific numerical scores or values, such as temperature. The ratio scales are similar to the interval scales, except the concept of the absolute zero point (e.g., speed has an absolute zero point) (Stevens, 1946). While interval or ratio scales inform the differences between scores, ordinal scales can only indicate whether scores are higher or lower compared to other scores (Sandham et al., 2019). In the fields of research and clinical practice related to cognitive ageing, assessments play a key role in the detection of cognitive decline, which can then inform diagnosis and facilitate appropriate treatment. While assessments of predictors related to bio-physiological indices have high precision due to the usage of interval- and ratio-level data, most psychological assessments are conducted using psychometric instruments relying largely on ordinal-level data (e.g., self-report questionnaires and cognitive tests), which might contribute to assessment bias, and compromise their precision, thus limiting their usage.

One limitation of ordinal scales is that the distances between subsequent categories are treated as having the same values. This implies that the difference between "*strongly disagree*" and "*disagree*" is 1 and the difference between "*disagree*" and "*slightly disagree*" is also 1, which is not an assumption of the ordinal scale (Jamieson, 2004). By definition of the ordinal

scale, we can only assume that the subsequent response category is higher than the previous one, which does not imply that the differences are the same (Merbitz et al., 1989). Consequently, ordinal data cannot be added, subtracted, multiplied or divided meaning that means and standard deviations are meaningless in this case (Merbitz et al., 1989). However, the assessment scores are normally computed using numerical values assigned to categories while ignoring the fact that the differences between categories may not be the same. Even so, clinicians and researchers have most widely used ordinal scales to detect risk and protective factors.

Another limitation is that scores collected from ordinal measurements are considered inappropriate to carry out statistical parametric tests and thus it is inappropriate to conduct comparative analyses with interval level variables, such as neuroimaging and interval-transformed scores (Linacre, 2004; Norquist et al., 2004). Ideally, all measures should be at least interval-level in order to meet the basic assumptions of parametric tests. For example, Rasch methodology (Rasch, 1960, 1961) is a relevant solution to improve psychometric properties of an ordinal level scale up to an interval-level measure if the strict assumptions of the unidimensional Rasch model, which defines parameters of an interval scale, are met (Hobart & Cano, 2009; Merkin et al., 2020; Tennant & Conaghan, 2007). This method can be used to develop ordinal-to-interval conversion algorithms and transformed interval-level data can be used for conducting parametric statistical analyses (Linacre, 2004).

Lack of Research Investigating the Complex Interactions Across Components of Predictors of Cognitive Ageing

Predictors of cognitive ageing and/or their components interact with each other, and these interactions should be taken into consideration. Specific predictors and their components (e.g., assessment domains or items) are more or less relevant to specific aspects of cognition

(e.g., memory or language) and there is also a complex interaction between them that may impact on the outcome. Correlational studies in cognitive ageing typically focus on the associations between specific predictors and outcomes but there is lack of research investigating the complex interactions across components of predictors of cognitive ageing. Such complex interactions can be investigated by network analysis methodology (Borsboom & Cramer, 2013).

Network analysis provides a robust and comprehensive statistical approach to examine the validity of a measure and its components of cognitive ageing (e.g., the MMSE and its cognitive domains) in relation to a more comprehensive test assessments (e.g., neuropsychological test performance) (Borsboom & Cramer, 2013; Chalmers et al., 2022). Network analysis can provide a clearer depiction of the global network of unique associations between a set of components, thereby enhancing understanding of the complex relationships among these components (Borsboom & Cramer, 2013; Cholerton et al., 2016). All associations in the network, estimated through network analysis, are not classified as either dependent variables (i.e., caused by other variables) or independent variables (i.e., affecting the dependent variable) (Åkerblom et al., 2021). By estimating the unique associations between variables in a network, the unique links between different domains can be outlined, thereby leading to a clearer understanding of the validity of psychometric measures in the field of cognitive ageing.

In summary, it is important to estimate sources of measurement error and differentiate between state (dynamic) and trait (enduring) patterns of cognitive assessments as it helps in monitoring changes of these patterns over time and enhancing the reliability and validity of psychometric measures used to assess cognitive ageing. However, limited evidence exists to comprehensively differentiate between enduring and dynamic patterns in the psychometric measures used in the field of cognitive aging. Moreover, psychometric assessment tools relying on ordinal scales present limitations, such as the equal distances between subsequent categories

of ordinal scales restrict the meaningful application of mathematical operations and statistical parametric tests. Furthermore, predictors of cognitive ageing or their components interact with each other, and such interactions may then impact on the outcomes measured at the same time, as well as their ability to predict the future outcomes of cognitive health. Together, these issues challenge the reliability and validity of scales used to assess cognitive ageing and can be addressed using advanced theories of measurement, including G-Theory (G-Theory; (Cronbach et al., 1963), Rasch analysis (Rasch, 1960, 1961) and Network analysis (Borsboom & Cramer, 2013). These extended statistical methodologies offer promising approaches to enhance the psychometric properties of scales used for assessment of cognitive ageing and advance our knowledge and understanding of cognitive ageing processes.

Theories of Measurement

Investigating Enduring and Dynamic Patterns of Measures of Cognitive Ageing Using G-Theory

Even though most widely used psychometric measures of cognitive ageing achieve adequate internal consistency, they do not reflect their ability to differentiate between enduring and dynamic patterns or aspects of a variable they supposed to measure (e.g., Subjective Cognitive Complains [SCC]). Internal consistency coefficients, such as Cronbach's alpha or McDonalds omega, are not able to reflect temporal reliability of psychometric measures as they estimate the internal consistency of measures at only one time point. Similarly, the intraclass correlation coefficient (ICC) and test-retest correlations that often used to estimate reliability over time have limited accuracy because they do not account for multiple potential sources error variance (i.e., individual scale items, assessment occasions and/or their interactions with the object of measurement or person) (Bloch & Norman, 2012; Medvedev, Krägeloh, et al., 2017).

The traditional Classical Test Theory (CTT) approach to the distinction between enduring (trait) and dynamic (state) patterns through examining test-retest reliability of a scale over time, where the coefficients tend to be lower for measures of state (e.g. <0.70) (Ramanaiah et al., 1983; Spielberger, 1999; Spielberger et al., 1970). However, these reliability coefficients (test-retest coefficients) have limited accuracy in differentiation between dynamic and enduring patterns of the scale because these are merely based on correlations between total scale scores at two separate time points (e.g., Occasion 1 and Occasion 2). For instance, if an individual shows improvement in one symptom but experiences deterioration in another, the overall score stays constant, failing to reflect clinically significant changes. Therefore, these coefficients do not account for the measurement error caused by individual scale items over

time, nor do they consider other potential sources such as occasions (time points), person, and/or their interactions with each other (Bloch & Norman, 2012; Medvedev, Berk, et al., 2020). For example, a response to an individual item of scale might be contingent on the assessment occasion rather than actual changes in the individual's performance. Hence, it is essential to have a comprehensive estimation of reliability, and G-Theory has been recommended as the most appropriate approach to examine both dynamic and enduring patterns in an assessment tool, while simultaneously examining reliability and generalisability of assessment scores (Medvedev, Krägeloh, et al., 2017; Truong et al., 2020).

G-Theory is an advanced extension of CTT, and is a comprehensive approach to evaluate the overall reliability of psychometric instruments (Brennan, 2009; Shavelson & Webb, 1991). CTT assumes that a measurement consists of true score and a single error variance, while G-Theory utilises ANOVA to estimate all potential error variance of multiple sources that may affect the precision of the measurement (Allen & Yen, 2001; Cronbach et al., 1963). Furthermore, CTT only examine one aspect of the reliability of a measure, for example, internal consistency or test-retest reliability coefficients. As G-Theory is an extension of CTT, it simultaneously examines all involved sources of error variance that may affect reliability (e.g., person, scale items, occasion, and their interactions to each other) (Medvedev, Krägeloh, et al., 2017; Shavelson et al., 1989). Numerous studies have shown the suitability of G-theory as the most appropriate approach for evaluating the overall reliability and generalisability of assessment scores and differentiation between dynamic and enduring patterns in a measure (Arterberry et al., 2014; Medvedev, Krägeloh, et al., 2017; Paterson et al., 2018; Truong et al., 2020). Therefore, it is undeniable that G-Theory can be valuable in evaluating and improving the precision of a psychometric instrument by identifying and modifying sources of measurement error, as well as differentiating between enduring and dynamic patterns in such

a measure. Nevertheless, there have been limited studies employing this method in the field of cognitive aging to date.

Evaluating and Enhancing Ordinal Scales Using Rasch Methodology

Ordinal scales only indicate whether scores are higher or lower than compared to other scores (Sandham et al., 2019). For example, if an ordinal scale uses the four-point Liker scale (e.g., 1 is "*strongly disagree*", 2 is "*disagree*", 3 is "*agree*", and 4 is "*strongly agree*"), it means that a subsequent response category (e.g., "*disagree*") is higher than the previous one (e.g., "*strongly disagree*"). It does not indicate the precise interval between two response categories of the ordinal scale. Therefore, if we assume that the difference between "*strongly agree*" and "*agree*" is considered the same as the difference between *"disagree"* and *"strongly disagree"* (i.e., 4-3 = 2-1), this would violate fundamental mathematical and statistical assumptions as we only know that 1 < 2 < 3 < 4, but we do not know the real differences between these categories (Jamieson, 2004; Merbitz et al., 1989), meaning that the scores collected from ordinal scales are not suitable to perform mathematical operations such as adding, subtracting, dividing and multiplying making computing means and standard deviations meaningless (Merbitz et al., 1989). For this reason, the ordinal scale estimates are not suitable to conduct statistical parametrical tests as the scales based on Likert type yield only ordinal raw scores (Merbitz et al., 1989). Questions also arise about whether normal distribution assumptions are met to allow a number of well-used statistical tests, such as *t*-tests and ANOVA. If the number of Likert scale points is small (i.e., three or four points), such scales tend to violate the assumptions of normality (Hodge & Gillespie, 2007; Leung, 2011).

In a clinical context, all items of an ordinal scale are displayed as having the same level of contribution to the overall construct and the sum score is calculated by adding individual item scores, which can bias the diagnostic results (Tennant & Conaghan, 2007). Rasch analysis

permits conversion of ordinal scores into interval level data by estimating person ability and item difficulty parameters, if the data fits Rasch model, which defines parameters of an interval scale (Hobart & Cano, 2009; Medvedev et al., 2019). Rasch analysis has demonstrated its advanced theoretical and empirical application in examining and enhancing psychometric properties of ordinal scales, as it improves accuracy of assessment and provides better covering of the sample population (Christensen et al., 2013). Literature reviews have revealed that the Rasch model has been widely employed to evaluate and enhance psychometric properties of a number of ordinal measures used in health-related research and clinical practice fields (Hobart & Cano, 2009; Medvedev et al., 2019; Medvedev, Krägeloh, et al., 2020; Medvedev, Pratscher, et al., 2020; Medvedev, Siegert, Mohamed, et al., 2017; Medvedev, Turner-Stokes, et al., 2018; Merkin et al., 2020).

The Rasch model is one of the strictest methods to examine item parameters and scale properties (Hobart & Cano, 2009). The basic Rasch model is unidimensional, probabilistic and logistic, which permits to accurately estimate scale ability to differentiate at individual levels of an assessment (Christensen et al., 2013; Rasch, 1960, 1961). According to the model, the response to a particular item of an instrument is influenced by individual ability and item difficulty. In the literature, the Rasch model is considered identical to the one parameter Item Response Theory (IRT) model (Hobart & Cano, 2009). However, the Rasch model was composed before the first publication of IRT (Lord & Novick, 1968); thus, Rasch and IRT may be considered different paradigms.

The aim of IRT is to find a suitable model which explains the data most precisely, whereas the Rasch model is defined by principles of fundamental measurement (Thurstone, 1931), which are similar to the laws of physics (Rasch, 1960). Fundamental measurement means that a measurement scale should ‘transcend’ sample diversity so that a scale must equally fit for every particular person regardless of individual factors (e.g., ages, sexes). The

assessment instrument should measure one parameter of the object, which refers to the unidimensionality of the Rasch method; this is also a universal criterion for all interval- or ratio-level measurements. A unit of the measurement should remain the same at all parts of the continuum of a scale, which indicates that scales are at an interval level as a requirement of the model.

When achieving the Rasch model fit, the ordinal raw scores of a measures can be converted into interval-level data (Linacre, 2004; Norquist et al., 2004). This interval-transformed data will reflect precise changes on a latent trait which is similar to other interval level scales (e.g., temperature or height) and hence, increase the precision of a scale. This is the most advanced benefit of the Rasch model over CTT methods because the classical statistic techniques are not capable to account the unique contribution of each scale individual item to the overall latent trait. Also, CTT methods do not differentiate between item difficulty based on individual ability (Fox & Jones, 1998). The Rasch analysis method has the ability to estimate the difficulty of each item in a scale, assess the appropriateness of response options used in polytomous items and finally convert an ordinal level measurement into an interval measure (Hobart & Cano, 2009; Rasch, 1960; Tennant & Conaghan, 2007; Wilson, 2004; Wright & Stone, 1979). Thus, Rasch methodology can precisely evaluate reliability and validity of a measurement as well as improve its psychometric properties.

Using Network Analysis to Investigate Cognitive Assessments

Contemporary research in cognitive ageing typically focuses on specific measures that measure the particular aspects of cognitive decline in ageing populations (Brasure et al., 2018; Peters et al., 2019). There have been many measures used in the assessments of cognitive ageing and each measure usually consists of a number of components (i.e., assessment items or domains). However, the complex interactions between different measures and their

components in the assessments of cognitive ageing are not yet well-understood (Cholerton et al., 2016; Peters et al., 2019). For instance, there is a lack of research examining unique links between different cognitive ageing measures (e.g., global cognitive domains) and the major outcomes (e.g., neuropsychological domains).

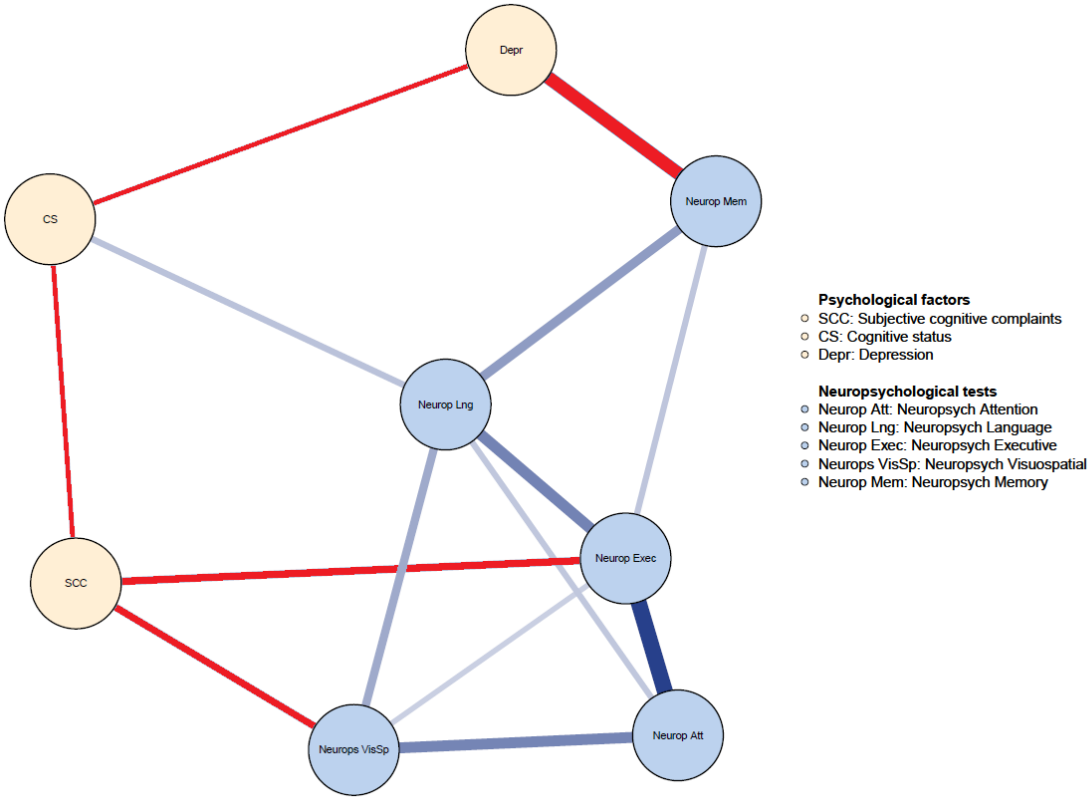
Network analysis is a modern statistical approach that was developed to study and enhance the understanding of psychopathology (Borsboom & Cramer, 2013; Contreras et al., 2019). Network analysis was designed to examine the interplay between measures as well as sub-measures (e.g., sets of components, individual subscales, symptoms, or domains). From a statistical perspective, all associations in the network, estimated through network analysis, are not classified as either dependent or independent variables (Åkerblom et al., 2021). This approach is a preferred analysis to determine reciprocal or mutually maintaining mechanisms within a system when no transparent causal pathways. By estimating the unique associations between variables in a network, the unique links between different domains can be outlined. Therefore, network analysis can provide a clearer depiction of the global network of unique associations between symptoms or domains of measures, thereby enhancing understanding of the complex relationships between variables (Borsboom & Cramer, 2013; Cholerton et al., 2016).

Another advantage of network analysis is its visual representation of the existence and intensity of associations between various variables using “nodes” and “edges” (Borsboom et al., 2021). The “nodes” signify the variables within the network, while the “edges” represent distinct statistical relationships between the nodes, often in the form of partial correlations. Figure 1.1 displays a demonstration of the network between global cognition or cognitive status (TICS-M), subjective complaints (IQCODE-16), depression (Geriatric Depression Scale [GDS]) and neuropsychological cognitive domains. As can be seen in Figure 1.1, the nodes represent the relevant variables, and the edges represent unique associations between the

nodes/variables. The strength of each relationship is reflected by the thickness of the edge (i.e., the line), with blue edges representing positive correlations between measures, while red edges represent negative correlations, and node colour indicates a family of nodes (i.e., light yellow for psychological factors and light blue for neuropsychological tests). This hypothetical network illustrates the complex interactions between psychometrics measures used to assess cognitive ageing (TICS-M, IQCODE-16, and GDS) with neuropsychological domains and hence revealed the construct validity, particularly the convergent and divergent validity, of each measure.

Figure 1.1

Demonstration of the correlational network between cognitive status (TICS-M), subjective complaints (IQCODE-16), depression (GDS) and neuropsychological/cognitive domains



Notes. This figure is used for demonstration purposes. Each node represents a variable. Lines between nodes are edges. Blue and red edges reflect positive and negative associations, respectively. Wider and darker edges are represented stronger associations.

The demonstration network was based on the Bayesian Gaussian Graphical Models (*BGGM*; Williams & Mulder, 2020) which graphically present the multiple statistically significant associations between nodes/variables simultaneously. To select which unique relations were statistically significant, the *BGGM* estimates a 95% credible interval (*CI*) for each variable-to-variable association and a CI that excluded zero is considered to indicate a statistically significant edge between nodes. A CI is a range of scores where a specific relation will fall 95% of the time (Hespanhol et al., 2019). These features makes network analysis more advanced than "traditional" methods (e.g., regression and mediation/moderation analyses), making it easier to identify important variables and understand the structure of a set of variables. Thus, network analysis is a promising method for evaluating the validity of psychometric measures in the field of cognitive ageing (Åkerblom et al., 2021; Chalmers et al., 2022).

Aims and Objectives

This thesis aimed to investigate reliability, validity, and enhance psychometric properties of key assessment instruments used to measure cognitive ageing by harnessing the pre-existing longitudinal dataset of the Memory and Ageing Sydney Study (MAS; Australia) study (Sachdev et al., 2010). All research included in this thesis were approved by the University of Waikato Human Research Ethics Committee, ethics approval application code "HREC(Health)2020#41: Protective and Risk Factors of Mentally Healthy Ageing" (see Appendix 1.1).

Table 1.1 shows assessments of cognitive ageing relevant to the current work, which were extracted from the MAS study (Sachdev et al., 2010). There were various assessment tools used to evaluate the cognitive and psychological predictors of cognitive ageing. Firstly, to diagnose dementia or MCI during each biennial follow-up, individuals attended a consensus review meeting. During this meeting, a panel of at least three clinicians (i.e., neuropsychologists, neuropsychiatrists, and psychogeriatricians) discussed all the accessible laboratory, neuropsychological, clinical, and imaging data to reach a collective diagnosis. The diagnosis of dementia was determined according to the DSM-IV criteria (Sachdev et al., 2010).

To assess subjective cognitive complaints, two psychometric measurement tools were used: the Memory Assessment Clinic Questionnaire (MAC-Q; Crook et al., 1992) and the 16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16; Jorm, 1994). In addition, in order to measure cognitive status in older people, the Mini Mental State Examination (MMSE; Folstein et al., 1975) and the Modified Telephone Interview for Cognitive Status (TICS-M; Brandt et al., 1993) were used. It is noteworthy that this thesis focused specifically on these four measures of subjective cognitive complaints and global cognition due to their wide usage and significance as assessment tools for diagnosing and monitoring cognitive ageing.

Table 1.1

Predictor and outcome assessments in Sydney Memory and Ageing Study

Factor assessments	Follow-ups					
	1	**2**	**3**	**4**	**5**	**6**
Demographics	x	x	x	x	x	x
Clinical Status						
Clinical diagnosis (MCI/Dementia)	x	x	x	x	x	x
Cognitive Complaints						
Memory Assessment Clinic Questionnaire (MAC-Q)	x	x	x	x		x
16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16)	x	x	x	x	x	x
Cognitive Status						
Mini Mental State Examination (MMSE)	x	x	x	x		x
Modified Telephone Interview for Cognitive Status (TICS-M)	x	x	x	x	x	x
Neuropsych/cognitive domains						
Wechsler Adult Intelligence Scale-III (WAIS-III) Digit Symbol-Coding	x	x	x	x		
Trail Making Test A	x	x	x	x		
Boston Naming Test	x	x	x	x		
Animal Fluency	x	x	x	x		
FAS Fluency	x	x	x	x		
Trail Making Test B	x	x	x	x		
Wechsler Adult Intelligence Scale-Revised (WAIS-R) Block Design	x	x	x	x		
Wechsler Memory Scale-III (WMS-III) Logical Memory Story A delayed recall	x	x	x	x		
Rey Auditory Verbal Learning Test	x	x	x	x		
The Benton Visual Retention Test	x	x	x	x		

Notes. Highlighted texts present factors; x indicates the times of follow-up which assessments were administrated.

Cognitive performance was assessed using a comprehensive neuropsychological test battery which comprised 10 tests that measured cognitive domains of language, visuospatial ability, memory, attention/processing speed, and executive function. Language was measured using the Boston Naming Test (Kaplan et al., 2001) and Animal Fluency (Spreen, 1977). Visio-spatial ability was measured using the Wechsler Adult Intelligence Scale-Revised (WAIS-R) Block Design (Wechsler, 1981b). Memory was measured using the Wechsler Memory Scale-III (WMS-III) Logical Memory Story A delayed recall (Wechsler, 1997), Rey Auditory Verbal Learning Test (Strauss et al., 2006) total learning, short-term and long-term recall scores, and the Benton Visual Retention Test (Benton et al., 1996). Attention/processing speed was measured using the Wechsler Adult Intelligence Scale-III (WAIS-III) Digit Symbol-Coding (Wechsler, 2003) and the Trail Making Test A (Tombaugh, 2004). Executive function was measured using FAS Fluency (Benton, 1967) and Trail Making Test B (Reitan & Wolfson, 1993).

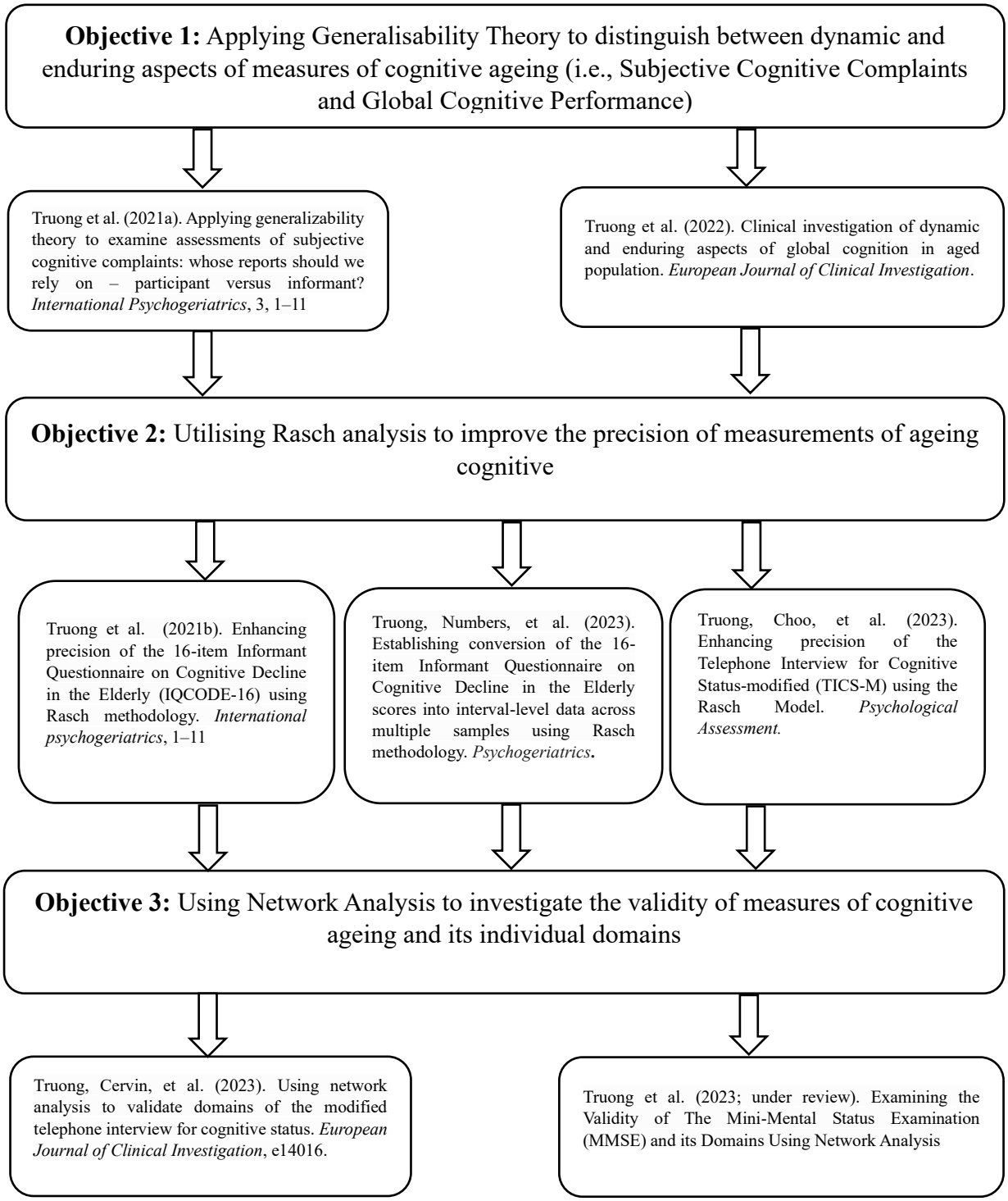
Two more longitudinal datasets were also used in this thesis: the Older Australian Twins Study (OATS) (OATS; Sachdev et al., 2009), and the Canberra Longitudinal Study (CLS; Christensen et al., 2004; Jorm et al., 2000) (*see* Chapter 5 for more details of these studies). This work was conducted to further enhance the validity of one of the most important cognitive ageing assessments, that is, the IQCODE.

Figure 1.2 presents the flow diagram of the thesis objectives and studies conducted to fulfil these objectives. As can be seen, the first objective was to apply G-theory to investigate dynamic and enduring aspects of assessments of cognitive ageing. The clear differentiation of enduring and dynamic aspects of assessments of cognitive ageing would help to monitor the changes in both these aspects and was necessary to develop and implement the most appropriate diagnostic tools, as well as to facilitate interventions for conditions caused by ageing. Previous works showed that G-theory is a reliable and useful statistical approach to

distinguish dynamic and enduring aspects of measures. To satisfy this objective, two studies investigated measurement of enduring and dynamic patterns of subjective cognitive complaints (also known cognitive changes in older people) and global cognitive performance respectively.

Figure 1.2

Flow diagram of the thesis "Assessment of Cognitive Ageing: Applying Generalisability Theory, Rasch and Network Analyses to Establish Reliability and Validity of Measures"



The second objective extended the work under the first objective. Assessment tools which were found reliable in measuring enduring aspects of cognitive ageing were selected as subjects for this objective, and Rasch methodology was applied to investigate and improve the precision of the scales used to assess cognitive ageing. The Rasch method was selected because this model was designed for assessment of latent traits. Finally, the third objective of this thesis was to, for the first time, investigate the validity of cognitive ageing measurements using network analysis. This final objective evaluated the validity of widely used measures of cognitive ageing and their individual domains (e.g., TICS-M and MMSE domains) in relation to neuropsychological test performance (Borsboom & Cramer, 2013). This work advanced the understanding of global network mechanisms involved in assessments of cognitive ageing, thereby allowing us to evaluate and enhance assessments of cognitive ageing.

Chapter 2 Applying Generalisability Theory to Examine Assessments of Subjective Cognitive Complaints: Whose Reports Should We Rely on - Participant versus Informant?

Truong, Q. C., Choo, C., Numbers, K., Merkin, A. G., Brodaty, H., Kochan, N. A., Sachdev, P. S., Feigin, V. L., & Medvedev, O. N. (2021). Applying generalizability theory to examine assessments of subjective cognitive complaints: whose reports should we rely on – participant versus informant? *International Psychogeriatrics*, 1–11. https://doi.org/doi.org/10.1017/s1041610221000363

Declaration

This study is co-authored work with C. Choo, K. Numbers, A. G. Merkin, H. Brodaty, N. A. Kochan, P. S. Sachdev, V. L. Feigin, and O. N. Medvedev (see Appendix P1 for the original article). I was the first author and made the major contribution to this study that included designing and conducting the study, completing data analysis, writing the manuscript for submission to *International Psychogeriatrics* journal, and responding to reviews/addressing their comments during peer-review process. All co-authors have signed to the co-authorship form (see Appendix 2.1).

Abstract

***Objectives*:** This study aimed to apply the generalizability theory (G-Theory) to investigate dynamic and enduring patterns of subjective cognitive complaints (SCC), and reliability of two widely used SCC assessment tools.

***Design*:** G-Theory was applied to assessment scales using longitudinal measurement design with five assessments spanning 10 years of follow-up.

***Setting*:** Community-dwelling older adults aged 70-90 years and their informants, living in Sydney, Australia, participated in the longitudinal Sydney Memory and Ageing Study.

***Participants*:** The sample included 232 participants aged 70 years and older, and 232 associated informants. Participants were predominantly White Europeans (97.8%). The sample of informants included 76 males (32.8%), 153 females (65.9%), and their age ranged from 27 to 86 years, with a mean age of 61.3 years (*SD* = 14.38).

***Measurements*:** The Memory Complaint Questionnaire (MAC-Q) and the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16).

***Results*:** The IQCODE-16 demonstrated strong reliability in measuring enduring patterns of SCC with G = 0.86. Marginally acceptable reliability of the 6-item MAC-Q (*G* = 0.77-0.80) was optimized by removing one item resulting in *G* = 0.80-0.81. Most items of both assessments were measuring enduring SCC with exception of one dynamic MAC-Q item. The IQCODE-16 significantly predicted global cognition scores and risk of dementia incident across all occasions, while MAC-Q scores were only significant predictors on some occasions.

***Conclusions*:** While both informants' (IQCODE-16) and self-reported (MAC-Q) SCC scores were generalizable across sample population and occasions, self-reported (MAC-Q) scores may be less accurate in predicting cognitive ability and diagnosis of each individual.

***Keywords*:** ageing; cognitive assessment; dementia; generalizability theory; longitudinal design; mental capacity.

Introduction

Older adults often report subjective cognitive complaints (SCC), which relate to an individual's self-experience of cognitive deterioration (Hildreth & Church, 2015). Currently, SCC contribute to the criteria for a diagnosis of mild cognitive impairment (MCI) (Petersen, 2016; Winblad et al., 2004), and may be considered as the earliest detectable stage of pre-clinical dementia (Jonker et al., 2000; Mitchell et al., 2014). SCC can be self-reported or reported by informants (e.g. family member or friend) with the advantage of capturing daily cognitive and memory changes that standardized neuropsychological tests may not detect (Brodaty et al., 2002; Jorm et al., 1991; Numbers et al., 2020).

Despite the potential benefits of SCC assessments, it remains questionable as to whether self-reported SCC reliably predict objective cognitive performance and/or dementia incident. A relationship between self-reported SCC and cognitive impairment ranges from negligible (e.g., Burmester et al., 2016; Crumley et al. 2014) to none (e.g., Lenehan et al., 2012; Reid & MacLullich, 2006) . One explanation for such inconsistency may be the influence of mood and certain personality traits on complaining behaviours (Ponds & Jolles, 1996). It is well established that subjective impressions of decline are exacerbated by depression and anxiety, as well as personality traits such as neuroticism and conscientiousness (Reid & MacLullich, 2006). Therefore, subjective reports of cognitive ability provided by close informants may present more reliable approximation of objective cognitive performance (Slavin et al., 2015), and future cognitive decline (Caselli et al., 2014). Furthermore, in the clinical setting, informant-reported SCC are often increasingly relied upon as individuals progress through pre-clinical stages of dementia, and begin losing insight into their cognitive changes over the debilitating course of dementia (American Psychiatric Association, 1994). However, no empirical examination was conducted to date using an appropriate methodology to investigate whose reports (i.e., participants' or informants') are more reliable, and at what stage researchers

and clinicians should rely on which reports. As dementia is typically marked by insidious onset and gradual progression, a longitudinal design will be ideal in tracking any cognitive change over time (American Psychiatric Association, 1994).

Moreover, it is important to differentiate reliably between dynamic and enduring SCC patterns over longer time. A reliable trait measure would reflect enduring changes overtime (e.g., alterations in long-term subjective cognition) and remain unaffected by individual's transient changes (e.g., mood or current stress level). Conversely, a state measure would be sensitive to dynamic changes, which may confound assessment of long-term subjective cognition. While the widely used SCC measures such as the Memory Complaint Questionnaire (MAC-Q; Crook et al., 1992) and the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16; Jorm, 1994) have good internal consistency, it does not support their ability to distinguish between enduring and dynamic patterns of SCC. Internal consistency coefficients (e.g., Cronbach's alpha) are not appropriate to estimate temporal reliability of scales because they only estimate consistency or inter-correlations between individual items at one time point. Moreover, test-retest reliability coefficients often used to distinguish between dynamic and enduring patterns have limited accuracy because these are merely correlations between total scale scores at two different times (e.g., Time 1 and Time 2). For example, if a person improves on one symptom but gets worse on another, the total score remains the same without reflecting clinically important changes. Therefore, these coefficients do not account for variability of individual items over time and other sources of measurement error such as the effects of item, occasion, person, and their interactions (Bloch & Norman, 2012; Medvedev, Berk, et al., 2020). For instance, a response to an item may depend on assessment occasion rather than changes in individual's performance. A comprehensive estimation of reliability is therefore required and Generalisability Theory (G- Theory) was advocated as the most appropriate method to investigate dynamic and enduring patterns in a measure, and examine

reliability and generalisability of assessment scores (Medvedev, Krägeloh, et al., 2017; Truong et al., 2020).

G-Theory is a successor of Classical Test Theory (CTT), and is particularly well suited to examine the overall reliability of psychometric instruments (Brennan, 2009; Shavelson & Webb, 1991). While CTT postulates that any measurement consists of true variance and error variance presented as a single factor, G-Theory utilizes ANOVA to estimate all possible sources of error variance that may affect the main outcome variable, as well as the accuracy of the measurement itself (Allen & Yen, 2001; Cronbach et al., 1963). Furthermore, CTT evaluates the reliability of a measure at only one aspect (e.g., internal consistency) at a time, or examines the distinction between dynamic and enduring patterns of a measure using test-retest reliability coefficients. G-Theory extends CTT and simultaneously examines all potential sources of error variance that may influence reliability such as person, scale items, occasion, and all their interactions (Medvedev, Krägeloh, et al., 2017; Shavelson et al., 1989). Many studies have demonstrated applicability of G-theory as the most appropriate method for estimating the overall reliability and generalisability of assessment scores and distinction between dynamic and enduring patterns in a measure (Arterberry et al., 2014; Medvedev, Krägeloh, et al., 2017; Paterson et al., 2018; Truong et al., 2020). Therefore, applying G-Theory can be useful to examine and improve the precision of a psychometric instrument, as well as differentiate between enduring and dynamic patterns reflected by such measure.

The aim of the current study was to apply G-Theory to examine reliability, and distinguish between dynamic and enduring patterns in the self-report MAC-Q and informants IQCODE-16 SCC assessment tools. A longitudinal design was utilized with participants assessed at five occasions, separated by 2-4 years intervals. Application of G-Theory involved two parts: a Generalisability study (G-study) and a Decision study (D-study). The purpose of the G-study was to examine the overall generalisability of the MAC-Q and IQCODE-16 and

evaluate sources of error variance in each measure. D-study aimed to subsequently evaluate psychometric properties of individual items of these two scales, and to manipulate measurement design to optimize the reliability of measurement (Cardinet et al., 2011; Shavelson et al., 1989). This study also aimed to evaluate the utility of these scales for predicting incident dementia and global cognition scores.

Method

Participants

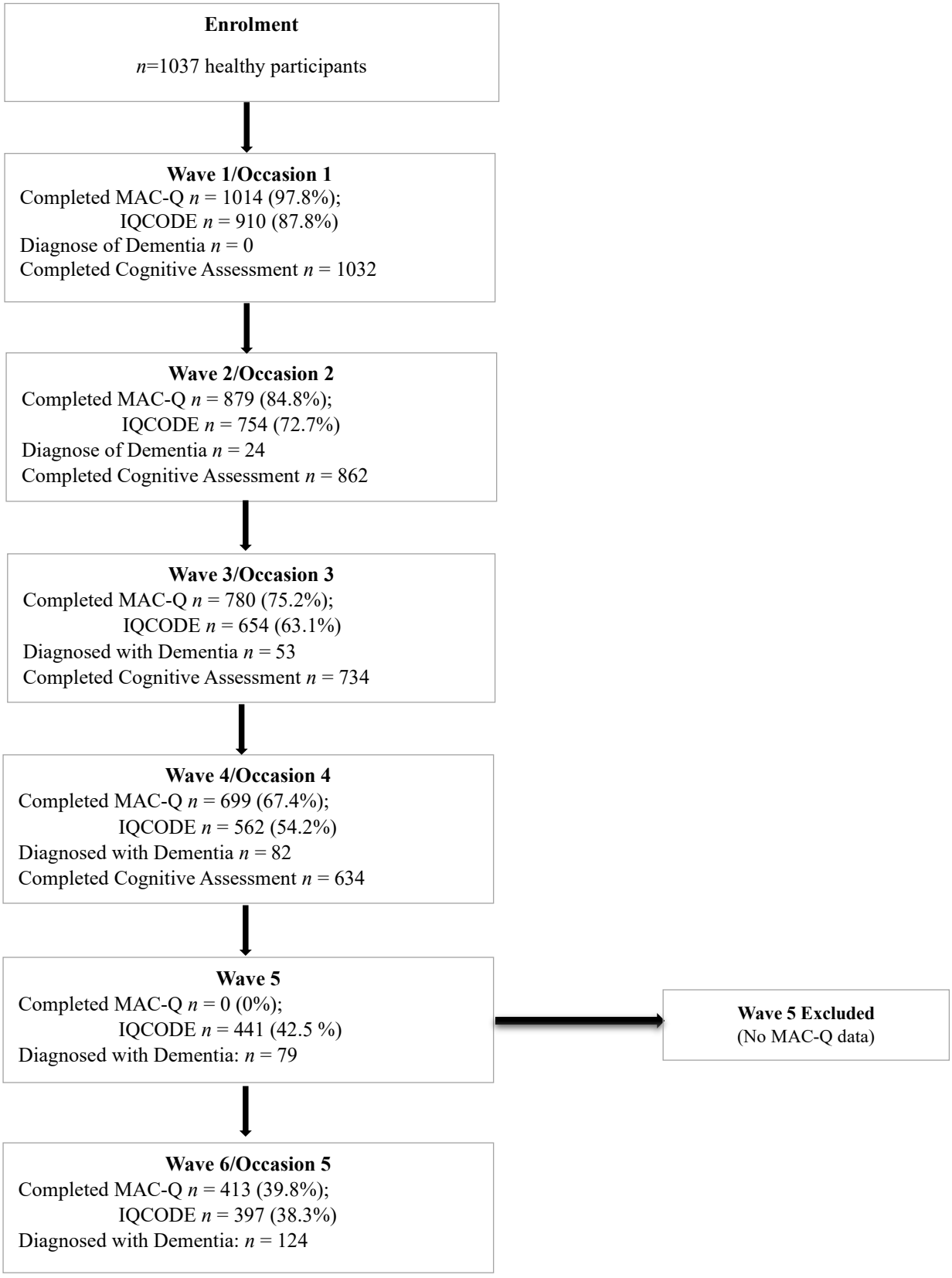
Community-dwelling older adults aged 70 – 90 years, living in the Eastern Suburbs of Sydney, Australia, were selected via the electoral roll and invited to participate in the Sydney Memory and Ageing Study (MAS; Sachdev et al., 2010). Of 8,914 individuals invited to participate, 1,037 participants were included in the baseline sample (occasion 1). Inclusion criteria were the ability to speak and write English sufficiently well to complete a psychometric assessment and self-report questionnaires. Exclusion criteria were any major psychiatric diagnoses, acute psychotic symptoms, current diagnosis of multiple sclerosis, motor neuron disease, developmental disability, progressive malignancy, and/or dementia. More detailed methods of recruitment and baseline demographics have been previously described by Sachdev and colleagues (Sachdev et al., 2010). Of the 1037 participants included in the present study, 1,009 (97.3%) had an informant. Informants were selected by nominations of participants. Informants answered questions relating to the participant's memory, thinking, and daily functioning. Qualified informants were those who had at least 1 hour of contact with the participant per week; on average they had 8.3 hours of weekly contact. All participants and informants provided written consent to participate in this study, which was approved by the University of New South Wales Human Ethics Review Committee (HC 05037, 09382, 14327).

Figure 2.1 presents consort diagram including the number of MAC-Q and IQCODE-16 reports, and the number of participants diagnosed with dementia along with computed their global cognition scores, for each occasion (wave). Of the 1,009 participants with informants, 232 (23%) had reports of MAC-Q and IQCODE-16 at all five occasions and were included in the G-analyses. This study excluded participants with informants (77%) whose MAC-Q and IQCODE-16 were incomplete at one or more occasions. The MAC-Q or IQCODE-16 data were missing at some waves because either the participant or informant was not contactable or was not able to do the assessment at that wave. In some instances, participants were too ill or

advanced in dementia to answer questions in later waves (Informant Only), and in others, participants simply did not have an informant who was willing to complete an interview or questionnaire on their behalf. The ethnicity of the extracted sample was predominantly white Europeans (97.8%); the remaining sample was 0.4% other, and 1.7 % unrevealed. Informants from the extracted sample were 76 males (32.8%), 153 females (65.9%), and their age ranged from 27 to 86 years, with a mean age of 61.3 years (*SD* = 14.38). Missing responses per item of either the MAC-Q or the IQCODE-16 in the extracted sample comprised less than 0.05% which were negligible and thus substituted by mean imputation at each respective wave (Huisman, 2000). This sample size of 232 participants exceeded the required sample size of 84 participants for repeated measures ANOVA over five occasions needed to accomplish the power (1-β) of 0.95 to detect effect size of 0.15 under *p* value of 0.05.

Figure 2.1

CONSORT diagram for participants who completed cognitive assessments inclusive of dementia-diagnosed cases at each wave/occasion.



Measures

The MAC-Q (Crook et al., 1992) is a well-validated unidimensional 6-item questionnaire. Internal consistency Cronbach's alpha of the MAC-Q was reported in the range from 0.57 to 0.88 with most studies indicating acceptable values and confirming unidimensionality of the scale (Buckley et al., 2013; Crook et al., 1992; Reid et al., 2012). The MAC-Q asked participants to rate themselves compared to how they previously performed on several everyday memory tasks (e.g., difficulty remembering names). At occasion 1, participants received the conventional MAC-Q wording "How would you rate yourself compared to 5 years ago", but for each subsequent occasion, the wording was changed to "How would you rate yourself compared to 2 years ago" to capture the intervening time between assessments. Participants rated themselves for each item on a scale of 1 to 5; total score range from 5 to 30, with higher scores indicating greater subjective memory loss.

The IQCODE-16 (Jorm, 1994) consists of 16 items that asks informants to report on their perceived changes of the participant's cognition and functioning. Each item is scored on a 5-point Likert scale with options ranging from 1 = "much improved" to 5 = "much worse". The IQCODE-16 is completed by informants who are well-known to the individual (Harrison et al., 2016) and has been shown to reliably predict incident dementia (Numbers et al., 2020). The original Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) consists of 26 questions/items. An abbreviated version, the IQCODE-16, has been found to perform as reliably as the original version (Jorm, 1994; Jorm, 2004), with a number of studies confirming high internal consistency (Cronbach's alpha = 0.93 to 0.97) (Harrison et al., 2015; Phung et al., 2015; Tang et al., 2004) and a superior ability to predict incident dementia (Park, 2017; Perroco et al., 2008).

Dementia Diagnosis

Clinical diagnoses were performed for all occasions (10-year follow-up). At occasion 1, and at each two-year follow-up, individuals were brought to a consensus review meeting where at least three clinicians from a panel of neuropsychiatrists, psychogeriatricians, and neuropsychologists discussed all available clinical, neuro-psychological, laboratory and imageing data to reach a consensus diagnosis. A diagnosis of dementia was based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (American Psychology Association, 1994) - that is, the development of one or more cognitive deficit(s) that represent a decline from a previous level of performance and were sufficiently severe as to cause impairment in functioning (Bayer IADL scale score $\geqslant$ 3.0). Individuals who did not receive a dementia diagnosis were classified as "not dementia" at each occasion, and no dementia cases were present at occasion 1 as this was an exclusionary criterion (Sachdev et al., 2010).

Objective Cognitive Performance

Comprehensive cognitive data were available for occasion 1 to occasion 4 (6-year follow up), only. Cognitive performance over these first four occasions was assessed using a comprehensive neuropsychological test battery which comprised 10 tests that measured the domains of attention/processing speed, language, executive function, visuospatial ability, and memory. The 10 tests were categorized into domains on a priori basis according to the principal cognitive function they represented. Attention/processing speed was measured using the Wechsler Adult Intelligence Scale-III (WAIS-III) Digit Symbol-Coding (Wechsler, 2003) and the Trail Making Test A (Tombaugh, 2004). Language was measured using the Boston Naming Test (Kaplan et al., 2001) and Animal Fluency (Spreen, 1977). Executive function was measured using FAS Fluency (Benton, 1967) and Trail Making Test B (Reitan & Wolfson, 1993). Visio-spatial ability was measured using the Wechsler Adult Intelligence Scale-Revised

(WAIS-R) Block Design (Wechsler, 1981b). Memory was measured using the Wechsler Memory Scale-III (WMS-III) Logical Memory Story A delayed recall (Wechsler, 1997), Rey Auditory Verbal Learning Test (Strauss et al., 2006) total learning, short-term and long-term recall scores, and the Benton Visual Retention Test (Benton et al., 1996). Domain and global cognition composites were computed as standardized z-scores as follows. Firstly, raw test scores were converted to z-scores using the baseline means and standard deviations (SDs) of a reference group which comprised 732 MAS participants classified as cognitively healthy at occasion 1 (native English speakers with a Mini-Mental State Examination score of 24 or above, no evidence of dementia or current depression, no history of delusions or hallucinations, and no major neurological disease, significant head injuries, progressive malignancies or CNS medications). Of the 732 participants (ages ranged from 70.29 to 90.80 years with M=78.57, SD=4.72), 219 (29.9%) completed tertiary qualification, 128 (17.5%) completed high school and/or diploma, 350 (47.8%) were not completed high school/diploma, 23 (3.2%) incomplete tertiary qualification, and 12 (1.6%) completed primary school. Secondly, composite domain scores were formed by averaging the z-scores of the component tests (as defined above), apart from the visuospatial domain which was represented by a single test. Each domain composite was standardized by transforming so that the mean and standard deviation of the baseline cognitively healthy group were 0 and 1, respectively. Finally, global cognition scores for each occasion were calculated by averaging the domain z-scores, and again transforming these scores so that the means and standard deviations for the baseline reference group were 0 and 1, respectively.

Data Analyses

EduG 6.1-e software (Swiss Society for Research in Education Working Group, 2006) was used to conduct Generalisability analyses by following the guidelines described in Truong et al. (2020). Both G-study and D-study used two-facet design (person by item by occasion)

which item (I) and occasion (O) were two facets of interest (instrumentation facets) and person (P) was the object of measurement (differentiation facet), expressed as P x I x O (Cardinet et al., 2011; Truong et al., 2020; Vispoel et al., 2018). The facet of I was fixed because the same items of assessments were used across all participants and all occasions, whereas the P and O facets were infinite. Besides that, the facet P was not a source of error and in a study employing G-Theory method, all error variances are counted as 100% after controlling for person variance (P), which reflects true differences between persons (Cardinet et al., 2011). G-Theory estimates for the design of person by item by occasion, express as P x I x O were calculated using formulae included in Table 2.1 (Shavelson et al., 1989).

Table 2.1

G-Theory model estimates' calculation including component variances, formula for the design of person by item by occasion, express as P x I x O.

Indices	Characteristic (Formula)
$X =$	observed score of a person on a particular item across occasions
μ	grand mean of X
$+X_p$	person effect ($\mu_p - \mu$)
$+X_i$	item effect ($\mu_i - \mu$)
$+X_o$	occasion effect ($\mu_o - \mu$)
$+X_{pi}$	person x item effect ($\mu_{pi} - \mu_p - \mu_i + \mu$)
$+X_{po}$	person x occasion effect ($\mu_{po} - \mu_p - \mu_o + \mu$)
$+X_{io}$	item x occasion effect ($\mu_{io} - \mu_i - \mu_o + \mu$)
$+X_{pio}$	residual/person x item x occasion effect ($\mu_{pio} - \mu_{pi} - \mu_{po}$ - $\mu_{io} + \mu_p + \mu_i + \mu_o$ - μ)
σ^2_p	person variance component ($MS_p - MS_{pi} - MS_{po} + MS_{pio}$)/$n_i n_o$
σ^2_i	item variance component ($MS_i - MS_{pi} - MS_{io} + MS_{pio}$)/$n_p n_o$
σ^2_o	occasion variance component ($MS_o - MS_{io} - MS_{po} + MS_{pio}$)/$n_i n_o$
σ^2_{pi}	person x item variance component ($MS_{pi} - MS_{pio}$)/n_o
σ^2_{po}	person x occasion variance component ($MS_{po} - MS_{pio}$)/n_i
σ^2_{io}	item x occasion variance component ($MS_{io} - MS_{pio}$)/n_p
σ^2_{pio}	residual/ person x item x occasion variance component: (MS_{pio})
σ^2_δ	relative error variance ($\frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{pio}}{n_i n_o}$)
σ^2_Δ	absolute error variance ($\frac{\sigma^2_0}{n_o} + \frac{\sigma^2_i}{n_i} + \frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{io}}{n_i n_o} + \frac{\sigma^2_{pio}}{n_i n_o}$)
Gr	relative G-coefficient ($\frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta}$)
Ga	absolute G-coefficient ($\frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Delta}$)
SCI	state component index ($\frac{\sigma^2_{po}}{\sigma^2_{po} + \sigma^2_p}$)
TCI	trait component index ($\frac{\sigma^2_p}{\sigma^2_{po} + \sigma^2_p}$)

Note: MS stands for the mean of effect square; n_i: number of items; n_o: number of occasions; n_p: number of persons/participants

There are two reliability coefficients, relative G-coefficient (*Gr*) and absolute G-coefficient (*Ga*), for the object of measurement (person) in a generalisability study. The relative model of measurement is based on a norm-referenced manner in which a person's assessment score is compared against the scores of others (Vispoel et al., 2018). *Gr* accounts for a relative error variance which is related to the I facet (object of measurement) that may affect a relative measurement (e.g., interaction between person and occasion- PxQ, and interaction between person and item- PxI) and includes divisions by desired sample sizes (Shavelson and Webb, 1991). Both *G* coefficients are estimating reliability of an enduring pattern of a measurement

if the person (P) is differentiation facet. Specifically, *Gr* of 0.80 or higher is determined as good reliability of assessment score (Cardinet et al., 2011) while *Ga* above 0.70 is considered as acceptable reliability (Truong et al., 2020; Arterberry et al., 2014). Both a state component index (*SCI*) and trait component index (*TCI*) were obtained, which represent the variance proportion attributed to a dynamic (state) and an enduring (trait) pattern in a measure (Medvedev, Krägeloh, et al., 2017). SCI of 0.60 or higher (*TCI* < 0.40) would indicate that variance is reflecting a dynamic pattern. On the contrary, TCI above 0.60 (*SCI* < 0.40) would signify a variance is reflecting an enduring pattern. In the D-study, variance components were computed for each individual item and effects of removing facets levels were examined to optimize the reliability of the MAC-Q and the IQCODE-16.

IBM SPSS Statistics 25 software was used to compute estimates that related to CTT approaches and descriptive statistics including means, standard deviation (*SD*), Cronbach's alpha, and intraclass correlation coefficient (*ICC*) for the IQCODE-16 and the MAC-Q. Logistic regression analyses were conducted to examine how the IQCODE-16 and the MAC-Q were able to predict the incidence of dementia across occasions 2-5. Three logistic regression models were carried out for each occasion. Each logistic regression model involved the outcome variable of dementia diagnose with one predictor of either MAC-Q scores or IQCODE-16 scores at the same occasion. Moreover, linear regression analyses were also used to estimate ability of these measures to predict global cognition scores across the first four occasions. Three independent linear regression models were conducted at each occasion with the outcome variable of global cognition scores and either MAC-Q scores or IQCODE-16 scores as a predictor. Prior to all regression analyses, assumption tests were conducted to screen for potential violations. It is worthy to note that three decimal points were used for variance components and error estimates in G-analyses due to the need for precision. Two decimal

points were used for other estimates, such as reliability, *G*-coefficients, *TCI*, and *SCI* estimates. Additionally, one decimal point was used for percentages.

Results

G-Study

The variance components attributed to person (P), item (I), and occasion (O), and their interactions (PxI, PxO, IxO, PxIxO) together with generalisability coefficients and state and trait component indices are presented in Table 2.2 for the MAC-Q and the IQCODE-16. The IQCODE-16 showed better reliability and generalisability of scores across persons and occasions, with both relative and absolute *G* coefficients of 0.86, 95% *CI* [0.84; 0.88]. Measurement error was predominantly explained by PxI and PxO interactions for the IQCODE-16, which together explained 77.1% of the total error variance after accounting for the true person variance. Slightly lower, but still acceptable values *Gr* = 0.80, 95%CI [0.77; 0.83]; *Ga* = 0.77, 95%CI [0.73; 0.81]) were observed for the 6-item MAC-Q, with the main source of error variance due to the PxI interaction explaining 35.7% of the total error variance. Consistent with reliability estimates, *TCI* values were 0.96 for the MAC-Q and 0.95 for the IQCODE-16 indicating that both instruments reliably assess enduring patterns of SCC.

Table 2.2

G-study estimates for the MAC-Q and IQCODE-16 including standard errors of the grand mean (SE), Coefficient G relative (Gr), Coefficient G absolute (Ga), Trait Component Index (TCI), State Component Index (SCI), grand mean (GM), variance components (in %), and for the Person (P) × Occasion (O) × Item (I) design including interactions with subtracting one item at a time (n = 232).

	MAC-Q		IQCODE-16	
Facets	σ^2	%	σ^2	%
P	0.044		0.019	
I	0.000	2.9	0.000	2.2
O	0.001	5.5	0.000	2.1
PI	0.005	35.7	0.001	39.6
PO	0.002	12.6	0.001	37.5
IO	0.000	3.5	0.000	1.5
PIO	0.005	39.7	0.001	17.1
GM	3.224		3.155	
SE	0.042		0.137	
Gr	**0.80**		**0.86**	
Ga	**0.77**		**0.86**	
TCI	0.96		0.95	
SCI	0.04		0.05	

D-Study

A series of generalisability analysis were conducted to obtain variance components for each individual items of the MAC-Q and IQCODE-16. The item-level estimates for variance of person, occasion, and person-occasion interaction, together with computed *SCI*, are included in Table 2.3. There was only one MAC-Q item reflecting high sensitivity for transitory changes in SCC patterns over time; “item e: Remembering the item(s) you intended to buy when you arrive at the supermarket store or pharmacy?”, which had the highest *SCI* of 0.66. The other five items of the MAC-Q revealed an *SCI* range from 0.15 to 0.45 indicating a lower proportion of variance associated with dynamic changes in SCC over time. However, all IQCODE-16 items reflected predominantly enduring patterns of SCC.

Six additional generalisability analyses were conducted by excluding one item at a time for the MAC-Q, as expected that this may result in improving the reliability of the scale in measuring enduring patterns of SCC (Table 2.4). The first analysis involved removing the first item (item a), with subsequent analyses removing one item at a time and examining reliability. Removing the final item (f) of the MAC-Q "In general, how would you describe your memory as compared to 10 years ago?" was the only analysis that resulted in improvement of both relative and absolute G coefficients above 0.80 benchmark suggesting that the 5-item MAC-Q (i.e., MACQ without item f) has better reliability compared to the 6-item MAC-Q. Next analyses involved removing one occasion at a time for the MAC-Q (Table 2.5) to examine how this affects the reliability of the scale. Removing any occasion only slightly decreased both G coefficients, which remained in the acceptable range.

Additional G-analyses were conducted on the IQCODE-16 (Table 2.6), which involved removing items more sensitive to dynamic SCC with $SCI \geqslant 0.40$ in attempt to optimize reliability. This resulted in lower G-coefficients compared to the original IQCODE-16. Additionally, removing one occasion at a time for the IQCODE-16 only slightly decreased G-coefficients. These findings together support reliability of the IQCODE-16 with the current measurement design.

Table 2.3

Variance components of Person (P), Occasion (O) and P x O interaction together with state component index (SCI) for each individual item of the MAC-Q and the IQCODE-16 (n = 232x5).

Items	P	O	PxO	*SCI*
MAC-Q				
a: Remembering the name of a person just introduced to you?	0.092	0.001	0.040	0.30
b: Recalling telephone numbers or post codes that you use on a daily or weekly basis?	0.088	0.010	0.054	0.38
c: Recalling where you have put objects (such as keys) in your home or office?	0.095	0.002	0.027	0.22
d: Remembering specific facts from a newspaper or magazine article you have just finished reading?	0.104	0.001	0.018	0.15
e: Remembering the item(s) you intended to buy when you arrive at the supermarket store or pharmacy?	0.022	0.003	0.042	**0.66**
f: In general, how would you describe your memory as compared to 10 years ago?	0.073	0.007	0.059	0.45
IQCODE-16				
1. Remembering things about family and friends (e.g. occupations, birthdays, addresses)	0.112	0.003	0.024	0.18
2. Remembering things that have happened recently	0.121	0.001	0.021	0.15
3. Recalling conversations a few days later	0.072	0.005	0.026	0.27
4.Remembering his/her address and telephone number	0.083	0.006	0.054	0.39
5. Remembering what day and month it is	0.080	0.003	0.019	0.19
6. Remembering where things are usually kept	0.096	0.001	0.026	0.21
7. Remembering where to find things which have been put in a different place from usual	0.069	0.001	0.013	0.16
8. Knowing how to work familiar machines around the house	0.059	0.002	0.038	0.39
9. Learning to use a new gadget or machine around the house	0.150	0.005	0.063	0.30
10. Learning new things in general	0.129	0.002	0.021	0.14
11. Following a story in a book or on TV	0.155	0.001	0.027	0.15
12. Making decisions on everyday matters	0.054	0.003	0.037	0.41
13. Handling money for shopping	0.063	0.007	0.061	0.49
14. Handling financial matters e.g. the pension, dealing with the bank	0.058	0.003	0.022	0.28
15. Handling other everyday arithmetic problems (e.g. knowing how much food to buy, knowing how long between visits from family or friends)	0.060	0.002	0.027	0.31
16. Using his/her intelligence to understand what's going on and to reason things through	0.048	0.000	0.012	0.20

Note. Number in bold reflects item measuring dynamic (state) pattern

Table 2.4

D-study reliability estimates and variance components for the Person (P) × Occasion (O) × Item (I) design including interactions for MAC-Q with subtracting one item at a time

	MAC-Q without item a		MAC-Q without item b		MAC-Q without item c		MAC-Q without item d		MAC-Q without item e		MAC-Q without item f	
Facets	σ^2	%	σ^2	%	σ^2		σ^2	%	σ^2	%	σ^2	%
P	0.049		0.036		0.040		0.038		0.050		0.053	
I	0.001	3.9	0.001	4.1	0.001	3.6	0.000	0.0	0.001	3.9	0.000	3.3
O	0.001	5.8	0.001	4.9	0.001	4.3	0.001	4.5	0.001	5.6	0.000	1.3
PI	0.005	32.6	0.007	41.9	0.006	36.3	0.006	35.9	0.007	39.8	0.005	35.3
PO	0.001	7.2	0.002	10.4	0.002	10.6	0.002	12.8	0.002	9.4	0.001	10.8
IO	0.001	4.6	0.000	2.2	0.001	3.8	0.001	4.1	0.001	3.3	0.001	4.3
PIO	0.007	46.0	0.006	36.5	0.007	41.3	0.007	42.7	0.006	37.9	0.006	45.0
GM	3.213		3.208		3.232		3.249		3.232		3.207	
SE	0.048		0.046		0.047		0.041		0.050		0.038	
Gr	0.80		0.71		0.73		0.71		0.77		0.81	
Ga	0.77		0.69		0.70		0.69		0.75		0.80	
TCI	0.83		0.95		0.95		0.95		0.96		0.98	
SCI	0.17		0.05		0.05		0.05		0.04		0.02	

Table 2.5

D-study reliability estimates and variance components for the Person (P) × Occasion (O) × Item (I) design including interactions for MAC-Q with subtracting an occasion

	Without occasion 1		Without occasion 2		Without occasion 3		Without occasion 4		Without occasion 5	
Facets	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%
P	0.041		0.045		0.041		0.047		0.048	
I	0.000	2.0	0.000	2.1	0.000	2.9	0.000	1.2	0.001	4.3
O	0.001	7.6	0.001	5.0	0.000	2.7	0.001	7.4	0.001	6.8
PI	0.006	37.1	0.005	31.3	0.004	31.6	0.004	25.6	0.005	29.1
PO	0.002	14.8	0.002	11.6	0.002	13.6	0.002	14.9	0.002	13.3
IO	0.000	2.4	0.001	4.7	0.000	3.8	0.001	4.4	0.001	3.8
PIO	0.006	36.0	0.007	45.3	0.006	45.4	0.007	46.5	0.007	42.5
GM	3.208		3.240		3.198		3.235		3.237	
SE	0.046		0.046		0.038		0.047		0.051	
Gr	0.75		0.77		0.77		0.79		0.78	
Ga	0.72		0.75		0.76		0.76		0.75	
TCI	0.95		0.96		0.95		0.96		0.96	
SCI	0.05		0.04		0.05		0.04		0.04	

Table 2.6

D-study reliability estimates and variance components for the Person (P) × Occasion (O) × Item (I) design including interactions for IQCODE-16 with subtracting items with different SCI benchmarks, and occasions

	Items with *SCI*≤0.30		Items with *SCI*<0.40		Without occasion 1		Without occasion 2		Without occasion 3		Without occasion 4		Without occasion 5	
Facets	σ^2	%	σ^2	%	σ^2		σ^2	%	σ^2	%	σ^2	%	σ^2	%
P	0.023		0.019		0.021		0.019		0.020		0.019		0.019	
I	0.000	2.2	0.000	2.3	0.000	2.7	0.000	1.3	0.000	2.0	0.000	1.7	0.000	2.0
O	0.000	1.4	0.000	2.7	0.000	2.7	0.000	1.9	0.000	0.3	0.000	3.5	0.000	3.1
PI	0.004	60.2	0.002	48.3	0.001	37.1	0.001	34.1	0.001	32.1	0.001	33.3	0.001	36.1
PO	0.001	15.6	0.001	27.3	0.001	37.1	0.002	42.3	0.002	43.7	0.002	41.3	0.002	40.2
IO	0.000	1.6	0.000	1.5	0.000	1.7	0.000	1.7	0.000	1.9	0.000	1.5	0.000	1.4
PIO	0.001	19.0	0.001	17.9	0.001	18.8	0.001	18.7	0.001	20.1	0.001	18.6	0.001	17.2
GM	3.127		3.146		3.160		3.149		3.164		3.152		3.151	
SE	0.021		0.019		0.019		0.017		0.016		0.018		0.019	
Gr	0.79		0.83		0.86		0.85		0.85		0.85		0.84	
Ga	0.78		0.82		0.85		0.84		0.84		0.84		0.83	
TCI	0.96		0.95		0.95		0.90		0.91		0.90		0.90	
SCI	0.04		0.05		0.05		0.10		0.09		0.10		0.10	

CTT Analyses

Descriptive statistics for the 5-item MAC-Q, the 6-item MAC-Q, and the IQCODE-16 at five occasions are presented in Table 2.7. The internal consistency Cronbach's alpha of the 6- item MAC-Q was fair to good over five occasions and ranged between 0.57 and 0.76, which was consistent with values reported by other studies (Buckley et al., 2013, Crook et al., 1992, Reid et al., 2012). Temporal stability was supported by intraclass correlation coefficient (ICC) of 0.84 across all occasions. The mean scores of the 5-item MAC-Q were only significantly different between occasion 1 and 2, while that of the 6-item MAC-Q were significantly different between occasion 1 and occasions 2 and 3. The IQCODE-16 demonstrated higher internal consistency with Cronbach's alphas ranging from 0.84 to 0.95, though the ICC of 0.70 was lower than both MAC-Q scales. Overall, the MAC-Q and IQCODE-16 scales showed acceptable to high internal reliability and acceptable temporal reliability for a measure of enduring patterns over time, which is consistent with G-study results.

Table 2.7

Means, standard deviation (SD), Cronbach's alpha, and intraclass correlation coefficient (ICC) for the MAC-Q 5 and 6 item versions and the IQCODE-16 (n=232)

	Occasion 1	**Occasion 2**	**Occasion 3**	**Occasion 4**	**Occasion 5**	**ICC(95%*CI*)**
MAC-Q (6-item)						
Mean (*SD*)	20.14 (2.45)	18.64* (2.00)	19.98 (2.03)	20.32 (2.16)	20.31 (2.29)	0.84(0.80-0.87)
Cronbach's alpha	0.73	0.57	0.76	0.76	0.75	
MAC-Q (5-item)						
Mean (*SD*)	16.51 (1.84)	16.09* (1.50)	16.04* (1.69)	16.36 (1.80)	16.36 (1.85)	0.81(0.77-0.84)
Cronbach's alpha	0.67	0.63	0.72	0.72	0.69	
IQCODE-16						
Mean (*SD*)	49.61 (3.32)	49.10 (3.49)	49.51 (4.38)	49.08 (5.88)	52.45* (6.82)	0.70(0.63-0.75)
Cronbach's alpha	0.84	0.88	0.92	0.95	0.95	

Note: *Mean differences are significant compared to occasion 1 (Bonferroni corrected); There were significant main effects of occasion observed on the MAC-Q (6-item) scale scores ($F(231,4)=50.58$, $p<0.001$; on the MAC-Q (5-item) scale scores ($F(231,4)=5.25$, $p<0.001$; and on the IQCODE-16 scale scores ($F(231,4)=27.63$, $p<0.001$ with significance increase at the last occasion (5) compare to previous occasions.

To evaluate predictive validity of the original MAC-Q (6-item), the shortened MAC-Q (5-item) and the IQCODE-16, a series of binary logistic regression analyses were conducted to predict risk of incident dementia. Table 8 presents coefficients for the models' predictors, as well as model percentage correct (MPC), for the three SCC measures across the 4 follow-up occasions (occasion 1 was excluded because all participants were initially healthy). Prior to the analyses, assumption testing was conducted for all the models and did not indicate any violations. The Hosmoer and Lemeshow tests indicated good-fit for these logistic regression models (all *p*'s>0.05). Accuracy of all models across occasions were ranging from 78.2% to 98.2% in their predictions of incidence of dementia. The IQCODE-16 significantly predicted incident dementia at all examined occasions, with all *p*'s < 0.001. Whereas the two versions of the MAC-Q (i.e., 5-item vs. 6-item) were only significant predictors of dementia incident on occasions 3 and 4 with *p*'s $\leqslant$ 0.02 but not at occasions 2 and 5 (*p*'s $\geqslant$ 0.30).

Table 2.8

Logistic Regression Model Coefficients for the MAC-Q and IQCODE-16 variables across occasions 2-5 predicting the incidence of dementia

Predicting Diagnosis	**MPC**	***β***	***SE (β)***	***p***	***Exp(β) [95% CI]***
Occasion 2:					
MAC-Q (5-item)	97.6	1.21	0.13	0.35	1.13 [0.88, 1.45]
MAC-Q (6-item)	96.9	-0.10	0.11	0.37	0.91 [0.73, 1.13]
IQCODE-16	98.2	5.73	0.90	<0.001	308.80 [53.24, 1791.26]
Occasion 3					
MAC-Q (5-item)	93.9	0.28	0.07	<0.001	1.32 [1.14, 1.53]
MAC-Q (6-item)	94.0	0.30	0.06	<0.001	1.35 [1.19, 1.53]
IQCODE-16	94.8	4.05	0.53	<0.001	57.13 [20.07,162.66]
Occasion 4					
MAC-Q (5-item)	90.8	0.29	0.06	<0.001	1.34 [1.20, 1.50]
MAC-Q (6-item)	90.9	0.25	0.05	<0.001	1.29 [1.17, 1.42]
IQCODE-16	92.3	3.58	0.39	<0.001	36.01 [16.74, 77.48]
Occasion 5					
MAC-Q (5-item)	80.8	0.16	0.07	0.02	1.17 [1.03, 1.33]
MAC-Q (6-item)	78.2	0.05	0.30	0.30	1.05 [0.96, 1.14]
IQCODE-16	85.3	3.65	0.41	<0.001	38.60 [17.31, 86.108]

Note: MPC: Model Percentage Correct; *Exp(β)*: the exponentiation of the β coefficient.

Table 2.9 presents a series of linear regression analyses conducted to determine the relationship between predictors such as the two MAC-Q scales and the IQCODE-16, and the outcome measured as participants' global cognition scores at the first four occasions. Several assumptions were evaluated for these linear regression models prior to the interpretation of the results. The data was distributed close to normal with skewness values for all variables in the models ranging from -1.03 (dementia diagnose at occasion 4) to 2.12 (5-item MAC-Q scores at occasion 5). Inspection of normal probability plots of regression standardized residuals also indicated that these variables were normally distributed. The scatterplots of standardized residuals were compared against standardized predicted values and also revealed that these variables met the assumptions of linearity, homoscedasticity of residuals, and were free from univariate outliers. Finally, there were no multicollinearity issue because only one predictor was independently tested for each model. The IQCODE-16 significantly predicted global cognition scores for all four occasions, with all p's < 0.01 while both MAC-Q versions were not significantly associated with global cognition at the first occasion (p's > 0.05).

Table 2.9

Linear Regression Model Coefficients for the MAC-Q and IQCODE-16 variables across occasions 1-4 predicting the global cognition scores

Predicting Cognition	R^2	β	*SE (β)*	*p*	*Standardised β [95% CI]*
Occasion 1					
MAC-Q (5-item)	0.002	-0.04	0.03	0.13	-0.05 [-0.11, 0.01]
MAC-Q (6-item)	0.004	-0.04	0.02	0.06	-0.06 [0.10, 1.16]
IQCODE-16	0.030	-1.02	0.19	<0.001	-0.17 [-0.24, -0.11]
Occasion 2					
MAC-Q (5-item)	0.022	-0.14	0.03	<0.001	-0.15 [-0.22, -0.08]
MAC-Q (6-item)	0.012	-0.08	0.03	<0.01	-0.11 [-0.18, -0.04]
IQCODE-16	0.055	-1.31	0.20	<0.001	-0.24 [-0.31, -0.17]
Occasion 3					
MAC-Q (5-item)	0.011	-0.10	0.04	<0.01	-0.11 [-0.18, -0.03]
MAC-Q (6-item)	0.014	-0.09	0.03	0.00	-0.12 [-0.18, -0.05]
IQCODE-16	0.040	-0.94	0.19	<0.001	-0.20 [-0.28, -0.12]
Occasion 4					
MAC-Q (5-item)	0.050	-0.21	0.04	<0.001	-0.23 [-0.31, -0.15]
MAC-Q (6-item)	0.047	-0.16	0.03	0.00	-0.22 [-0.29, -0.14]
IQCODE-16	0.085	-1.12	0.16	<0.001	-0.29 [-0.37, -0.21]

Note: *SE (β)*: standard error of the β coefficient.

Discussion

The aim of the current study was to evaluate and optimize reliability, and distinguish between dynamic and enduring patterns in the MAC-Q (self-report) and IQCODE-16 (informants) SCC measures using G-Theory. The results showed that the optimized 5-item MAC-Q and the IQCODE-16 were reliable in measuring enduring pattern of SCC with *G*-coefficients of 0.80 and higher, and index of trait (*TCI*) above 0.94, suggesting that their scores are generalizable across sample population and occasions. In line with previous research (Slavin et al., 2015), this study found that the IQCODE-16 SCC scores significantly predicted risk of dementia incident and global cognition across all occasions while the MAC-Q scores were only significant predictors on some occasions. However, these results should be interpreted with caution due to differences in length and format between the IQCODE-16 and the MAC-Q. Together our findings suggest that the MAC-Q reliably measures individual levels of SCC, but these self-reported SCC may be less accurate in reflecting cognitive abilities and diagnosis of each individual. It is possible that the MAC-Q tend to reflect individual tendencies to report complaints (e.g., about their self-perceived memory errors) rather than their actual cognitive capacities. In other words, some people may have stronger tendency to ruminate on everyday memory errors or lean towards complaining behaviour, which may not necessary reflect their actual cognitive abilities. Consistent with the recent clinical literature, SCC may be related to anxiety and stress in individuals with normal cognition (Chin et al., 2019). The outcome of our study has clinical implications, which underscore the importance for clinicians to seek corroboratory evidence from knowledgeable informants in their follow-up of ageing patients. This in turn, could help in detection of Mild Cognitive Impairment, which is the pre-clinical stage of the trajectory of cognitive decline, and would assist in ongoing clinical management and planning, as once dementia is diagnosed, it runs a debilitating course (Langa & Levine, 2014).

A D-study was conducted to examine psychometric properties of individual items of the MAC-Q and IQCODE-16 in an effort to optimize reliability of the measurement. Results showed that most individual items of the IQCODE-16 and MAC-Q measured enduring patterns of SCC, except item-e of the MAC-Q (item e: “Remembering the item(s) you intended to buy when you arrive at the supermarket store or pharmacy”). However, removing this item did not improve the reliability of the MAC-Q in measuring enduring pattern of SCC. Similar results were found when removing each item of the MAC-Q one at a time, with the exception of the last item (item-f: “In general, how would you describe your memory as compared to 10 years ago?”). Removing this last item boosted the marginally acceptable reliability of the 6-item MAC-Q (G = 0.77-0.80) up to G = 0.80-0.81 in the optimized 5-item MAC-Q. No reliability improvements were achieved by manipulating measurement design of the IQCODE-16 suggesting optimal reliability of the scale in the current measurement design.

Strengths and Limitations

The main strength of the study was to apply the comprehensive methodology of G-Theory to a relatively large sample in a longitudinal study spanning over 10 years. However, limitations need to be acknowledged. The study was conducted with participants recruited from a relatively small catchment area in Sydney, Australia. Moreover, the participants belonged to a predominantly White (European) ethnic group, and the generalisability to other ethnicities is questionable. Recent research suggests that cultural variations contribute to vulnerabilities and resilience across a range of health issues (Choo et al., 2017). As such, it would be beneficial to replicate these analyses on samples comprising other ethnicities, including culturally and linguistically diverse groups. This study aimed to analyse data from 5 occasions with equal intervals of 2 years; however, the interval between occasion 4 and occasion 5 was 4 years, as MAC-Q data was missing for wave 5 (8-year follow-up) assessments. Future studies should endeavour to replicate these analyses using equal intervals between occasions.

The findings of this study added evidence supporting the benefits of using the informant SCC report. However, due to differences between the IQCODE-16 and the MAC-Q format, more accurate comparison between informants' reports measured by the IQCODE-16 and self-reports could be achieved using the self-report version of the IQCODE- the Informant Questionnaire on Cognitive Decline- Self-report (IQCODE-SR) (Jansen et al., 2008). Therefore, further studies are warranted to compare the IQCODE-16 with the IQCODE-SR.

In addition, this study did not control for demographic variables (e.g., age, sex, socioeconomic and education status), mood or personality in both sets of regression analyses. Nevertheless, the results of G-analyses indicated that the IQCODE-16 and both versions of MAC-Q were measuring enduring pattern of SCC and were less affected by dynamic and transient conditions such as mood. Notably, less than 20% of variance was explained by error due to temporal factor and interactions (G=0.80-0.86).

Conclusion

The findings of this study indicated that the IQCODE-16 and MAC-Q assessment scores were generalizable across sample population and occasions and captured enduring patterns of SCC over 10-years. The optimized 5-item MAC-Q was superior to the original 6-item scale when assessing SCC over time. While clinicians and researchers could rely on both participants and informants' SCC reports of the IQCODE-16 and the MAC-Q, self-reported (MAC-Q) scores may be less accurate in predicting cognitive ability and diagnosis of each individual.

Chapter 3 Clinical investigation of dynamic and enduring aspects of global cognition in aged population

Truong, Q. C., Choo, C., Numbers, K., Merkin, A. G., Brodaty, H., Kochan, N. A., Sachdev, P. S., Feigin, V. L., & Medvedev, O. N. (2021). Clinical investigation of dynamic and enduring aspects of global cognition in aged population. *European Journal of Clinical Investigation, 51* (10). https://doi.org/10.1111/eci.13681

Declaration

This study is co-authored work with C. Choo, K. Numbers, A. G. Merkin, H. Brodaty, N. A. Kochan, P. S. Sachdev, V. L. Feigin, and O. N. Medvedev (see Appendix P2 for the original article). I was the first author and had the major contribution to this study that included designing and conducting the study, completing data analysis, writing the manuscript for submission to the *European Journal of Clinical Investigation*, and responding to reviews/addressing their comments during peer-review process. All co-authors have signed to the co-authorship form (see Appendix 3.1).

Abstract

Background: A major issue in evaluating the cognitive status of ageing populations is a clear distinction between enduring and dynamic aspects of global cognition necessary for evaluating risks of dementia and effectiveness of preventive interventions.

***Materials and methods*:** Generalizability Theory was applied to investigate dynamic and enduring aspects of global cognition using longitudinal data over 10 years of follow-up. Measures included the Mini-Mental Status Examination (MMSE) and the Telephone Interview for Cognitive Status-modified (TICS-M). The sample (n = 238) included 154 females, mean age = 76.54 years, SD = 3.94 from the Sydney Memory and Ageing Study.

Results: The MMSE measured dynamic and enduring aspects of cognition to a comparable degree with 56% of variance explained by enduring aspects and 44% by dynamic aspects and showed low sensitivity/high specificity in detecting dementia. A shortened version of the MMSE (MMSE-D8) better captured dynamic aspects of cognition after removing three items less sensitive to change. The TICS-M predominantly measured enduring aspects of cognition (72%) with the remaining 28% due to dynamic aspects and displayed high sensitivity/high specificity for dementia screening.

Conclusions: The MMSE measures both dynamic and enduring cognitive aspects and is suitable for general clinical assessments, while the MMSE-D8 can be used to monitor transitory changes of global cognition over time. The TICS-M is more useful for measuring enduring features of cognition and screening for dementia. Our findings highlight the value of generalisability theory to distinguish dynamic and enduring features of cognition, which may contribute to preventive interventions and monitoring cognitive ability over time.

***Keywords*:** Generalisability Theory; Global Cognition; Dynamic Aspects; Enduring Aspects; Longitudinal Data.

Introduction

The estimated prevalence of cognitive impairment and dementia amongst people aged 65 years and older is 11.7%, with the rate increasing with age (CDC, 2020). A clinical assessment usually includes a medical history and/or exam, subjective questionnaires and objective cognitive testing (Chin et al., 2019). The Mini-Mental State Examination (MMSE; Folstein et al., 1975) is a widely used standardised cognitive assessment tool in psychogeriatrics to screen for dementia. The modified Telephone Interview for Cognitive Status (TICS-M; Brandt et al., 1993) is an alternative tool for cognitive screening and is well-validated for telephone or face-to-face administration. The TICS-M has demonstrated strong correlations with the MMSE and may be preferable for large epidemiological studies as it is time and cost efficient (Fong et al., 2009; Wilson & Bennett, 2005).

There are some features in cognitive functions that are relatively stable (e.g., vocabulary) and more resistant to neurodegeneration, whereas others are more variable (e.g., attention) (Rocca et al., 2019). Cognitive functions can also be affected by temporary mental conditions, such as stress that may influence cognitive performance on a particular occasion (Maloney et al., 2014). Although the MMSE and the TICS-M are widely used and validated instruments for measuring cognitive conditions (Fong et al., 2009; Wilson & Bennett, 2005), there is no research investigating their ability to distinguish between enduring and dynamic aspects of global cognition. In both clinical and research contexts, it is critical to differentiate between dynamic and enduring aspects of global cognition and cognitive domains as this directly impacts on accuracy of diagnosis and validity of research results (Medvedev, Krägeloh, et al., 2017).

Traditionally, the differentiation between enduring and dynamic aspects of a measure was determined using classical test theory (CTT) and involved correlation of test scores at two different time points (e.g., occasion 1 and occasion 2) (Spielberger, 1999). The major limitation

of this method is an inability to accurately distinguish between dynamic and enduring aspects of a measure. This is because CTT does not consider variability of individual assessment items and other relevant sources of measurement error (i.e., effects of item, occasion, and person) nor their interactions with each other (Medvedev, Krägeloh, et al., 2017; Truong et al., 2020). In addition, CTT method uses test-retest coefficients to evaluate the steadiness of a measure over short intervals (e.g., few days, weeks, or months) (Feeney et al., 2016; Folstein et al., 1975). Generalizability Theory (G-Theory) is considered the most robust methodology for distinguishing between dynamic and enduring aspects in a measure, evaluating the overall reliability and generalisability of assessment scores, and identifying multiple sources of error variance (Medvedev, Krägeloh, et al., 2017). G-Theory is an extension of CTT, which closely examines all possible sources of error variance that can simultaneously affect reliability (e.g., person, scale items, situation/occasion, and their interactions) (Medvedev, Krägeloh, et al., 2017; Shavelson et al., 1989).

The primary aim of the current study was to employ G-Theory to investigate the distinction between enduring and dynamic aspects of cognitive status as measured by MMSE and the TICS-M in a large longitudinal sample assessed on five occasions over 10-years. This study also aimed to evaluate sensitivity and specificity of the MMSE and the TICS-M in detecting dementia. Based on available evidence, this study hypothesised that the MMSE and the TICS-M measured dynamic and enduring aspects of global cognition to a different extent, and would have varying sensitivity and specificity for detecting dementia.

Method

Participants

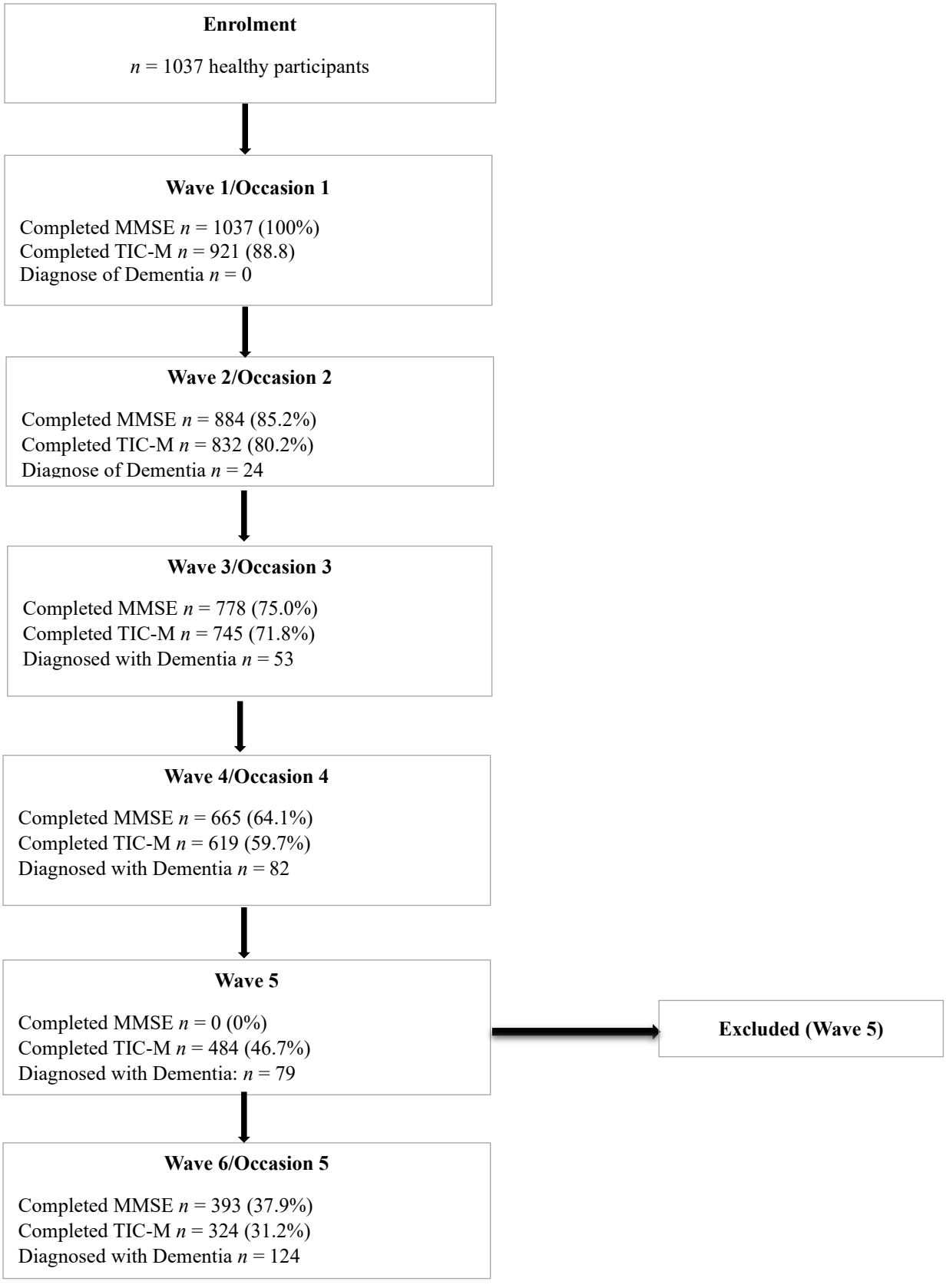
A sample of 8,914 community-dwelling older adults aged 70-90 years, living in the Eastern Suburbs of Sydney, Australia, was selected via the electoral roll and invited to participate in the Sydney Memory and Ageing Study (MAS) (Sachdev et al., 2010). The baseline sample (occasion 1) included 1,037 participants who met inclusion criteria, which were English language ability to complete psychometric assessment and self-report questionnaires. Baseline exclusion criteria included major psychiatric diagnoses, acute psychotic symptoms, current diagnoses of multiple sclerosis, motor neuron disease, developmental disability, progressive malignancy, or dementia. This study was approved by the University of New South Wales Human Ethics Review Committee (HC: 05037, 09382, 14327).

Figure 3.1 presents a consort diagram that includes the number of participants who completed both the MMSE and TICS-M, as well as number of participants diagnosed with dementia for each occasion (called a 'wave'). G-analyses included 238 (23%) participants who completed the MMSE and TICS-M at all five occasions. This study excluded 799 participants (77%, 98% Europeans, 384 males) whose MMSE and TICS-M were incomplete at one or more occasions. There were no significant demographic differences between the original retained and excluded samples ($p > 0.05$; see Appendix 3.2 for more demographic details). The interval between occasions was generally 2 years, except for occasion 4 and 5, which was 4 years. The ethnicity of the selected sample was predominantly Europeans (98.2%); the remaining sample was 0.3% Asian, 0.8% others and 0.8% unrevealed. This sample included 84 males and 154 females. Participants were aged 70-88 years (M = 76.54 years, SD = 3.94) at occasion 1. This sample size of 238 participants exceeded the required sample size of 84 participants for

repeated measures ANOVA over five occasions needed to accomplish the power (1-β) of 0.95 to detect effect size of 0.15 under p value of 0.05 (Faul et al., 2007).

Figure 3.1

CONSORT diagram for participants who completed MMSE, TICS-M and cognitive assessment inclusive of dementia-diagnosed cases at each wave/occasion.



Measures

The MMSE (Folstein et al., 1975) is a 30-point screening tool used to detect cognitive impairment (Appendix 3.3). This scale consists of 11 items designed to capture global cognition. Score range from 0 to 30, and the total scores are calculated by adding responses of individual items together (Folstein et al., 1975). Previous studies have shown that an MMSE ≤ 24 had a sensitivity ranging from 66% to 83.3% and a specificity ranging from 96.7% to 99% in identifying dementia (O'Bryant et al., 2008; Weston, 1987).

The TICS-M (Brandt et al., 1993) is a 13-item test of cognitive functioning conducted via telephone or face-to-face, with total scores ranging from 0 to 39 (Appendix 3.3). Compared to its original scale- the 11-item Telephone Interview for Cognitive Status (TICS) (Brandt et al., 1988), most TICS items are retained but a delayed verbal recall item is added in the TICS-M(Lacruz et al., 2013). The TICSs had been validated against neuropsychological measures (e.g., the MMSE), and other cognitive screens (de Jager et al., 2003), and demonstrated high diagnostic efficiency for identifying individuals with dementia (Crooks et al., 2005). Particularly, a recent study indicated that the TICS-M had a sensitivity of 77% and specificity of 88% in detecting those with dementia with the cut-off score of 21 (Bentvelzen et al., 2019).

Dementia Diagnosis

Clinical diagnoses were made by consensus panel consisting of neuropsychiatrists, psychogeriatricians, and neuropsychologists who considered all available cognitive data. The Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (American Psychiatric Association, 1994) criteria was used to diagnose dementia. Participants who were not diagnosed with dementia were coded as "not dementia" on each occasion.

Data Analyses

Generalizability analyses (G-analyses) were conducted using EduG 6.1 software(Swiss Society for Research in Education Working Group, 2006) by following the guidelines described in Truong et al. (2020) and Cardinet et al. (2011). Application of G-Theory requires a Generalizability study (G-study) and a Decision study (D-study). All G-analyses that were involved in the G-study and D-study used a two-facet design (person by item by occasion, express as PxIxO) (Cardinet et al., 2011). Person (P) was the object of measurement, while the facets of item (I) and occasion (O) were the two instrumentation facets. The P and O facets were infinite, while the I facet was fixed, or finite, and was not derived from the infinite populations. Whimbey's correction coefficient was adapted to ANOVA estimates, which had no effect on facets derived from infinite populations (i.e., the P facet) (Ye et al., 2020). In this study, the P facet was not a source of error utilizing G-Theory method, meaning that all error variances were 100% accounted for after controlling for person variance (P), which reflects true differences between persons (Cardinet et al., 2011). The G-study was conducted to examine the overall generalizability of the MMSE and the TICS-M and to identify sources of error variance in each measure. The D-study was subsequently used to evaluate psychometric properties of individual items of the MMSE and the TICS-M. The D-study also included generalizability analyses which removed MMSE and TICS-M item(s) in an effort to optimize reliability of each measure, as well as analyses that removed occasion(s) at a time for the MMSE and the TICS-M to examine how this affected reliability and measurement errors for both scales (Cardinet et al., 2011). G-Theory estimates for the current study were obtained using the formulas presented in Table 3.1 (Shavelson et al., 1989).

Table 3.1

G-Theory model estimates' calculation including component variances, formula for the design of person by item by occasion, express as P x I x O.

Indices	Characteristic (Formula)
$X =$	observed score of a person on a particular item across occasions
μ	grand mean of X
$+X_p$	person effect ($\mu_p - \mu$)
$+X_i$	item effect ($\mu_i - \mu$)
$+X_o$	occasion effect ($\mu_o - \mu$)
$+X_{pi}$	person x item effect ($\mu_{pi} - \mu_p - \mu_i + \mu$)
$+X_{po}$	person x occasion effect ($\mu_{po} - \mu_p - \mu_o + \mu$)
$+X_{io}$	item x occasion effect ($\mu_{io} - \mu_i - \mu_o + \mu$)
$+X_{pio}$	residual/person x item x occasion effect ($\mu_{pio} - \mu_{pi} - \mu_{po} - \mu_{io} + \mu_p + \mu_i + \mu_o - \mu$)
σ^2_p	person variance component ($MS_p - MS_{pi} - MS_{po} + MS_{pio}$)/$n_i n_o$
σ^2_i	item variance component ($MS_i - MS_{pi} - MS_{io} + MS_{pio}$)/$n_p n_o$
σ^2_o	occasion variance component ($MS_o - MS_{io} - MS_{po} + MS_{pio}$)/$n_i n_o$
σ^2_{pi}	person x item variance component($MS_{pi} - MS_{pio}$)/n_o
σ^2_{po}	person x occasion variance component ($MS_{po} - MS_{pio}$)/n_i
σ^2_{io}	item x occasion variance component ($MS_{io} - MS_{pio}$)/n_p
σ^2_{pio}	residual/ person x item x occasion variance component: (MS_{pio})
σ^2_δ	relative error variance ($\frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{pio}}{n_i n_o}$)
σ^2_Δ	absolute error variance ($\frac{\sigma^2_o}{n_o} + \frac{\sigma^2_i}{n_i} + \frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{io}}{n_i n_o} + \frac{\sigma^2_{pio}}{n_i n_o}$)
Gr	relative G-coefficient ($\frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta}$)
Ga	absolute G-coefficient ($\frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Delta}$)
SCI	state component index ($\frac{\sigma^2_{po}}{\sigma^2_{po} + \sigma^2_p}$)
TCI	trait component index ($\frac{\sigma^2_p}{\sigma^2_{po} + \sigma^2_p}$)

Note: MS stands for the mean of effect square; n_i: number of items; n_o: number of occasions; n_p: number of persons/participants

There are two reliability coefficients for the object of measurement (P) in a generalizability study: the relative G-coefficient (*Gr*) and the absolute G-coefficient (*Ga*). *Gr* accounts for relative error variance (σ_{δ}^{2}), which only involves error variance due to the interaction between the object of measurement (facet P) and all other facets (i.e., interaction between person, occasion and item) that may affect a relative measurement(Shavelson et al., 1989). *Ga* accounts for an absolute error variance (σ^{2}_{Δ}), which involves variance of all instrumentation facets (i.e., facet I and facet O) and their interaction, as well as their interaction with the object of measurement that may indirectly affect an absolute measurement (Cardinet et al., 2011). *G* coefficients of 0.70 or higher are expected for a measure of enduring patterns (Cardinet et al., 2011; Medvedev, Krägeloh, et al., 2017).

This study also computed a state component index (*SCI*) and trait component index (TCI) for the MMSE and the TICS-M, and their individual items in the G-study and D-study. These indices represent the variance proportion attributed to a dynamic (state) and an enduring (trait) aspect in a measure(Medvedev, Krägeloh, et al., 2017). *SCI* above 0.60 ($TCI \leq 0.40$) reflects assessment of predominantly dynamic aspects (Truong et al., 2020).

Estimates related to CTT approaches and descriptive statistics including mean, *SD*, intraclass correlation coefficient (*ICC*), and Cronbach's alpha for the total MMSE and TICS-M were obtained using IBM SPSS Statistics 27 software. The analyses of ROC curves were produced to evaluate diagnostic accuracy of both scales in detecting dementia. Sensitivity and specificity were computed using the cut-off point of 21 for the TICS-M (Bentvelzen et al., 2019), and that of 24 for the MMSE (Folstein et al., 1975) across four follow-up occasions. It should be noted that three decimal points were used for variance components and error estimates in G-analyses due to the need for precision. Two decimal points were used for other

estimates, such as reliability, *G*-coefficients, *TCI*, and *SCI* estimates. Additionally, one decimal point was used for percentages.

Results

Descriptive Statistics

Table 3.2a and 3.2b present descriptive statistics for the MMSE and TICS-M, as well as their individual items on the five occasions. Intraclass correlation coefficient of the MMSE and TICS-M across all occasions were 0.72 (95% *CI* [0.66, 0.77]) and 0.81 (95% *CI* [0.77, 0.84]), respectively. Skewness and kurtosis values for the MMSE scores across occasions were distributed close to normal, except MMSE scores at occasion 5. However, inspection of the normal probability plots for this variable supported normality. Degrees of freedom in repeated ANOVA were corrected using Greenhouse-Geisser due to the estimate of sphericity < 0.75 (ε = 0.63) (Field, 2013). The results showed that MMSE scores were significantly different across the five occasions, $F(52.16, 2.80) = 18.66$, $p \leq 0.001$. Post-hoc tests using Bonferroni correction revealed that MMSE scores significantly increased at occasion 3 ($M = 28.82$, $SD = 1.20$, $p < 0.001$), and decreased at occasion 5 ($M = 27.87$, $SD = 2.52$, $p < 0.001$) compared to occasion 1 ($M = 28.50$, $SD = 1.21$). Skewness and kurtosis values for TICS-M scores supported normality of distribution. Degrees of freedom were corrected using Huynh-Feldt due to the estimate of sphericity >0.75 (ε = 0.91)(Field, 2013). The results showed that TICS-M scores were significantly different across the five occasions, $F(863.36, 3.64) = 38.11$, $p \leq 0.001$. Post-hoc tests using Bonferroni correction revealed that TICS-M scores were only significantly lower at occasion 5 ($M = 23.08$, $SD = 5.20$) compared to occasion 1 ($M = 25.15$, $SD = 3.35$), $p < 0.001$.

Table 3.2a

Descriptive statistics including means, standard deviation (SD), Skewness, Kurtosis and test-retest coefficients for the MMSE and its individual items across occasions (n = 238). Cronbach alpha (α) at the latest occasion are presented in parentheses

	Occasion 1		**Occasion 2**		**Occasion 3**		**Occasion 4**		**Occasion 5**	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
MMSE (α = 0.72)										
1. Orientation time	4.85	0.37	4.84	0.40	4.84	0.42	4.77	0.57	4.67	0.86
2. Orientation place	4.70	0.53	4.85	0.44	4.97	0.18	4.97	0.18	4.90	0.39
3. Listen to the names of three objects and then recall them	2.98	0.13	2.97	0.18	3.00	0.00	2.98	0.14	2.84	0.47
4. Subtract 7 from 90 and keep subtracting 7 from each new number (If there are one or more errors on the serial 7s then ask participants to spell world backwards)	4.89	0.41	4.97	0.17	4.96	0.23	4.89	0.37	4.87	0.46
5. Recall three objects read out earlier	2.73	0.51	2.78	0.47	2.75	0.52	2.71	0.61	2.46	0.76
6. What is this called?	2.00	0.00	2.00	0.00	1.99	0.11	2.00	0.06	1.99	0.09
7. Repeat a phrase: “No ifs, ands, or buts”	0.61	0.49	0.59	0.49	0.56	0.50	0.44	0.50	0.50	0.50
8. Take this piece of paper in your (non-dominant) hand, fold it in half using both hands, and put the piece of paper on the floor	2.81	0.41	2.83	0.40	2.85	0.40	2.74	0.51	2.80	0.49
9. Read the words on this page and then do what it says	0.99	0.11	0.95	0.21	0.96	0.19	0.99	0.09	0.98	0.13
10. Write any complete sentence	0.99	0.11	0.99	0.11	0.97	0.18	0.98	0.14	0.94	0.24
11. Copy this design (pentagons)	0.95	0.22	0.95	0.22	0.97	0.16	0.93	0.25	0.91	0.29
Total MMSE score	28.50	1.21	28.72	1.30	28.82*	1.20	28.40	1.55	27.87*	2.52
Skewness	-0.79		-1.21		-1.83		-1.67		-3.26	
Kurtosis	0.27		1.24		7.10		4.91		17.16	
Test-retest coefficients	…		0.61		0.53		0.57		0.32	

Note. *SD*: standard deviation; *Mean differences are significant compared to occasion 1 (Bonferroni corrected)

Table 3.2b

Descriptive statistics including means, standard deviation (SD), Skewness, Kurtosis and test-retest coefficients for the TICS-M and its individual items across occasions (n = 238). Cronbach alpha (α) at the latest occasion are presented in parentheses

	Occasion 1		**Occasion 2**		**Occasion 3**		**Occasion 4**		**Occasion 5**	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
TICS-M (α = 0.74)										
1. What is today's date? What season are we currently in?	4.79	0.47	4.75	0.61	4.75	0.50	4.71	0.61	4.32	1.05
2. Can you tell me your age?	0.99	0.09	0.97	0.16	0.97	0.16	0.99	0.11	0.91	0.29
3. Do you know your telephone number?	0.99	0.11	0.98	0.14	1.00	0.06	0.97	0.18	0.87	0.34
4. Listen to a list of ten words carefully and then recall as many words as can be in any order.	4.62	1.44	4.85	1.45	5.01	1.70	4.79	1.69	4.11	1.65
5. Take 7 away from 100 and continue to take 7 away from the answer	4.23	1.07	4.45	1.01	4.39	1.00	4.39	1.04	4.26	1.24
6. Can I ask you now to count backwards from 20 to 1	0.94	0.24	0.96	0.20	0.96	0.19	0.95	0.23	0.91	0.28
7. What do people normally use to cut paper?	0.96	0.19	0.98	0.14	0.99	0.11	0.99	0.09	0.98	0.13
8. What is the name of the prickly green plant that grows in the desert?	0.94	0.24	0.93	0.26	0.94	0.24	0.94	0.24	0.88	0.33
9. Who is Prime Minister of Australia now?	0.98	0.14	0.97	0.16	0.94	0.24	0.97	0.18	0.94	0.24
10. Do you know the name of the Premier of NSW?	0.89	0.31	0.80	0.40	0.73	0.45	0.70	0.46	0.62	0.49
11. What is the opposite of "east"?	0.99	0.09	1.00	0.06	1.00	0.00	0.99	0.11	0.98	0.13
12. Can you say the words: "British constitution"	0.98	0.13	0.99	0.11	1.00	0.00	0.99	0.09	1.00	0.06
13. Recall any of the words from the list read out earlier	2.85	1.54	3.19	1.45	3.29	1.71	3.12	1.87	2.29	1.76
Total TICS-M score	25.15	3.35	25.82	2.98	25.96	3.75	25.48	4.22	23.08*	5.20
Skewness	0.74		-0.03		-0.17		-0.59		-0.85	
Kurtosis	0.45		0.13		1.19		1.26		1.23	
Test-retest coefficients	…		0.57		0.60		0.57		0.59	

Note. *SD*: standard deviation; *Mean differences are significant compared to occasion 1 (Bonferroni corrected)

G-Study

Table 3.3 presents the variance components attributed to person (P), item (I), and occasion (O), and their interactions (PxI, PxO, IxO, PxIxO) together with generalizability coefficients and state and trait component indices for the MMSE and the TICS-M. Both relative and absolute *G* coefficients of the MMSE were 0.56, meaning that 56% of variance were explained by enduring aspects of global cognition while the remaining 44% were attributed to dynamic aspects as reflected by *SCI* of 0.40 (Truong et al., 2020). These findings indicated that MMSE scores measure both dynamic and enduring aspects of cognition. Conversely, TICS-M scores predominantly measures enduring aspects of cognitive status as indicated by *G* coefficients of 0.72 and *TCI* of 0.71. The interaction between person and occasion (PxO), which represents dynamic aspect of cognition, accounted for 99.5% of the MMSE and 99.7% of the TICS-M total error variance, suggesting both scales were not affected by measurement error due to other sources. Together, these results illustrate that the TICS-M assesses enduring aspects of global cognition to a larger extent compared to the MMSE.

Table 3.3

ANOVA and G-study estimates for the MMSE and TICS-M including standard errors of the grand mean (SE), the sum of squares (SS), degrees of freedom (df), mean squares (MS), Coefficient G relative (Gr), Coefficient G absolute (Ga), Trait Component Index (TCI), State Component Index (SCI), grand mean (GM), variance components (in %), and for the Person (P) × Occasion (O) × Item (I) design including interactions (n = 238).

	ANOVA						G-study	
Source of variance	SS	df	MS	Components			σ^2	%
				Random	Mixed	Corrected*		
MMSE								
P	65.261	237	0.275	0.002	0.003	0.003	0.003	
I	9728.958	10	972.896	0.317	0.317	0.288	0.000	0.0
O	0.612	4	0.153	0.000	0.000	0.000	0.000	0.5
PI	443.551	2370	0.187	0.012	0.012	0.011	0.000	0.0
PO	115.388	948	0.122	0.000	0.011	0.011	0.002	99.5
IO	23836.81	40	595.92	2.503	2.504	2.504	0.000	0.0
PIO	1209.594	9480	0.128	0.128	0.128	0.128	0.000	0.0
			Grand mean				2.588	
			SE				0.003	
			Gr [95%CI]				0.56[0.65,0.47]	
			Ga [95%CI]				0.56[0.65,0.47]	
			TCI				0.60	
			SCI				0.40	
TICS-M								
P	370.546	237	1.563	0.016	0.017	0.017	0.017	
I	7706.705	12	642.225	0.034	0.034	0.032	0.000	0.0
O	1.237	4	0.309	0.000	0.000	0.000	0.000	0.3
PI	1811.387	2844	0.637	0.014	0.014	0.013	0.000	0.0
PO	415.194	948	0.438	0.000	0.034	0.034	0.007	99.7
IO	28874.85	48	601.559	2.525	2.528	2.528	0.000	0.0
PIO	6465.52	11376	0.568	0.568	0.568	0.568	0.000	0.0
			Grand mean				1.931	
			SE				0.004	
			Gr [95%CI]				0.72[0.80,0.64]	
			Ga [95%CI]				0.72[0.80,0.64]	
			TCI				0.71	
			SCI				0.29	

Note: *Whimbey's corrected estimates

D-Study

The item-level estimates for variance of person (P), occasion (O), and person-occasion interaction (PxO), together with computed SCI, are presented in Table 3.4. Most individual items of the MMSE reflected high sensitivity for transitory changes over time with *SCIs* ranging from 0.66 to 0.96. Only one MMSE item (item 1. Orientation time), which had the lowest *SCI* of 0.30, reflected predominantly enduring aspects of global cognition. Similarly, 11 out of 13 individual items of the TICS-M reflected transient changes over time with *SCIs* ranging from 0.73 to 1.00. The two remaining items of the TICS-M (item 1. What is today's date? What season are we currently in?; and item5. Take 7 away from 100 and continue to take 7 away from the answer) revealed *SCIs* of 0.55 and 0.56, respectively, and thus cannot be clearly classified as reflecting either dynamic or enduring aspects of cognition as they measure both to a comparable extent.

Table 3.4

Variance components of Person (P), Occasion (O) and P x O interaction together with state component index (SCI) for each individual item of the MMSE and TICS-M (n = 238x5)

Items	P	O	PxO	*SCI*
MMSE				
1. Orientation time	0.086	0.247	0.037	0.30
2. Orientation place	0.006	0.166	0.021	**0.78**
3. Listen to the names of three objects and then recall them	0.002	0.622	0.018	**0.90**
4. Subtract 7 from 90 and keep subtracting 7 from each new number (If there are one or more errors on the serial 7s then ask participants to spell *world* backwards)	0.002	0.203	0.031	**0.94**
5. Recall three objects read out earlier	0.001	0.754	0.026	**0.96**
6. What is this called?	0.020	0.489	0.039	**0.66**
7. Repeat a phrase: "No ifs, ands, or buts"	0.001	0.891	0.020	**0.95**
8. Take this piece of paper in your (non-dominant) hand, fold it in half using both hands, and put the piece of paper on the floor	0.012	0.519	0.024	**0.67**
9. Read the words on this page and then do what it says	0.003	0.591	0.020	**0.87**
10. Write any complete sentence	0.010	0.356	0.027	**0.73**
11. Copy this design (pentagons)	0.005	0.170	0.017	**0.77**
TICS-M				
1. What is today's date? What season are we currently in?	0.132	0.796	0.162	0.55
2. Can you tell me your age?	0.006	0.002	0.015	**0.71**
3. Do you know your telephone number?	0.006	0.547	0.130	**0.96**
4. Listen to a list of ten words carefully and then recall as many words as can be in any order.	0.052	0.732	0.138	**0.73**
5. Take 7 away from 100 and continue to take 7 away from the answer	0.008	0.002	0.010	0.56
6. Can I ask you now to count backwards from 20 to 1	0.060	0.693	0.222	**0.79**
7. What do people normally use to cut paper?	0.010	0.458	0.057	**0.85**
8. What is the name of the prickly green plant that grows in the desert?	0.007	0.590	0.136	**0.95**
9. Who is Prime Minister of Australia now?	0.054	0.777	0.151	**0.74**
10. Do you know the name of the Premier of NSW?	0.002	0.002	0.013	**0.87**
11. What is the opposite of "east"?	0.000	0.561	0.138	**1.00**
12. Can you say the words: "British constitution"	0.041	0.748	0.154	**0.79**
13. Recall any of the words from the list read out earlier	0.016	0.159	0.126	**0.89**

Note: Number in bold reflects item measuring dynamic (state) aspects

Table 3.5 presents generalizability analyses, which involved removing MMSE item(s) with different SCI benchmarks in an effort to optimize reliability of the MMSE. However, all *G* coefficients for these analyses were below 0.50, while the acceptable generalizability for a measure of enduring aspects requires *G* of at least 0.70 (Medvedev, Krägeloh, et al., 2017). After removing MMSE items with *SCI* < 0.70 (i.e., items 1, 6, and 8), the measure become suitable for assessing dynamic aspects of cognition (*G* < 0.40 and *SCI* > 0.60) (Medvedev, Krägeloh, et al., 2017; Truong et al., 2020). This modified 8-item MMSE (MMSE-D8) included: items 2, 3, 4, 5, 7, 9, 10, and 11.

Table 3.5

D-study reliability estimates and variance components for the Person (P) × Occasion (O) × Item (I) design including interactions for the MMSE with subtracting item(s)

	Item 5 removed		SCI>0.90 items removed		SCI>0.80 items removed		SCI>0.70 items removed		Item 1 removed		SCI<0.70 items removed	
Facets	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%
P	0.002		0.003		0.004		0.009		0.002		0.001	
I	0.004	26.8	0.035	33.4	0.070	38.3	0.091	27.5	0.001	13.1	0.011	24.7
O	0.003	23.0	0.045	43.0	0.085	46.3	0.139	41.8	0.000	0.0	0.012	27.2
PI	0.000	0.8	0.001	0.9	0.002	0.9	0.008	2.5	0.000	0.5	0.000	0.2
PO	0.002	15.9	0.002	2.2	0.003	1.4	0.003	0.8	0.002	23.9	0.002	4.5
IO	0.005	31.7	0.020	19.0	0.022	12.0	0.082	24.7	0.005	59.7	0.019	41.5
PIO	0.000	1.7	0.002	1.4	0.002	1.2	0.009	2.7	0.000	2.8	0.001	1.9
GM	2.556		2.504		2.570		3.077		2.443		2.404	
SE	0.109		0.316		0.421		0.559		0.080		0.206	
Gr	0.48		0.40		0.36		0.30		0.43		0.30	
Ga	0.14		0.03		0.02		0.03		0.17		0.03	
TCI	0.50		0.60		0.57		0.75		0.50		0.33	
SCI	0.50		0.40		0.43		0.25		0.50		0.67	

Likewise, additional generalizability analyses were carried out for the TIC-M, which involved removing item(s) with different *SCI* benchmarks (see Table 3.6). However, these modifications did not yield an improvement of the TIC-M for measuring more dynamic or more enduring aspects of cognition.

Table 3.6

D-study reliability estimates and variance components for the Person (P) × Occasion (O) × Item (I) design including interactions for the TIC-M with subtracting item(s)

	Item 11 removed		*SCI*>0.80 items removed		*SCI*>0.90 items removed		Item 1 removed		*SCI*<0.60 items removed		*SCI*<0.80 items removed	
Facets	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%
P	0.017		0.020		0.015		0.014		0.016		0.008	
I	0.001	4.7	0.011	14.3	0.005	18.9	0.000	0.0	0.000	0.0	0.000	0.0
O	0.001	5.5	0.000	0.0	0.001	3.6	0.002	18.4	0.000	0.0	0.000	0.0
PI	0.000	0.9	0.002	3.0	0.001	2.1	0.000	0.4	0.000	0.6	0.000	0.6
PO	0.007	54.6	0.009	11.9	0.007	24.9	0.006	49.6	0.007	40.6	0.007	11.9
IO	0.003	28.1	0.044	59.0	0.012	41.3	0.003	25.6	0.008	47.9	0.043	71.7
PIO	0.001	6.3	0.009	11.8	0.003	9.2	0.001	5.9	0.002	11.0	0.010	16.3
GM	1.916		2.129		1.878		1.831		1.913		1.699	
SE	0.069		0.233		0.135		0.075		0.089		0.207	
Gr	0.69		0.50		0.58		0.67		0.66		0.31	
Ga	0.58		0.21		0.34		0.53		0.50		0.11	
TCI	0.71		0.69		0.68		0.70		0.70		0.53	
SCI	0.29		0.31		0.32		0.30		0.30		0.47	

Next, analyses were performed that removed one and two random occasions at a time for the MMSE and the TICS-M to examine how these affected the ability of both scales to discriminate dynamic and enduring aspect of cognitive functions, as well as reliability and measurement errors for both scales. These G-analyses resulted in minor changes in *G*-coefficients and *TCI* values for both the MMSE and the TICS-M compared to the original *G*-analyses at all five occasions (Table 3.7). *G*-coefficients and *TCI* in these analyses were consistent, but slightly lower compared to the initial analysis, suggesting that our results do not depend on errors associated with specific assessment occasion(s).

Table 3.7

D-study reliability estimates and variance components for the Person (P) × Occasion (O) × Item (I) design including interactions for the MMSE and the TICS-M with subtracting occasions

		Without occasion 1		Without occasion 2		Without occasion 3		Without occasion 4		Without occasion 5		Without occasions 1&2		Without occasions 1&4		Without occasions 3& 5	
	Facets	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%	σ^2	%
MMSE																	
	P	0.002		0.003		0.003		0.003		0.003		0.002		0.002		0.003	
	I	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	O	0.000	0.6	0.000	0.3	0.000	0.7	0.000	0.6	0.000	0.5	0.000	0.3	0.000	0.8	0.000	0.6
	PI	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	PO	0.003	99.4	0.003	99.7	0.003	99.3	0.003	99.4	0.003	99.5	0.004	99.7	0.004	99.2	0.004	99.4
	IO	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	PIO	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	GM	2.589		2.585		2.588		2.586		2.589		2.586		2.588		2.591	
	SE	0.004		0.003		0.004		0.004		0.004		0.004		0.005		0.005	
	Gr	0.46		0.47		0.51		0.55		0.51		0.35		0.41		0.44	
	Ga	0.46		0.47		0.51		0.54		0.51		0.35		0.40		0.43	
	TCI	0.40		0.50		0.50		0.50		0.50		0.33		0.33		0.43	
	SCI	0.60		0.50		0.50		0.50		0.50		0.67		0.67		0.57	
TICS-M																	
	P	0.015		0.018		0.018		0.016		0.019		0.017		0.012		0.018	
	I	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	O	0.000	0.4	0.000	0.1	0.000	0.3	0.000	0.4	0.000	0.3	0.000	0	0.000	0.5	0.000	0.3
	PI	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	PO	0.008	99.6	0.008	99.9	0.009	99.7	0.009	99.6	0.009	99.7	0.010	100	0.012	99.5	0.012	99.7
	IO	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	PIO	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	GM	1.932		1.934		1.928		1.929		1.929		1.938		1.931		1.925	
	SE	0.005		0.003		0.005		0.005		0.005		0.001		0.007		0.006	
	Gr	0.65		0.69		0.67		0.65		0.68		0.64		0.51		0.60	
	Ga	0.65		0.69		0.67		0.65		0.68		0.64		0.51		0.60	
	TCI	0.65		0.69		0.67		0.64		0.68		0.63		0.50		0.60	
	SCI	0.35		0.31		0.33		0.36		0.32		0.37		0.50		0.40	

Sensitivity and Specificity

Table 3.8 and Figure 3.2 display the results of ROC analyses and the ROC curves respectively. All area under the curve of the ROC analysis (*AUC*) values for the MMSE (0.84-0.96) and TICS-M (0.85-0.95) indicated good to excellent overall accuracy in detecting dementia across follow-up occasions (Field, 2013). The sensitivity and specificity of the MMSE and TICS-M considered incident dementia for four follow-up occasions (occasion 1 was excluded because no participant had a diagnosis of dementia (Sachdev et al., 2010). Sensitivity and specificity were computed based on the number of participants who were diagnosed with or without dementia and their MMSE and TICS-M scores at each follow-up occasion (Figure 1). The TICS-M demonstrated a high sensitivity and high specificity for dementia with sensitivity ranging from 80.0% to 89.5%, and specificity ranging from 81.0% to 87.3%, for all four follow-up occasions. The MMSE displayed dramatically higher specificity, ranging from 98.8% to 99.5%; however, sensitivity for detecting dementia was moderately low ranging from 35.4% to 55.2% across the same periods.

Table 3.8

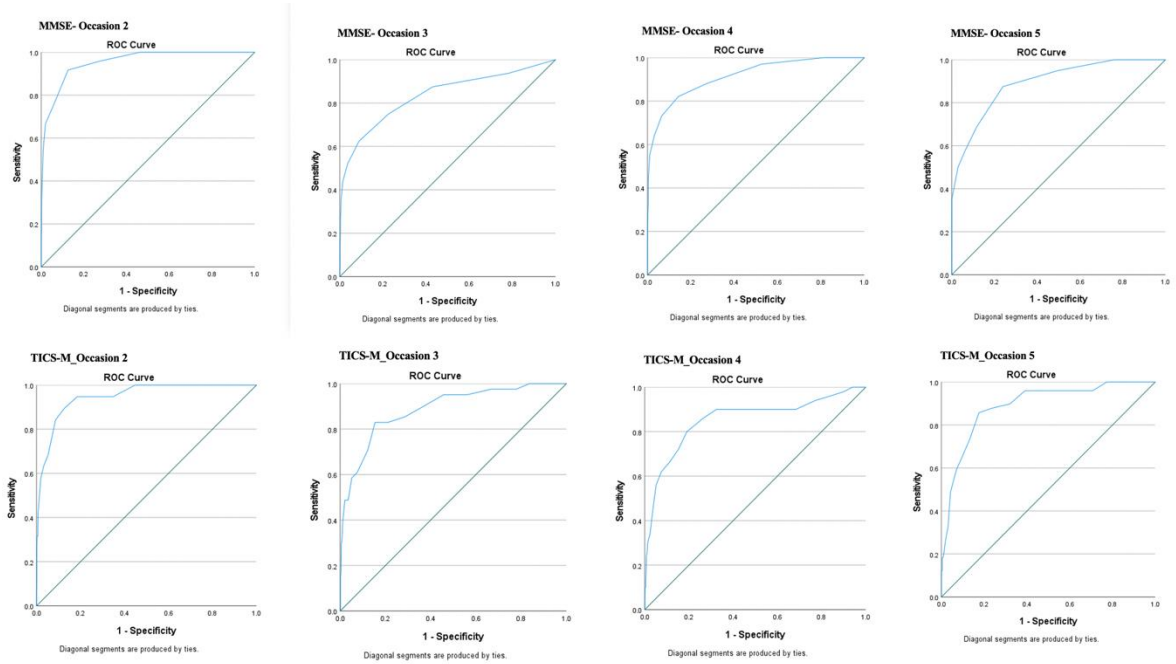
ROC analysis of the MMSE and the TICS-M for detecting dementia across follow-up occasions

		AUC	95% *CI*	Sensitivity	Specificity
MMSE					
	Occasion 2	0.96*	0.92-0.99	54.2	99.1
	Occasion 3	0.84*	0.76-0.91	35.4	99.4
	Occasion 4	0.92*	0.88-0.95	55.2	99.0
	Occasion 5	0.89*	0.85-0.93	40.0	99.0
TICS-M					
	Occasion 2	0.95*	0.91-0.99	89.5	87.3
	Occasion 3	0.89*	0.83-0.94	82.9	84.6
	Occasion 4	0.85*	0.79-0.92	80.0	81.0
	Occasion 5	0.89*	0.84-0.94	85.7	82.4

Note: *$p<0.001$

Figure 3.2

ROC curves for the MMSE (top panel) and for the TICS-M (bottom panel) across follow-up occasions



Discussion

The aim of this study was to investigate dynamic and enduring aspects of global cognition measured by the MMSE and the TICS-M, and to examine sources of measurement error using G-Theory. Our results demonstrated that the MMSE captures both dynamic and enduring aspects of cognitive status, which makes it suitable for general clinical assessments. This study has also developed a shorter version of the MMSE (MMSE-D8), which is suitable for monitoring transitory changes of cognition over time. The TICS-M demonstrated acceptable reliability and generalisability of assessment scores across the sample population at five occasions and appears to capture more enduring aspects of cognitive status. That is, enduring aspects captured 72% of assessments scores with the remaining 28% attributed to dynamic aspects. The TICS-M also demonstrated high sensitivity and specificity for dementia screening across four follow-up occasions. Therefore, the TICS-M may be useful as a measure of enduring features of cognition and as a screen for dementia but may be less sensitive to transitory changes in cognition. Besides measuring enduring and dynamic aspects of cognition, both measures were not affected by measurement error to any significant degree.

Repeated ANOVA confirmed that the TICS-M measures more enduring aspects of global cognition as there was no significant change in the TICS-M scores across occasions 1 to 4. In contrast, the MMSE appears to equally capture variance for both enduring (56%) and dynamic (44%) aspects of cognitive performance. The ANOVA results also showed that the MMSE scores were essentially dynamic due to the significant fluctuation of MMSE scores across five occasions. Given these results, this study has developed a version of the MMSE designed to capture dynamic aspects of cognitive performance by removing the three items least sensitive to change. This resulted in the modified MMSE-D8, which satisfied the criteria of reliable and valid state (i.e., dynamic) measures ($G < 0.40$, $SCI > 0.60$) and can be used to measure transitory changes of cognitive function over time.

Moreover, our D-study was conducted to examine psychometric properties of individual items of the MMSE and the TICS-M to optimise the measurements. It is noteworthy that most individual items of both scales measured dynamic aspects of global cognition, except for MMSE item 1 and TICS-M items 1 and 5. These findings highlight that some aspects of cognition (e.g., memory) may be more adaptable overtime and can possibly be enhanced by means of interventions. For example, our findings are consistent with some other studies showing that there is a relationship between cognitive training and cognitive gains in specific domains such as memory in healthy older adults (Mahncke et al., 2006; Shatil et al., 2010).

We found that the TICS-M is both highly sensitive and highly specific in detecting dementia, which is aligned with previous research (Bentvelzen et al., 2019). High sensitivity and high specificity of the TICS-M in detecting dementia is also consistent with measuring enduring aspects of global cognition. Therefore, this study adds evidence to the benefits of using the TICS-M for large epidemiological studies (Fong et al., 2009; Wilson & Bennett, 2005) as it is time and cost efficient and demonstrates high sensitivity and specificity for identifying dementia. In contrast, the MMSE demonstrated low sensitivity and high specificity in identification of dementia, which is consistent with our findings that it mostly captures dynamic aspects of global cognition. Our findings are consistent with previous studies demonstrating high specificity of the MMSE, but are less consistent with results other studies that have found low sensitivity for the MMSE (O'Bryant et al., 2008; Weston, 1987). These differences may be due to heterogeneity in study designs, such as different diagnosis criteria of dementia, different sample sizes, or highly homogeneous sample of participants (e.g., highly educated participants).

The main strength of this study is that it is the first to apply the robust methodology of G-Theory to both MMSE and TICS-M in the same large longitudinal sample over a relatively long period of 10 years. However, the limitations need to be acknowledged. The sample was

recruited from a small catchment area in Sydney, Australia, which is highly educated, upper-middle class and predominantly European. Further, not all participants were native speakers of English, which precluded them from our analyses. Besides that, it should be noted that the attrition rate for follow-up in the G-analyses was high, which may have resulted non-random attrition bias. Therefore, it would be beneficial to replicate these analyses on a more representative sample with a lower attrition over time. Moreover, this study did not distinguish between different subtypes of dementia, which might impact the results given research has shown that some dementia subtypes are less sensitivity to the changes in individual cognitive performance. For example, individuals with the behavioural variant of frontotemporal dementia (bvFTD) are less sensitive to cognitive changes captured by the MMSE (Piguet et al., 2009). In addition, although this study aimed to analyse data from occasions with equal intervals of 2 years the interval between occasion 4 and occasion 5 was 4 years, as there were no MMSE data collected for wave 5 assessments (Figure 1). Future studies should replicate these analyses using equal intervals between occasions and samples with clarifications of dementia subtypes.

Conclusion

The findings of this study indicated that TICS-M assessment scores are reliable and generalizable across the same sample population and across occasions, and that they capture enduring aspects of global cognition. The TICS-M also demonstrated its superior sensitively for screening incidence of dementia. Further, the MMSE measured both enduring and dynamic aspects of cognitive status, and our results suggest that a modified version of the MMSE (MMSE-D8) is a reliable assessment of transitory changes in cognitive functions over time. Our findings highlight the importance of distinguish dynamic and enduring features of cognition using techniques like G-Theory, which may be beneficial for developing and evaluating preventive interventions and monitoring global cognition overtime.

Chapter 4 Enhancing precision of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16) using Rasch Model

Truong, Q. C., Choo, C., Numbers, K., Merkin, A. G., Brodaty, H., Kochan, N. A., Sachdev, P. S., Feigin, V. L., & Medvedev, O. N. (2021). Enhancing precision of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16) using Rasch Model. *International Psychogeriatrics*, 1–11. https://doi.org/doi.org/10.1017/S1041610221002568

Declaration

This study is co-authored work with C. Choo, K. Numbers, A. G. Merkin, H. Brodaty, N. A. Kochan, P. S. Sachdev, V. L. Feigin, and O. N. Medvedev (see Appendix P3 for the original article). I was the first author and had the major contribution to this study that included designing and conducting the study, completing data analysis, writing the manuscript for submission to *International Psychogeriatrics* journal, and responding to reviews/addressing their comments during peer-review process. All co-authors have signed to the co-authorship form (see Appendix 4.1).

Abstract

***Objective*:** This study aimed to investigate psychometric properties and enhance precision of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16) up to interval-level scale using Rasch methodology.

Design: Partial Credit Rasch model was applied to the IQCODE-16 scores using longitudinal measurement design spanning 10 years of biennial follow-up

Setting: Community-dwelling older adults aged 70-90 years and their informants, living in Sydney, Australia, participated in the longitudinal Sydney Memory and Ageing Study (MAS).

Participants: The sample included 400 participants of the MAS aged 70 years and older, 109 out of those were diagnosed with dementia 10 year after the baseline assessment.

Measurements: The IQCODE-16

***Results*:** Initial analysis indicated excellent reliability of the IQCODE-16, Person Separation Index (*PSI*) = 0.92, but there were four misfitting items and local dependency issues. Combining locally dependent items into four super-items resulted in the best Rasch model fit with no misfitting or locally dependent items, strict unidimensionality, strong reliability, and invariance across person factors such as participants' diagnosis and relationship to their informants, as well as informants' age and sex. This permitted the generation of conversion algorithms to transform ordinal scores into interval data to enhance precision of measurement.

***Conclusions*:** The IQCODE-16 demonstrated strong reliability and satisfied expectations of the unidimensional Rasch model after minor modifications. Ordinal-to-interval transformation tables published here can be used to increase accuracy of the IQCODE-16 without altering its current format. These findings could contribute to enhancement of precision in assessing clinical conditions such as cognitive decline in older people.

***Keywords*:** Subjective Cognitive Complaints, Measurement, Informant Questionnaire on Cognitive Decline in the Elderly, Rasch Analysis, Reliability

Introduction

Older adults frequently report subjective cognitive complaints (SCC), which may reflect self-estimations of changes in cognitive functions . SCC can be reported by an individual or by their close kin, friends or caregivers, referred to as informants (Brodaty et al., 2002; Jorm et al., 1991). SCC contributes to a diagnosis of Subjective Cognitive Decline (SCD), which are cognitive complaints in the absence of impaired cognitive performance (Centers for Disease Control and Prevention, 2019). SCD is thought to be a pre-Mild Cognitive Impairment (MCI) stage of dementia (Jessen et al., 2014). Therefore, SCC also contribute to screening and diagnosis of MCI.

Recent studies have shown that self- and informant-reported SCC, in the absence of impaired performance, predict steeper rates of cognitive decline and incident dementia six years out (Numbers et al., 2020), and are related to the presence of Alzheimer's Disease-dementia biomarkers such as amyloid plaques in the brain, tau proteins in cerebral spinal fluid (Amariglio et al., 2015; Sierra-Rio et al., 2016), and atrophy and/or hypometabolism (Jessen et al., 2014; Striepens et al., 2010). Together, these data suggest that SCC may be the earliest detectable stage of dementia (Skoog et al., 2017). Moreover, SCC are quick and easy to capture and may provide insight into cognitive changes over and above those captured by more time-consuming and costly formal standardised neuropsychological testing (Numbers et al., 2020).

More recently, researchers have begun examining the predictive utility of self- versus informant-reported SCC, and their association with cognitive performance and/or future decline. In the MAS study, SCC were measured over 10 years using two scales: the MAC-Q and IQCODE (Sachdev et al., 2010). Research showed that the informant IQCODE scores may be a more reliable approximation of actual cognitive performance than self-reported SCC (Rami et al., 2014; Slavin et al., 2015), and are more predictive of future cognitive decline (Caselli et al., 2014). Moreover, self-report SCC are highly correlated with an individual’s

mood, personality traits (under/over-complaining behaviors), life events, and medications (Buckley et al., 2015; Ponds & Jolles, 1996). Therefore, informant-reported SCC may be more accurate representations of cognitive decline as they are not subject to such biases. Indeed, in clinical contexts, informant SCC reports are reliably used as individuals begin losing insight into their cognitive changes over the progress of pre-clinical stages and the debilitating course of dementia (American Psychology Association, 1994).

The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE; Jorm et al., 1991) is a well-established psychometric informant-reported measure of subjective cognitive complaints with 26 items. The short 16-item scale version of the IQCODE (IQCODE-16; Jorm, 1994) is a widely used SCC assessment instrument, which has demonstrated comparable reliability to the original version with Cronbach's alpha ranging from 0.93 to 0.97 (Harrison et al., 2015; Jorm, 1994; Phung et al., 2015). The IQCODE-16 is completed by informants who are caregivers, close kin or friends of the individual and know them well enough to comment on their memory (Jorm, 1994). Studies have shown that the IQCODE-16 predict incident dementia and can be used as a screening tool for dementia (Park, 2017; Perroco et al., 2008). However, the differences between response options of an ordinal scale (e.g., 1 and 2 vs 2 and 3) may not reflect the same amount of clinical change compared to the interval scale and individual items contribute different amount of information about the latent trait to the overall assessment score (Hobart & Cano, 2009). Therefore, psychometric properties of the IQCODE-16 should be thoroughly examined in order to improve the precision of the instrument up to an interval measure by utilising an advanced methodology such as Rasch analysis (Rasch, 1960; Tennant & Conaghan, 2007).

Rasch analysis is a powerful statistical method used to examine reliability and internal validity of psychometric instruments, as well as their specific psychometric properties such as functioning of individual items (Rasch, 1960, 1961) that has increasingly become a gold

standard in clinical assessments (Hobart & Cano, 2009; Lundgren-Nilsson & Tennant, 2011). Rasch model is unidimensional and based on assumptions that the response to any specific item of a scale is determined by both an individual's ability and an item's difficulty (Rasch, 1960, 1961). While every ordinal item in an instrument usually has the same categorical values, the total scale score may be biased because each individual item has a different contribution to the overall latent trait (i.e. SCC levels) represented by items of a scale (Rasch, 1960). Moreover, the differences between response options of individual items reflect varying levels of clinically important change (Masters, 1982). Rasch analysis can reduce these biases by estimating precise thresholds between response categories of individual items and the unique contribution of each individual item to the overarching trait being measured (Stucki et al., 1996). Such precise estimations are possible because most individuals get higher scores on the easy items while only a few score highly on difficult items (Tennant & Conaghan, 2007). When data fit to the Rasch model, ordinal scores can be converted into interval level data by accounting for item difficulty and person ability, hence increasing the precision of measurement that was demonstrated in a number of studies across different areas of clinical assessments at both the group and individual level (Norquist et al., 2004).

Other statistical methods such as Classical Test Theory and its extension, Generalisability Theory (Cronbach et al., 1963) are not able to estimate the contribution of each individual item of the scale to the overall latent trait because they do not differentiate between item difficulty based on individual ability (Fox & Jones, 1998). An item-person threshold distribution is another advanced feature of the Rasch analysis that can graphically show how the range of item difficulties covers range of persons' ability. This graphic is useful to detect potentially significant ceiling or floor effects.

However, no Rasch analysis has been conducted on the English scale version of the IQCODE-16 to further enhance its accuracy up to interval-level scale more suitable for

parametric statistics and clinical evaluations. In fact, only one study to date has applied Rasch analysis to investigate the psychometric properties of the 26-item Chinese version of IQCODE in a Chinese sample (Tang et al., 2004). Although this study concluded that four out of the 26 items had statistically inadequate fit and should be removed from the scale, the authors did not modify the scale to enhance its psychometric properties. These four misfitting items were excluded in the 16-item English version of the IQCODE, which was developed earlier (Jorm, 1994).

Nonetheless, Tang et al.'s (2004) study findings are not applicable to the English version of the IQCODE and may not be generalizable to older adults with and without cognitive impairment as the sample was selectively comprised of Chinese stroke survivors and informants. Moreover, Person Separation Index (PSI) was not reported in this study, even though it is an essential reliability estimate in Rasch analysis (Tennant & Conaghan, 2007). PSI is considered a reliable alternative to sensitivity analysis because it reflects the IQCODE's abilities to distinguish between individual SCC levels (Fisher, 1992). Finally, Tang et al. (2004) did not generate converging tables to transform raw scores into interval level data. Such converging tables have been recommended in the recent reporting guidelines as essential for Rasch analytic studies (Leung et al., 2014) as they allow clinicians and researchers to transform raw scores into interval-level data to achieve higher precision of the scale without modifying the scale's original response format (Merkin et al., 2020).

Given no study to date has applied Rash analysis to the English version of the IQCODE-16, and several shortcomings of the Tang et al study's methodology, the aim of the current study was to apply Rasch methodology to investigate and enhance psychometric properties of the 16-item IQCODE. This is also important because cross-sectional subjective cognitive complaints provided by a patient may be different from informant-based scores that considered as more reliable. This study has also considered that the longitudinal (10 years) informant-

based history captured by the IQCODE may have impact on assessment scores and include two independent samples from the first and the last wave (10 years apart) in the analysis to address this issue. A secondary aim was to produce convention tables that can be used to convert raw scores into interval level data to enhance the precision of the scale, if an acceptable fit to the Rasch model is achieved.

Method

Participants

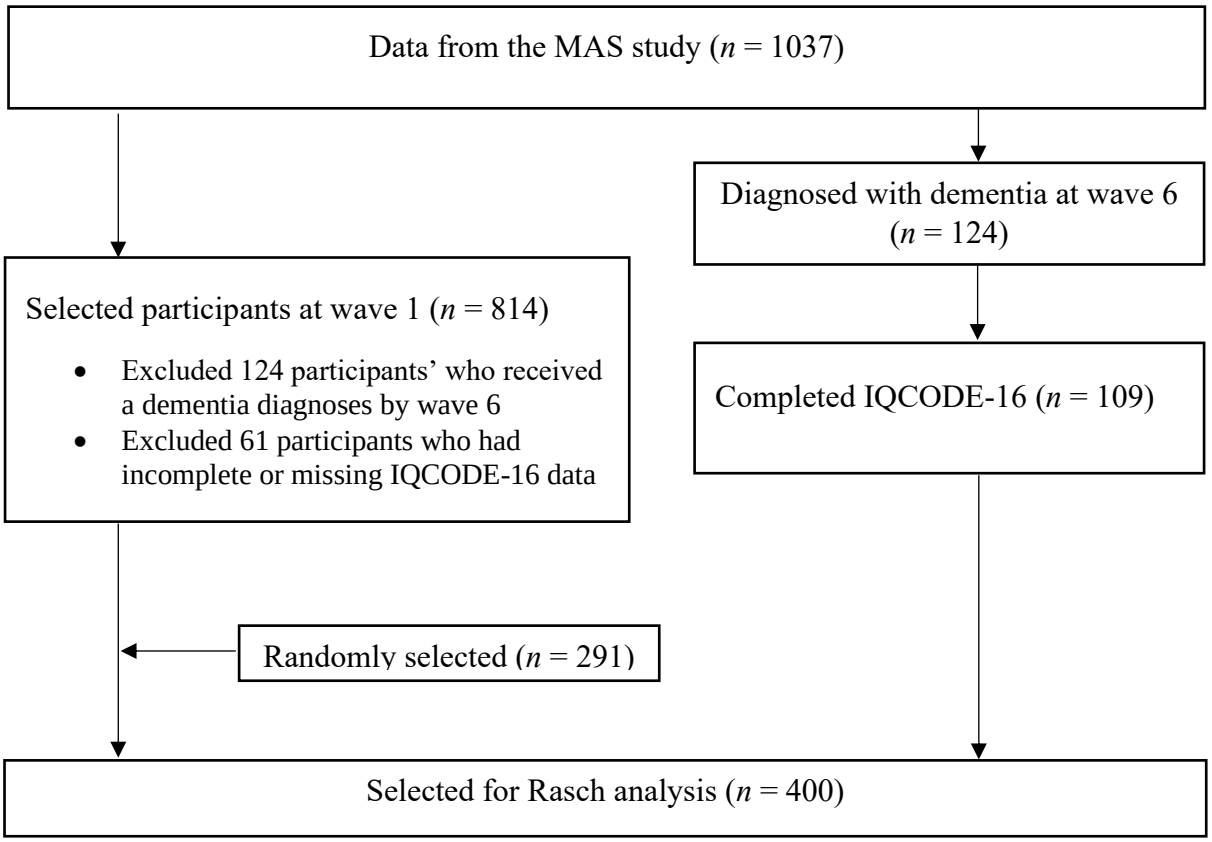
Participants were drawn from the Sydney Memory and Ageing Study (MAS), a longitudinal study of cognitive ageing that included a baseline sample of 1037 community-dwelling older adults aged between 70-90 years, without dementia. MAS participants were recruited from the Eastern suburbs of Sydney, Australia between 2005 and 2007 (Sachdev et al., 2010). The vast majority of participants identified as European (98%), with 1% identifying as Asian, 0.3% as Mixed-race, and 0.8% as other or not revealed. Most participants (97.3%) had informants, who were a close friend or family member that knew the participant sufficiently well to answer questions relating to the participant's memory, thinking, and daily functioning. Informants were required to have at least 1 hour of contact with the participant per week; on average they had 8.3 hours of weekly contact. Participants and informants were interviewed biennially from Wave 1 (baseline) to Wave 6 (10-year follow-up). All participants and informants provided written consent to participate in this study, which was approved by the University of New South Wales Human Ethics Review Committee (HC 05037, 09382, 14327).

Figure 4.1 presents the consort diagram of how participants were selected for Rasch analysis. To control for differential item functioning (DIF) between participants who went on to be diagnosed as normal versus dementia at the latest wave, this study included all IQCODE-16 reports of participants who were diagnosed with dementia at wave 6 (n = 109). Studies indicated that the appropriate sample size for Rasch analysis is between 250 and 500 participants because it allows a researcher to minimise Type I and Type II errors with questionnaires consisting of 15 to 20 items in Rasch analysis (Azizan et al., 2020; Hagell & Westergren, 2016). Therefore, to insure unique participants in the sample and an appropriate sample size for Rasch analysis, 291 participants were then randomly selected at wave 1, after

excluding data from participants who were diagnosed with dementia at wave 6. Informants in the selected sample were 121 males (30.3%) and 277 females (69.3%). Informant ages ranged from 24 to 95 years, with a mean age of 62.02 years (standard deviation *SD* = 14.48). Ninety-eight informants were participants' spouses, 125 were participants' children, 2 were grandchildren, 7 were siblings, 16 were other relatives (e.g., niece), 64 were close friends and 88 were "other". Missing data in the extracted sample comprised less than 0.01% and were completely at random.

Figure 4.1

CONSORT flow diagram of participants selected for Rasch analysis of the IQCODE-16



Measure

The IQCODE-16 (Jorm, 1994) consists 16 items that informant-reported SCC levels of participants. To complete the IQCODE-16, informants are asked questions about how the participants' memory and cognitive function have changed, regardless of their (participants') premorbid intelligence or education (Park, 2017). Items are scored on a 5-point Likert scale with options ranging from 1 = "much improved" to 5 = "much worse". For example, item 1 is "Remembering things about family and friends (e.g., occupations, birthdays, addresses)".

Data Analyses

IBM SPSS v.27 was used to compute descriptive statistics including means, standard deviation (SD), and Cronbach's alpha for the IQCODE-16. RUMM2030 software package (Andrich et al., 2009) was used to conduct Rasch analyses by following the recommendations of Tennant and Conaghan (2007) and standardised criteria to evaluate the Rasch model fit as described elsewhere (Leung et al., 2014).

As individual IQCODE-16 items were polytomous, a likelihood-ratio test was conducted to identify an appropriate polytomous-item Rasch model should be applied, the Partial Credit model (Masters, 1982) or the Rating Scale model (Andrich, 1978). While the rating scale model implies that response categories across all individual scale items have the same rating scale structure, the partial credit model is unrestricted and assumes that each individual item has its own unique response options structure (Linacre, 2000). The likelihood-ratio test examined distances between thresholds of individual items. If threshold distances across individual items were significantly different across individual items, the partial credit model should be used (Lundgren-Nilsson & Tennant, 2011). Otherwise, the rating scale model would be suitable (Tennant & Conaghan, 2007).

Rasch analysis was iterative and continued until the best model fit was achieved. The overall model fit requires the estimate of item-trait interaction to be not significant, which is reflected by chi square index ($p > .05$). The fit residuals for individual items were evaluated to detect item misfitting (i.e. item fit residuals should be between −2.50 and +2.50). The residual correlations between individual items were then examined and those had values above 0.20 indicated local dependency (Christensen et al., 2013). DIF due to relevant individual characteristics (i.e., personal factors) were also tested to identify whether all items were invariant across different groups (e.g., age, sex). Generally, the scale meets expectations of the Rasch model if there are no significant interactions between items and the latent trait, no misfitting items, no local dependency and/or DIF, and unidimensionality is evident by examining the principal components analysis (PCA) of the residuals and the equating *t*-test (Leung et al., 2014). In addition, *PSI* is used to evaluate reliability in Rasch analysis, which is not the Rasch model fit criteria but reflects how well the scale discriminates between individuals with different levels of the latent trait (e.g., SCC). *PSI* is interpreted somewhat similar to Cronbach's alpha with values above 0.70 indicating acceptable reliability for group assessments and 0.80 and higher for individual assessments (Medvedev, Turner-Stokes, et al., 2018).

In this study, super-items were created by combining locally dependent items together to improve the Rasch model fit (Lundgren-Nilsson et al., 2013; Medvedev, Turner-Stokes, et al., 2018). When the Rasch model fit was satisfactory, the person-item thresholds distribution was examined showing how well items thresholds of the IQCODE-16 cover SCC levels of the sample. Lastly, the transformation table was produced to convert raw scores into interval level data to increase the precision of assessment. Statistical significance was estimated using the conventional cut-off point of *p*-value >0.05.

Results

The IQCODE-16 (*M* = 52.45, *SD* = 7.98) showed satisfactory internal reliability, with Cronbach's alpha value of 0.96, which is consistent with previous validation reports (Harrison et al., 2015; Phung et al., 2015).

A likelihood-ratio test showed significant differences between thresholds across individual IQCODE-16 items, $\chi^2(44) = 150.57$, $p < .001$, which means that the unrestricted Partial Credit model was more appropriate to use for the data of this study. Table 4.1 displays the overall model fit estimates of the initial, second and third Rasch analyses of the IQCODE-16. As can be seen, initial analysis (A1) indicated excellent reliability PSI = 0.92 for the full IQCODE-16 and the overall fit to the Rasch model was acceptable as indicated by non-significant chi-square, $\chi^2(80) = 92.23$, $p = .16$. Table 4.2 presents estimates of the initial Rasch analysis for individual items including item location, fit residual, and chi-square values. There are four items (i.e., item 3, 10, 12, and 15) with significant misfit to the model.

Table 4.1

Summary of fit statistics for the initial and the final Rasch analyses of the IQCODE-16 (n=400)

	Person mean		Goodness of fit		PSI	Significant *t*-tests (Unidimensionality)	
Analyses	**Value / SD**		**χ2 (df)**	***p***		**%**	**Lower bound**
Initial (A1)	1.41	2.26	92.23(80)	0.16	0.92	5.5	3.4 (YES)
Second (A2)	1.36	2.22	85.70(86)	0.49	0.92	6.8	4.6 (YES)
Final (A3)	1.26	2.11	81.32 (78)	0.38	0.92	4.3	2.1 (STRICT)

Table 4.2

Rasch model fit statistics including item locations, fit residuals, and Chi-square for the initial analysis of the IQCODE-16 individual items

Item	Item location	Item-fit residual	Chi-square
1. Remembering things about family and friends (e.g. occupations, birthdays, addresses)	-0.01	-1.66	5.58
2. Remembering things that have happened recently	-0.03	-2.52	6.72
3. Recalling conversations a few days later	-0.91	-4.09*	6.02
4. Remembering his/her address and telephone number	0.59	0.48	8.87
5. Remembering what day and month it is	0.15	-2.15	3.48
6. Remembering where things are usually kept	-0.61	-1.45	1.11
7. Remembering where to find things which have been put in a different place from usual	-1.04	-2.23	20.47
8. Knowing how to work familiar machines around the house	0.04	-2.07	3.88
9. Learning to use a new gadget or machine around the house	-0.34	-2.35	7.52
10. Learning new things in general	-0.30	-4.05*	4.60
11. Following a story in a book or on TV	0.13	-2.32	1.38
12. Making decisions on everyday matters	<0.01	-3.67*	2.68
13. Handling money for shopping	0.55	-0.51	7.43
14. Handling financial matters e.g. the pension, dealing with the bank	0.76	-1.84	2.98
15. Handling other everyday arithmetic problems (e.g. knowing how much food to buy, knowing how long between visits from family or friends)	0.28	-2.84*	7.05
16. Using his/her intelligence to understand what's going on and to reason things through	0.72	-1.71	2.43

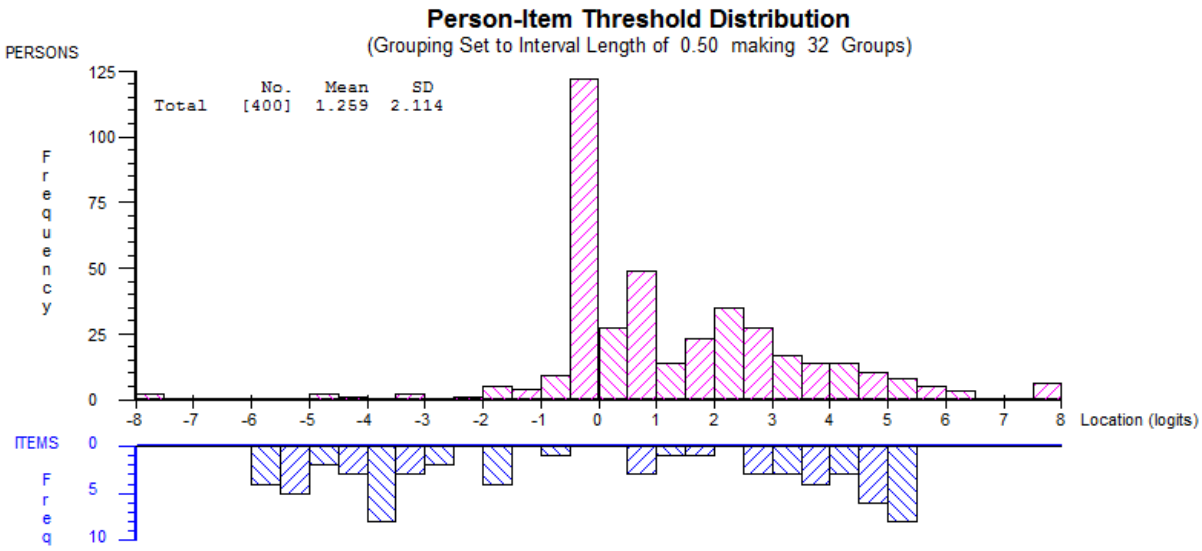
Note: * Significant misfit to the Rasch model

The residual correlation matrix was also examined and showed local dependency between several items was above 0.20. For example, the residual correlation between item 9 and 10 was 0.52, and between item 1 and 2, and 2 and 3, were both 0.32. As mentioned previously, combining locally dependent items into super-items can reduce measurement error as well as improve individual items and the overall model fit (Lundgren-Nilsson et al., 2013; Medvedev, Turner-Stokes, et al., 2018; Merkin et al., 2020). Therefore, five super-items comprised of locally dependent items were created: super-item 1 (Items 1&9); super-item 2 (2&13); super-item 3 (3&14); super-item 4 (6&7); and super-item 5 (10&11). This improved the overall model fit (see Table 4.1, analysis A2) with no change of reliability and acceptable unidimensionality (4.6% of significant t-tests). However, the residual correlation between super-item 1 and super-item 3 was still above 0.20, which indicated local dependency. To address this issue, the third analysis (A3) was conducted with 4 super-items (i.e. super-item 1: item 2,3,13, &14; super-item 2: item 1&9; super-item 3: item 6&7, super-item 4: items 10&11). This analysis resulted in strict unidimensionality and excellent reliability (see Table 4.1). Besides that, examining the residual correlations for this analysis indicated no local dependency and DIF amongst items/super-items. Therefore, this analysis (A3) achieved the best Rasch model fit for the IQCODE.

Figure 4.2 presents person-item threshold distribution from the analysis of the best model fit (analysis A3) for the IQCODE-16. It shows that the scale's thresholds satisfactory cover SCC levels of the sample and there are no significant ceiling or floor effects. The sample mean is higher than the item means, reflecting the overall higher SCC levels of the sample due to inclusion of the subsample of participants with dementia diagnosis, though the mode was just below zero mark.

Figure 4.2

Person-item threshold distribution of the final analysis of the IQCODE-16



The third analysis (A3) for the IQCODE-16 demonstrated the best Rasch model fit that permitted conversion of raw ordinal IQCODE-16 scores into interval data. Table 4.3 displays Rasch ordinal-to-interval transformation table developed for the IQCODE-16 based on person estimates.

Paired sample *t*-test were used to examine the difference between IQCODE-16 raw scores and interval-transform scores interval-transform scores at wave 6 because at this wave more participants were diagnosed with dementia (n=124). The results revealed that for the full sample, ordinal raw scores (M = 53.22, SD = 7.79) were significantly lower compared to interval level Rasch scores (M = 53.53, SD = 8.24) as evidenced by the test statistics ($t(337)$ =-3.21, p =0.001). Similarly, ordinal raw scores (M = 61.12, SD = 10.13) were significantly lower compared to interval transformed scores (M = 61.71, SD = 9.53) in the subsample consisted of those who were diagnosed with dementia at wave 6 ($t(81)$ = -2.52, p =0.01) but not in the subsample of those who were not dementia diagnosed ($p > 0.05$).

Table 4.3

Converting ordinal scores into interval level scores for the IQCODE-16

Raw scores	Logits	Interval scores	Raw scores	Logits	Interval scores
16	-7.65	16.00	49	0.15	48.58
17	-6.93	19.01	50	0.58	50.38
18	-6.45	21.01	51	0.99	52.09
19	-6.13	22.34	52	1.38	53.69
20	-5.89	23.37	53	1.73	55.18
21	-5.68	24.24	54	2.06	56.53
22	-5.49	25.01	55	2.35	57.76
23	-5.32	25.72	56	2.61	58.86
24	-5.17	26.38	57	2.85	59.86
25	-5.02	27.01	58	3.07	60.76
26	-4.87	27.60	59	3.27	61.59
27	-4.74	28.17	60	3.45	62.36
28	-4.61	28.72	61	3.62	63.07
29	-4.48	29.25	62	3.78	63.74
30	-4.35	29.78	63	3.93	64.36
31	-4.23	30.30	64	4.07	64.96
32	-4.10	30.83	65	4.21	65.53
33	-3.97	31.37	66	4.34	66.07
34	-3.84	31.93	67	4.47	66.61
35	-3.70	32.51	68	4.60	67.13
36	-3.55	33.12	69	4.72	67.65
37	-3.40	33.76	70	4.85	68.18
38	-3.23	34.45	71	4.97	68.71
39	-3.06	35.18	72	5.11	69.26
40	-2.87	35.98	73	5.25	69.85
41	-2.66	36.86	74	5.40	70.50
42	-2.42	37.83	75	5.57	71.22
43	-2.16	38.92	76	5.78	72.07
44	-1.86	40.16	77	6.03	73.13
45	-1.53	41.56	78	6.37	74.55
46	-1.15	43.14	79	6.89	76.71
47	-0.74	44.87	80	7.68	80.00
48	-0.29	46.72			

Discussion

The IQCODE-16 is one of the most widely used tools to assess SCC levels in older persons. This study applied Rasch analysis to investigate psychometric properties of the IQCODE-16 and derived ordinal-to-interval conversion tables to enhance precision of the instrument. Achieving excellent reliability (*PSI* = 0.92) in this study provides further empirical evidence to support the robust psychometric properties of the 16-item IQCODE and suitability of scale for both individual and group assessment. Other findings of the current study demonstrated that the IQCODE-16, reorganized into super-items, achieved the best fit to the unidimensional Rasch model, which then permitted us to improve precision of the measure using the ordinal-to-interval conversion algorithms presented in Table 4.3. Such transformation is important because individual items of ordinal scales such as the IQCODE-16 have varying degrees of difficulty and thus, each item contributes uniquely to the total score, which should be accounted for (Stucki et al., 1996). As such, using Rasch transformation tables decreases measurement error associated with ordinal scales scores (Medvedev, Siegert, Kersten, et al., 2017).

This study has also demonstrated that the ordinal scale bias was statistically significant in the current study as evident by paired t-tests comparing raw scores and interval-transformed scores converted to the same scale range in the total sample and dementia diagnosed subsample. The ordinal IQCODE raw scores were significantly lower than interval level scores in the sample of participants who were diagnosed with dementia. Although the effect size of this difference was relatively small, this illustrates that using IQCODE-16 ordinal scores may impact on the results if conducting parametric statistical tests. An extensive literature on Rasch methodology suggests that the interval-level scores derived from Rasch analyses may more accurately reflect an individual's level of SCC, and could thus be used to conduct parametric statistics without violating their arithmetic assumptions (Hobart & Cano, 2009; Leung, 2011).

Interval-level data of the IQCODE-16 can also be used for valid statistical comparisons with other interval measures like electrophysiological and neuroimageing data and biomarkers and earlier collected data can be reanalysed/replicated to increase reliability and validity of the results. Further, transforming the IQCODE-16 ordinal scores into interval data is user-friendly and does not require expertise in statistics, meaning that researchers can easily determine an individual's interval score based on their raw IQCODE-16 score simply by referring to the data in Table 4.3.

Ordinal-to-interval transformation tables also have important clinical implications as our study offers a parsimonious tool to assist clinicians in determining presence and severity of SCC. Although original IQCODE scores are reliable and have demonstrated good validity in the clinical context, these scores constitute an ordinal measure. For example, if participant A has an IQCODE score of 32, they are less risk to have cognitive impairment (e.g., dementia) than participant B, who scores 59. Both participants A and B are then engaging in the therapeutic intervention, the participant A's score reduces to 22 and the participant B's reduces to 49. As can be seen, both participants have decreased by ten points on the IQCODE ordinal scores. When using the Rasch interval transform scores, participant A has reduced by 5.82 points, while participant B has reduced by 13.01. This shows that participant B's reduction in SCC level is more considerable than participant A, meaning that one is more able to detect real change when using measurement with interval transform data. When using the IQCODE to identify who does or does not meet screening criteria (e.g., cognitive decline), clinicians do not need to transform the scores because ordinal scores are worked well for this purpose. The transformed scores should be used in clinical context where the IQCODE is used both for screening and outcome monitoring, but it should be noted that it is important to convert the criterion being used.

Strengths

The main strength of this study was the application of modern and robust Rasch methodology to an adequate sample size within an optimal range ($250 < n \leqslant 500$) that permits minimising Type I and Type II errors with questionnaires consisting of 15 to 20 items (Azizan et al., 2020; Hagell & Westergren, 2016). Besides that, this study is also novel in several ways. First, to date there have been no studies investigating the psychometric properties of the 16-item English version of the IQCODE using Rasch methodology. Combining locally dependent items into four super items resolved local dependency problems that can produce spurious correlations affecting scale accuracy and the Rasch model fit (Medvedev, Turner-Stokes, et al., 2018). Appropriateness of using super-items in the current study was further supported by strict unidimensionality and invariance of the modified 16-item IQCODE, because research has shown that multidimensional scales cannot generate super-items with an adequate Rasch model fit (Mitchell‐Parker et al., 2018). In addition, the achievement of strict unidimensionality and invariance across all personal factors of the modified IQCODE-16 also indicates that the modified 16-item IQCODE works equally well for both heathy older persons and those with a diagnosis of dementia, and is not impacted by the informant's sexes, age, or relationship to the individual. It showed be noted that the IQCODE-16's modifications were implemented internally and work if the ordinal-to interval conversion tables are applied, which does not require modifications of the original administration format of the scale.

Limitations

This study is not without some limitations, which should be acknowledged. Data used for the current study's analyses are highly homogenous and not representative of all older adults given MAS study participants and informants were recruited from an affluent area of Sydney, Australia with a predominantly white European ethnic group. Clinical research from

culturally diverse data provides support for use of the IQCODE-16 to detect early stages of dementia, there could be possible cultural variations across a range of health issues (Choo et al., 2017). Besides that, it is worthy to note that there was unbalance between informant genders, 693 (68.7%) and 277 (69.3%) females in the original sample and in the sample selected for Rasch analysis. Although Rasch methodology tends to be robust and less affected by sampling bias compared to other methods (Hobart and Cano, 2009), the results of this study should be replicated in a more diverse sample to investigate potential DIF across sample groups unrepresented in the current study. If DIF would be found for a specific group (e.g., other English-speaking countries, less affluent groups, or samples that are more balanced genders or reflect other ethnic groups), an additional conversion table could be produced for such group to permit valid score comparisons across a wider population.

Conclusion

In conclusion, the findings of this study indicate that the IQCODE-16 is a reliable and valid assessment tool for measuring SCC among older adults. Our adjustments of the IQCODE-16, made by Rasch analyses, resolved local dependency issues permitting transformation of raw scores into interval-level data, which improve the precision of measurement. The interval-transform data table allows both clinicians and researchers to apply this sound psychometric measurement in a variety of contexts with higher precision, without needing any modification to the original IQCODE-16 administration format.

Chapter 5 Establishing conversion of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly scores into interval-level data across multiple samples using Rasch methodology

Truong Quoc Cuong, Katya Numbers, Carol C. Choo , Adam C. Bentvelzen , Vibeke S. Catts , Matti Cervin , Anthony F. Jorm , Nicole A. Kochan, Henry Brodaty , Perminder . Sachdev & Oleg N. Medvedev (2023). Establishing conversion of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly scores into interval-level data across multiple samples using Rasch methodology. *Psychogeriatrics*. https://doi.org/10.1111/psyg.12946

Declaration

This study is co-authored work with Katya Numbers, Carol C. Choo, Adam C. Bentvelzen, Vibeke S. Catts, Matti Cervin, Anthony F. Jorm, Nicole A. Kochan, Henry Brodaty, Perminder S. Sachdev& Oleg N. Medvedev (see Appendix P4 for the original article). I was the first author and had the major contribution to this study that included designing and conducting the study, completing data analysis, writing the manuscript for submission to *Psychogeriatrics* journal, and responding to reviews/addressing their comments during peer-review process. All co-authors have signed the co-authorship form (see Appendix 5.1).

Abstract

***Background*:** The 16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16) is a well-validated and widely used measure of cognitive changes among older adults. This study aimed to use Rasch methodology to establish psychometric properties of the IQCODE-16 and validate the existing ordinal-to-interval transformation algorithms across multiple large samples.

Methods: A Partial Credit Rasch model was employed to examine psychometric properties of the IQCODE-16 using data (n = 918) from two longitudinal studies of participants aged 57 to 99 years: the Older Australian Twins Study (n = 450) and the Canberra Longitudinal Study (n = 468), and re-using the Sydney Memory and Ageing Study (MAS) sample (n = 400).

***Results*:** Initial analyses indicated good reliability for the IQCODE-16 (Person Separation Index ranged 0.82 to 0.90). However, local dependency was identified between items, with several items showing misfit to the model. Replicating the existence Rasch solution could not reproduce the best Rasch model fit for all samples. Combining locally dependent items into three testlets resolved all misfit and local dependency issues and resulted in the best Rasch model fit for all samples with evidence of unidimensionality, strong reliability, and invariance across person factors. Therefore, new ordinal-to-interval transformation algorithms were produced to convert the IQCODE-16 ordinal scores into interval data to improve the accuracy of its scores.

***Conclusions*:** The findings of this study support the reliability and validity of the IQCODE-16 in measuring cognitive changes among older adults. New ordinal-to-interval conversion tables generated using samples from multiple independent datasets are more generalizable and can be used to enhance precision of the IQCODE-16 without changing its original format. An easy-to-use converter has been made available for clinical and research use.

***Keywords*:** Cognitive Changes, Measurement, Informant Questionnaire on Cognitive Decline in the Elderly, Rasch Analysis, Reliability

Introduction

Cognitive changes (CCs) reported by older adults (self-reported), or their informants (informant-observed) reflect an individual's personal experience of deterioration in cognitive functions (Brodaty et al., 2002; Crook et al., 1992; Jorm, 2004). CCs are an essential criterion for the diagnosis of Mild Cognitive Impairment (MCI), which may in turn lead to dementia (Petersen et al., 1999). Recent studies have shown that both self-reported and informant-observed CCs significantly predicted incident dementia (Numbers et al., 2020; Truong et al., 2021a). Therefore, CCs may be considered the earliest stage of pre-clinical dementia (Mitchell et al., 2014; Skoog et al., 2017). Furthermore, CCs can be conveniently captured by self-reported or informant-observed questionnaires that are difficult to detect otherwise in the absence of comprehensive neuropsychological testing (Numbers et al., 2020).

Researchers have recently examined the advantages of self-reported versus informant-observed CCs (Nosheny et al., 2022). Several studies have shown that informant-observed CCs may be a more reliable approximation for predicting concurrent objective cognitive performance and/or future decline in comparison to self-reported CCs (Caselli et al., 2014; Rami et al., 2014; Slavin et al., 2015; Truong et al., 2021a). Particularly, in the longitudinal Memory and Ageing Study (MAS), cognitive changes were measured by the self-reported Memory Complaint Questionnaire (MAC-Q; Crook et al., 1992) and the informant-observed Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16; Jorm, 1994) over 10 years (Sachdev et al., 2010). Results from the MAS study showed that informant IQCODE scores were more accurate in predicting cognitive performance and incident dementia compared to self-reported MAC-Q scores (Truong et al., 2021a). Moreover, self-reports are subject to various biases such as the individual's mood (e.g., depression, anxiety), personality traits (e.g., under/over complaining behaviors, neuroticism, or conscientiousness), life events, and medications (Buckley et al., 2015; Ponds & Jolles, 1996; Reid & MacLullich, 2006).

However, informant CCs may also be affected by the same factors, although perhaps to a lesser extent (Numbers et al., 2020; Truong et al., 2021a). Therefore, informant-observed CC reports are increasingly used in clinical settings as individuals with incipient dementia may lose insight about their cognitive changes (American Psychology Association, 1994).

The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16; Jorm, 1994) is a widely-used 16-item CC scale with virtually equivalent psychometric properties to the original 26 item version (Jorm, 1994). The IQCODE-16 has demonstrated excellent internal consistency with Cronbach's alphas ranging from 0.93 to 0.97 (Harrison et al., 2015; Jorm, 1994; Phung et al., 2015; Truong et al., 2021b) and its scores have been shown to significantly predict incident dementia (Park, 2017; Perroco et al., 2008; Truong et al., 2021a). However, the IQCODE-16 is a well-validated measure, but its scores constitute an ordinal scale. The differences between response options of individual IQCODE-16 items (e.g., 1 and 2 vs 2 and 3) may not as accurately reflect the same amount of clinical change as an interval scale whereby the difference between 1 and 2 is the same as the difference between 2 and 3 (Masters, 1982). All 16 ordinal items of the IQCODE-16 are scored using Likert-scale response values ranging from 1 (much improved) to 5 (much worse) and the total score of IQCODE is the mean of the items. However, each individual item can reflect a symptom with different severity or difficulty levels, and may contribute differently to the overall latent assessment score (i.e., CC levels) (Hobart & Cano, 2009; Truong et al., 2021b).

Another issue related to the ordinal nature of the IQCODE scores is conducting parametric statistics as they violate arithmetic assumptions of such tests (Jamieson, 2004; Merbitz et al., 1989). Rasch analysis (Rasch, 1960) can help overcome these limitations. It is a specialized statistical approach that is increasingly used in evaluating reliability and internal validity of clinical psychometric instruments (Christensen et al., 2013; Hobart & Cano, 2009; Lundgren-Nilsson & Tennant, 2011; Medvedev, Berk, et al., 2020; Medvedev, Krägeloh, et al.,

2020; Merkin et al., 2020; Truong et al., 2021b). Rasch methodology has numerous advantages compared to other widely used statistical methods such as Classical Test Theory (Lord & Novick, 1968) and Generalizability Theory (Cronbach et al., 1963), which are unable to account for the different contributions of individual items to the overall latent trait (Fox & Jones, 1998). Specifically, these methods do not distinguish between item difficulty based on individual ability (Rasch, 1960, 1961). A Rasch model is strictly unidimensional and assumes that a person's response to a particular scale item is determined by that item's difficulty and the capacity of the person on the measured trait (Rasch, 1960, 1961). Rasch analysis can precisely estimate thresholds between response options of the individual scale items and the contribution of each individual scale item to the overall latent trait (Medvedev, Titkova, et al., 2018; Truong et al., 2021b). Such estimations are possible because most respondents score higher on easy items while only a few score high on more difficult items (Tennant & Conaghan, 2007). Therefore, Rasch analysis can eliminate the biases caused by arbitrarily assigning categorical options to individual items, which are then used to compute total scores, as such scores are not mathematically equal to the total (Stucki et al., 1996).

The Rasch model also involves invariance testing, meaning it checks for Differential Item Functioning (DIF) as a function of personal factors (e.g., age, sex). DIF is useful to test whether a scale will fit equally for every person regardless of individual factors. If DIF is detected, it indicates that the personal characteristics of a respondent is influencing their responses, meaning that the scale items do not work equally well across different groups (Hagquist & Andrich, 2004). Moreover, the results of Rasch analysis can be presented graphically as an item-person threshold distribution that illustrates how the range of a person's ability is covered by the range of item difficulties, which can be useful for detecting potentially significant ceiling or floor effects (Medvedev & Krägeloh, 2022). In additions, when the sample data is fitted to the Rasch model, the raw scores of an ordinal instrument can be

transformed into interval-level data (Linacre, 2004; Norquist et al., 2004). Such interval-transformed data reflect changes on a latent trait the same way as any other interval level scale (e.g., Celsius temperature). Increased measurement precision from ordinal-to-interval Rasch transformation has been demonstrated in several studies using different clinical measures (Medvedev, Berk, et al., 2020; Merkin et al., 2020; Norquist et al., 2004). This is the major benefit of the Rasch model over classical test theory methods, which are unable to produce a genuine interval scale.

This methodology has been recently applied to investigate psychometric properties of the IQCODE-16 in a study where ordinal-to-interval transformation algorithms were proposed that enhanced precision up to that of an interval-level scale (Truong et al., 2021b). However, this study had limited generalizability as the sample was selectively comprised of participants from a small area of Sydney, Australia. Moreover, while the study had the novel feature of testing the DIF according to the informants' age and sex, it has been more common in ageing research to examine the age and sex of the assessed participants (e.g., Tang and colleagues (Tang et al., 2004)). These limitations need to be addressed to ensure that transformed scores are generalizable across a wider population of older adults. Therefore, the purpose of the current study was to address these limitations by systematically replicating this earlier study within a larger, more diverse sample to extend the robustness and generalizability of these previous findings.

The first aim of the current study was to replicate Truong and colleagues' previous work applying Rasch methodology to IQCODE-16 scores from informants participant in the Sydney Memory and Ageing Study (Truong et al., 2021b), as replication studies are important to scientific inquiry and can further enhance the precision of the IQCODE-16. The current study will also allow us to investigate whether the recently developed Rasch model for the IQCODE-16 is generalizable across other samples of participants. Other aims were to

investigate invariance (i.e., potential DIF) across the different cohorts and participants and, if necessary, to convert the IQCODE-16 scores into interval-level data. Additionally, we aimed to examine the difference between the new and old conversion algorithms of the IQCODE-16.

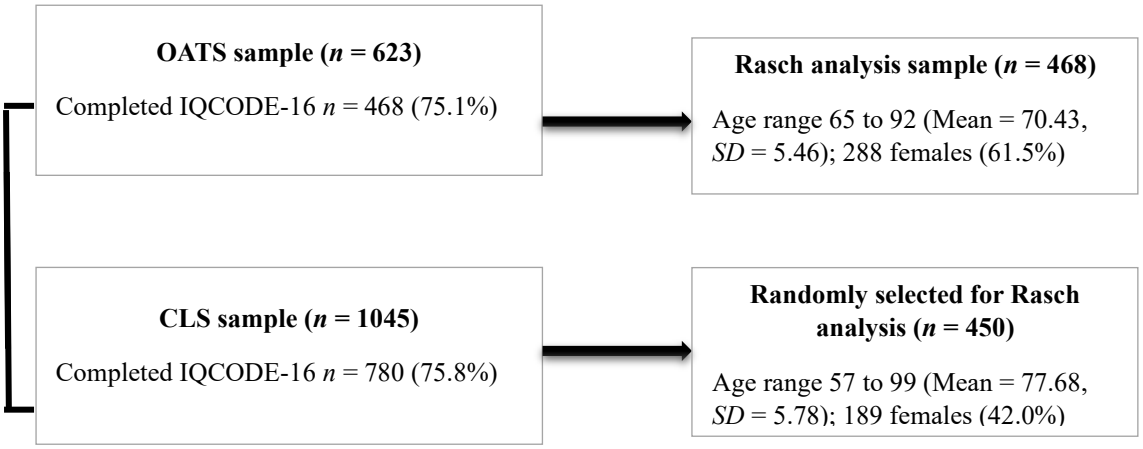
Method

Participants

Participants' scores for the current study were extracted from two longitudinal studies: the Older Australian Twins Study (OATS; Sachdev et al., 2009), and the Canberra Longitudinal Study (CLS; Christensen et al., 2004; Jorm et al., 2000). Both studies received ethical approvals permitted secondary analyses of their data from the University of New South Wales and the Australian National University. Figure 5.1 presents the consort diagram of how participants were selected for Rasch analyses. All participants and informants provided written consent for their participation; this study was completed according to the guidelines of the contributing studies' institutions, which are based on the internationally accepted ethical standards.

Figure 5.1

CONSORT diagram for participants selected for Rasch analysis of the IQCODE-16



The OATS sample included 623 participants (600 twins and 23 siblings) aged above 65 years at the wave 1/baseline assessment. Of these, 468 completed the IQCODE-16 and were thus included in the current study. The OATS sample was aged from 65 to 92 (*M* = 70.43, *SD* = 5.46), and comprised 288 females (62%) and 180 males (39%).

The original CLS sample included 1045 participants aged 57 or above at the wave 1/baseline assessment, of these 780 completed the IQCODE-16, and 450 were randomly drawn for Rasch analysis. This selection made the number of participants in this sample roughly similar to the OATS sample and satisfied the optimal sample size to minimize Type I and Type II errors of 250 to 500 for Rach analysis (Azizan et al., 2020; Hagell & Westergren, 2016). This selected CLS sample was aged 57 to 99 (*M* = 77.68, *SD* = 5.78), and comprised 189 females (42%) and 261 males (58%).

Moreover, we included the sample from the Sydney Memory and Ageing Study (MAS) that was used in Truong and colleagues' original study (Truong et al., 2021b). The baseline (wave 1) MAS sample included 1037 participants aged between 70 and 90. MAS participants were predominantly European (98%) and were recruited from the Eastern suburbs of Sydney, Australia between 2005 and 2007 (Sachdev et al., 2010). The MAS sample selected for Rasch analysis in Truong and colleagues' (Truong et al., 2021b) study comprised 400 participants.

We then randomly selected 150 participants from each of three original samples (OATS, CLS, and MAS) for invariance (DIF) testing across samples and to determine the overall model fit, as well as to produce a conversion table that is more representative of the general older Australian population. This random sample included participants aged 65 to 95 (*M* = 75.36, *SD* = 6.27), and comprised 254 females (56.4%) and 196 males (43.6%). Demographic details of participants in each sample used for the Rasch analysis were presented in Appendix 5.2. The

results of chi-square and ANOVA tests showed that there were significances differences between age and sexes among three Rasch samples (p's < .001).

It should be noted that the full CLS sample (n = 468) was used for Rasch analysis because this sample size satisfied requirements of the Rasch model to reduce Type I and Type II error, which is between 250 and 500 cases. To ensure valid comparisons with CLS sample and to satisfy Rasch model requirements, two comparable samples with n = 450 each were extracted using Simple Random Sampling (e.g., computer generated randomization). The first comparison sample was comprised of 450 OATS participant and the second comparison sample was comprised of 450 participants randomly selected from all three original datasets (i.e., OATS, CLS, and MAS).

Measure

Informant-observed CCs were measured using the IQCODE-16 (Jorm, 1994), which consists of 16 individual items that ask informants about how the participant's memory and cognitive function have changed (e.g., "Using his/her intelligence to understand what's going on and to reason things through"). Each item is scored on a 5-point Likert-scale with response options ranging from 1 = "much improved" to 5 = "much worse", where 3 = "no change".

Data Analyses

Descriptive statistics including mean, standard deviation (SD), Cronbach's alpha, and McDonald's omega for the IQCODE-16 across the four samples were computed using IBM SPSS v.27. Rasch analyses were conducted using the RUMM2030 software package (Andrich et al., 2009) replicating the analytical solutions in Truong and colleagues (Truong et al., 2021b), which were based on the standardized criteria for the Rasch model fit as recommended elsewhere (Leung et al., 2014; Tennant & Conaghan, 2007).

Replications involved the use of unrestricted Partial Credit models (Masters, 1982) and examined whether the Rasch model fit solution in Truong and colleagues' (Truong et al., 2021b) study was appropriate for all samples in the current study. The previous solution combined individual items into super-items to reduce measurement error and improve the Rasch model fit (Lundgren-Nilsson et al., 2013; Medvedev, Turner-Stokes, et al., 2018; Merkin et al., 2020; Truong et al., 2021b). Generally, the scale satisfies expectations of the Rasch model if it is not significantly different from characteristics of an interval measure. This is reflected by non-significant item-trait interactions, no misfitting items, no local dependency, no DIF, and evidence of unidimensionality (Leung et al., 2014; Tennant & Conaghan, 2007). Firstly, the overall Rasch model fit requires a non-significant chi square index of the estimate of item-trait interaction ($p > 0.05$) (Linacre, 2002). Secondly, no misfitting items can be identified, meaning fit residuals for individual items are in the range of ±2.50 (Lundgren-Nilsson & Tennant, 2011). Thirdly, there can be no local dependency detected when observing the residual correlations between individual items, meaning values are below 0.20 (Christensen et al., 2013). Fourthly, no DIF due to personal factors (e.g., age, sex, sample) can be detected, which suggests the scale items work equally well across different groups of people (Hagquist & Andrich, 2004). Lastly, the achievement of unidimensionality is evidenced by a non-significant principal components analysis (PCA) of the residuals and the equating *t*-test (Leung et al., 2014). In addition, the reliability coefficient PSI used in Rasch analysis is not a criterion of the Rasch model fit but it reflects how well the scale discriminates between individuals with different levels of the latent trait (e.g., CC severity). PSI ranging from 0.70 to 0.80 indicate acceptable reliability and a *PSI* above 0.80 indicates good to excellent reliability (Andrich et al., 2009; Medvedev, Turner-Stokes, et al., 2018).

The best model fit for this study involved a solution that worked equally well across all included samples. When the best Rasch model fit was achieved, the person-item thresholds

distribution for each sample was evaluated to examine how well items thresholds of the IQCODE-16 cover the sample's CC levels. A transformation table was then generated to transform the IQCODE-16 raw scores into interval transformed scores to increase the precision of assessment.

Results

The IQCODE-16 showed excellent internal consistency, with both Cronbach's alpha and McDonald's omega ranging from 0.93 to 0.96, which is consistent with previous reports (Harrison et al., 2015; Phung et al., 2015; Truong et al., 2021a, 2021b). Table 5.1 presents estimates of item-fit residuals for the initial Rasch analyses for each sample. As can be seen, most items with significant misfit to the Rasch model in the MAS sample from the Truong et al.'s study (Truong et al., 2021b) (i.e., item 3, 10, 12, and 15) were consistently found to be misfitting in all samples in the present analyses, the exception being item 15, which was not misfitting to the random sample. Besides that, other items misfit to the Rasch model were observed across the OATS, CLS and random samples. These misfitting items were 2, 5, 9, 13, and 16, though, only items 1 and 8 were only significantly misfit to the OATS sample. Table 5.2 displays the overall model fit estimates of the Rasch analyses for each IQCODE-16 items across all samples. All initial analyses for the OATS, CLS and random samples (A1's) resulted in good to excellent reliability with *PSIs* ranging from 0.84 to 0.90. However, the overall fit to the Rasch model was not satisfactory as indicated by significant chi-square indexes (p's $<$ 0.004), which reflects deviation of the scale from the Rasch model expectations across samples.

Table 5.1

Rasch model fit statistics of item-fit residuals for the initial analyses (A1's) of the IQCODE-16 individual items across four samples

Items	Item-fit residuals			
	MAS‡	**OATS**	**CLS**	**Random**
1. Remembering things about family and friends (e.g. occupations, birthdays, addresses)	-1.66	-2.64*	-1.10	-2.61*
2. Remembering things that have happened recently	-2.52	-4.14*	-3.39*	-3.81*
3. Recalling conversations a few days later	-4.09*	-3.56*	-2.54*	-3.63*
4. Remembering his/her address and telephone number	0.48	-0.85	-1.30	-1.94
5. Remembering what day and month it is	-2.15	-2.54*	-3.32*	-2.74*
6. Remembering where things are usually kept	-1.45	-0.43	-2.11	-2.46
7. Remembering where to find things which have been put in a different place from usual	-2.23	-0.92	-2.02	-1.69
8. Knowing how to work familiar machines around the house	-2.07	-4.42*	-2.12	-2.32
9. Learning to use a new gadget or machine around the house	-2.35	-5.19*	-3.98*	-4.08*
10. Learning new things in general	-4.05*	-6.16*	-4.67*	-5.42*
11. Following a story in a book or on TV	-2.32	-2.04	-1.49	-2.31
12. Making decisions on everyday matters	-3.67*	-3.93*	-6.73*	-4.08*
13. Handling money for shopping	-0.51	-4.03*	-3.87*	-3.74*
14. Handling financial matters e.g. the pension, dealing with the bank	-1.84	-4.06*	-4.19*	-3.42*
15. Handling other everyday arithmetic problems (e.g. knowing how much food to buy, knowing how long between visits from family or friends)	-2.84*	-3.97*	-5.22*	-1.95
16. Using his/her intelligence to understand what's going on and to reason things through	-1.71	-5.78*	-4.35*	-3.16*

Note: ‡ Results for MAS was reproduced with permission from Truong and colleagues (Truong et al., 2021b) Table 2; *Significant misfit to the Rasch model

Table 5.2

Summary of fit statistics for the Rasch analyses of the IQCODE-16 across samples

Sample	Analysis	Person mean		Goodness of fit		PSI	Significant *t*-tests (Unidimensionality)	
		Value	SD	$\chi 2$	*p*		%	Lower bound
MAS	A1[§]	1.41	2.26	92.23	0.16	0.92	5.5	3.4 (Acceptable)
	A2[§]	1.26	2.11	81.32	0.38	0.92	4.3	2.1 (Strict)
	A3[¶]	1.14	1.93	46.66	0.40	0.92	6.3	4.2 (Acceptable)
OATS	A1	0.02	1.58	191.51	<0.001	0.84	5.6	3.6 (Acceptable)
	A2	-0.08	1.52	93.13	<0.001	0.83	11.0	9.0 (Not acceptable)
	A3	-0.05	1.63	68.42	0.50	0.83	6.5	4.5 (Acceptable)
CLS	A1	0.76	1.69	138.94	0.003	0.90	4.9	2.8 (Strict)
	A2	1.04	1.82	30.99	0.06	0.91	8.1	6.0 (Not acceptable)
	A3	0.70	1.78	22.71	0.20	0.91	6.9	4.8 (Acceptable)
Random	A1	0.39	1.50	58.10	0.003	0.86	11.3	9.3 (Not acceptable)
	A2	0.30	1.65	92.33	0.16	0.85	8.1	6.1 (Not acceptable)
	A3	0.30	1.48	19.48	0.36	0.85	3.9	1.9 (Strict)

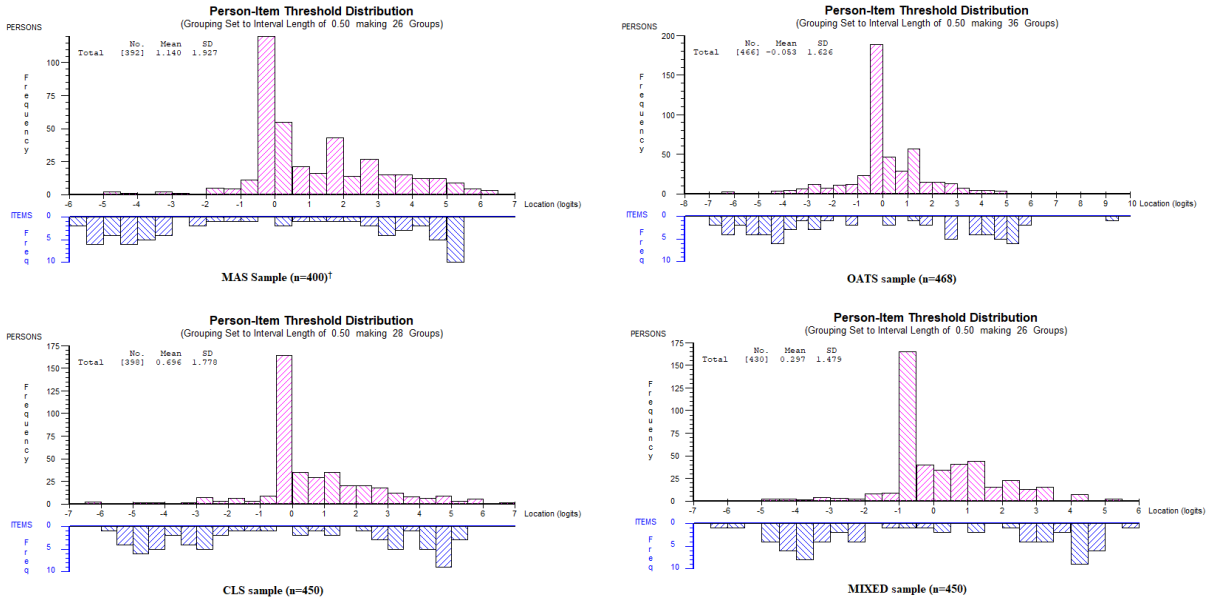
Note. [§] Results for MAS was reproduced with permission from Truong and colleagues (Truong et al., 2021b) Table 1; [¶] Analysis used MAS data with permission from Truong and colleagues (Truong et al., 2021b); A1: initial analysis; A2: replicating analysis; A3: Rasch model fit analysis

The overall model fit estimates using the Rasch model fit solution originally suggested by Truong and colleagues (Truong et al., 2021b) across four samples are also presented in Table 5.2 (analyses A2). This solution included four super-items: super-item 1 (Items 2, 3, 13, & 14); super-item 2 (1 & 9); super-item 3 (6 & 7); and super-item 4 (10 & 11). These analyses demonstrated good to excellent reliability with PSIs ranging from 0.83 to 0.91 for the IQCODE-16, with an acceptable Rasch model fit as evident by non-significant chi-square indexes ($p > 0.05$) for the CLS and random samples. However, the OATS sample did not achieve the Rasch model fit with this solution (chi-square index $p < 0.001$). The residual correlation matrix in the analysis for the OATS sample also showed local dependency between super-item 2 and super-item 4, and between super-item 2 and regular item 8 as reflected by significant residual correlations of 0.27 and 0.31, respectively.

To address these issues, minor modifications were made that involved combining locally dependent items in the OATS sample into three super-items: super-item 1 (Items 2, 3, 13, &14); super-item 2 (1 & 9); and super-item 3 (6, 7, 10 & 11). This new Rasch model provided not only a better fit for the OATS sample, but also for the CLS and random samples, and the MAS sample used in Truong and colleagues' (Truong et al., 2021b) study (Table 5.1, analyses A3). This solution resolved item misfit and local dependency while achieving acceptable or strict unidimensionality, good reliability, and invariance across person factors such as participants' age and sex and the study sample. Figure 5.2 presents person-item threshold distributions from the analyses of the best fitting model for the samples in the current study as well as the MAS sample used in Truong and colleagues' (Truong et al., 2021b) study. Interestingly, they show that thresholds of the IQCODE scale satisfactorily cover CC levels with no significant ceiling or floor effects in each sample.

Figure 5.2

Person-item threshold distribution of the model fit analyses of the TICS-M across four samples



Note. † Analysis used MAS data with permission from Truong and colleagues (21)

We generated a new conversion algorithm to transform raw IQCODE-16 scores into interval data using the best model fit analysis allowing us to enhance precision of measurement. Table 5.3 presents a Rasch ordinal-to-interval conversion table which was developed based on person estimates for the IQCODE-16. Pearson correlation between conversion scores in the current study and Truong and colleagues' (Truong et al., 2021b) study indicated that the two sets of Rasch conversion scores were strongly correlated (r = 0.998).

Table 5.3

Converting ordinal scores into interval level scores for the IQCODE-16

Raw scores	Logits	Interval scores	Raw scores	Logits	Interval scores
1.00	-6.02	1.00	3.06	-0.10	2.79
1.06	-5.57	1.14	3.13	0.33	2.92
1.13	-5.28	1.22	3.19	0.73	3.04
1.19	-5.10	1.28	3.25	1.11	3.16
1.25	-4.96	1.32	3.31	1.46	3.26
1.31	-4.84	1.35	3.38	1.78	3.36
1.38	-4.74	1.39	3.44	2.07	3.45
1.44	-4.65	1.41	3.50	2.33	3.53
1.50	-4.56	1.44	3.56	2.56	3.60
1.56	-4.48	1.47	3.63	2.77	3.66
1.63	-4.39	1.49	3.69	2.97	3.72
1.69	-4.31	1.52	3.75	3.15	3.78
1.75	-4.23	1.54	3.81	3.32	3.83
1.81	-4.15	1.56	3.88	3.47	3.87
1.88	-4.07	1.59	3.94	3.61	3.92
1.94	-3.99	1.61	4.00	3.75	3.96
2.00	-3.90	1.64	4.06	3.87	3.99
2.06	-3.81	1.67	4.13	3.99	4.03
2.13	-3.71	1.70	4.19	4.11	4.06
2.19	-3.60	1.73	4.25	4.22	4.10
2.25	-3.48	1.77	4.31	4.32	4.13
2.31	-3.35	1.81	4.38	4.43	4.16
2.38	-3.21	1.85	4.44	4.54	4.19
2.44	-3.06	1.90	4.50	4.65	4.23
2.50	-2.89	1.95	4.56	4.77	4.27
2.56	-2.71	2.00	4.63	4.90	4.31
2.63	-2.50	2.06	4.69	5.06	4.35
2.69	-2.27	2.14	4.75	5.24	4.41
2.75	-2.00	2.22	4.81	5.48	4.48
2.81	-1.69	2.31	4.88	5.81	4.58
2.88	-1.33	2.42	4.94	6.34	4.74
2.94	-0.94	2.54	5.00	7.20	5.00
3.00	-0.52	2.66			

Discussion

This study applied Rasch analysis to investigate psychometric properties of a widely used informant-reported CC measure, the IQCODE-16, and re-evaluated the generalizability of the Rasch conversion algorithms proposed in a recent study by Truong and colleagues (Truong et al., 2021b). Good reliability of the IQCODE-16 was consistently obtained across analyses in this study and adds further empirical evidence supporting the robust psychometric properties of the scale. Interestingly, the Rasch solution suggested by Truong and colleagues (Truong et al., 2021b) did not work equally well for all samples in the current study, potentially due to the previous study's biased sample. Therefore, an alternative solution was found in the current study, which involved reorganizing super-items of the IQCODE-16 to achieve the best Rasch model fit across all samples. This allowed us to generate ordinal-to-interval conversion algorithms to enhance the accuracy of the IQCODE-16 across more diverse samples with higher generalizability of assessment scores.

Evidence shows that the self-reported CCs can be significantly affected by individual person factors such as mood, personality, life events, and medications, which can also impact on informants' reports. However, to the best of our knowledge, no interval-transformed algorithms have been previously established for self-reported CC assessments (e.g., MAC-Q). This is important as studies have shown that interval-transformed data can more accurately reflect real clinical changes (e.g., CC levels) compared to ordinal scales (Hobart & Cano, 2009; Merkin et al., 2020; Truong et al., 2021b). Moreover, using interval-transformed data can also decrease measurement errors associated with ordinal scale scores (Medvedev, Berk, et al., 2020). Therefore, the ordinal-to-interval transformation algorithms developed by this study contribute to the higher precision of the IQCODE-16 by lessening the influences of personal and affective factors on the informant's observation.

It should be noted that the Rasch solution fit in Truong and colleagues' (Truong et al., 2021b) study was not replicated in this study, possibly because we used different DIF factors in the current study. Also, the differences between the original conversion algorithm generated by Truong and colleagues (Truong et al., 2021b) and the one generated in the current study are marginal, meaning both conversion tables can be used for ordinal-to-interval transformations of the data. However, the current conversion table has a higher degree of generalizability across different samples and is more robust to personal factors such as participants' age and sex. Moreover, the Rasch solution used in this study also worked well with the MAS sample used in Truong and colleagues' (Truong et al., 2021b) study, even with different personal factors (i.e., informant age and sex), suggesting the conversion table produced in the current study is a superior option.

Each individual item of the original IQCODE-16 varies by its degree of difficulty and hence, contributes uniquely to the raw score, which should be considered when computing the overall score (Stucki et al., 1996). Traditional methods (e.g., classical test theory or generalizability theory) do not account for unequal contribution of items to the total score, while our interval-transformed scores generated by Rasch analyses can precisely estimate the unique contribution of each item to the overall assessment score, therefore contributing to higher accuracy of assessment. Studies suggest that Rasch interval-transformed data are more likely to reflect the CC levels of an individual accurately (Leung, 2011; Truong et al., 2021b). As such, using Rasch interval-transformed data is crucial because it decreases measurement error associated with raw ordinal scores (Medvedev, Siegert, Kersten, et al., 2017). Besides that, such interval level data is also suitable for conducting parametric statistics or statistical comparisons against other interval measurements (e.g., biomarkers, electrophysiological and neuroimageing data) and may increase reliability and validity of the results because using

interval data can avoid the violations of arithmetic assumptions inherent to ordinal data (Leung, 2011).

The main strength of this study was the application of Rasch analysis methodology to the IQCODE-16 across several cohorts using an appropriate sample size that allowed us to minimize both Type I and Type II errors. Type I error occurs due to inflated chi square statistics in RUMM2030 if the sample size exceeds 500 cases while Type 2 error is common in smaller samples below 250 cases, as this limited robustness of items calibration (Azizan et al., 2020; Hagell & Westergren, 2016). In addition, this study is novel as, to date, there have been no replication studies using Rasch analysis to re-evaluate the psychometric properties of the 16-item version of the IQCODE. Moreover, this study found that the modified 16-item IQCODE works equally well across all samples and personal factors of the participants (i.e., participants' age and gender) which were not investigated in the previous study.

However, there are limitations which should be acknowledged. Although we considered older adults across three cohort studies, they were all Australian studies and may therefore not be representative of older adults in other regions, especially those in lower income countries. Therefore, this study should be replicated in different samples with different personal factors to investigate potential DIFs. For example, older adults from non-English speaking countries, or in low- and middle-income countries. It should be also noted that this study inclusively focused on reliability and internal validity because external validity of the IQCODE-16 is well established by other studies (Park, 2017; Perroco et al., 2008; Truong et al., 2021a).

In conclusion, the findings of this study demonstrated the reliability and internal validity of the informant-reported IQCODE-16 measure of CCs across older Australians. Our modification of the IQCODE-16, made using Rasch analyses, resolved local dependency issues

and established the scale invariance across different samples and personal factors. This allowed us to generate transformation tables to convert raw ordinal scores into interval-level data, which improves the precision of measurement. The interval-transformation table is more robust compared to a corresponding table generated in a previous study by our group. Clinicians and researchers can employ the IQCODE-16 in a variety of contexts with higher precision by using the conversion table published here, without needing any modification to the original IQCODE-16 administration format.

Chapter 6 Enhancing precision of the Telephone Interview for Cognitive Status-modified (TICS-M) using Rasch Model

Truong, Q. C., Choo, C., Numbers, K., Bentvelzen, A., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N.A.[3] & Medvedev, O. N. (2023). Enhancing precision of the Telephone Interview for Cognitive Status-modified (TICS-M) using the Rasch Model. *Psychological Assessment*. https://doi.org/10.1111/eci.13681

Declaration

This study is co-authored work with C. Choo, K. Numbers, A. Bentvelzen, A. G. Merkin, H. Brodaty, N. A. Kochan, P. S. Sachdev, V. L. Feigin, and O. N. Medvedev (see Appendix P5 for the original article). I was the first author and had the major contribution to this study that included designing and conducting the study, completing data analysis, writing the manuscript for submission to the *Psychological Assessment*, and responding to reviews/addressing their comments during peer-review process. All co-authors have signed the co-authorship form (see Appendix 6.1).

Abstract

The Telephone Interview for Cognitive Status-modified (TICS-M) is a well-established and widely used screening instrument for dementia and assessment of global cognitive function in older people. This study aimed to evaluate the psychometric properties of the TICS-M and to enhance the accuracy of the instrument using Rasch methodology. Partial Credit Rasch model was applied to the TICS-M scores. The sample selected for Rasch analysis consisted of 432 participants aged 70 to 90 years (M = 78.85, SD = 4.73) including 195 males (237 females), and 132 (30.56%) of whom were diagnosed with dementia after the baseline assessment. Initial analysis indicated good reliability of the TICS-M assessment scores, but there were three misfitting items and local dependency issues. Combining locally dependent and misfitting items into super-items achieved the best Rasch model fit for the TICS-M. This modification improved reliability of the assessment scores and resulted in no misfitting items, no local dependency, strict unidimensionality, and invariance across individual factors such as participants' age, sex, diagnosis and in-person neuropsychological assessment scores. Satisfying Rasch model expectations allowed for creation of a transformation table to convert raw TICS-M scores into interval-level data, which improves precision of the instrument. In summary, the TICS-M assessment scores demonstrated excellent reliability as reflected by Person Separation Index (PSI = 0.86) and met expectations of the unidimensional Rasch model after minor adjustments. The ordinal-to-interval transformation table can be used to increase accuracy of the TICS-M without altering its current format. These findings contribute to more accurate assessments of cognitive decline in older people and screening for conditions such as dementia.

Keywords**:** Cognitive Status, Measurement, Telephone Interview for Cognitive Status-modified, Rasch Analysis, Reliability

Introduction

Globally, the prevalence of dementia is about 1 in 10 adults over age of 65 (Centers for Disease Control and Prevention, 2019). An objective clinical assessment is essential to distinguish between normal cognition and cognitive impairment (Chin et al., 2019; Sperling et al., 2011). Brief mental status examinations, such as the Mini-Mental State Exam (MMSE; Folstein et al., 1975), or the Telephone Interview for Cognitive Status (TICS; Brandt et al., 1988), are used to assess global cognitive function and screen for dementia. The TICS and its modified versions are well-validated screeners for global cognitive function over time (Brandt et al., 1988; Fong et al., 2009). However, the TICS subscales may not be considered valid tools to assess individual cognitive domains. The TICSs also offer an alternative to face-to-face assessments and has demonstrated strong correlations with the widely used and validated MMSE, therefore making it preferable for large epidemiological studies due to its time and cost efficiency (Fong et al., 2009; Wilson & Bennett, 2005).

The modified Telephone Interview for Cognitive Status (TICS-M; Brandt et al., 1993) is a well-established screener of global cognitive functioning and includes 13 assessment items compared to 11 items in the original TICS. Both the TICS and TICS-M have demonstrated high internal consistency with Cronbach alphas of above 0.80, and high test-retest reliability for assessment scores (Castanho et al., 2016; Plassman et al., 1994; Seo et al., 2011). Both TICS versions have been validated against other cognitive screeners such as the MMSE (de Jager et al., 2003). In addition, studies indicated that there are correlations between the TICS total score and tests of different cognitive domains captured by a battery of standardized neuropsychological tests (Crooks et al., 2006; Van Den Berg et al., 2012) such as the Logical Memory test from the Wechsler Memory Scale-III (Wechsler, 1997), the California Verbal Learning Test (Delis et al., 1987), the 30-item Boston Naming Test (Kaplan et al., 2001), the Animal Naming test (Goodglass & Kaplan, 1983), the Trail Making Test B (US Army, 1944)

and the Controlled Oral Word Association tests (Spreen, 1977). The TICS-M also demonstrated high sensitivity and high specificity in detecting individuals with dementia from the longitudinal Sydney Memory and Ageing Study (MAS) (Bentvelzen et al., 2019) using a cut-off score of 21. Using the same sample, a recent study indicated that the TICS-M had higher sensitivity compared to the MMSE for detecting dementia (Truong et al., 2022).

Even though, the TICS-M has been well-validated, such validation studies were based merely on the Classical Test Theory (e.g., Brandt et al., 1993) or Generalisability Theory (e.g., Truong et al., 2022) methodologies, which are usually sample dependent (Cohen et al., 1996). The raw TICS-M scores have demonstrated good validity in clinical settings but the accuracy of these scores might be a concern as they constitute an ordinal measure. In an ordinal scale such as the TICS-M, the differences between ordinal response options of individual items (e.g., 0 and 1 versus 1 and 2) may not reflect the same clinical changes (Tennant & Conaghan, 2007). Moreover, each individual TICS-M item contributes differently to the total scale score, which is not accounted for in the ordinal measurement and needs to be adjusted using psychometric methods suitable for such modifications (Hobart & Cano, 2009).

To better explain the limitations of ordinal scales, Sandham et al. (2019) used different amount of Vitamin C contained in different fruits such as kiwifruits, bananas, and strawberries, as an example of a latent trait (e.g., global cognition) reflected by different scale items to a different extent. As the amounts of Vitamin C vary by different fruits, they contribute different amounts of Vitamin C to the juice squeezed out of these fruits. For example, a lemon would contain more Vitamin C compared to an apple, while an apple would have more Vitamin C compared to a banana. In the same way, different items reflect cognitive ability to different extent, which refers to an item difficulty or location in Rasch analysis. By estimating a unique contribution of Vitamin C by each fruit we can accurately measure its overall content. Another example is mathematical operators such as adding and subtracting represent less difficulty (i.e.,

require less mathematical abilities) compared to determining a square root or estimating a log reflecting higher abilities. In the same way, some items contribute more to the overall construct operationalized by the scale (e.g., recalling a list of words), while other items contribute only minor amounts (e.g., reporting one's own age). For these reasons, psychometric properties of the TICS-M should be thoroughly examined with a view to enhance the accuracy of the screener up to an interval-level scale by employing an appropriate and robust statistical method such as Rasch analysis (Rasch, 1960; Tennant & Conaghan, 2007).

Rasch analysis is an advanced statistical method used to investigate the reliability of assessment scores and the internal validity of psychometric tools (Rasch, 1960, 1961). This approach also evaluates specific psychometric properties of scales, such as functioning of individual items. A Rasch model is unidimensional and postulates that a response to a specific item of an assessment tool is determined by both an item difficulty and an individual ability (Rasch, 1960, 1961). As such, Rasch method has become a gold standard in examining reliability of clinical assessment scores and validity of measures (Hobart & Cano, 2009; Lundgren-Nilsson & Tennant, 2011).

Common methods for determining reliability of assessment scores and internal validity of a measure such as Classical Test Theory and Generalisability Theory (Cronbach et al., 1963) cannot discriminate between item difficulty based on individual ability (Fox & Jones, 1998). Normally, the distances between subsequent categories of psychometric scales are treated as having the same values and total scores are computed using numerical values assigned to categories. Such total scores are naturally biased because each individual item contributes to a different degree to the overall latent trait (i.e., cognitive ability; (Rasch, 1960). Rasch analysis reduces such bias by estimating precise thresholds for each response option of each individual item while accounting for unique contribution of each item to the overall assessment score (Stucki et al., 1996; Tennant & Conaghan, 2007). This method can also be used to create a

graphical distribution of item-person thresholds for visual examination of a scale coverage and identifying ceiling and floor effects.

Moreover, the Rasch model also examines invariance of a scale by testing for Differential Item Functioning (DIF) across individual factors such as age, sex, diagnosis, and cognitive performance. DIF checks are conducted to identify whether there is a circumstance where participants from different groups (e.g., male vs female), with the same ability on the latent trait, respond differently to an assessment item. When DIF occurs, it means that the scale and its items do not work equally well for people in different groups because their personal characteristics (e.g., diagnosis, sex, age) can influence their responses (Hagquist & Andrich, 2004). However, to the best of our knowledge, there have been no rigorous invariance testing of the TICS-M reported that used a rigorous appropriate methodology. Besides that, fitting data to the Rasch model allows for the transformation of raw scores into interval-level data, which improves the precision of measurement without altering the original response format of a scale (Norquist et al., 2004).

Rasch analysis precisely estimates the difficulty of each item in a scale using the universal log-odd metrics and tests the appropriateness of response options used in polytomous items. This permits transformation of ordinal scale scores into a genuine interval measure if scale parameters comply with parameters of an interval scale defined by the Rasch model (Hobart & Cano, 2009; Rasch, 1960; Tennant & Conaghan, 2007; Wilson, 2004; Wright & Stone, 1979). The interval-transformed data is more accurate compared to the ordinal raw scores because it reflects changes on a latent trait using the same measurement unit across continuum of a scale similar to physical measures (e.g., temperature or height) (Linacre, 2004; Merbitz et al., 1989; Norquist et al., 2004). An interval-transformation table has been recommended as an essential standard for studies utilizing Rasch methodology (Leung et al., 2014). The benefits of Rasch transformation have been also demonstrated empirically in many

studies across disciplines using different populations and clinical assessment tools (Leung et al., 2014; Merkin et al., 2020; Norquist et al., 2004).

To date, Rasch analysis has not been conducted on the TICS-M even though this could potentially enhance its precision up to an interval-level scale that permits parametric statistical analyses without violation of the fundamental mathematical and statistical assumptions (Jamieson, 2004; Merbitz et al., 1989), also allowing for more accurate comparisons with other interval measures (e.g., neuroimaging data; (Hobart & Cano, 2009). The aim of the current study was to employ Rasch analysis to investigate and enhance the psychometric properties of the TICS-M. If the data would fit to the Rasch model, a secondary aim was to generate a table that can be used to transform TICS-M raw scores into interval-level data to improve the accuracy of measurement.

Method

Participants

Participants were drawn from the MAS dataset consisting of 1037 community-dwelling older adults aged 70 – 90 years old, without dementia at baseline, living in the Eastern suburbs of Sydney, Australia (Sachdev et al., 2010). The vast majority of MAS participants were European (98%), with the remaining sample identifying as Asian (1%) or other/not revealed (1%). Participants were interviewed biennially from wave 1 to wave 6 (across 10-years of follow-up). The MAS was approved by the University of New South Wales Human Research Ethics Committee (HC 05037, 09382, 14327) and all participants provided written consent to participate in the study prior to data collection.

Figure 6.1 presents the consort diagram of how participants were selected for Rasch analysis. To control for DIF, we created three groups of participants representing three different levels of global cognition based on the distribution of neuropsychological assessment *z*-scores (*n*=300). We randomly selected 100 cases for each of these groups from wave 1 as follows: group 1 with *z*-scores ≤-0.41 (33.8%), group 2 with *z*-scores ranging from -0.42 to 0.58 (34.3%), and group 3 with *z*-scores ≥ 0.59 (32.3%). We also included 132 independent participants excluded from randomization who were diagnosed with dementia at follow-up waves (specifically, 19 from wave 2, 34 from wave 3, 25 from wave 4, 16 from wave 5, and 38 from wave 6). The final sample included 432 participants to satisfy the sample size requirement for Rasch analysis. Previous studies suggested that a sample size between 250 and 500 cases is the most appropriate for the unidimensional Rasch model because it minimizes both Type I and Type II errors (Hagell & Westergren, 2016). The results of an independent samples *t*-test indicated that averaged neuropsychological assessment *z*-scores at wave 1 for participants who were diagnosed with dementia from follow-up waves in the selected sample were significantly lower (*z* = -0.80) in comparison to participants who were not diagnosed with

dementia (z = -0.00), $t(429) = 6.81$, $p < 0.01$. Participants in the final Rasch sample were 195 males (45.1%) and 237 females (54.9%) with a mean age of 78.85 years (SD = 4.73).

Figure 6.1

CONSORT flow diagram of participants selected for Rasch analysis of the TICS-M

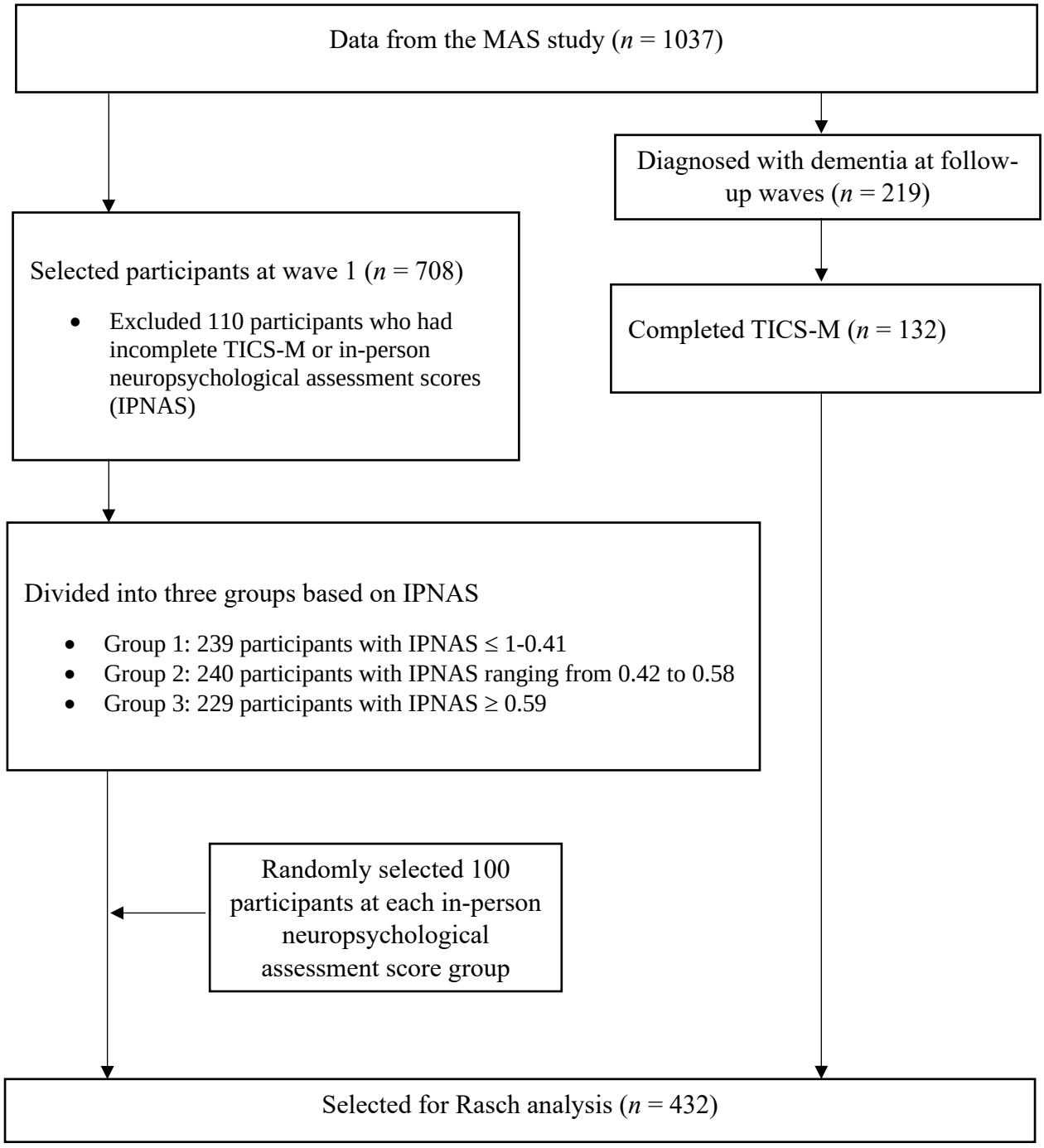


Table 6.1 displays demographic details of participants from the original sample and sample selected for Rasch analyses. There were no significant demographic differences between the original sample and those selected for Rash analysis by age, sex, and education (*p's* > 0.05). However, the results of *t*-tests showed that the original sample had significantly higher TICS-M scores (M = 23.79, SD = 3.87) compared to the Rasch sample (M = 21.77, SD = 5.82), $t(613.01) = 6.42$, $p < 0.01$. This significant difference was expected because the Rasch sample included participants with dementia. Appendix 6.2 presents demographic details of participants who were diagnosed with dementia at follow-up waves and those who were not diagnosed with dementia in the Rasch sample. There were no significant differences between these subsamples on any demographic variables including age, sex and education. However, it is noticeable that participants who diagnosed with dementia had significantly lower TICS-M scores compared to not dementia participants with a large effect size (Cohen's d = 1.68) in the Rasch sample. This significance was expected as this Rasch sample was intentionally chosen for the DIF testing across the individual factor of diagnosis.

Table 6.1

Demographic details of participants from the original and Rasch samples including statistical comparisons using Chi-square (p-values) for categorical variables and t-tests (p-values and Cohen's d if significant) for continuous variables

Demographic details		Original sample (n = 1037)	Rasch sample (n = 432)	p
Age	Mean (SD)	78.84 (4.82)	78.85 (4.73)	0.98
Sex				0.92
	Male	465	195	
	Female	572	237	
Education				0.85
	Primary school	26	13	
	Incomplete high school	411	163	
	Completed high school	142	65	
	Incomplete tertiary	35	17	
	Complete tertiary	311	122	
	Incomplete high school +certificate/diploma	57	30	
	Complete high school +certificate/diploma	55	22	
TICS-M scores	Mean (SD)	23.79 (3.87)	21.77 (5.82)	< 0.001 (d = 0.41)
Diagnosis				< 0.001
	Dementia	0	132	
	MCI	446	128	
	Healthy	591	172	

Measure

The TICS-M (Brandt et al., 1993) is a telephone administered screening tool that is comprised of 13 items. Individual items are scored differently based on their content. For example, item 4 ("Listen to a list of ten words carefully and then recall as many words as can in any order") has a score ranging from 0 to 10, while item 2 ("Can you tell me your age?") has a score of 0 or 1. Scores across all 13 items are summed and range from 0 – 39, with higher scores indicating better cognitive ability. A full list of individual items and their maximum score can be found in Table 6.3.

Dementia Diagnosis

Consensus dementia diagnoses were completed for all waves. At wave 1, and at each follow-up, there was a consensus review meeting for each participant where at least three clinicians (i.e., neuropsychiatrists, psychogeriatricians, and neuropsychologists) discussed all available clinical data, which included neuropsychological tests, informant reports, and neuroimaging data (where available) to reach a consensus diagnosis. Participants were diagnosed with dementia based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (American Psychiatric Association, 1994). Participants who did not meet these criteria were classified as "not dementia" at each wave. No participants had a diagnosis of dementia at wave 1 as this was an exclusion criteria for MAS participants at baseline (Sachdev et al., 2010).

In-person Neuropsychological Assessment

In-person neuropsychological assessments were conducted at wave 1 using a comprehensive test battery that assesses the major cognitive domains of attention/processing speed, language, executive function, visuospatial ability, and memory (Sachdev et al., 2010). Ten neuropsychological tests were included according to the principal cognitive function they are meant to represent according to consensus, acknowledging that neuropsychological

measures are multifactorial and there is no complete consensus (Lezak et al., 2004; Strauss et al., 2006). Attention/processing speed was measured using the Wechsler Adult Intelligence Scale-III (WAIS-III), Digit Symbol-Coding (Wechsler, 2003) and Trail Making Test A (Tombaugh, 2004). Language was measured using the 30-item Boston Naming Test (Kaplan et al., 2001) and Animal Fluency (Spreen, 1977). Executive function was measured using the FAS Fluency (Benton, 1967) and Trail Making Test B (Reitan & Wolfson, 1993). Visuo-spatial ability was measured using the Wechsler Adult Intelligence Scale-Revised (WAIS-R) Block Design (Wechsler, 1981b). Memory was measured using the Wechsler Memory Scale-III (WMS-III), Logical Memory Story A delayed recall (Wechsler, 1997), Rey Auditory Verbal Learning Test (Strauss et al., 2006), total learning, short-term and long-term recall scores, and the Benton Visual Retention Test (Benton et al., 1996).

To ensure that the results of statistical analyses were not affected by standardization of variables, the raw domain scores were standardized using participants who were classified as healthy at wave 1. This was an important consideration because not all MAS participants were cognitively healthy at baseline. Therefore, domain and in-person neuropsychological assessment scores were computed as quasi *z*-scores accounting for age as follows. First, raw test scores were converted to *z*-scores using baseline means and standard deviations (*SDs*) of a reference group comprised of 723 MAS participants classified as cognitively healthy at baseline. Second, composite domain scores were formed by averaging z-scores of the component tests, apart from the visuospatial domain which was represented by a single test. Each domain composite was standardised by transforming it so that the mean and *SD* of baseline for cognitively healthy group were 0 and 1, respectively. Finally, in-person neuropsychological assessment scores were computed by averaging z-scores of domains, and again transforming these scores so that means and SDs for baseline reference group were 0 and 1, respectively. The higher *z*-scores represented a better cognitive performance.

Table 6.2 includes means, standard deviations (SDs), or median of MAS participants' neuropsychological component test scores and their normative data compared to population norms. Both the full MAS sample and Rasch sample means and SD or median for component test raw scores were mostly in the range of normative data found in the literature. Specifically, the Trail Making Test A and B (Tombaugh, 2004), 30-item Boston Naming Test (Fastenau et al., 1998), FAS and Animal Fluency (Tombaugh et al., 1999), Wechsler Memory Scale-III Logical Memory Story A (Smith et al., 1997), Benton Visual Retention Test (Lechevallier-Michel et al., 2004), and Rey Auditory Verbal Learning Test (Magalhães & Hamdan, 2010), were in the normative range. The exceptions were the Wechsler Adult Intelligence Scale-III Digit Symbol-Coding, which was out of the upper range of normative data (Ryan et al., 2020) and Wechsler Adult Intelligence Scale-Revised Block Design, which was marginally below its lower normative range (Scott et al., 2021) for the MAS sample. There were no statistically significant differences on these assessment scores between the full MAS and Rasch samples as indicated by *t*-test comparisons (all p's > 0.05). Moreover, it should be noted that we used the MAS sample to compute standardized scores for participants on the neuropsychological tests instead of the published normative data. This is due to that for all neuropsychological battery tests there were not available suitable normative data from an appropriate reference population, or they have been adjusted for only some, but not all, of the demographic variables, age, sex and education.

Table 6.2

Mean (SD) or median of neuropsychological raw test scores compared to population norms

Neuropsychological Domain/ component tests	Full MAS sample Mean (SD)	Rasch sample Mean (SD)	Normative data Mean (SD)
Attention/processing speed (in seconds)			
Wechsler Adult Intelligence Scale-III Digit Symbol-Coding	47.76 (12.21)	47.85 (13.04)	28.90 (9.00) – 40.00 (9.10)
Trail Making Test A	46.96 (16.63)	47.93 (17.78)	41.74 (15.32) – 50.81 (17.44)
Language			
30-item Boston Naming Test	24.23 (4.10)	24.17 (4.29)	26.30 (3.50) – 26.40 (2.70)
Animal Fluency	15.59 (4.37)	15.30 (4.64)	13.10 (3.80) – 18.20 (4.20)
Executive function (in seconds)			
FAS Fluency	36.57 (12.65)	36.51 (12.49)	22.40(8.20) – 42.00(12.10)
Trail Making Test B	120.5 (54.60)	124.92 (59.93)	100.68 (44.16) – 152.74 (65.68)
Visuo-spatial ability			
Wechsler Adult Intelligence Scale-Revised (WAIS-R) Block Design	21.33 (8.22)	21.11 (8.58)	21.58 (8.48) – 26.77 (9.89)
Memory			
Rey Auditory Verbal Learning Test	40.43 (9.57)	39.74 (9.30)	35.60 (7.70) - 46.60 (11.60)
Benton Visual Retention Test	12 (Median)	12 (Median)	12 – 15 (Median)
Wechsler Memory Scale-III (WMS-III) Logical Memory Story A	11 (Median)	11 (Median)	10 – 12 (Median)

Data Analyses

Descriptive statistics including means, standard deviations (SD), and Cronbach's alphas for the TICS-M were computed using IBM SPSS v.27 software. Rasch analyses were conducted using the RUMM2030 software package (Andrich et al., 2009) and by following the standard criteria and recommendations used to evaluate the Rasch model fit as described elsewhere (Leung et al., 2014; Tennant & Conaghan, 2007).

As TICS-M items differ in their response options structure, the Partial Credit model (Masters, 1982) was used (Lundgren-Nilsson & Tennant, 2011; Linacre, 2002). Rasch analysis begins with examination of the overall data fit and the individual item fit to the Rasch model, which involves the following steps. First, the overall model fit requires a non-significant chi-square index of item-trait interaction ($p > 0.05$). Second, item fit residuals for each scale item should range from −2.50 to +2.50. Third, the residual correlations between individual items should be evaluated to identify local dependency with estimates above 0.20 (Christensen et al., 2013). Fourth, category probability curves should also be examined for responses to individual items to determine disordered thresholds (Tennant et al., 2011), or instances where participants have difficulty discriminating between response categories for an item (Linacre, 2002). Finally, there should be no DIF for personal factors to ensure item invariance across different groups (e.g., different diagnoses, in-person neuropsychological assessment scores).

Next, the Person Separation Index (*PSI*) was used to evaluate reliability of the assessment scores in Rasch analysis. This reliability coefficient is interpreted similarly to Cronbach's alpha with values above 0.80 indicating good to excellent reliability of assessment scores (Andrich et al., 2009; Medvedev, Turner-Stokes, et al., 2018). *PSI* is not a criterion of the Rasch model fit but rather reflects how well a particular psychometric measure differentiates between individuals with different levels of the latent trait (i.e., cognition).

Rasch analysis is iteratively conducted until the best fitting model is achieved. Traditionally, to improve the Rasch model fit, misfitting items were removed. However, this technique may affect the construct validity of a scale and should be used as the last resort (Medvedev et al., 2018). Instead of deleting misfitting items, combining locally dependent items into super-items can potentially improve the Rasch model fit because it reduces spurious correlations and measurement errors associated with individual items (Lundgren-Nilsson et al., 2013; Merkin et al., 2020). Generally, data fit the Rasch model when there are no misfitting items, no local dependency and/or DIF, and unidimensionality is confirmed (Leung et al., 2014). In this study, deleting items was considered the last resort to boost the Rasch model fit because the TICS-M is a relatively short scale with well-established construct validity (Bentvelzen et al., 2019; Crooks et al., 2005).

It should be noted that super-items are created to resolve local dependency issues between individual scale items due to variety of reasons (e.g., method effect) that may be unrelated to the construct being measured or the factor structure. Local dependency between items is identified based on correlations between residual errors that exceed 0.20. Such issues can be addressed by combining dependent items to reduce unwanted error variances (Medvedev et al., 2018; Lundgren-Nilsson et al., 2013). Therefore, super-items are unrelated to conceptual similarities or differences between individual items and aim to reduce measurement error. If, after creating super-items, the best fit to the Rasch model is achieved, it is taken as evidence that measurement errors were successfully reduced using this technique.

To evaluate unidimensionality of the instrument using Rasch analysis, the equating *t*-test and principal components analysis (PCA) of the residuals are examined. Unidimensionality is confirmed when there are no significant *t*-test comparisons between person estimates calculated for the sets of items with the highest and the lowest loadings on the first principal component of residuals (< 5%). Alternatively, if the lower bound of the binominal confidence

interval calculated for significant *t*-tests overlaps at 5%, unidimensionality is also met (Smith Jr, 2002). When the data satisfactorily fits the Rasch model, the person-item thresholds distribution is inspected to evaluate the coverage of items' thresholds for the TICS-M across the sample. Finally, a transformation table can be generated to transform ordinal TICS-M scores into interval level data to enhance the accuracy of assessment.

Results

The TICS-M (M = 21.77, SD = 5.82) showed adequate internal consistency with a Cronbach's alpha value of 0.74, which is slightly lower in comparison to previous validation reports (Castanho et al., 2016; Plassman et al., 1994; Seo et al., 2011). The total scores of TICS-M in participants who were diagnosed with dementia at follow-up waves (n = 132, M = 16.11, SD = 5.56) were clearly below the established cut-off of 21 points for detecting risk of dementia and also significantly lower compared to scores of participants who were not diagnosed with dementia (n = 300, M = 24.21, SD = 3.93), $t(185.33) = 15.02$, $p < 0.001$.

Table 6.3 presents Rasch model estimates from the initial analyses for each individual item including item location, fit residual, and chi-square values. There were three items (i.e., item 4, 5, and 13) with significant misfit to the model. Table 6.4 displays the overall model fit estimates of the initial (A1), and final (A2) Rasch analyses of the TICS-M. As can be seen, initial analysis (A1) indicated good reliability of assessment scores (PSI = 0.83) for the TICS-M but the overall fit to the Rasch model was not satisfactory, as indicated by significant chi-square, $\chi^2(117) = 216.10$, $p < 0.001$, reflecting deviation of the scale from the Rasch model expectations and a violation of unidimensionality.

Table 6.3

Maximum score of TICS-M individual items and their Rasch model fit statistics including item locations, fit residuals, and Chi-square in the initial analysis

Item	Max score	Item location	Item-fit residual	Chi-square
1. What day of the week is it today? What is today's date? What season are we currently in?	5	-0.04	1.04	4.76
2. Can you tell me your age?	1	-1.10	-1.59	10.33
3. Do you know your telephone number?	1	-0.66	-1.79	19.23
4. Listen to a list of ten words carefully and then recall as many words as can be in any order.	10	2.62	-4.13*	8.54
5. Take 7 away from 100 and continue to take 7 away from the answer	5	0.76	3.95*	40.22
6. Can I ask you now to count backwards from 20 to 1	1	-0.65	1.87	13.51
7. What do people normally use to cut paper?	1	-1.22	0.98	15.06
8. What is the name of the prickly green plant that grows in the desert?	1	-0.08	-0.93	11.34
9. Who is Prime Minister of Australia now?	1	-0.65	-1.44	19.04
10. Do you know the name of the Premier of NSW?	1	1.11	0.15	15.53
11. What is the opposite of "east"?	1	-1.79	-0.33	5.95
12. Can you say the words: "British constitution"	1	-1.84	0.90	10.76
13. Recall any of the words from the list read out earlier	10	3.54	-5.59*	9.61

Note: * = Significant misfit to the Rasch model

Table 6.4

Summary of fit statistics for the initial, second and third Rasch analyses of the TICS-M (n=432)

	Person mean		Goodness of fit		PSI	Significant *t*-tests (Unidimensionality)	
Analyses	**Value / *SD***		**$\chi 2$ (*df*)**	***p***		**%**	**Lower bound**
Initial (A1)	1.79	1.09	216.10(117)	<0.001	0.83	7.2	5.1 (NOT ACCEPTABLE)
Final (A2)	-0.33	1.37	15.68(18)	0.61	0.86	3.7	1.7 (STRICT)

The residual correlation matrix was also examined because residual correlations can affect both the overall fit and individual item fit within the Rasch model, which can be improved by combining locally dependent and misfitting items into super-items (Lundgren-Nilsson et al., 2013; Medvedev, Turner-Stokes, et al., 2018; Merkin et al., 2020). Most of these misfitting items displayed residual correlations with other items that indicated local dependency. Therefore, two super-items comprised of locally dependent items were created: (i.e. super-item 1: item 1, 2, 3, 4 & 6; super-item 2: item 5, 7, 8, 9, 10, 11, 12 & 13). This analysis resulted in the best model fit with excellent reliability of assessment scores, no misfitting items, no local dependency amongst items/super-items, no DIF by personal factors, and strict unidimensionality (see Table 6.4, Final A2).

In addition, the final analysis also illustrates ordered thresholds for all super-items. Figure 6.2 shows the item category response probability curves for all items with disordered thresholds in the initial analysis (A1) as well as response probability curves for the super-items after adjustment in the final analysis (A2). As can be seen, the initial analysis identified two TICS-M items with disordered thresholds (item 5 and 13), but in the final analysis (A2), all items' thresholds were perfectly ordered.

Figure 6.2

Item category probability curves illustrating disordered thresholds for item 5 and item 13 (top panel) and orderly thresholds for super-items after adjustment (bottom panel)

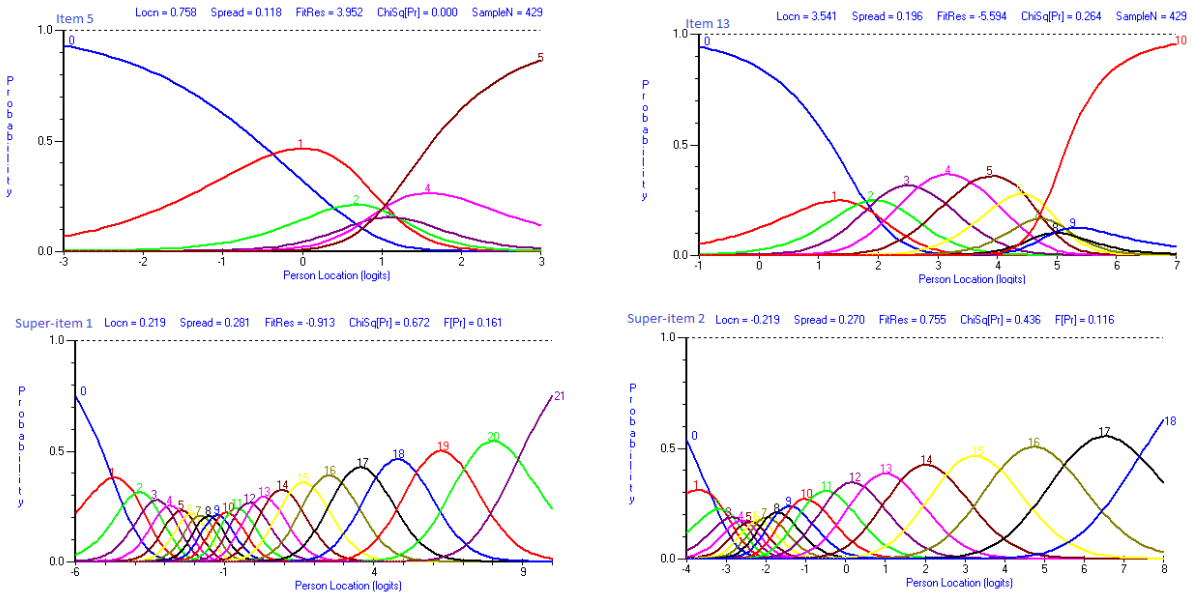
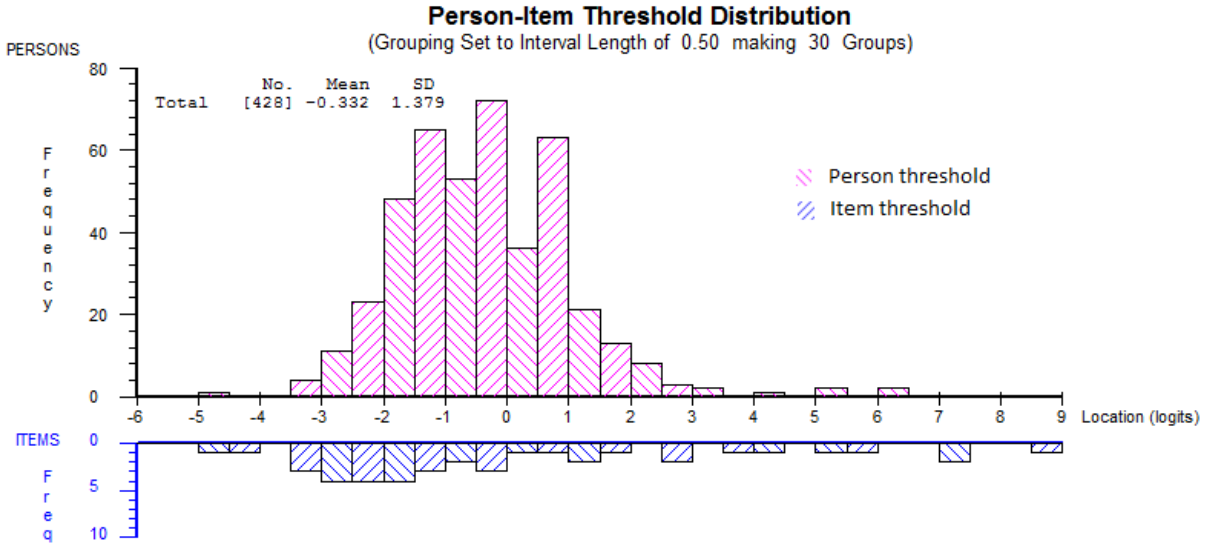


Figure 6.3 illustrates the person-item threshold distributions from the best model fit analysis (A2) for the TICS-M. As shown, the sample mean (*M* = -0.33) is slightly lower than the item mean, which is close to 0 and indicates good targeting of the sample. This suggests item thresholds account for cognitive status across participants in the sample, regardless of their diagnosis or in-person neuropsychological assessment scores, and that there was no evidence of significant ceiling or floor effects.

Figure 6.3

Person-item threshold distribution of the best model fit analysis (A2) of the TICS-M



As the final analysis (A2) indicated the best Rasch model fit for the TICS-M data, this was used to transform raw TICS-M scores into interval-level data based on person-estimates produced by the Rasch model. Table 6.5 presents the transformation table developed to convert TICS-M raw scores into interval data based on person estimates. Table 6.6 displays the results of a series of paired sample *t*-tests conducted to examine the differences between TICS-M raw scores and interval-transform scores. These *t*-tests revealed statistically significant differences between ordinal and transformed interval TICS-M scores using the same metric at all waves (all $p < 0.001$), where raw scores (*M* ranged from 23.75 to 24.11) were significantly higher than interval level scores (*M* ranged from 13.90 to 14.57). Estimates of an effect size Cohen's *d* ranged from 3.74 to 7.21 across waves indicated that the differences between the two measurement scores are large. This may reflect a gain in accuracy because Rasch transformation accounts for the unique contribution of each test item to the overarching construct of global cognition as operationalised by the TICS-M. Accuracy in the context of Rasch analysis refers to the scale compliance with the parameter of an interval level scale defined by the Rasch model and is consistent with the principles of fundamental measurement (Hobart & Cano, 2009; Medvedev & Krägeloh, 2022). This is important given a main goal of the study was to achieve the parameters of an interval scale for the TICS-M as opposed to examining the construct validity and sensitivity and specificity reported by other studies (Crooks et al., 2006; Van Den Berg et al., 2012).

Table 6.5

Converting raw scores into interval level scores for the TICS-M

Raw scores	Logits	Interval scores	Raw scores	Logits	Interval scores
0	-5.41	0.00	20	-1.00	11.51
1	-4.59	2.16	21	-0.81	12.01
2	-4.02	3.65	22	-0.59	12.57
3	-3.63	4.66	23	-0.36	13.19
4	-3.34	5.41	24	-0.09	13.88
5	-3.12	5.99	25	0.20	14.65
6	-2.94	6.46	26	0.53	15.51
7	-2.79	6.85	27	0.90	16.46
8	-2.65	7.21	28	1.30	17.52
9	-2.53	7.53	29	1.75	18.69
10	-2.41	7.85	30	2.25	19.98
11	-2.29	8.15	31	2.79	21.40
12	-2.17	8.46	32	3.38	22.95
13	-2.05	8.78	33	4.03	24.64
14	-1.92	9.11	34	4.74	26.48
15	-1.79	9.45	35	5.50	28.46
16	-1.65	9.82	36	6.32	30.61
17	-1.50	10.20	37	7.23	32.97
18	-1.35	10.61	38	8.29	35.74
19	-1.18	11.04	39	9.54	39.00

Table 6.6

Paired sample t-tests of the differences between TICS-M raw scores and interval-transform scores across waves

	Mean differences	Cohen's *d*	95% Confidence Interval of Cohen's *d*		*t*	*df*	*p*
			Lower	Upper			
Wave 1	9.50	7.21	6.87	7.54	218.79	920	<0.001
Wave 2	9.56	6.37	6.05	6.68	183.62	831	<0.001
Wave 3	9.26	5.27	4.89	5.44	141.01	744	<0.001
Wave 4	9.10	4.55	4.29	4.82	113.28	618	<0.001
Wave 5	8.91	3.95	3.56	4.08	84.07	483	<0.001
Wave 6	8.85	3.74	3.43	4.05	67.37	323	<0.001

Table 6.7 includes means and standard errors of TICS-M ordinal raw scores and interval-transformed Rasch scores across diagnoses groups for the 6 waves. It can be seen that all standard errors of ordinal TICS-M scores were consistently higher compared Rasch transformed scores across waves and diagnostic subgroups. These differences are especially apparent for those diagnosed with dementia at later waves. Overall, these findings indicate that Rasch transformed scores are more accurate compared to the original TICS-M scores because Rasch scores are less affected by measurement error in comparison to ordinal scores. Therefore, using interval-level scores in clinical assessment and research reduces both Type I and Type II errors and increases statistical power of parametric tests as the probability of these errors is directly linked to noise in the data.

Table 6.7

Means and standard errors (SE) of TICS-M ordinal raw scores and interval-transformed scores across diagnoses of six waves

Diagnosis	Scores	Wave 1		Wave 2		Wave 3		Wave 4		Wave 5		Wave 6	
		Mean	*SE*	Mean	*SE*	Mean	*SE*	Mean	*SE*	Mean	*SE*	Mean	*SE*
Normal	Ordinal	23.22	0.93	23.84	0.47	22.81	0.56	25.25	1.05	23.92	0.22	24.18	0.27
	Rasch	14.13	0.76	14.18	0.38	13.59	0.44	15.35	0.81	14.56	0.17	14.69	0.21
MCI	Ordinal	22.61	0.21	23.06	0.21	22.47	0.31	22.22	0.33	-	-	-	-
	Rasch	13.49	0.15	13.7	0.14	13.56	0.23	13.4	0.23	-	-	-	-
Dementia	Ordinal	-	-	14.58	1.28	13.1	0.89	16.26	0.85	12.91	1.02	15.92	0.80
	Rasch	-	-	9.5	0.48	10.14	0.41	10.4	0.43	8.73	0.42	10	0.36

Note. Diagnosis at wave 5 and 6 only included dementia and not dementia only.

Discussion

The TICS-M is a widely used measure of cognitive functioning in older persons and a screen for dementia. Determining cognitive function and/or screening for dementia requires high measurement accuracy that can be achieved using Rasch analysis. The aim of this study was to examine and improve psychometric properties of the TICS-M using Rasch methodology using a statistically appropriate sample. Initial analysis indicated that the TICS-M did not fully comply with principles of fundamental measurement defined by the Rasch model and modifications were required to improve the scale. We created super-items by combining original TICS-M items, which resolved problems associated with item misfit, local dependency, and disordered thresholds resulting in both better scale accuracy for assessing global cognition and enhanced Rasch model fit (Medvedev, Krägeloh, et al., 2020). These new super-items demonstrated their appropriateness by achieving strict unidimensionality and invariance in the adjusted TICS-M final analysis. Research has shown that super-items can only be generated in scales measuring one overarching latent variable (e.g., cognition or perfectionism), and that adequate Rasch model fit cannot be achieved when extraneous factors are present in a measure (Mitchell-Parker et al., 2018). Additionally, achieving strict unidimensionality and invariance across diagnostic groups in the current sample confirms that the adjusted scale works equally well for older persons with normal cognition and those with dementia, and is not impacted by their sex, age range (70 to 90 years old), or global cognitive performance.

The adjusted TICS-M also showed excellent reliability of the assessment scores (*PSI* = 0.86) for both individual and group assessments, which allowed us to generate a transformation table that converts ordinal raw scores into interval-level data to improve the accuracy of TICS-M without altering its current format (Merkin et al., 2020; Truong et al., 2021b). Importantly, the transformation table created here is user friendly, which allows for quick conversion of raw

scores into interval scores and is a brief adjunct tool that researchers and clinicians can use when screening people at risk of dementia or monitoring people's cognitive performance over time. This is important because detecting dementia at its earliest stage could facilitate targeted interventions before cognitive and functional decline is progressed too far (Langa & Levine, 2014). To use Table 6.5, after computing the total ordinal score of the TICS-M (labelled "Raw scores"), administrators can use such raw scores to find the corresponding interval score on the column labelled "Interval scores". For example, if the computed total Raw score is 21 (cut-off point for dementia risk), the corresponding interval score is 12.01, which is the cut-off point on the interval scale. We provided interval-level conversion scores that increase the precision of the ordinal TICS-M scores, which have been standardised against a normative older population. To interpret interval-level scores in our table, clinicians can use ordinal TICS-M scores found on the left-hand side and find corresponding interval scores on the right-hand side of Table 6.5.

The results of this study add evidence for the benefits of using a super-item approach in Rasch analysis (Lundgren-Nilsson et al., 2013). Combining items into super-items reduces measurement error due to individual item errors and counterbalances error variance across items (Lundgren-Nilsson et al., 2013; Medvedev, Turner-Stokes, et al., 2018). Creating super-items is similar to the averaging technique used in studies of event related potentials utilizing electroencephalogram (EEG), where the test is repeated several times until error/noise variances cancel each other out and the relevant neural response can be identified (Luck, 2014).

In addition, as raw scores of TICS-M constitute an ordinal measure we cannot warrant that the distances between response categories of its items have the same values (Merbitz et al., 1989). In clinical contexts, all items of a scale are displayed as having the same level of contribution to the overall construct and the sum score is calculated by adding individual items scores, which can bias the diagnostic results (Tennant & Conaghan, 2007). Achieving the best

Rasch model fit in this study eliminates TICS-M assessment bias by accounting for precise contributions of each individual item to the overall score of the scale in the ordinal-to-interval conversion algorithm (Christensen et al., 2013; Rasch, 1960, 1961). These findings are also useful to illustrate robustness of Rasch methodology.

Even though previous studies have demonstrated the original TICS-M scores are reliable and valid in clinical settings these ordinal scores may be less accurate reflections of cognitive ability compared to the interval transformed scores as demonstrated by numerous Rasch studies (Mayhew et al., 211; Norquist et al. 2004; Hobart & Cano, 2009). For example, if individual A has a TICS-M ordinal score of 15 they are at higher risk of having dementia compared to individual B who scores 25. Assume, then, that both individuals A and B receive a specific treatment and individual A's ordinal score increases to 20 and participant B's ordinal score increased to 30. Both individuals have increased their ordinal TICS-M scores five points suggesting the same amount of clinical improvement. Importantly, though, when transforming these original TICS-M scores into Rasch interval scores, individual A's interval score has increased by 2.06 points, while individual B's interval score has increased by 5.33 points. This illustrates that individual B's improvement in global cognition level is approximately 2.5 times greater than individual A's. Therefore, using Rasch interval transformed data may more accurately capture true changes in cognitive performance associated with dementia. When using the TICS-M for screening purposes (e.g., screening dementia), clinicians should not transform the original TICS-M scores because they work well for this purpose. Clinicians and researchers should use interval transformed TICS-M scores for screening purposes and monitoring the outcomes simultaneously, and it should be noted that converting the criterion for relative evaluation in the interval metrics is important in this case.

Strengths and Limitations

Major strengths of our study include an optimal sample size range ($250 < n \leqslant 500$), which reduces both Type I and Type II errors (Hagell & Westergren, 2016). Secondly, this study is novel because, to the best of our knowledge, no studies have examined the psychometric properties of the TICS-M using Rasch methodology. In our sample, the overall Wechsler Adult Intelligence Scale-III Digit Symbol-Coding scores were above the upper range of normative data (Ryan et al., 2020), which might be explained by sampling bias. The sample for the current study was drawn from the MAS, which recruited participants from an affluent area of Sydney, Australia, who are predominately white, European, and well-educated. Thus, our sample is not representative of older adults in general, even within Sydney, Australia. This speaks to a larger issue around neuropsychological tests which are largely normed against non-culturally or linguistically diverse samples. We also found that the Wechsler Adult Intelligence Scale-Revised Block Design scores were marginally below its lower normative range (Scott et al., 2021). This might be explained by the fact that our samples included participants with MCI and dementia. Although, Rasch methodology is generally considered as less sample dependent (Hobart & Cano, 2009), future replications of our results are warranted to confirm our results in a more culturally and linguistically diverse (CALD) sample of older adults. Future studies should use a more culturally, linguistically, and educationally diverse sample of older adults to investigate invariance (i.e., potential DIF) across different CALD samples (e.g., English speakers from Anglosphere versus non-Anglosphere countries, or those use English as second language; English versus translated versions) and different educational backgrounds. In addition, the MAS data included 43% of participants who were diagnosed with MCI at wave 1 (nearly 50% in our Rasch sample), which may explain lower mean TICS-M scores for both the total sample and our Rasch sample compared to population norms. However, this may be an advantage for calibrating TICS-M items in Rasch analyses because a large range of cognitive

abilities was represented in our sample. This also means that our transformation table would be suitable for both healthy and clinical populations.

Conclusion

In conclusion, the findings of this study demonstrate that the TICS-M is a reliable and valid assessment tool for capturing cognitive functioning among older adults. The adjustments of the TICS-M based on Rasch analysis resolved psychometric issues associated with individual items and permitted transformation of raw scores into interval-level data, which enhanced the precision of measurement. Clinicians and researchers can apply this interval-transform data table in a variety of contexts to improve the scale accuracy without altering the original format of the TICS-M.

Chapter 7 Using Network Analysis to Validate Domains of the Modified Telephone Interview for Cognitive Status

Truong, Q. C., Cervin, M., Choo, C. C., Numbers, K., Bentvelzen, A. C., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N. A., & Medvedev, O. N. (2023). Using network analysis to validate domains of the modified telephone interview for cognitive status. *European Journal of Clinical Investigation*, e14016. Advance online publication. https://doi.org/10.1111/eci.14016

Declaration

This study is co-authored work with Cervin, M., Choo, C. C., Numbers, K., Bentvelzen, A. C., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N. A., and Medvedev, O. N. (see Appendix P6 for the original article). I was the first author and had the major contribution to this study that included designing and conducting the study, completing data analysis, writing the manuscript for submission to the *European Journal of Clinical Investigation*, and responding to reviews/addressing their comments during peer-review process. All co-authors have signed the co-authorship form (see Appendix 7.1).

Abstract

***Background*:** The modified Telephone Interview for Cognitive Status (TICS-M) is a widely used tool for assessing global cognitive functions and screening for cognitive impairments. The tool was conceptualised to capture various cognitive domains, but the validity of such domains has not been investigated against comprehensive neuropsychological assessments tools. Therefore, this study aimed to explore the associations between the TICS-M domains and neuropsychological domains to evaluate the validity of the TICS-M domains using network analysis.

***Materials and methods*:** A longitudinal research design was used with a large sample of older adults (aged above 70 years; n = 1037 at the baseline assessment) who completed the TICS-M and comprehensive neuropsychological assessments biennially. We applied network analysis to identify unique links between the TICS-M domains and neuropsychological test scores.

***Results*:** At baseline, there were weak internal links between the TICS-M domains. The TICS-M memory and language domains were significantly related to their corresponding neuropsychological domains. The TICS-M attention domain had significant associations with executive function and visuospatial abilities. The TICS-M orientation domain was not significantly associated with any of the five neuropsychological domains. Despite an attrition of almost 50% at wave four, weak internal links between the TICS-M domains and most associations between TICS-M and neuropsychological domains that were found initially, remained stable at least over two waves within the six-year period.

***Conclusions*:** This study supports the overall structural validity of the TICS-M screener in assessing enduring global cognitive function. However, separate TICS-M cognitive domains should not be considered equivalent to the analogous neuropsychological domains.

Keywords: Global Cognition, Network Analysis, modified Telephone Interview for Cognitive Status, Neuropsychological Domains

Introduction

The modified Telephone Interview for Cognitive Status (TICS-M) (Welsh et al., 1993) is a well-validated assessment tool used to assess global cognitive function in older adults. The TICS-M is suitable for screening for cognitive impairment (Crooks et al., 2005; Truong et al., 2022) and monitoring changes in cognition over time such as due to an effective intervention (Brandt et al., 1988; Brandt et al., 1993; Fong et al., 2009). This tool, which can be administrated by telephone or face-to-face, is often used in large epidemiological studies by the virtue of its brevity (Fong et al., 2009; Wilson & Bennett, 2005).

The TICS-M, while brief, consists of 13 items grouped into four cognitive domains, broadly including orientation, memory, attention, and language according to factor analytic data from the original authors and a subsequent key paper (Welsh et al., 1993) (Brandt et al., 1993; de Jager et al., 2003). Appendix 7.2 displays a detailed description of the TICS-M questionnaire, including its domains and individual items as well as the maximum score of each item. Previous studies have demonstrated high internal and temporal reliability of this measure (Castanho et al., 2016; Plassman et al., 1994; Seo et al., 2011; Truong et al., 2022), while its validity was supported by strong correlations with the widely used Mini-Mental State Examination (de Jager et al., 2003; Folstein et al., 1975). In addition, the TICS-M has been validated against more comprehensive neuropsychological assessment tools, such as tests of general intellectual function, memory, and language (Crooks et al., 2006; Van Den Berg et al., 2012). However, to the best of our knowledge, such validation studies have only focused on the total score of the TICS-M rather than the scores of its cognitive domains.

Network analysis offers a rigorous and comprehensive statistical framework to evaluate the validity of the TICS-M cognitive domains in relation to neuropsychological test performance (Borsboom & Cramer, 2013; Chalmers et al., 2022). Network analysis can be used to understand unique associations between a set of components by integrating single

components into a global network, providing a clearer picture to enhance understanding of the complex relationship between components (Borsboom & Cramer, 2013; Cholerton et al., 2016). Network analysis graphically illustrates the presence and strength of associations between different variables using “nodes” and “edges” (Borsboom et al., 2021). The nodes represent variables in the network and the edges represent unique statistical associations between the nodes (often in the form of partial correlations). Network analysis estimates all associations within the network without dividing variables into dependent (i.e., caused by other variables) and independent (i.e., affecting the dependent variable) variables (Åkerblom et al., 2021). This is preferable when no clear causal pathways are evident, making it a preferred method of analysis to determine the associations between the TICS-M domains and neuropsychological test performance. Furthermore, both the TICS-M domains and various cognitive tests capture different aspects of a person’s broader cognitive functioning and thus are expected to be interrelated. By estimating their unique associations within a single network, these complex interrelations can be disentangled and unique links between different domains outlined.

This study aimed to explore the associations between the TICS-M domains and neuropsychological domains to evaluate the convergent validity of the TICS-M domains in a large sample of older adults over a six-year period. Exploratory analyses using data collected at baseline, also called “wave 1,” aimed to explore unique associations between TICS-M domains and their associated cognitive domains as captured by standardised neuropsychological testing. Such analyses could provide preliminary evidence about convergent validity of individual TICS-M domains. Confirmatory network analyses were then conducted using data collected after two, four, then six years, or three follow-up waves (i.e., waves 2, 3, and 4), which allowed us to discard spurious exploratory findings and identify whether associations between domains captured by the TICS-M and neuropsychological tests

remained stable over the six-year period. To do this, we derived hypotheses after performing exploratory analyses, which were then tested in confirmatory analyses across waves.

Method

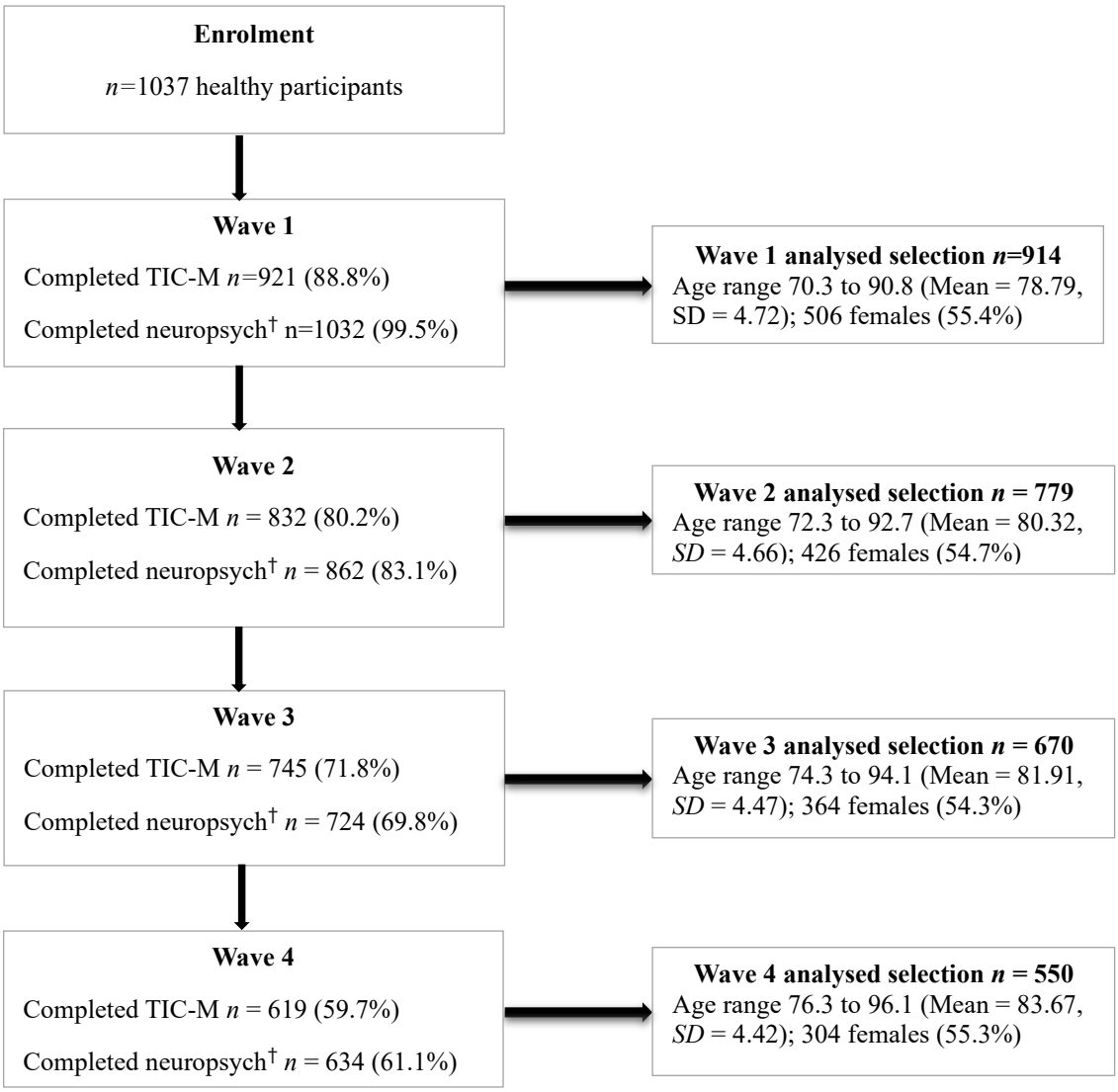
Participants

Participants were from the Sydney Memory and Ageing Study (MAS) which included 1037 older adults (572 females) aged 70 to 90 (M_{age} = 78.84, *SD* = 4.82) at the baseline assessment (i.e., wave 1) (Sachdev et al., 2010). Participants were living in the Eastern suburbs of Sydney, Australia, and were fluent in English such that they could provide written consent and complete psychometric assessments and self-report questionnaires (Sachdev et al., 2010). The ethnicity of the MAS participants was predominantly European (98%). Participants were followed-up every two years after the baseline assessment over the six-year period. Exclusion criteria were prior dementia diagnosis, major psychological or neurological disorder, or progressive malignancy at baseline. More details about inclusion and exclusion criteria for MAS, as well as participants' baseline demographics, have been previously published elsewhere (Sachdev et al., 2010). It should be noted that no inclusion/exclusion criteria were applied to participants at follow-up waves. Missing data at each followed wave were due to participants being unwell, having passed away, or being unreachable at that wave. In some instances, participants were unable to participate in the assessment or were in advanced stages of dementia, rendering them unable to answer questions during the follow-up waves. All participants provided written informed consent to participate in this study, which was approved by the University of New South Wales Human Ethics Research Committee (HC: 05037, 09382, 14327, 190962).

Figure 7.1 presents the consort diagram of how participants were selected at each wave for data analyses. We excluded participants who did not have complete assessments for each wave. Figure 7.1 also displays some basic demographic details of participants (i.e., age and sex) at each wave. The results of independent samples *t*-tests indicated that there was no significant difference in the ages of female and male participants across all waves (all p's>.05).

Figure 7.1

CONSORT diagram for participants selected at each wave



Note: † neuropsychological assessments

Measures

TICS-M. The TICS-M (Welsh et al., 1993) is a 13-item test of cognitive functioning conducted via telephone or face-to-face. The scale theoretically captures four domains: orientation, memory, attention, and language (de Jager et al., 2003). Each TICS-M item is scored differently with scores ranging from 1 to 10 and higher scores indicate better cognition function. Individual domain scores are computed by adding responses of relevant item(s) together, and thus their maximum scores are different (see Appendix 7.2). For example, the memory domain has a maximum score of 22, while attention has a maximum score of six. It should be noted that the TICS-M assessments were conducted one year after neuropsychological assessments were done at each wave.

Neuropsychological Cognitive Assessment. At each wave, participants completed a full neuropsychological test battery that captured the domains of attention/processing speed, language, executive function, visuospatial ability, and memory. These cognitive domains were evaluated because they assess cognitive abilities directly relevant for the diagnosis of dementia and predementia syndromes in the MAS study (Sachdev et al., 2010). Domains were formed based on participants' scores across ten neuropsychological tests according to the principal cognitive function each test represented (Sachdev et al., 2010). The Wechsler Adult Intelligence Scale-III Digit Symbol-Coding (Wechsler, 2003) and the Trail Making Test A (Tombaugh, 2004) were used to assess attention/processing speed. The Boston Naming Test (Kaplan et al., 2001) and Animal Fluency (Spreen, 1977) assessed language. The FAS Fluency (Benton, 1967) and Trail Making Test B (Reitan & Wolfson, 1993) assessed executive function. The Wechsler Adult Intelligence Scale-Revised (WAIS-R) Block Design (Wechsler, 1981b) assessed visuo-spatial ability. Memory was assessed by the Wechsler Memory Scale-III (WMS-III) Logical Memory Story A delayed recall (Wechsler, 1997), Rey Auditory Verbal Learning Test (Strauss et al., 2006) total learning, short-term and long-term recall scores, and

the Benton Visual Retention Test (Benton et al., 1996). Neuropsychological domain scores were computed as quasi *z*-scores as follows. First, raw test scores were converted to *z*-scores using baseline means and standard deviations (*SDs*) of a reference group comprised of 723 MAS participants classified as cognitively healthy at baseline. Second, composite domain scores were formed by averaging the z-scores of the component tests, apart from the visuospatial domain which was represented by a single test. Each domain composite was standardised by transforming it so that the mean and *SD* of the baseline cognitively healthy group were 0 and 1, respectively. Higher *z*-score represented better performance/ability.

Data Analyses

IBM SPSS Statistics 28 software was used to compute descriptive statistics including mean, standard deviation, skewness and kurtosis for the domain/node scores. We estimated a network that included the TICS-M domains and the neuropsychological test domains. Each variable/domain was represented as a node in the graphical network and unique associations between variable pairs were represented as edges (i.e., lines). The networks were Gaussian graphical models (GGM). In a network, nodes are connected with edges, and edges represent statistically significant associations. In this study, edges were estimated using partial correlations which can range from -1 (perfect negative association) to +1 (perfect positive association). The strength of each edge is reflected by the thickness of the line. Blue lines represent positive associations and red lines represent negative associations. We coloured the nodes according to whether they were included in the TICS-M or whether they were a cognitive domain measured using neuropsychological tests.

The network analyses were conducted in R (version 4.0.4; R Core Team, 2021) using the package *BGGM* which can estimate Bayesian Gaussian Graphical Models (BGGM) (Williams & Mulder, 2020). We used Copula Gaussian graphical model estimation, which can estimate linear associations (edges) between a set of variables with different distributions

because some variables were expected to be non-normally distributed. Unique edges between nodes were estimated as partial correlations, that is, each partial correlation represented the edge while accounting for all other linear relationships in the full set of nodes. Using a Bayesian theory approach, credible intervals (*CIs*) for each edge were used to control for false positive rate (Edwards et al., 1963). Edges whose 95% CI did not include zero were considered statistically significant. A 95% CI indicates the lower and upper limits of an interval where an unobserved parameter is expected to fall 95% of the times (Hespanhol et al., 2019).

The R library *qgraph* and the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991) were used to plot the networks. This algorithm allows the most interconnected nodes to be central in the diagram as well as placing strongly related nodes closer to each other whilst avoiding overlap of nodes (Epskamp et al., 2012). To facilitate comparisons, this study used an average network layout that was based on estimates from all four waves. That is, the placement of each node was kept the same in all networks while the edge thickness differed (Fruchterman & Reingold, 1991). Moreover, predictability of nodes represented by Bayes R^2 were also conducted to indicate which nodes can be influential in the network (Haslbeck & Waldorp, 2018). A node with higher predictability means that such a node is more central or influential in comparison to other nodes in their relationships (Chalmers et al., 2022; Haslbeck & Waldorp, 2018).

This study involved exploratory and confirmatory analyses. The exploratory analyses were carried out with the sample of participants at baseline (wave 1) to explore the network relations between nodes. In other words, this analysis was used to explore edges (associations) between nodes (domains) of the TICS-M and neuropsychological assessments. Post hoc tests were subsequently conducted to further test whether network patterns were statistically significant (e.g., that one edge is statistically significantly stronger than another edge), again using wave 1 data. Confirmatory analyses were then conducted with samples of participants at

each follow-up wave to examine whether the network patterns that had been established with the exploratory sample at baseline could be confirmed using new data at future waves. The posterior probability (*PP*) method was used to test hypotheses (Lambert, 2018). PP is the probability of a pre-specified event occurring; for instance, that the edge between nodes A and B will be larger than between nodes A and C. To calculate PPs and CIs, we used 5,000 posterior estimates, which aims to approximate the posterior distribution of parameters. A *PP* above 0.95 (95%) was used as an indicator of statistical significance. That is, if the *PP* for edge A-B > edge A-C, is above 0.95, we consider this difference to be statistically significant and the hypothesis confirmed.

Results

Table 7.1 presents descriptive statistics for the domains of TICS-M and the neuropsychological assessments domains across the four waves. Because skewness and kurtosis values for some domains/nodes indicated non-normality (Muthén & Kaplan, 1985), associations were estimated using Copula Gaussian graphical model estimation.

Exploratory Analyses

Figure 7.2 shows the network for the TICS-M domains/nodes and neuropsychological assessment domains/nodes at wave 1. The neuropsychological nodes were highly interconnected (i.e., showing many thick edges between nodes with edge weights between 0.25 and 0.59) while the TICS-M nodes were not (i.e., edge weights between nodes ranged from 0.10 to 0.27). The TICS-M attention node had statistically significant edges (i.e., the 95% *CI* excluded zero) with three neuropsychological nodes: executive function (edge weight=0.26, 95% *CI* [0.061, 0.228]), visuospatial ability (edge weight=0.25, 95% *CI* [0.117, 0.265]), and attention (edge weight =-0.15, 95% *CI* [-0.159, -0.003]). Post hoc tests were used to examine whether these edges were confirmed. That is, whether the edges between the TICS-M attention node and the neuropsychological executive function and visuospatial ability nodes would be significantly stronger than between the TICS-M attention node and the other neuropsychological nodes. Also, whether the edge between the TICS-M attention node the neuropsychological attention node would be significantly weaker than between the TICS-M attention domain and the two other neuropsychological nodes (i.e., memory and language). The results confirmed the edges between the TICS-M attention node and the neuropsychological executive function and visuospatial ability nodes were statistically significant with all *PPs*>98.8%, but not the edge between the TICS-M attention node and the neuropsychological attention node.

Table 7.1

Descriptive statistics including means, standard deviation (SD), skewness, and kurtosis for the domain scores of neuropsychological assessments, and the TICS-M on the four waves.

	Wave 1 (*n* = 914)				**Wave 2 (*n* = 779)**				**Wave 3 (*n* = 670)**				**Wave 4 (*n* = 550)**			
Scale/Domains	*M*	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis
TICS-M																
Orientation	6.61	0.69	-2.39	8.98	6.59	0.75	-2.68	10.54	6.58	0.71	-2.30	7.83	6.48	0.96	-2.92	10.96
Memory	8.35	2.99	0.32	0.56	8.61	3.07	-0.81	0.01	8.72	3.60	-0.12	0.15	8.33	3.77	0.14	-0.10
Attention	5.05	1.30	-1.45	1.48	5.13	1.29	-1.66	2.31	5.13	1.28	-1.69	2.40	5.09	1.34	-1.59	1.75
Language	3.78	0.52	-3.17	13.61	3.80	0.49	-3.13	13.14	3.84	0.42	-2.84	9.27	3.82	0.44	-2.69	8.40
Neuropsychology																
Attention/processing speed	-0.12	1.09	-0.84	1.90	-0.17	1.23	-1.89	12.12	-0.29	1.23	-1.70	9.52	-0.49	1.27	-1.35	5.64
Language	-0.16	1.11	-0.59	0.92	-0.29	1.11	-0.62	0.22	-0.26	1.13	-0.63	0.72	-0.36	1.19	-0.71	0.68
Executive function	-0.05	1.02	-0.58	1.13	-0.18	1.18	-1.40	4.83	-0.20	1.11	-0.72	1.05	-0.40	1.39	-1.43	4.24
Visuospatial ability	-0.06	1.02	0.23	-0.04	-0.05	1.08	0.08	0.00	-0.01	1.07	0.17	0.12	-0.20	1.11	0.04	-0.14
Memory	-0.09	1.04	-0.03	-0.34	-0.15	1.06	-0.03	-0.29	-0.09	1.11	-0.23	-0.30	-0.21	1.18	-0.13	-0.58

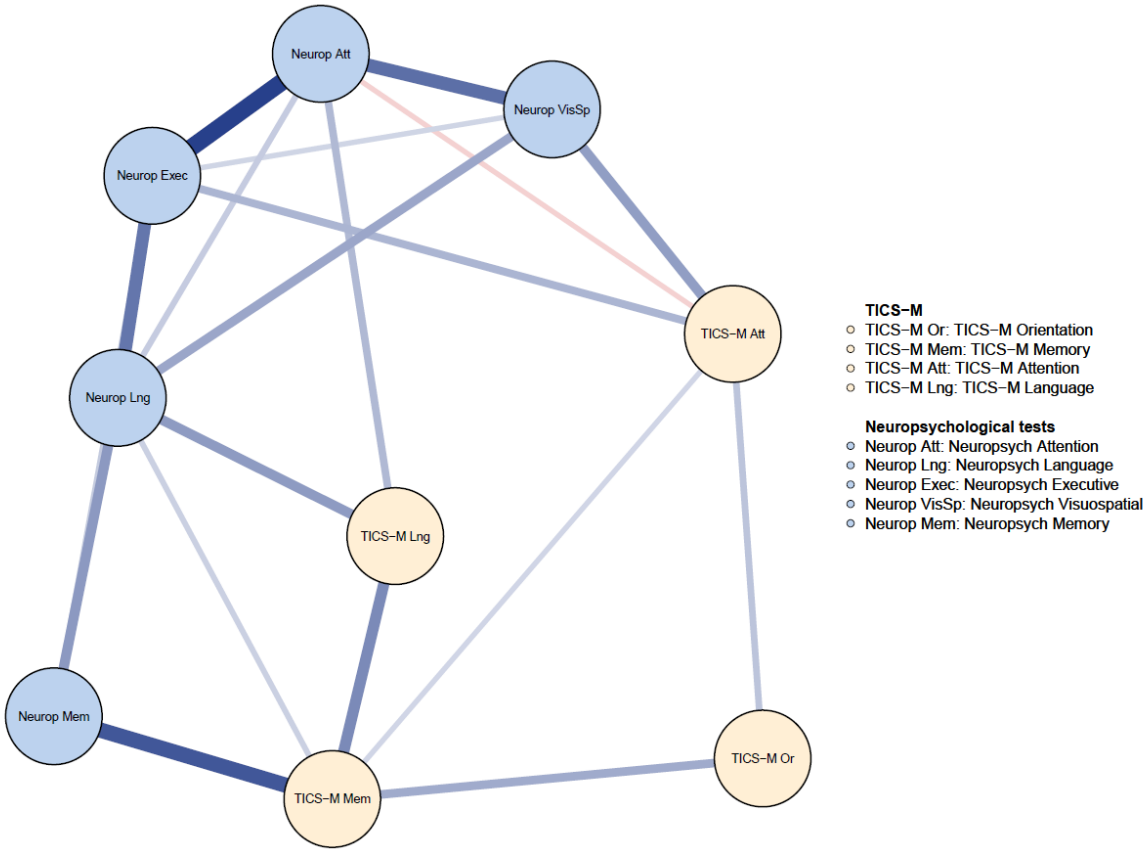
Note. *M*: mean, *SD*: standard deviation

The TICS-M language node had significant edges with the neuropsychological language (edge weight of 0.29, 95% *CI* [0.106, 0.288]) and attention nodes (edge weight of 0.26, 95% *CI* [0.041, 0.233]). Further tests were conducted to determine whether these edges were significantly stronger than edges between the TICS-M language node and the three remaining neuropsychological nodes (i.e., executive functions, visuospatial, and memory). The results confirmed this, with all *PPs* > 95.1%. In addition, the TICS-M memory node had a significant edge with neuropsychological memory (edge weight = 0.47, 95% *CI* [0.270, 0.395]) and this edge was significantly stronger than all other edges between TICS-M memory and the other neuropsychological domains (all *PPs*=100%).

In sum, the results from the exploratory network analyses allowed us to derive hypotheses that 1) there would be weak links between the TICS-M nodes internally; 2) the TICS-M attention node would be more strongly associated with the neuropsychological executive function and visuospatial nodes than other neuropsychological domain nodes; 3) the TICS-M memory node would be more strongly related to neuropsychological memory node than to the other neuropsychological domain nodes; 4) the TICS-M language node would be more strongly related to the neuropsychological language and attention nodes than to the other neuropsychological domain nodes; and 5) the TICS-M orientation node would not have an associations with any of the neuropsychological domain nodes. These hypotheses would be tested in confirmatory analyses using data collected at waves 2 – 4.

Figure 7.2

Exploratory network of neuropsychological domains and TICS-M domains between Wave 1



Confirmatory Analyses

Figure 7.3 displays the estimated networks for each follow-up wave with averaged network layouts. Table 7.2 illustrates the summary of the confirmations of hypotheses of associations between TICS-M and neuropsychological nodes formed in wave 1/exploratory analyses across three follow-up waves. As shown, even though new associations between the TICS-M and some neuropsychological nodes emerged in the followed-up waves as comparison to wave 1 (e.g., edges between TICS-M orientation and neuropsychological visuospatial and memory nodes), the relative differences between edges were not affected by the presence of these new edges. Specifically, the network for wave 2 partly confirmed our hypotheses that weak internal associations between TICS-M nodes would remain (i.e., edge weights ranged from 0.12 to 0.27), and that the edge between the TICS-M memory and neuropsychological memory nodes, and that the edge between TICS-M attention and neuropsychological visuospatial nodes, would remain significant. The hypothesis that no associations between TICS-M orientation and neuropsychological nodes was also confirmed at wave 2. The network at wave 3 confirmed most our hypotheses, except the hypothesis that no associations between TICS-M orientation and neuropsychological nodes was not confirmed as there was an association between TICS-M orientation and neuropsychological visuospatial nodes. The hypotheses that were supported at wave 4 were the edge between TICS-M memory and neuropsychological memory node, and the edge between TICS-M language and neuropsychological language nodes. Besides that, the network analysis at wave 4 revealed stronger internal associations between TICS-M nodes (i.e., some edge weights were 0.30 or above) compared to the wave 1 findings. However, these TICS-M node interconnections were incoherent, and thus this finding confirmed the TICS-M internal associations found at wave 1. Therefore, the confirmatory results indicated that the majority of associations between nodes that emerged during the exploratory network analyses at wave 1 were verified by confirmatory

network analyses with at least one follow-up wave during the six-year period. This also means that there were significant and enduring relationships between TICS-M nodes and neuropsychological nodes.

Predictability of nodes was also analysed across waves, which is presented in Figure 7.4. As can be seen, all neuropsychological nodes had high predictability, with the TICS-M memory node showing the highest predictability compared to other TICS-M nodes. This suggests that all neuropsychological domain nodes and the TICS-M memory node, specifically, reflected cognitive abilities that were more influential across all networks.

Figure 7.3

Confirmatory networks of neuropsychological domains and TICS-M domains at follow-up waves

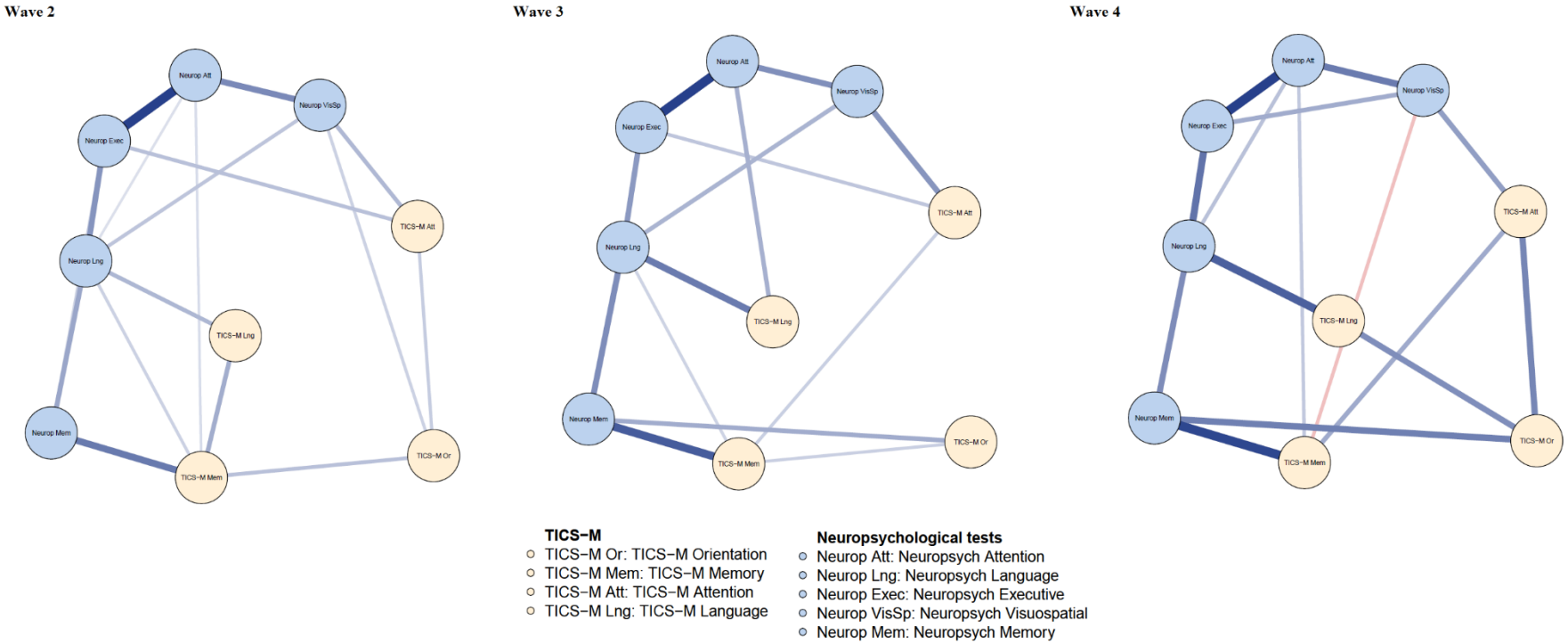


Table 7.2

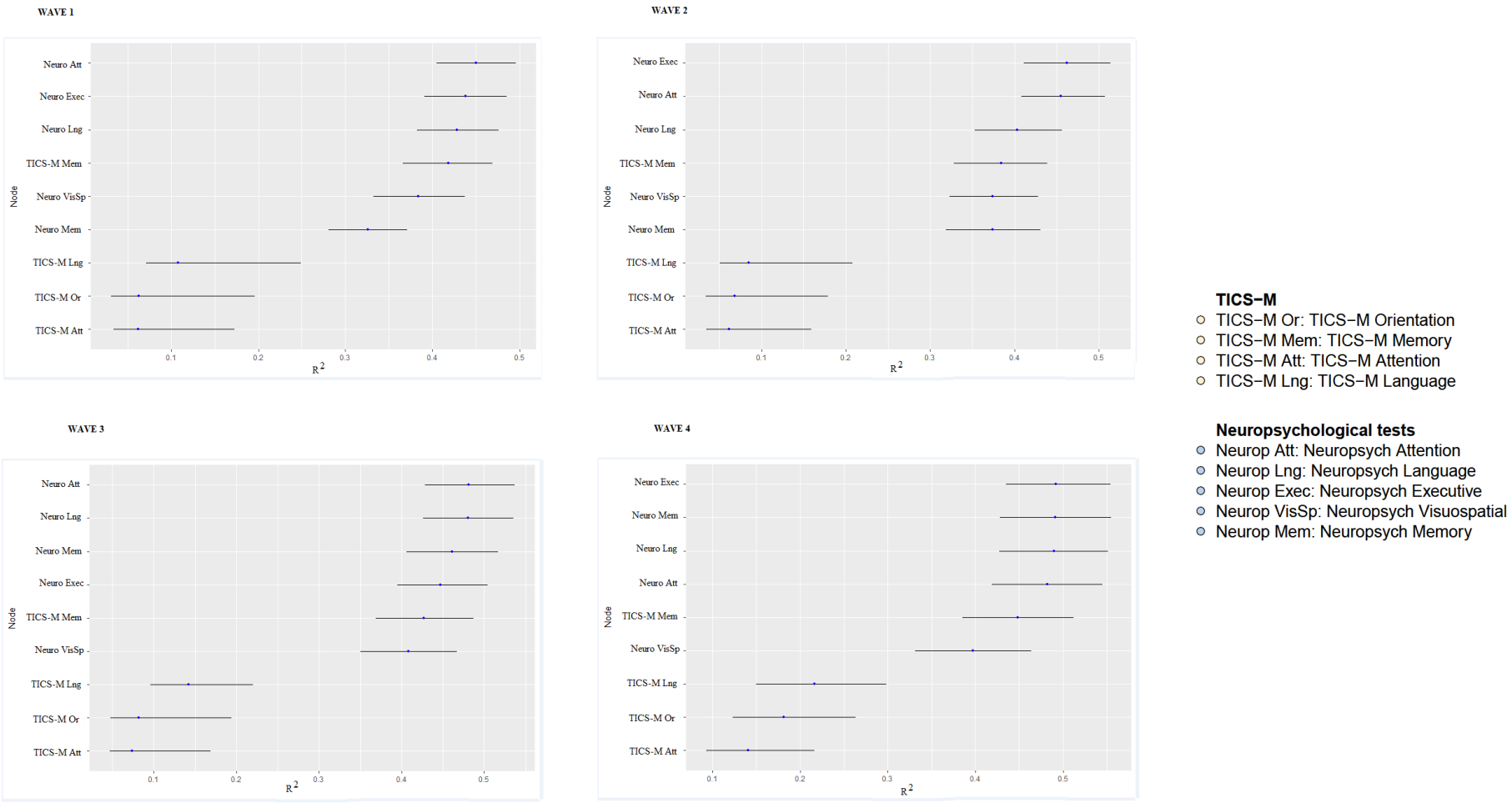
Summarised the Confirmations of Hypotheses of Associations between TICS-M and Neuropsychological Nodes formed in the Exploratory Analyses across Follow-up Waves

Hypotheses formed at wave 1/exploratory analyses	Confirmatory analyses		
	wave 2	wave 3	wave 4
Weak internal associations between TICS-M nodes	✓	✓	✓
Significant association between TICS-M attention and neuropsych visuospatial nodes	✓	✓	×
Significant association between TICS-M attention and neuropsych executive nodes	×	✓	×
Significant association between TICS-M language and neuropsych language nodes	×	✓	✓
Significant association between TICS-M language and neuropsych attention nodes	×	✓	×
Significant association between TICS-M memory and neuropsych memory nodes	✓	✓	✓
No associations between TICS-M orientation and neuropsych nodes	✓	×	×

✓confirmed with wave 1 hypothesis; × not confirmed with wave 1 hypothesis

Figure 7.4

Predictability of nodes measured by Bayes R^2 across waves



Discussion

The present study investigated the convergent validity of the domains of the TICS-M in a large ageing sample of older adults over a six-year period using network analyses. The networks showed consistent associations between particular TICS-M domains and neuropsychological domains across four waves of assessment. It might be expected that TICS-M domains would be significantly associated with their corresponding neuropsychological domains at baseline and over time. It could also be expected that the TICS-M orientation domain would not be associated with any of the neuropsychological nodes because the orientation domain was exclusive to the TICS-M and was added to increase the accuracy of the brief cognitive screen for detecting older people with cognitive impairments (Brandt et al., 1993; Folstein et al., 1975; O'Keeffe et al., 2011). However, the results did not fully meet these expectations. While the TICS-M memory and orientation domains remained significantly and consistently associated with their affiliated neuropsychological domains over six-years, the remaining TICS-M domains (i.e., language and attention domains) did not. For example, instead of being clearly linked to the neuropsychological domains of attention, the TICS-M attention domain was significantly related to the executive function and visuospatial neuropsychological domains instead and thus seems to capture important aspects of broader cognitive functioning. This pattern of results is not entirely surprising given attention, executive function and visuospatial ability are not mutually exclusive and do overlap to some extent for most tasks. These tasks are both numerical (which can overlap with spatial processing type functions mediated by the parietal lobe), and also involve some executive aspects (holding information in mind while manipulating it) where multiple circuits are exercised to complete each task (Cojan et al., 2021).

It is noteworthy that the majority of the associations that did emerge between TICS-M domains and neuropsychological domains were observed at least over two waves within the

six-year period. Specifically, the association between the TICS-M memory domain and the neuropsychological memory domain remained significant across all four waves. The associations between the TICS-M attention domain and neuropsychological visuospatial domain, and between the TICS-M language domain and the neuropsychological language domain, remained significant across three waves. The associations between the TICS-M attention domain and the neuropsychological executive function, and between the TICS-M language domain and the neuropsychological attention domain, remained significant across two waves. Interestingly, the TICS-M orientation domain did not have significant associations with most neuropsychological domains across all waves, except associations with neuropsychological visuospatial and memory at wave 3 and wave 4 respectively. Our findings regarding these consistent associations between the TICS-M domains internally and with neuropsychological domains over time can add more evidence to support the overall structural validity of the TICS-M screener in assessing enduring cognitive function, and are consistent with previous findings (Truong et al., 2022).

Moreover, the neuropsychological domains were interconnected to each other, and each node had high predictability within each wave, which was expected and in line with previous research (Harvey, 2022) and contrasts with the TICS-M domains, which were not. This finding may also raise concerns regarding internal consistency and validity of the individual domains of the TICS-M as these were expected to have stronger links with each other, and each TICS-M node had high predictability similar to neuropsychological tests. This is possibly due to the fact that the constructs of individual TICS-M domains were not cognitively comprehensive. For example, the TICS-M attention domain only has two items that are more likely focused on mathematical capacity rather than on the actual attention skill, and hence is only a proxy measure of such a skill. Therefore, the findings of this study revealed that global cognitive abilities captured by the TICS-M as a whole appeared more reliable compared to its individual

domains. These findings are in-line with a previous study, which suggested that the TICS-M may be more useful as a neuropsychological screener of global cognitive performance than as an indicator of domain-specific ability (Aiello et al., 2022). Thus, the individual TICS-M domains should be used cautiously as indicators of specific cognitive abilities.

We observed changes in the relationship between TICS-M domains and neuropsychological domains over the course of six years. Several factors could potentially explain these changes. Firstly, the aging process of participants during the MAS study may play a role, as cognitive abilities can decline or change with age (Folstein et al., 1975). Secondly, the presence of medical conditions or neurological disorders, which may emerge or progress over time, can impact the associations between cognitive measures (Knopman et al., 2009). Thirdly, differences between the assessment tools could contribute to these observed changes. For example, one assessment tool might be more sensitive to enduring aspects of cognition, while the other might capture more temporal fluctuations in cognitive performance. This would lead to differences in how the tools measure the associations between cognitive domains over time. Lastly, other external factors such as lifestyle changes, socio-economic status, and environmental factors may also influence the cognitive domains and their relationships over time. Taking these factors into account is important when interpreting the relationship between TICS-M domains and assessed neuropsychological domains at different time periods.

Our study is not without limitations. Our sample was not representative as participants were predominately Caucasian and well-educated and were recruited from a moderately affluent area of Sydney, Australia. Future studies should replicate our results using a more culturally, linguistically, and educationally diverse sample of older adults. In addition, the confirmatory analyses were conducted using the same sample but at later time points. It may be worth mentioning that there was non-random attrition between waves 1 and 4. That is,

participants who were younger, healthier, and more cognitively intact made it to wave 4 compared to those who dropped out, meaning, the confirmatory analyses could be biased by this. Therefore, it would be preferred to conduct the confirmatory analyses with independent samples of older persons. Moreover, we experienced a considerable decrease in sample size, with almost 50% of participants lost at the final wave. This reduction in sample size may have impacted the statistical power of our study, and hence caution should be exercised when interpreting our findings. Future research should consider strategies to enhance participant retention and minimize attrition, which could strengthen the validity and generalizability of the findings.

Conclusion

The weak internal links between the TICS-M domains remained across all four waves, and most associations between TICS-M and neuropsychological domains were observed at least over two waves within the six-year period. These results support the overall structural validity of the TICS-M screener in assessing enduring global cognitive function. However, the results also suggest that individual TICS-M domains should not be used as equivalent to the analogous neuropsychological domains.

Chapter 8 Examining the Validity of The Mini-Mental Status Examination (MMSE) and its Domains Using Network Analysis

Truong, Q. C., Cervin, M., Choo, C. C., Numbers, K., Bentvelzen, A. C., Kochan, N. A., Brodaty, H., Sachdev, P. S., & Medvedev, O. N. (2023). Examining the Validity of The Mini-Mental Status Examination (MMSE) and its Domains Using Network Analysis. Under review on *Psychogeriatrics.*

Declaration

This study is co-authored work with Cervin, M., Choo, C. C., Numbers, K., Bentvelzen, A. C., Kochan, N. A., Brodaty, H., Sachdev, P. S., and Medvedev, O. N. I was the first author and had the major contribution to this study that included designing and conducting the study, completing data analysis, writing the manuscript for submission to *Psychogeriatrics*, and responding to reviews/addressing their comments during peer-review process. All co-authors have signed the co-authorship form (see Appendix 8.1).

Abstract

***Background*:** The Mini-Mental Status Examination (MMSE) is the most widely used standardised screener for impairments across a range of cognitive domains. However, the degree to which its domains (orientation, registration, attention, recall, language, and visuospatial) capture cognitive functioning measured using standardized neuropsychological tests is unclear.

***Method*:** A longitudinal research design with four biannual assessments over a six-year period was used with an initial sample of 1037 older adults (aged above 70 years). Participants completed MMSE and neuropsychological tests at each assessment. Network analysis was utilised to investigate unique associations among the MMSE and its domains and neuropsychological test performance at each time point.

***Results*:** The total MMSE and two of its domains, language and recall, were associated with neuropsychological memory performance. The MMSE orientation, registration and visuospatial domains did not have any unique associations with neuropsychological performance. No stable internal interconnections between MMSE domains were found over time. The association of total MMSE as well as its recall domain with neuropsychological memory performance remained very similar over the six-year period.

***Conclusions*:** The present study adds evidence to the validity of the MMSE and support the clinical usage of the MMSE whereby the total score is used for screening patients with or without cognitive impairments; with repeated administration to monitor cognitive changes over time, to inform intervention. However, the tool is not able to diagnose the cases for changes in specific cognitive domains and as such, should not replace a complete neuropsychological assessment.

Keywords: cognitive domains, network analysis, the Mini-Mental Status Examination, neuropsychological assessments

Introduction

The Mini-Mental Status Examination (MMSE) is the most widely-used cognitive screener (Lacy et al., 2015). The MMSE is well-validated (Brayne, 1998; Norris et al., 2016) and relatively brief, which is suitable for screening cognitive impairments (e.g., dementia). The MMSE consists of 11 items grouped into six domains of orientation, working memory/registration, concentration/attention, recall, language, and visuospatial (Cameron et al., 2012; Ciolek & Lee, 2019; Folstein et al., 1975; Tierney et al., 1997). Previous studies have indicated that the MMSE achieved adequate internal consistency (Cronbach alphas of above 0.71), high test-retest coefficients (ranged from 0.80 to 0.89) and good inter-rater reliability (0.75) (El-Hayeck et al., 2019; Feeney et al., 2016; Folstein et al., 1975). It was also validated against more comprehensive cognitive assessments such as the Wechsler Adult Intelligence Scale – Revised (WAIS-R) verbal IQ (r = 0.84) and performance IQ (r = 0.51) (Folstein et al., 1975; Horton Jr et al., 1987; Mitrushina & Satz, 1991; Wechsler, 1981a).

More recently, a study conducted by Schmitt and colleagues (Schmitt et al., 2016) showed moderate correlations (0.41 to 0.49) between the MMSE total score and multiple cognitive domains indexed by the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph et al., 1998). Especially, the MMSE showed higher correlations with RBANS factor related to memory ability (r = 0.63) (Schmitt et al., 2016). Although the MMSE is a relatively brief assessment tool, it was originally designed to be a measure of global cognition covering a variety of cognitive domains (Cameron et al., 2012; Ciolek & Lee, 2019; Folstein et al., 1975; Ramirez et al., 2010). However, to the best of our knowledge, validations of the MMSE against neuropsychological tests have only considered the MMSE total score (Strauss et al., 2006).

Theoretical frameworks such as the Cattell-Horn-Carroll (CHC; McGrew, 2009) theory and the Hierarchical Model of Neuropsychological Functioning (Miyake & Friedman, 2012)

suggest that cognitive abilities can be organized into broad domains, with more specific abilities nested within each domain. These domains include attention/processing speed, language, executive function, visuospatial ability, and memory. Research has shown that these domains are not entirely independent and may share common underlying neural networks (Miyake & Friedman, 2012). For example, attention/processing speed and executive function have been linked to the prefrontal cortex, while language processing involves both temporal and frontal regions of the brain. Furthermore, deficits in one domain may affect performance in other domains (McGrew, 2009). For instance, individuals with executive function deficits may also have difficulties with attention and processing speed, while individuals with language deficits may struggle with both verbal memory and visuospatial ability. Thus, theoretical frameworks and empirical research has established organisation and links between neuropsychological cognitive domains. However, the validity of the MMSE domains is unclear and research on associations between its putative domains and domains captured by neuropsychological test performance is needed. A novel and suitable approach to examine unique association between the MMSE domains and performance-based measures of cognitive functions is provided by network analysis (Borsboom & Cramer, 2013).

Network analysis is an advanced statistical method that can provide a clearer understanding of unique associations among measured factors (e.g., cognitive domains/aspects) (Barcaccia et al., 2020; Borsboom & Cramer, 2013; Contreras et al., 2019). In network analysis, all factors are integrated into a single network, which provides a graphical representation of the network and the relations among the included factors (Chalmers et al., 2022; Cholerton et al., 2016). In a network, factors/variables are represented by “nodes” (circles) and relations between variables by “edges” (lines) (Borsboom et al., 2021). The thickness of an edge reflects the strength of each relation. Network analysis has several advantages over more "traditional" methods (e.g., regression and mediation/moderation

analyses). That is, it is more robust to outliers and can capture non-linear relationships that traditional correlation analysis may miss (Heeren & McNally, 2016). Network analysis also provides a graphic network includes multiple interactions between nodes/variables simultaneously without assigning dependent and independent variables, making it easier to identify important variables and understand the data's structure (Åkerblom et al., 2021; Chalmers et al., 2022).

The main aim of this study was to explore unique associations between the MMSE domains and comparable domains captured by neuropsychological test performance using a large six-year longitudinal sample of older adults. We would also examine unique associations between the total MMSE score and neuropsychological test performance. As to our knowledge there has been no previous studies on the validity of the MMSE domains, we would use a combination of exploratory and confirmatory analyses. First, baseline data (i.e., wave 1) would be used to conduct exploratory network analyses, which would be used to form hypotheses. Then, data collected at three follow-up biannual waves (i.e., waves 2, 3, and 4) would be used to carry out confirmatory network analyses to explore whether associations found in exploratory analyses are replicated across three follow-up waves.

Method

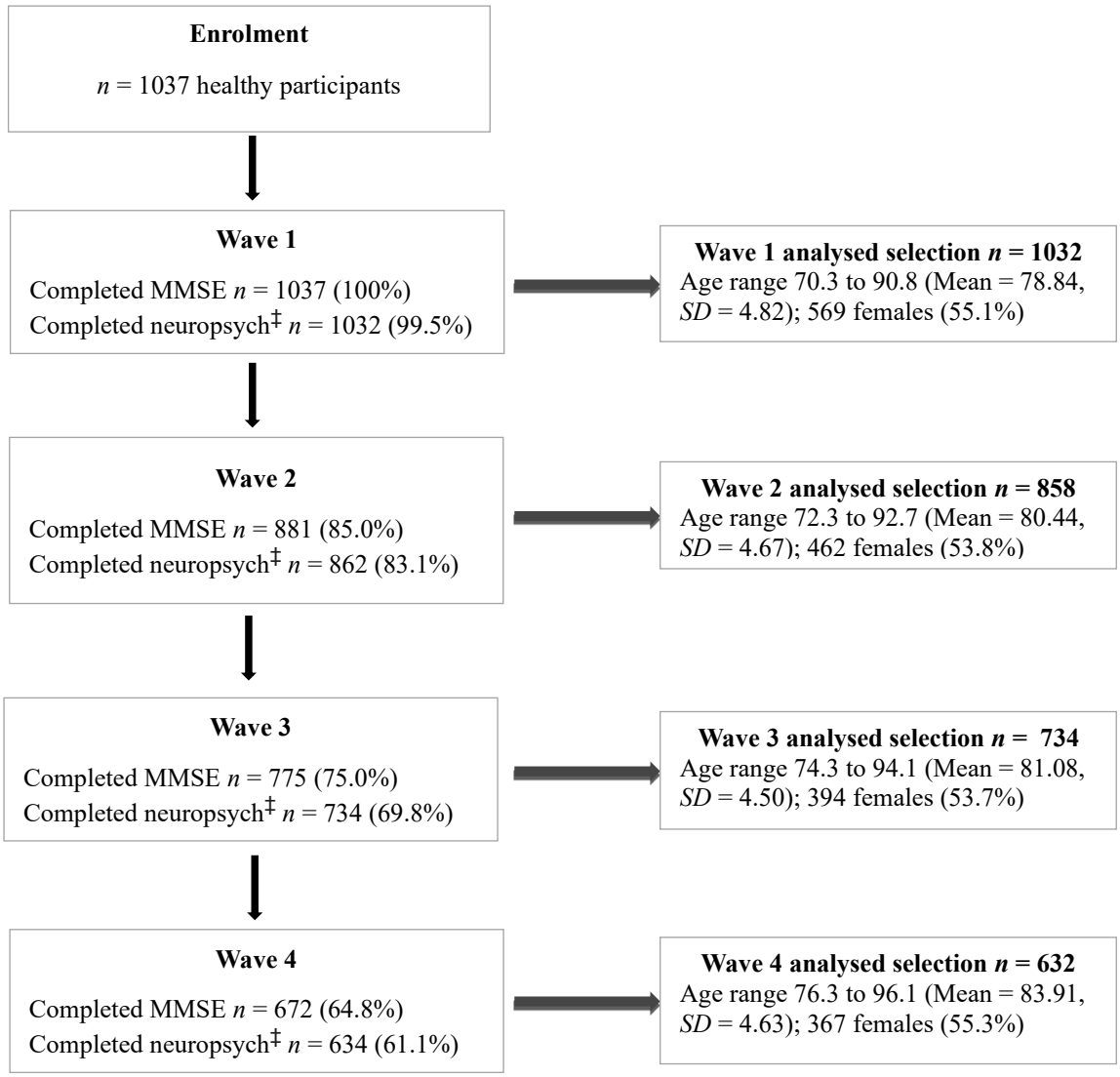
Participants

Participants were from the longitudinal Sydney Memory and Ageing Study (MAS) that included a baseline sample of 1037 older adults aged 70 to 90 (Sachdev et al., 2010). MAS participants were from the Eastern Suburbs of Sydney, Australia, and had appropriate language abilities to complete psychometric assessments (Sachdev et al., 2010). The major ethnicity of the MAS participants was European (98%); the remaining sample was 1.1% non-European and 0.8% not revealed. Participants were interviewed every two years from wave 1 to wave 4 (six-year follow-up). To be included in the MAS at the baseline assessment (wave 1), participants could not have a current or previous diagnosis of dementia, major psychological or neurological disorder, or progressive malignancy. More detailed information about how participants were recruited and their demographics in the MAS can be found in Sachdev and colleagues (Sachdev et al., 2010). All participants provided written consent to participate in this study, which was approved by the University of New South Wales Human Ethics Review Committee (HC 05037, 09382, 14327).

CONSORT diagram of how participants were selected at each wave for network analyses is presented in Figure 8.1. Demographic details of participants (i.e., age and sex) at each wave are also displayed in this figure. Independent samples t-tests revealed no significant difference in participant age between females and males across waves (all p's > 0.05). There were missing data at each wave if participants were not well, had passed away or otherwise unable to complete the assessment at that wave or were not contactable.

Figure 8.1

CONSORT diagram for participants selected at each wave



Note: ‡neuropsychological assessments

Measures

MMSE. The MMSE (Folstein et al., 1975) is an 11-item screening tool designed to capture global cognition and cognitive functions across six domains (i.e., orientation, registration, attention, recall, language, and visuospatial). Individual domain scores are calculated by adding responses of relevant item(s) together (Cameron et al., 2012; Ciolek & Lee, 2019; Tierney et al., 1997). The maximum scores for individual domains vary (Appendix 8.2), for example, the *language* domain has a maximum score of eight, while the visuospatial domain has a maximum score of one. Total MMSE scores are calculated by summing all MMSE items, with scores ranging from 0 to 30.

Neuropsychological Cognitive Assessment. Performance across five neuropsychological domains (i.e., attention/processing speed, language, executive function, visuospatial ability, and memory) was assessed at wave 1 to wave 4 using a comprehensive neuropsychological battery comprised of 10 tests (Sachdev et al., 2010; Trollor et al., 2012). The Wechsler Adult Intelligence Scale-III Digit Symbol-Coding (Wechsler, 2003) and the Trail Making Test A (Tombaugh, 2004) were used to assess attention/processing speed. The Boston Naming Test (Kaplan et al., 2001) and Animal Fluency (Spreen, 1977) assessed language. The FAS Fluency (Benton, 1967) and Trail Making Test B (Reitan & Wolfson, 1993) assessed executive function. The Wechsler Adult Intelligence Scale-Revised (WAIS-R) Block Design (Wechsler, 1981b) assessed visuo-spatial ability. Memory was assessed by the Wechsler Memory Scale-III (WMS-III) Logical Memory Story A delayed recall (Wechsler, 1997), Rey Auditory Verbal Learning Test (Strauss et al., 2006) total learning, short-term and long-term recall scores, and the Benton Visual Retention Test (Benton et al., 1996). Composite domain scores were computed for each cognitive domain as follows. Raw test scores were transformed into quasi z-scores by using baseline means and standard deviations (SDs) from a reference group comprised of 732 MAS participants classified as cognitively healthy at wave

1 (Sachdev et al., 2010). Z-scores from the neuropsychological tests relevant to each neuropsychological domain were then averaged to compute composite z-scores for each domain. The exception was the visuospatial domain, which was comprised of a single test. Higher z-scores indicated better cognitive function.

Data Analyses

Descriptive statistics, including mean, standard deviation (*SD*), skewness and kurtosis for the domain scores, were computed by IBM SPSS Statistics 28 software. Network analysis was conducted using R software (version 4.0.4; R Core Team, 2021) with the package *BGGM*, short for Bayesian Gaussian Graphical Models (Williams & Mulder, 2020). The Copula Gaussian graphical model estimation was selected to estimate unique associations between nodes because several variables in the study were non-normally distributed. The networks were illustrated by Gaussian graphical models (GGM), which graphically present the statistically significant relations between variables in the form of nodes and edges. The nodes represent the variables, and the edges represent unique associations in the form of partial correlations (accounting for all associations among the full set of variables) between nodes and range from -1 (perfect negative association) to +1 (perfect positive association). The strength of each relationship is reflected by the thickness of the edge. Blue lines represent positive relationships and red lines represent negative relationships. For clarity, the MMSE nodes (i.e., total MMSE node and MMSE domain nodes) and the neuropsychological nodes (i.e., attention/processing speed, language, executive function, visuospatial ability, and memory) were assigned different colours. To select which unique relations were statistically significant, we estimated a 95% credible interval (*CI*) for each association. A CI that excluded zero was considered to indicate a statistically significant edge between nodes. A CI is a range of scores where a specific relation will fall 95% of the time (Hespanhol et al., 2019).

The networks were plotted using the R library *qgraph*, which is based on the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991). This algorithm places nodes that are strongly connected with other nodes centrally in the network and nodes with strong connections to each other closely, whilst avoiding overlap of nodes and edges (Epskamp et al., 2012). To facilitate comparisons across all four waves in this study, the placement of each node remained the same in all networks by using the averaged network layout, though the inclusion/exclusion (i.e., contingent on whether CIs included zero) and thickness of edges differed (Fruchterman & Reingold, 1991). Strength centrality using the R library *qgraph* was also estimated, which identified how important a node was to each network by counting the number of both negative and positive edges a node had (Åkerblom et al., 2021; Robinaugh et al., 2016). Strength centrality has been argued to show the most robustness (Chalmers et al., 2022; Jones et al., 2017). A node with high centrality has strong links with other nodes in the network and thus acts as a more central node (Haslbeck & Waldorp, 2018; Robinaugh et al., 2016).

As introduced, we used a statistical framework that included both exploratory and confirmatory analyses. The exploratory analyses were conducted to explore the network relations between nodes by using the sample of participants at wave 1. These initial analyses allowed us to explore edges between the total MMSE node as well as its domain nodes (i.e., MMSE orientation, registration, attention, recall, language, and visuospatial nodes) and neuropsychological domains/nodes. Post hoc tests were then used (again using wave 1 data) to test whether graphical patterns in the network could be confirmed statistically. True confirmatory analyses were then subsequently conducted using data from the follow-up waves (i.e., wave 2, 3, and 4) by calculating the posterior probability (*PP*). A *PP* indicates the probability of a pre-specified event occurring, for example, that the edge between nodes A and B is larger than the edge between nodes A and C. *PPs* above 0.95 (95%) were used as indicators

of a confirmed hypothesis in the present study. To calculate *PPs*, 5,000 posterior samples of each edge were estimated and then used for each specific hypotheses (i.e., that the edge between nodes A and B is larger than the edge between nodes A and C).

Results

Descriptive Statistics

Table 8.1 presents descriptive statistics including mean, standard deviation, skewness and kurtosis for the MMSE and neuropsychological nodes across waves. As most MMSE and several other neuropsychological tests had high skewness and/or kurtosis that suggested non-normal distributions (Muthén & Kaplan, 1985), Copula Gaussian graphical model estimation were appropriate for estimating associations between nodes. Table 8.2 displays the means, standard deviations, and results of statistical comparisons using two-way repeated ANOVAs. The independent variables were the scores of MMSE node, its domain nodes, and neuropsychological nodes. The first dependent variable consisted of two subsamples of participants: those diagnosed with dementia after six years (n = 48) and those who remained healthy/not diagnosed with dementia (n = 584). The second dependent variable was the four assessment waves. The results indicate significant interaction effects between subsamples and waves on all neuropsychological nodes (all $p < 0.001$) and the majority of the MMSE nodes (p's ≤ 0.04), except for the scores of MMSE registration ($p = 0.37$) and language ($p = 0.84$). Post-hoc tests revealed that the dementia subsample scored significantly lower compared to the healthy subsample across all waves on all neuropsychological nodes ($p < 0.001$), as well as the scores of MMSE and its recall nodes (p's < 0.001). However, the MMSE orientation, attention, and visuospatial nodes were only significantly lower in the dementia subsample compared to the healthy subsample at waves 2, 3, and 4. Additionally, all nodes (MMSE node, its domain nodes, and neuropsychological nodes) generally showed a decrease over the six-year period within both subsamples. Notably, the decrease was more pronounced in the dementia subsample compared to the healthy subsample. These findings indicate significant differences between changes of cognitive scores over the six-year period of those diagnosed with dementia at wave 4 compared to healthy participants.

Table 8.1

Descriptive statistics including means (M), standard deviation (SD), Skewness, and Kurtosis for the domain scores of neuropsychological assessments, and the MMSE on the four waves.

	Wave 1				Wave 2				Wave 3				Wave 4			
Scale/Domains	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis
MMSE																
Orientation	9.50	0.71	-1.36	1.72	9.58	0.77	-2.43	8.36	9.68	0.65	-2.84	8.36	9.63	0.75	-2.93	11.90
Registration	2.97	0.22	-8.31	86.39	2.97	0.18	-6.01	38.25	2.99	0.15	-14.23	242.59	2.97	0.19	-6.03	39.20
Attention	4.85	0.52	-4.30	22.26	4.91	0.37	-5.40	34.83	4.93	0.34	-5.94	39.07	4.89	0.40	-4.19	18.79
Language	7.20	0.72	-0.53	-0.25	7.23	0.81	-0.96	0.79	7.20	0.78	-0.79	0.31	7.09	0.78	-0.70	0.56
Visuospatial ability	0.94	0.24	-3.73	11.94	0.94	0.24	-3.65	11.32	0.95	0.21	-4.27	16.29	0.91	0.28	-2.95	6.74
Recall	2.63	0.61	-1.54	1.90	2.63	0.64	-1.79	2.99	2.62	0.68	-1.91	3.44	2.55	0.74	-1.62	1.89
Total MMSE	28.00	1.51	-0.88	0.99	28.10	1.87	-1.78	5.89	28.04	1.77	-2.12	8.72	27.88	1.92	-1.44	2.39
Neuropsychology																
Attention/processing speed	-0.06	1.04	-0.67	1.71	-0.13	1.21	-1.97	12.94	-0.26	1.21	-1.71	1019	-0.46	1.27	-1.47	6.41
Language	-0.08	1.03	-0.49	0.64	-0.19	1.05	-0.63	0.54	-0.20	1.09	-0.60	0.75	-0.28	1.16	-0.77	0.93
Executive function	-0.06	1.02	-0.57	1.05	-0.18	1.19	-1.32	4.38	-0.19	1.09	-0.69	1.11	-0.41	1.47	-1.54	4.41
Visuospatial ability	-0.04	1.02	0.22	-0.07	-0.01	1.08	0.08	0.01	0.03	1.05	0.11	0.20	-0.18	1.14	-0.01	-0.11
Memory	-0.06	1.01	-0.04	-0.30	-0.11	1.06	-0.01	-0.33	-0.06	1.10	-0.27	-0.20	-0.19	1.17	-0.15	-0.54

Table 8.2

Descriptive statistics including means (M) and standard deviation (SD), and statistical comparisons using repeated ANOVAs (interactional effect p values) for the domain scores of neuropsychological assessments, and the MMSE on the four waves in the subsample of participants who were diagnosed with dementia after the six-year period (n = 48) and that of those who were healthy/not dementia (n = 584*)*

Scale/Domains	Subsample	Wave 1 M(SD)	Wave 2 M(SD)	Wave 3 M(SD)	Wave 4 M(SD)	*p*
MMSE						
Orientation						
	Dementia	9.35 (0.73)	9.29 (1.01)	9.16 (1.18)	8.25 (1.56)	<0.001
	Healthy	9.54 (0.69)	9.66 (0.66)	9.74 (0.49)	9.72 (0.55)	
Registration						
	Dementia	2.96 (0.20)	2.96 (0.20)	2.98 (0.15)	2.92 (0.28)	0.370
	Healthy	2.97 (0.21)	2.97 (0.18)	2.99 (0.09)	2.96 (0.21)	
Attention						
	Dementia	4.73 (0.61)	4.83 (0.63)	4.84 (0.43)	4.45 (0.9)	<0.001
	Healthy	4.85 (0.51)	4.93 (0.30)	4.94 (0.28)	4.90 (0.40)	
Language						
	Dementia	6.94 (0.81)	6.90 (0.78)	6.84 (0.91)	6.58 (1.09)	0.840
	Healthy	7.27 (0.72)	7.24 (0.82)	7.20 (0.77)	7.07 (0.79)	
Visuospatial ability						
	Dementia	0.92 (0.28)	0.90 (0.31)	0.91 (0.29)	0.77 (0.42)	0.04
	Healthy	0.93 (0.25)	0.95 (0.22)	0.95 (0.21)	0.91 (0.29)	
Recall						
	Dementia	2.42 (0.71)	2.23 (0.88)	1.86 (0.98)	1.40 (1.09)	<0.001
	Healthy	2.72 (0.53)	2.71 (0.54)	2.67 (0.58)	2.60 (0.65)	
Total MMSE						
	Dementia	27.31 (1.70)	27.10 (1.57)	26.59 (2.18)	24.27 (2.70)	<0.001
	Healthy	28.27 (1.40)	28.45 (1.47)	28.50 (1.33)	28.15 (1.56)	
Neuropsychology						
Attention/processing speed						
	Dementia	-0.54 (0.91)	-0.81 (1.08)	-1.58 (1.79)	-2.47 (1.99)	<0.001
	Healthy	0.10 (0.91)	0.01 (1.03)	-0.15 (1.04)	-0.40 (1.12)	
Language						
	Dementia	-0.73 (1.06)	-0.99 (1.17)	-1.54 (0.99)	-2.04 (1.18)	<0.001
	Healthy	0.06 (1.04)	-0.08 (1.00)	-0.12 (1.03)	-0.26 (1.09)	
Executive function						
	Dementia	-0.44 (1.08)	-1.02 (1.55)	-1.29 (1.41)	-2.82 (2.25)	<0.001
	Healthy	0.11 (0.95)	0.01 (0.98)	-0.08 (0.99)	-0.29 (1.27)	
Visuospatial ability						
	Dementia	-0.42 (0.88)	-0.45 (0.98)	-0.73 (1.03)	-1.39 (1.25)	<0.001
	Healthy	0.12 (1.00)	0.07 (1.05)	0.08 (1.02)	-0.16 (1.07)	
Memory						
	Dementia	-0.86 (1.02)	-1.24 (0.91)	-1.75 (0.86)	-2.18 (0.56)	<0.001
	Healthy	0.14 (0.96)	0.06 (0.97)	0.06 (1.01)	-0.10 (1.10)	

Exploratory Analyses

The exploratory network for the total MMSE node and neuropsychological assessment nodes at wave 1 are presented in Figure 8.2. The total MMSE node had statistically significant edges to most neuropsychological nodes (i.e., edge weights ranging from 0.34 to 0.42), except for the neuropsychological attention node. However, results from post hoc tests indicated that the edge between the total MMSE node and the neuropsychological memory node was stronger than between the total MMSE and all other neuropsychological nodes (all *PPs* = 100%). It means that only the significant edge between the total MMSE node and the neuropsychological memory node was confirmed.

Figure 8.3 displays the exploratory network for MMSE domain nodes and neuropsychological domains/nodes at wave 1. The neuropsychological nodes showed many edges between nodes internally were statistically significant (i.e., edge weights ranging from 0.27 to 0.59) meaning that they were highly interconnected. Conversely, the MMSE domain nodes were not interconnected to each other with most internal edge weights being close to zero, which means that domains are not sharing common variance as would be expected for cognitive domains based on theoretical and empirical evidence.

The MMSE language node had significant edges with neuropsychological language node (edge weight of 0.29, 95% *CI* [0.057, 0.213]) and neuropsychological memory node (edge weight of 0.27, 95% *CI* [0.049, 0.220]). However, post-hoc tests based on the posterior probability (*PP*) estimates only statistically confirmed the edge between the MMSE language node and the neuropsychological memory node (all *PPs* > 96.2%). In addition, the MMSE recall node had a significant edge with the neuropsychological memory node (edge weight = 0.37, 95% *CI* [0.253, 0.420]) and this edge was confirmed with all *PPs* > 99.9%. Furthermore, the MMSE orientation node had a significant edge with neuropsychological memory node (edge weight = 0.37, 95% CI [0.253, 0.420]) and the MMSE registration node had a significant

edge with neuropsychological attention node (edge weight = 0.37, 95% *CI* [0.253, 0.420]). However, post-hoc tests did not confirm these two edges as most *PPs* were below 95.0%. Moreover, Figure 8.3 also shows that no associations between the MMSE visuospatial node and other neuropsychological nodes. Furthermore, centrality of nodes was also estimated at this wave, which is presented in Appendix 8.3. As can be seen, all neuropsychological nodes had extensively higher centrality compared to all MMSE nodes. This suggests that all neuropsychological nodes reflected cognitive abilities that were more influential at wave 1.

Figure 8.2

Exploratory network of neuropsychological domains and the total MMSE at Wave 1

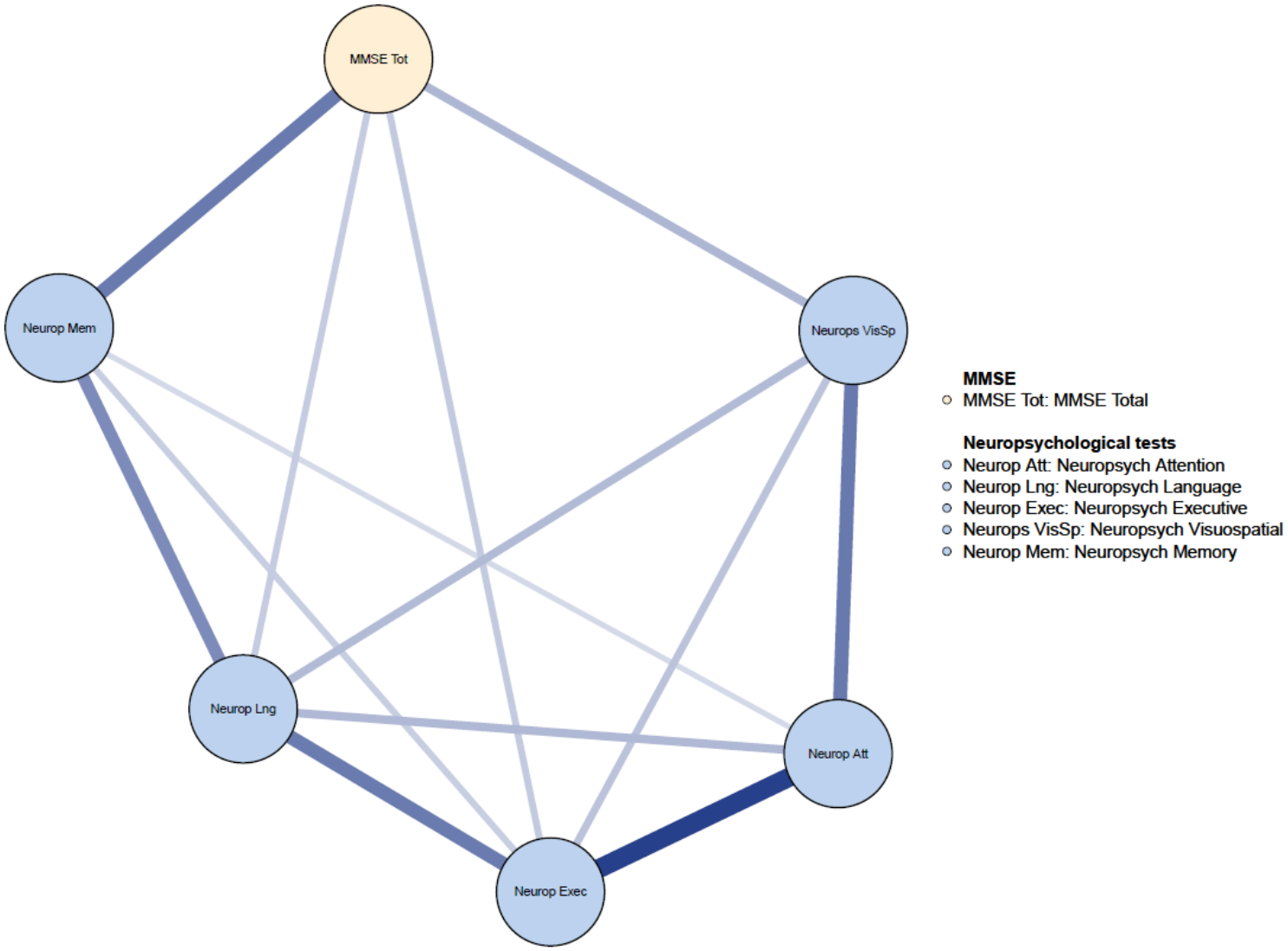
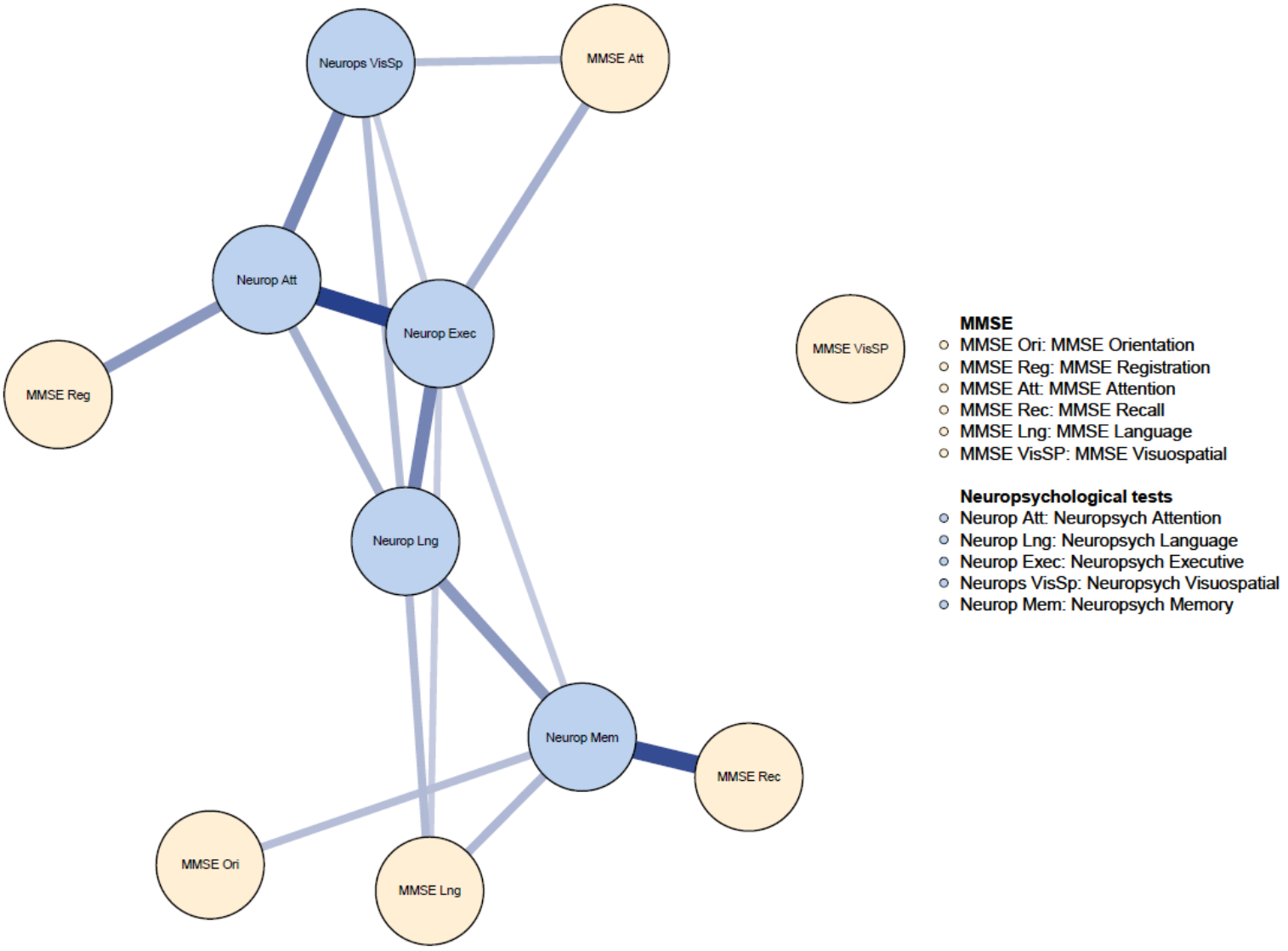


Figure 8.3

Exploratory network of neuropsychological domains and MMSE domains at Wave 1.



In sum, our exploratory network results allowed us to form hypotheses that 1) there would be a link between the total MMSE node and the neuropsychological memory node; 2) there would be no links between the MMSE nodes internally; 3) the MMSE language node would be more strongly related to neuropsychological memory node than to the other neuropsychological nodes; 4) the MMSE recall node would be more strongly related to neuropsychological memory node than to the other neuropsychological nodes; 5) the MMSE orientation, registration and visuospatial nodes would not have any associations with any of the neuropsychological nodes; and 6) the neuropsychological nodes would be more central or influential in comparison to the MMSE nodes on the network. Next, confirmatory analyses using data collected at waves 2 – 4 were conducted to test these hypotheses.

Confirmatory Analyses

Figures 8.4 and 8.5 present the estimated networks with averaged network layouts for the total MMSE node, as well as its domain nodes, and the neuropsychological nodes across the three follow-up waves. As shown, the networks for follow-up waves only partly confirmed our hypotheses formed at wave 1. More specifically, hypothesis 1 that the links between the total MMSE node and the neuropsychological memory node was confirmed in waves 2 to 4. Secondly, hypothesis 2 that there would be no links between the MMSE nodes internally was not confirmed in confirmatory analyses; rather, while MMSE nodes were internally associated across waves 2 to 4, these MMSE internal associations were unstable over time. Thirdly, hypothesis 3 that the MMSE language node would be more strongly related to neuropsychological memory node as compared to the other neuropsychological nodes was not confirmed across waves as there were no associations between the MMSE language nodes and the neuropsychological memory nodes at waves 3 and 4. However, hypothesis 4 was confirmed across waves as the association between the MMSE recall node and the neuropsychological memory nodes remained unchanged over the six-year period.

Furthermore, hypothesis 5 was not confirmed across follow-up waves as significant associations between the MMSE orientation, registration and visuospatial nodes emerged in regard to other neuropsychological nodes across waves. For example, there were significant associations between the MMSE visuospatial node and the neuropsychological visuospatial node at wave 2, and between the MMSE orientation node and the neuropsychological memory node at wave 3. Lastly, hypothesis 6 that the neuropsychological nodes would be more influential in comparison to the MMSE nodes on the networks was confirmed. In fact, most links between the neuropsychological nodes internally remained stable and most neuropsychological nodes had extensively higher centrality compared to all MMSE nodes (see Appendix 8.3) across wave 2-4.

In summary, the confirmatory results indicated that three out of six hypotheses, including hypothesis 1, 4 and 6, were fully confirmed across follow-up waves.

Figure 8.4

Confirmatory networks of neuropsychological domains and the total MMSE at follow-up waves.

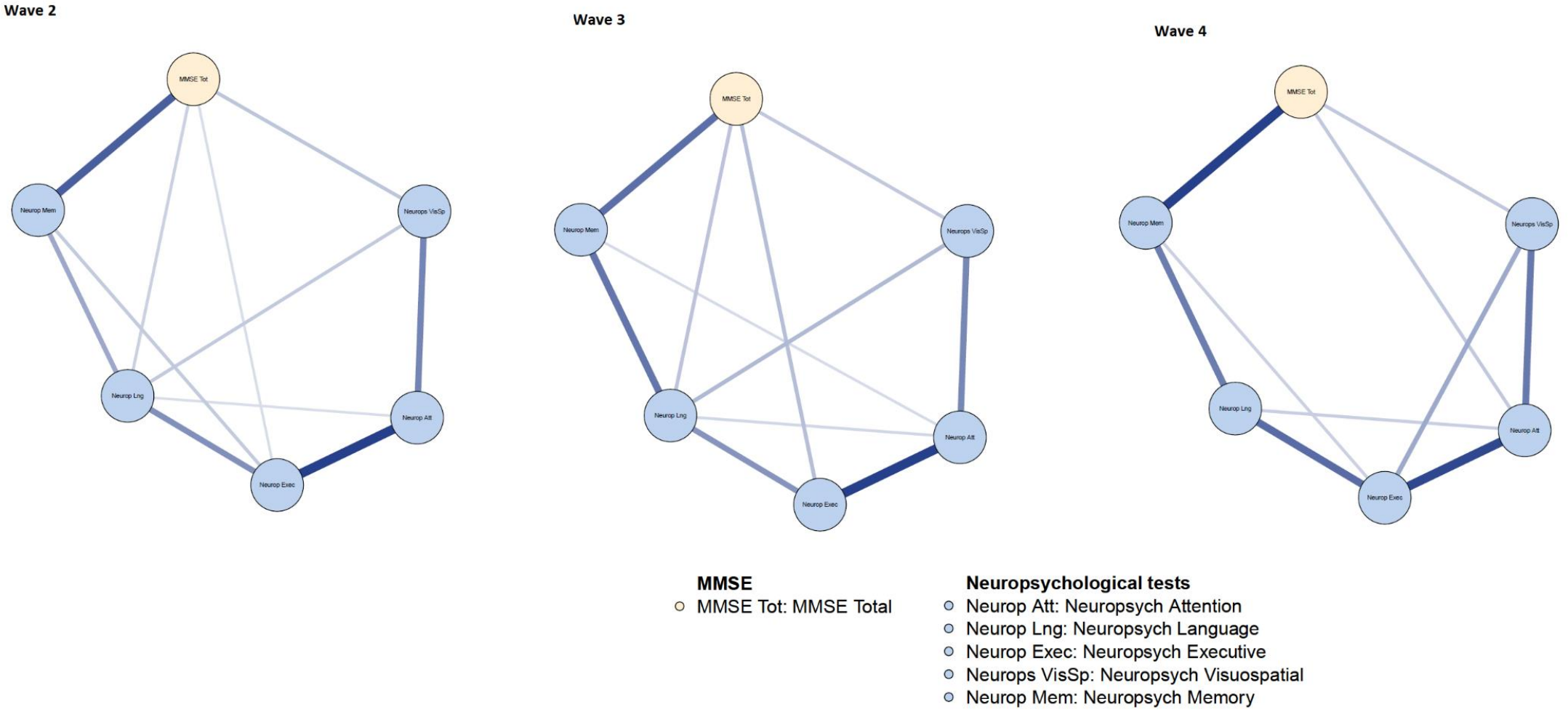
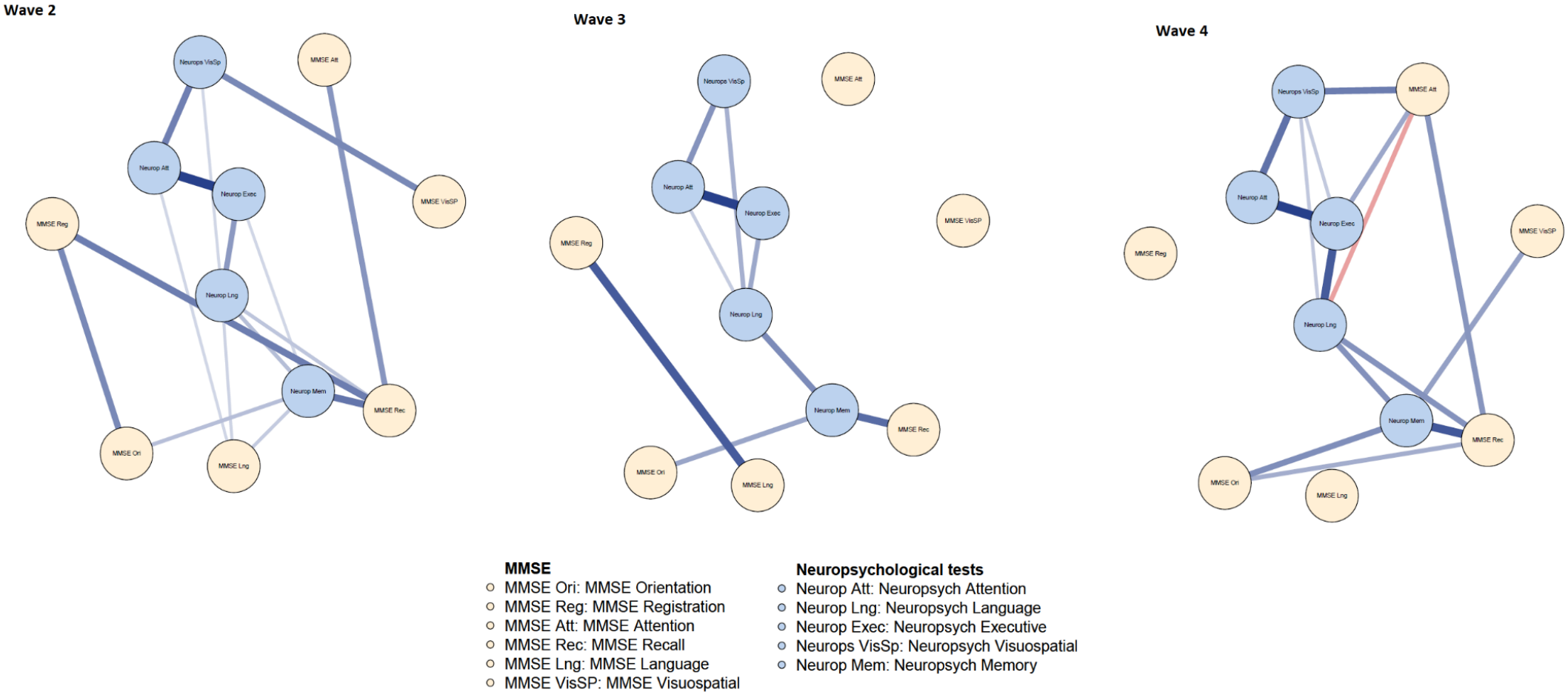


Figure 8.5

Confirmatory networks of neuropsychological domains and MMSE domains at follow-up waves.



Discussion

The present study applied network analysis in a novel way to investigate the validity of a standardised cognitive screener, the MMSE, in a large ageing sample using longitudinal data from four assessments collected over a six-year period. Overall, the total MMSE score was consistently associated with the neuropsychological memory node over time. In other words, our finding revealed that the total score of the MMSE appears to capture mostly the memory aspect of neuropsychological test performance. This is in line with previous research that the correlation between the MMSE and memory function captured by comprehensive neuropsychological tests was higher compared to that between the MMSE and other cognitive abilities (Schmitt et al., 2016). This indicates that the MMSE will continue to be the common and first-line screening tool for cognitive impairments in older adults because memory loss is commonly used as a criterion to detect abnormal cognitive ageing conditions such as dementia (American Psychiatric Association, 2013).

However, no robust relations between the individual MMSE domains and the neuropsychological domains appeared over the same period (6 years). Moreover, the individual neuropsychological domains showed consistent internal edges (i.e., significant associations) across waves, but the six domains of the MMSE were not consistently linked to each other internally across the four waves. The links found between the MMSE domains lacked stability over time and most of them (except the MMSE recall domain at waves 2 and 4) had considerably lower centrality within each wave as compared to the coherent and stable internal relations of the neuropsychological nodes. Since the MMSE is a brief cognitive screener, its individual domains might not represent cognitively comprehensive constructs, especially given each domain is comprised of very few items, and hence its domains cannot share common variance as would be expected for cognitive domains based on theoretical and empirical evidence (McGrew, 2009; Miyake & Friedman, 2012). For example, the MMSE visuospatial

domain had only one item requiring participants to copy a pentagon, which was only a proxy measure of visuospatial function rather than adequately represented the actual skill. Therefore, the MMSE domains should be interpreted cautiously, which is in line with previous recommendation that the domains of the MMSE are not empirically derived and may not be valid indicators of specific cognitive domains (Strauss et al., 2006).

It is noteworthy that the association between the MMSE recall domain and neuropsychological memory domain remained consistent over the six-year period. The other associations between the MMSE and neuropsychological domains found in wave 1 tended to change over time. In the same way, links between the MMSE domains changed across waves. For example, there were significant internal edges between MMSE registration, orientation, and recall domains at wave 2. However, these edges were changed at the followed-up waves (i.e., wave 3 and 4). At wave 3, the only internal association between MMSE domains that reached significance was between registration and language. At wave 4, two internal MMSE associations emerged between orientation and recall, and between attention and recall. These appear to confirm that the MMSE is capturing cognitive function that is sensitive to change over time, which is consistent with previous findings (Truong et al., 2022).

It should also be noted that the interconnections between neuropsychological domains remained stable over most waves. This result is consistent with theoretical frameworks (McGrew, 2009; Miyake & Friedman, 2012) and in line with a recent literature reviews (Harvey, 2022). These data add more empirical evidence that neuropsychological domains including attention/processing speed, language, executive function, visuospatial ability, and memory may share features (Miyake & Friedman, 2012). This finding also contributes to the body of knowledge supporting Spearman's two-factor theory of intelligence that suggests overarching general intelligence ("g") captures specific abilities ("s") in older individuals (Spearman, 1927). Specifically, high centrality of all neuropsychological domains and stable

links between the domains support the theory that older people with higher overall "g" – or high general intelligence – are more capable of other cognitive abilities. At the same time, different associations found between distinct domains support the presents of "s" abilities in older populations as well.

Moreover, our results indicate differences in score shift patterns between the healthy and dementia subsamples. This shows that the progression to dementia adversely influenced the cognitive performance of participants, as captured by the MMSE and neuropsychological tests. Therefore, this finding enhances our understanding of the significance of identifying potential differences that could highlight early signs of cognitive decline and, ultimately, dementia. However, these results should be interpreted cautiously as we experienced a high attrition rate in our sample size, with only 60% of participants having data for all four waves. This attrition may have impacted the statistical power of the results.

Our study had some limitations, which should be acknowledged. Participants were predominately white, European, and well-educated and recruited from an affluent area of Sydney, Australia, and hence our sample is not representative of the general population of older adults or older Australian adults specifically. This may bias the results of neuropsychological tests. Therefore, future studies should replicate this study with more diverse and representative samples of older adults. In addition, the confirmatory analyses were conducted entirely within the same sample that was used to derive our initial hypotheses (although at different time points). Confirmatory tests in true independent samples would have been a stronger source of evidence and should be explored by future studies.

In conclusion, the MMSE is a commonly used standardised cognitive screener that systematically assess cognitive function including: orientation, registration, attention and calculation, recall, and language. The results from the present study add evidence to the validity of the MMSE and support the clinical usage of the MMSE whereby the total score is used for

screening patients with or without cognitive impairments; with repeated administration to monitor cognitive changes over time, to inform intervention. However, the tool is not able to diagnose the cases for changes in specific cognitive domains and as such, should not replace a complete neuropsychological assessment.

Chapter 9 General Discussion

This thesis aimed to evaluate and enhance reliability and validity of key assessment instruments used to assess cognitive ageing. The research studies used advanced statistical and psychometric methodology and contributed to the improvement of measurement directly relevant to cognitive function in aging populations. Specifically, the first relevant variable used was subjective cognitive complaints (also known as cognitive change in older people) captured by the MAC-Q (Crook et al., 1992) and the 16-item IQCODE-16 (Jorm, 1994). Another key variable was global cognition, which was assessed in older people using the MMSE (Folstein et al., 1975) and the TICS-M (Brandt et al., 1993). Besides that, the data of dementia or MCI diagnoses and neuropsychological assessments over time were also frequently harnessed for examining predictive validity of psychometric instruments. Both the total neuropsychological scores, and composites scores of neuropsychological domains (i.e., attention/processing speed, language, executive function, visuospatial ability, and memory) were used for evaluating the predictive and construct validity of key measures used to assess cognitive ageing in older adults.

The studies included in this thesis used pre-existing datasets such as the longitudinal dataset of the Memory and Ageing Sydney Study (Sachdev et al., 2010). Moreover, two more longitudinal datasets, the Older Australian Twins Study (OATS; Sachdev et al., 2009) and the Canberra Longitudinal Study (CLS; Christensen et al., 2004; Jorm et al., 2000) were used in the study forming Chapter 5 (Truong, Numbers, et al., 2023), with a view to further enhance the validity of the IQCODE-16 that were established in the preceded study presented in Chapter 4 (Truong et al., 2021b).

Findings from G-Theory Studies and Their Implications

The first objective of this thesis was to apply G-Theory to evaluate sources of measurement error affecting reliability and investigate dynamic and enduring aspects of assessments of cognitive ageing. The clear differentiation of enduring and dynamic aspects of assessments of cognitively ageing would help to monitor changes in both these aspects, and was necessary to develop and implement the most appropriate diagnostic tools, as well as to facilitate preventive and targeted interventions for conditions caused by ageing. This objective was satisfactory achieved through two studies that estimated reliability, sources of measurement error, and enduring and dynamic patterns of subjective cognitive complaints (SCC; also known as cognitive changes in older people) and global cognitive performance, respectively (Truong et al., 2021a, 2022).

The first study (Chapter 2, Truong et al., 2021a) aimed to assess the reliability, sources of measurement error, and differentiate between dynamic and enduring patterns of the MAC-Q and IQCODE-16 measures of SCC using G-Theory. The findings indicated that measurement error was predominantly explained by person and item, and person and occasion interactions for the MAC-Q and the IQCODE-16, but measurement error was small after accounting for the true person variance. Therefore, both measures were reliable in capturing the enduring pattern of SCC, suggesting the generalizability of assessment scores across populations and occasions. However, compared to the MAC-Q scores, which were only significant predictors of dementia and global cognition on certain occasions, the IQCODE-16 scores were more consistent in predicting dementia risk and global cognition. It is possible that the self-report MAC-Q might only assess tendencies of individuals' complaints, such as self-perceived memory errors, and thus may not precisely capture individuals’ actual cognitive abilities. In contrast, the IQCODE-16 was reported by an individual’s close informant who can help to detect early cognitive decline, because as dementia progresses, individuals may lose awareness

of these changes (American Psychiatric Association, 1994). These results suggest that while the MAC-Q is a reliable measure of individual SCC levels, this self-reported SCC may be less accurate in capturing actual cognitive abilities and as such, may be less useful to facilitate diagnosis, as compared to the IQCODE-16. Therefore, this study adds evidence for the superiority of the IQCODE-16 over the MAC-Q in assessing SCC to screen for cognitive impairments and monitor cognitive changes over time. Moreover, this study also has clinical implications, emphasizing the importance for clinicians to seek information from knowledgeable informants in their evaluation of aging patients. This can aid in the early detection of Mild Cognitive Impairment and inform clinical management and planning.

The second study (Chapter 3, Truong et al., 2022) aimed to examine reliability, sources of measurement error, and distinguish between the dynamic and enduring aspects of global cognition as measured by the two major assessment tools, the MMSE and the TICS-M, by using G-Theory and to evaluate sensitivity and specificity of the MMSE and the TICS-M in detecting dementia. The MMSE (Folstein et al., 1975) is a widely used standardised cognitive assessment tool, while the TICS-M (Brandt et al., 1993) is a well-validated alternative tool suitable for administration through both for telephone or face-to-face methods. The results found that the TICS-M was highly sensitive and specific in detecting dementia, while the MMSE exhibited low sensitivity and high specificity. These results affirm the advantages of employing the TICS-M in extensive epidemiological studies due to its time and cost efficiency, as well as its high sensitivity and specificity in screening dementia, which aligned with previous research (Bentvelzen et al., 2019; Fong et al., 2009; Wilson & Bennett, 2005). In addition, these findings are also in line with previous research illustrating high specificity of the MMSE, but contradict research that found low sensitivity for the MMSE (O'Bryant et al., 2008; Weston, 1987). These contradict findings for the MMSE may be due to study design differences, such

as varying sample sizes, or sample containing highly homogeneous of participants (e.g., well-educated participants).

The second study also found that global cognition captured by the MMSE was explained by 56% of variance of enduring aspects and 44% of dynamic aspects. This indicates that the MMSE captures both dynamic and enduring aspects of cognition and making it suitable for clinical assessments. Additionally, a shorter version of the MMSE (MMSE-D8) was developed after removing three items, which was useful for capturing dynamic aspects of cognition. The TICS-M achieved acceptable reliability and generalizability of assessment scores across the sample population, primarily capturing enduring aspects of cognition with 72% of variance. Notably, most items in both scales assessed dynamic aspects of global cognition, suggesting that certain cognitive abilities, such as memory, may be more adaptable over time and could potentially be improved through interventions. The implications of this study were to help clinicians and researchers to be aware of the importance of differentiating dynamic and enduring cognitive features captured by brief cognitive assessment tools, as it can aid in the development and evaluation of preventive interventions and the monitoring of global cognition over time.

These two G-Theory studies provide evidence for the usefulness of applying the G-Theory methodology to differentiate between dynamic and enduring aspects of psychometric instruments used to assess cognitive ageing. Although G-Theory has become a standard for evaluating psychometric properties of measures (Brennan, 2001; Bloch & Norman, 2012), the number of studies conducted using this theory is still limited, especially in the field of cognitive ageing (Arterberry et al., 2014; Truong et al., 2020). This limitation may be due to the complexity of G-Theory calculations or data preparation for G-analysis, as it requires longitudinal data with at least three or more time points, and such data typically have high attrition rates. However, despite these disadvantages, the main advantage of G-Theory is its

comprehensive ability to accurately establish the differentiation between the dynamic and enduring aspects of psychometric measures and evaluate specific sources of error variance, instead of relying on test-retest correlations unable to differentiate between error sources.

Findings from Rasch Analysis Studies and Their Implications

The second objective of this thesis was to ultilise Rasch analysis methodology to evaluate psychometric properties and enhance assessments of cognitive ageing. To address this objective, three studies were conducted applying Rasch methodology to the informant SCC reports measure - the IQCODE-16, and the assessment of global cognition – the TICS-M (Truong, Choo, et al., 2023; Truong et al., 2021b; Truong, Numbers, et al., 2023). These two measurement tools were chosen due to their wide usage and ability to measure predominantly trait (enduring) aspects of cognition. The application of Rasch methodology in these studies resulted in generating conversion algorithms to transform the original ordinal scale scores of the IQCODE-16 and the TICS-M into interval-transformed data to enhance precision of their measurements and highlight the clinical utility of these assessments in detecting real changes in cognition in older adults.

The third study (Chapter 4; Truong et al., 2021b) employed Rasch analysis to examine psychometric properties of the IQCODE-16. The results indicate excellent reliability of the IQCODE-16, which provided additional evidence to support the strong psychometric properties of the IQCODE-16 and applicability of the measure for both individual and group assessment purposes. This study reorganized individual items of the IQCODE-16 into super-items, which achieved the best Rasch model fit and invariance across personal factors (i.e., age and sex of informants, and individual diagnosis). It means that the adjusted scale of the IQCODE-16 worked equally well for all individuals regardless of their individual factors (Hagquist & Andrich, 2004), and the ordinal-to-interval conversion algorithms can be used to improve precision of the IQCODE-16. Moreover, the use of Rasch conversion scores of the IQCODE-16 can also decrease measurement error linked with ordinal scores of the scale (Medvedev, Siegert, Kersten, et al., 2017).

Other findings from this study also suggested the ordinal scores of the IQCODE-16 exhibited bias. This was supported by the use of paired *t*-tests, which compared ordinal IQCODE-16 scores with interval Rasch scores in the total sample and the subsample diagnosed with dementia. The IQCODE-16 interval-transformed scores were significantly higher than the ordinal scores in the subsample diagnosed with dementia, illustrating that using IQCODE-16 ordinal scores may have an adverse effect on the results of parametric statistical tests conducted. Research suggested that the interval-level scores obtained from Rasch analyses may more precisely reflect the level of an individual's SCC. Hence, these interval scores could be employed for conducting parametric statistics without violating their arithmetic assumptions (Hobart & Cano, 2009; Leung, 2011).

The fourth study (Chapter 5; Truong, Numbers, et al., 2023) aimed to replicate the study in Chapter 4 (Truong et al., 2021b), acquiring more data and conducting confirmatory analyses to establish invariance across different personal factors and the high generalizability of assessment scores of the IQCODE-16. The same model used in the previous work (Truong et al., 2021b), which did not remove any items from the IQCODE-16, was applied to the OATS sample (n = 450), the CLS (n = 468), and the random sample (n = 450). This random sample comprised 450 participants selected from all three original datasets (i.e., OATS, CLS, and MAS). All Rasch analyses for the samples demonstrated the strong reliability of the IQCODE-16.

Moreover, additional findings indicate that the Rasch model fit solution originally proposed by the previous work was attained for the CLS and random samples, but not for the OATS sample. This conflicting outcome for the OATS sample might stem from sample bias, given that a significant number of participants in this sample were twins with a high level of homogeneity. However, minor modifications were applied to the OATS sample by combining three super-items instead of the suggested four super-items, resulting not only in an improved

fit for the OATS sample but also for the remaining samples. As such, a new conversion table was generated to transform ordinal raw scores of the IQCODE-16 into interval scores. However, both conversion tables are reliable and valid, as the differences between these conversion tables are marginal, evident by the extremely strong correlation with a Pearson *r* value of 0.998. This implies that both conversion tables can be utilised for transforming the IQCODE-16 from ordinal to interval data. The new table may be preferable, offering a slight increase in precision, which might not be noticeable when conducting research with groups but could make a discernible difference in precision at the individual level.

Another Rasch analysis study (Chapter 6; Truong, Choo, et al., 2023) aimed to examine and improve psychometric properties of the TICS-M. The result of the initial Rasch analysis demonstrated the strong reliability with *PSIs* of 0.83, which provided evidence to support the robust psychometric properties of the TICS-M. However, this finding indicated the original ordinal TICS-M did not fully comply with principles of fundamental measurement defined by the Rasch model and modifications were required to improve the scale. Finally, the TICS-M met expectations of the unidimensional Rasch model after creating two super-items by combining original TICS-M items. This modification increased reliability of the TICS-M with *PSI* of 0.86, as well as achieved the best fit to the Rasch model and invariance across personal factors (i.e., age, sex, diagnosis, and neuropsychological assessment scores). This means that the modified TICS-M scale works equally well for different groups of participants regardless of their personal characteristics (Hagquist & Andrich, 2004). Additionally, achieving the Rasch model fit with the modifications also allowed to generate a transformation table that converts ordinal raw scores of the TICS-M into interval-level data to improve the precision of TICS-M without changing its original format (Merkin et al., 2020; Truong et al., 2021b).

Furthermore, the results of these Rasch analytic studies added empirical evidence for the benefits of using a novel super-item approach in Rasch analysis (Lundgren-Nilsson et al.,

2013). Combining items into super-items can reduces measurement error due to individual item errors and counterbalances error variance across items (Lundgren-Nilsson et al., 2013; Medvedev et al., 2018). The overall effect of creating super-items is similar to the signal averaging technique used in studies of event related potentials utilizing electroencephalogram (EEG), where the test is repeated several times until error/noise variances cancel each other out and the relevant neural response can be identified (Luck, 2014). Here, instead of repatriation of a test, several items measuring the same construct are combined, which similarly cancel unwanted variances unrelated to the construct been measured.

These Rasch analytic studies have important implications for researchers and clinicians as they provide transformation algorithms, which can convert ordinal scores into interval-transformed scores that are able to reflect a true level of an individual cognition (i.e., SCC or global cognition) more accurately, and thus could be used to conduct parametric statistical analysis without violating their arithmetic assumptions (Hobart & Cano, 2009; Leung, 2011). Moreover, these Rasch studies included in the thesis have crucial implications for clinicians as they provide clinicians a simple tool in monitoring the real changes of cognition in older adults. For example, if individual A has a TICS-M ordinal score of 15, this indicates a higher risk of having dementia, as compared to individual B who scores 25. Assume, then, they both received the same clinical intervention, and individual A's ordinal score increases to 20 and participant B's ordinal score increases to 30. As can be seen, both individuals have increased by five points on their ordinal TICS-M scores, suggesting the same amount of clinical improvement. Importantly, though, when transforming these ordinal TICS-M scores into Rasch interval scores, individual A's interval score has increased by 2.06 points, while individual B's interval score has increased by 5.33 points. This indicates that individual B's improvement in global cognition level is approximately 2.5 times greater than individual A's. This example illustrates

that one is more able to detect real cognitive changes when using measurement with interval transform data.

Findings from Network Analysis Studies and Their Implications

The third objective of this thesis was to investigate the validity of cognitive ageing measurements in a novel way, using network analysis. To address this objective, two studies were conducted to evaluate the validity of widely used measures of cognitive abilities in older adults (i.e., TICS-M and MMSE domains) in relation to neuropsychological test performance. The advantage of network analysis is its modern statistical approach designed to examine unique associations among measures as well as sub-measures (i.e., individual subscales or domains), enhancing the understanding of cognitive ageing process holistically (Borsboom & Cramer, 2013; Contreras et al., 2019). Additionally, all associations in the network, estimated through network analysis, are not classified as either dependent or independent variable (Åkerblom et al., 2021). This made network analysis a preferred analysis to determine reciprocal or mutually maintaining mechanisms within a system when no transparent causal pathways. By estimating the unique associations between variables in a network, the unique links between different domains can be outlined (Borsboom & Cramer, 2013; Cholerton et al., 2016). Therefore, network analysis can provide a clearer depiction of the global network of unique associations between symptoms or domains of measures, thereby enhancing understanding of the complex relationships between variables.

The first network analytic study in Chapter 7 (Truong, Cervin, et al., 2023) examines the structure validity of the domains of the TICS-M, particularly convergent validity, in a substantial group of older adults across four separate assessments/waves over six years, using network analyses (four separate assessments/waves). The networks showed weak associations between the TICS-M domains internally over the four waves, which were conducted every two years, and the majority of the associations between TICS-M and neuropsychological domains were remained at least over two waves over the six-year period. Particularly, the association between the TICS-M memory domain and the neuropsychological memory domain remained

significant over all waves across six years. Two other associations, the TICS-M attention domain and neuropsychological visuospatial domain, and the TICS-M language domain and the neuropsychological language domain, were observed over three assessments. The remaining associations, between the TICS-M attention domain and the neuropsychological executive function, and between the TICS-M language domain and the neuropsychological attention domain, remained significant over two assessments. The consistency of associations between the TICS-M domains internally and with neuropsychological domains across four assessments indicates that the global cognitive function assessed by the TICS-M is stable or enduring over the six-year period, aligning with prior research (Truong et al., 2022). In other words, these findings provide further support for the TICS-M's structural validity in capturing enduring global cognition.

Even though the TICS-M was supported by four cognitive domains (Brandt et al., 1993; de Jager et al., 2003; Welsh et al., 1993), the results of this network analytic study raise concerns about the validity of individual domains of the TICS-M. These concerns arise when the domains are used as equivalents to the analogous neuropsychological domains. Specifically, individual neuropsychological domains were supported by the findings to be used separately because they exhibited interconnectedness across all assessments, and each node demonstrated high predictability within each wave. In contrast, the same level of support was not observed for the TICS-M domains. In order to use the individual domains of the TICS-M as separate measures for specific cognitive abilities, the results should align with the expectation that the TICS-M domains would display stronger links with each other as neuropsychological domains did. Additionally, each TICS-M domain/node must exhibit high predictability within each wave, and there should be strong associations between each TICS-M node and its analogous neuropsychological domain. This could be because the individual TICS-M domains may not have been sufficiently comprehensive in terms of cognition. For instance, the attention domain

of TICS-M consists of only two items that seem to emphasize mathematical ability rather than actual attention skills, making it more of an indirect measure of this skill. In sum, this study is in-line with a prior research that suggested the potential of TICS-M as a neuropsychological screening tool for global cognitive functioning rather than as an indicator of specific cognitive abilities (Aiello et al., 2022).

The study conducted in Chapter 8 (Truong, Cervin, et al, 2023b to be published) investigated the validity of a standardised cognitive screener, the MMSE, and its domains using network analysis. This study utilized a six-year longitudinal dataset of older adults in the MAS study. The findings showed a consistent association between the total MMSE score and the neuropsychological memory domain score over the six-year period. This indicates that the total MMSE score appears to mainly capture the memory aspect of neuropsychological test performance. This is in line with previous research, where the correlation between the MMSE score and neuropsychological memory function was higher as compared to that between the MMSE score and other neuropsychological domains (Schmitt et al., 2016). Therefore, this finding adds further evidence to support the common use of the MMSE in screening for cognitive impairments in older adults, as learning and memory is one of the six cognitive domains which may be affected in neurocognitive disorders (American Psychiatric Association, 2013)

A similar pattern as in the first network analytic study investigating the TICS-M domains was observed for the internal associations between the domains of the MMSE and between the MMSE and neuropsychological domains. While individual neuropsychological domains showed consistent internal edges (i.e., significant associations) across four assessments, the six domains of the MMSE were not consistently linked to each other internally across the same period. Furthermore, the majority of associations between the MMSE domains and neuropsychological domains found in wave 1 tended to change over time. These findings

suggest that the individual MMSE domains should be interpreted cautiously. This is in line with previous recommendation that the domains of the MMSE are not empirically derived and may not be valid indicators of specific cognitive domains (Strauss et al., 2006).

It should also be noted that the interconnections between neuropsychological domains remained stable over most waves in both network analytic studies. These results are consistent with theoretical frameworks, suggesting that neuropsychological domains including language, visuospatial ability, memory, processing speed/attention, and executive function may share underlying features (McGrew, 2009; Miyake & Friedman, 2012), which is also in line with a recent literature review (Harvey, 2022). This finding contributes to the body of knowledge supporting Spearman's two-factor theory of intelligence that suggests that overarching general intelligence (“g”) captures specific abilities (“s”) in older individuals (Spearman, 1927). Specifically, stable links between neuropsychological domains support the theory that older people with higher overall “g” – or high general intelligence – are more capable of other cognitive abilities. At the same time, different associations found between distinct domains support the presents of “s” abilities in older populations as well.

Overall, these network analytic studies have important implications that can assist clinicians and researchers in understanding the applications of the TICS-M and the MMSE for screening cognitive decline in older adults. The results support the overall structural validity of both measures in capturing global cognition. Specifically, the TICS-M captures broader cognitive functioning beyond any specific cognitive abilities, suggesting its potential as a more reliable measure of global cognitive abilities rather than domain-specific abilities. On the other hand, the total MMSE appears to predominantly assess the memory aspect of neuropsychological test performance. This indicates that the MMSE is a valuable screening tool for cognitive impairments in older adults, given that memory loss is often utilized as an important indicator to detect abnormal cognitive aging conditions. Furthermore, it is important

for researchers and clinicians to note that the separate cognitive domains of the TICS-M and MMSE should not be used as equivalent to similar neuropsychological domains.

Limitations and Directions for Further Research

There are general limitations that apply to all studies involved in this thesis. Firstly, all studies were carried out using data from the MAS study, where participants were selected from a small geographical area in Sydney, Australia. They were mostly from a White (European) ethnic background, with a high level of education and belonging to the upper-middle class. therefore, the generalisability and applicability to other ethnicities is questionable. Secondly, some participants were not native English speakers, which was not considered in analyses. Thus, our sample is not representative of all older adults in general, even within Sydney, Australia. Recent studies indicate that cultural differences play a role as both vulnerabilities and resilience in various health conditions (Choo et al., 2017). Therefore, future studies should duplicate these analyses using different samples that comprise culturally and linguistically diverse groups of communities or ethnicities. Besides that, each study also has its own unique limitations.

G-Theory Studies

The first G-Theory study examined enduring and dynamic patterns of subjective cognitive complaints assessed by the MAC-Q and IQCODE-16 (Truong et al., 2021a). This study aimed to examine data from five occasions with equal two-year intervals, but the interval between the last two occasions (i.e., occasions 4 and 5) extended to 4 years. Therefore, these analyses should be replicated using occasions with equal intervals. Moreover, this study's findings provided further support for the advantages of using the informant SCC report of the IQCODE-16. Nonetheless, due to format differences between the MAC-Q and the IQCODE-16, a more precise comparison between the informant and self-report of the IQCODE, the IQCODE-16 and the Informant Questionnaire on Cognitive Decline-Self-report (IQCODE-SR; Jansen et al., 2008), is necessary.

The second G-Theory study investigated enduring and dynamic patterns of global cognitive performance captured by the TICS-M and the MMSE (Truong et al., 2022). The rate of participant drop-off during the follow-up for the G-analyses was considerably high, potentially leading to biased non-random attrition. Hence, it would be advantageous to reproduce the results of this study in a more representative sample that experiences lower drop-off rates over time. Additionally, this study did not differentiate between various subtypes of dementia that could influence the results. Prior studies have indicated that certain dementia subtypes are less receptive to fluctuations in individual cognitive abilities. For instance, those with the behavioural variant of frontotemporal dementia (bvFTD) may be less responsive to cognitive shifts as measured by the MMSE (Piguet et al., 2009). Moreover, as the first G-Theory study, this study also aimed to examine data from five occasions with equal two-year intervals, but the interval between the last two occasions (i.e., occasions 4 and 5) extended to 4 years. Therefore, these analyses should be replicated using occasions with equal intervals as well as samples that provide specifications regarding dementia subtypes.

Rasch Analysis Studies

The two Rasch studies investigated psychometric properties of the IQCODE-16 and generated conversion tables to transform ordinal scores into interval data to enhance precision of the measurement (Truong et al., 2021b; Truong, Numbers, et al., 2023). In the first Rasch study (Truong et al., 2021b), it is noteworthy that there was a disparity in the genders of the informants, with a significant majority of females (nearly 70%) in both the original sample and the sample chosen for Rasch analysis. Moreover, other sample groups, such as those from other English-speaking nations, less financially privileged or affluent groups, or different ethnic groups, were unrepresented in the examination of invariance of the scale in the study. Therefore, To confirm the findings of this study, it is necessary to reproduce the analyses using a more diverse sample, thereby exploring invariance through different participant groups. This

approach could lead to the development of an additional conversion table for these diverse groups, enabling valid score comparisons across a wider population.

The second Rasch study (Truong, Numbers, et al., 2023) replicated Truong et al.'s (2021b) using multiple samples. Even though this study analysed data from three cohort studies with older adult participants, all participants were from Australia, and as such, they may not be representative of older adult population in different regions, such as those who are from developing countries or lower socioeconomic backgrounds. Hence, it is beneficial to replicate this study using diverse samples with various personal factors, such as older adults from low- and middle-income nations or those from non-English speaking countries.

In the third Rasch study investigating psychometric properties of the TICS-M, despite strengths of this study including an optimal sample size range ($250 < n \leq 500$) and its novelty, the study sample is not representative of all older adults. This speaks to a larger issue around neuropsychological tests which are largely normed against non-culturally or linguistically diverse samples. The Wechsler Adult Intelligence Scale-Revised Block Design scores were marginally below its lower normative range (Scott et al., 2021). This might be explained by the fact that the samples included participants with MCI and dementia. Although Rasch methodology is commonly known as less dependent on the sample (Hobart & Cano, 2009), future studies should investigate culturally and linguistically diverse (CALD) samples of older adults. Future studies could examine potential DIF or invariance in various CALD samples (e.g., English speakers from Anglosphere versus non-Anglosphere countries, or those using English as second language; or English versions versus translated versions) and different educational backgrounds. In addition, 43% of participants in the MAS data were diagnosed with MCI at the baseline or wave 1 (nearly 50% in our Rasch sample), which may explain lower mean TICS-M scores for both the total sample and our Rasch sample as compared to population norms.

Network-Analytic Studies

Both studies utilizing network analysis performed the confirmatory examinations with identical samples, although at subsequent time intervals. Importantly, it is crucial to recognize the non-random drop-off observed between the baseline and fourth waves. That is, individuals who were younger, in better health, and displayed higher cognitive abilities were more frequently involved in the fourth wave of participation, resulting in potential sampling bias for the confirmatory analyses. Therefore, confirmatory tests in true independent samples would have been a stronger source of evidence and should be explored in future studies. Moreover, these studies encountered a substantial reduction in the number of participants. Especially, nearly half of the individuals lost by the forth wave in the first network analytic study (Truong, Cervin, et al., 2023). This decrease in the sample size might have influenced the statistical strength of these studies, underscoring the need for prudence when interpreting the results. Subsequent research should explore methods to improve participant retention and reduce attrition, ultimately improving the generalizability and validity of the findings. Moreover, participants were well-educated, which may bias the results of neuropsychological tests. Therefore, future studies should replicate this study with more diverse and representative samples of older adults.

Conclusions

The current thesis applied G-Theory, Rasch analysis, and network analysis methodologies to evaluate and improve the reliability and validity of four psychometric measures used to assess cognitive aging in older adults. The analyses involved three datasets comprising a total of 2705 participants. The studies presented in this thesis made significant contributions to various aspects of cognitive aging assessments. Firstly, the use of G-Theory allowed for the differentiation between enduring and dynamic patterns of measures of cognitive abilities in older adults. This analysis revealed the reliability and validity of these measures for both research and clinical purposes. Secondly, Rasch analysis studies provided algorithms to transform ordinal data into interval-level data. This transformation enhanced the precision of measurements of cognitive abilities in older adults, enabling more accurate assessment of cognitive changes over time. Thirdly, network analytic studies evaluated evaluate the validity of measures used to assess cognitive ageing, advancing the understanding of the global network mechanisms involved in assessments of cognitive aging. In summary, the findings of this thesis have important implications for the evaluation and enhancement of assessments of cognitive aging. By improving our understanding of cognitive aging, these findings contribute to the development of more accurate cognitive assessment tools for older adults.

References

Aiello, E. N., Esposito, A., Giannone, I., Diana, L., Appollonio, I., & Bolognini, N. (2022). Telephone Interview for Cognitive Status (TICS): Italian adaptation, psychometrics and diagnostics. *Neurological Sciences*, *43*(5), 3071-3077. https://doi.org/10.1007/s10072-021-05729-7

Akbaraly, T. N., Portet, F., Fustinoni, S., Dartigues, J. F., Artero, S., Rouaud, O., Touchon, J., Ritchie, K., & Berr, C. (2009). Leisure activities and the risk of dementia in the elderly: results from the Three-City Study. *Neurology*, *73*(11), 854-861. https://doi.org/10.1212/WNL.0b013e3181b7849b

Åkerblom, S., Cervin, M., Perrin, S., Rivano Fischer, M., Gerdle, B., & McCracken, L. M. (2021). A network analysis of clinical variables in chronic pain: a study from the Swedish quality registry for pain rehabilitation (SQRP). *Pain Medicine*, *22*(7), 1591-1602. https://doi.org/10.1093/pm/pnaa473

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., & Petersen, R. C. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Ageing-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 270-279. https://doi.org/10.1016/j.jalz.2011.03.008

Alladi, S., Bak, T. H., Duggirala, V., Surampudi, B., Shailaja, M., Shukla, A. K., Chaudhuri, J. R., & Kaul, S. (2013). Bilingualism delays age at onset of dementia, independent of education and immigration status. *Neurology*, *81*(22), 1938-1944. https://doi.org/10.1212/01.wnl.0000436620.33155.a4

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, Illinois: Waveland Press.

Alzheimers New Zealand. (2020). https://www.alzheimers.org.nz/our-voice/new-zealand-data.

Amariglio, R. E., Donohue, M. C., Marshall, G. A., Rentz, D. M., Salmon, D. P., Ferris, S. H., Karantzoulis, S., Aisen, P. S., & Sperling, R. A. (2015). Tracking early decline in cognitive function in older individuals at risk for Alzheimer disease dementia: the Alzheimer's Disease Cooperative Study Cognitive Function Instrument. *JAMA Neurology*, *72*(4), 446-454.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (DSM-IV®)*. Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. Washington, DC: American Psychiatric Association.

American Psychology Association, A. (1994). *Diagnostic and statistical manual of mental disorders.*

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573. https://doi.org/10.1007/BF02293814

Andrich, D., Sheridan, B., & Luo, G. (2009). *RUMM 2030 (Beta Version for Windows)* In Perth, Western Australia: RUMM Laboratory Pty Ltd.

Anttila, T., Helkala, E.-L., Viitanen, M., Kåreholt, I., Fratiglioni, L., Winblad, B., Soininen, H., Tuomilehto, J., Nissinen, A., & Kivipelto, M. (2004). Alcohol drinking in middle age and subsequent risk of mild cognitive impairment and dementia in old age: a prospective population based study. *BMJ*, *329*(7465), 539. https://doi.org/10.1136/bmj.38181.418958.BE

Arterberry, B. J., Martens, M. P., Cadigan, J. M., & Rohrer, D. (2014). Application of generalizability theory to the big five inventory. *Personality and Individual Differences*, *69*, 98-103. https://doi.org/10.1016/j.paid.2014.05.015

Azizan, N. H., Mahmud, Z., & Rambli, A. (2020). Rasch Rating Scale Item Estimates using Maximum Likelihood Approach: Effects of Sample Size on the Accuracy and Bias of the Estimates. *International Journal of Advanced Science and Technology*, *29*(4), 2526-2531. https://doi.org/10.1007/s12144-022-02789-6

Barcaccia, B., Cervin, M., Pozza, A., Medvedev, O. N., Baiocco, R., & Pallini, S. (2020). Mindfulness, Self-Compassion and Attachment: a Network Analysis of Psychopathology Symptoms in Adolescents. *Mindfulness*, *11*(11), 2531-2541. https://doi.org/10.1007/s12671-020-01466-8

Barnes, D. E., Kaup, A., Kirby, K. A., Byers, A. L., Diaz-Arrastia, R., & Yaffe, K. (2014). Traumatic brain injury and risk of dementia in older veterans. *Neurology*, *83*(4), 312-319. https://doi.org/10.1212/WNL.0000000000000616

Beam, C. R., Kaneshiro, C., Jang, J. Y., Reynolds, C. A., Pedersen, N. L., & Gatz, M. (2018). Differences Between Women and Men in Incidence Rates of Dementia and Alzheimer's

Disease. *Journal of Alzheimers Disease*, *64*(4), 1077-1083. https://doi.org/10.3233/jad-180141

Beaudreau, S. A., & O'Hara, R. (2008). Late-life anxiety and cognitive impairment: a review. *The American Journal of Geriatric Psychiatry*, *16*(10), 790-803. https://doi.org/10.1097/JGP.0b013e31817945c3

Beeri, M. S., Goldbourt, U., Silverman, J. M., Noy, S., Schmeidler, J., Ravona-Springer, R., Sverdlick, A., & Davidson, M. (2004). Diabetes mellitus in midlife and the risk of dementia three decades later. *Neurology*, *63*(10), 1902-1907. https://doi.org/10.1212/01.wnl.0000144278.79488.dd

Benton, A. L. (1967). Problems of test construction in the field of aphasia. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*.

Benton, A. L., Sivan, A. B., & Spreen, O. (1996). *Der Benton Test*. Huber.

Bentvelzen, A. C., Crawford, J. D., Theobald, A., Maston, K., Slavin, M. J., Reppermund, S., Kang, K., Numbers, K., Brodaty, H., Sachdev, P., & Kochan, N. A. (2019). Validation and normative data for the modified telephone interview for cognitive status: The Sydney memory and ageing study. *Journal of the American Geriatrics Society*, *67*(10), 2108-2115. https://doi.org/10.1111/jgs.16033

Bickel, H., & Kurz, A. (2009). Education, occupation, and dementia: the Bavarian school sisters study. *Dementia & geriatric cognitive disorders*, *27*(6), 548-556. https://doi.org/10.1159/000227781

Blazer, D. G., Yaffe, K., & Liverman, C. T. (2015). *Cognitive Ageing: Progress in Understanding and Opportunities for Action*. Washington, DC: National Academies Press

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Medical teacher*, *34*(11), 960-992. https://doi.org/10.3109/0142159X.2012.703791

Bonaiuto, S., Rocca, W. A., Lippi, A., Giannandrea, E., Mele, M., Cavarzeran, F., & Amaducci, L. (1995). Education and occupation as risk factors for dementia: a population-based case-control study. *Neuroepidemiology*, *14*(3), 101-109. https://doi.org/10.1159/000109785

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., & Costantini, G. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 1-18. https://doi.org/10.1038/s43586-021-00055-w

Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology & Behavioral Neurology*, *1*(2), 111-118.

Brandt, J., Welsh, K. A., Breitner, J. C., Folstein, M. F., Helms, M., & Christian, J. C. (1993). Hereditary influences on cognitive functioning in older men: a study of 4000 twin pairs. *Archives of Neurology*, *50*(6), 599-603. https://doi.org/10.1001/archneur.1993.00540060039014

Brasso, C., Giordano, B., Badino, C., Bellino, S., Bozzatello, P., Montemagni, C., & Rocca, P. (2021). Primary Psychosis: Risk and Protective Factors and Early Detection of the Onset. *Diagnostics (Basel)*, *11*(11). https://doi.org/10.3390/diagnostics11112146

Brasure, M., Desai, P., Davila, H., Nelson, V. A., Calvert, C., Jutkowitz, E., Butler, M., Fink, H. A., Ratner, E., & Hemmy, L. S. (2018). Physical activity interventions in preventing cognitive decline and Alzheimer-type dementia: a systematic review. *Annals of Internal Medicine*, *168*(1), 30-38. https://doi.org/10.7326/M17-1528

Brayne, C. (1998). The mini-mental state examination, will we be using it in 2001? *International Journal of Geriatric Psychiatry*. https://doi.org/10.1002/(sici)1099-1166(199805)13:5<285::aid-gps753>3.0.co;2-v

Brennan, R. L. (2009). *Generalizability theory*. New York, USA: Springer-Verlag.

Brodaty, H., Pond, D., Kemp, N. M., Luscombe, G., Harding, L., Berman, K., & Huppert, F. A. (2002). The GPCOG: a new screening test for dementia designed for general practice. *Journal of the American Geriatrics Society*, *50*(3), 530-534. https://doi.org/10.1046/j.1532-5415.2002.50122.x

Brown, R. C., Lockwood, A. H., & Sonawane, B. R. (2005). Neurodegenerative diseases: an overview of environmental risk factors. *Environmental Health Perspectives*, *113*(9), 1250-1256. https://doi.org/10.1289/ehp.7567

Buckley, R., Saling, M. M., Ames, D., Rowe, C. C., Lautenschlager, N. T., Macaulay, S. L., Martins, R. N., Masters, C. L., O'Meara, T., & Savage, G. (2013). Factors affecting subjective memory complaints in the AIBL ageing study: biomarkers, memory, affect, and age. *International Psychogeriatrics*, *25*(8), 1307-1315. https://doi.org/10.1017/S1041610213000665

Buckley, R. F., Ellis, K. A., Ames, D., Rowe, C. C., Lautenschlager, N. T., Maruff, P., Villemagne, V. L., Macaulay, S. L., Szoeke, C., Martins, R. N., & ...Saling, M. M. (2015). Phenomenological characterization of memory complaints in preclinical and prodromal Alzheimer's disease. *Neuropsychology*, *29*(4), 571-581. https://doi.org/10.1037/neu0000156

Budhiraja, R., Budhiraja, P., & Quan, S. F. (2010). Sleep-disordered breathing and cardiovascular disorders. *Respiratory Care*, *55*(10), 1322-1332.

Burmester, B., Leathem, J., & Merrick, P. (2016). Subjective cognitive complaints and objective cognitive function in ageing: a systematic review and meta-analysis of recent cross-sectional findings. *Neuropsychology Review*, *26*(4), 376-393. https://doi.org/10.1007/s11065-016-9332-2

Buss, A. H. (1989). Personality as traits. *American Psychologist, 44*, 1378-1388. https://doi.org/10.1037/0003-066X.44.11.1378

Cameron, J., Worrall-Carter, L., Page, K., Stewart, S., & Ski, C. (2012). Screening for mild cognitive impairment in patients with heart failure: Montreal Cognitive Assessment versus Mini Mental State Exam. *European Journal of Cardiovascular Nursing*, *12*(3), 252-260. https://doi.org/10.1177/1474515111435606

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG*. Taylor & Francis.

Caselli, R. J., Chen, K., Locke, D. E., Lee, W., Roontiva, A., Bandy, D., Fleisher, A. S., & Reiman, E. M. (2014). Subjective cognitive decline: self and informant comparisons. *Alzheimer's & Dementia*, *10*(1), 93-98. https://doi.org/10.1016/j.jalz.2013.01.003

Castanho, T. C., Portugal-Nunes, C., Moreira, P. S., Amorim, L., Palha, J. A., Sousa, N., & Correia Santos, N. (2016). Applicability of the Telephone Interview for Cognitive Status (Modified) in a community sample with low education level: association with an extensive neuropsychological battery. *International Journal of Geriatric Psychiatry*, *31*(2), 128-136. https://doi.org/10.1002/gps.4301

Centers for Disease Control and Prevention. (2019). Subjective cognitive decline—a public health issue. In *Alzheimer's Disease and Healthy Ageing*. https://www.cdc.gov/ageing/data/subjective-cognitive-decline-brief.html

Centers for Disease Control and Prevention. (2020). What is Alzheimer's Disease? In *Alzheimer's Disease and Healthy Ageing*. https://www.cdc.gov/ageing/index.html

Chalmers, R. A., Cervin, M., Choo, C., Baune, B. T., Trollor, J. N., Numbers, K., Sachdev, P. S., Brodaty, H., Kochan, N. A., & Medvedev, O. N. (2022). Networks of inflammation, depression, and cognition in ageing males and females. *Ageing Clinical and Experimental Research*, 1-12. https://doi.org/10.1007/s40520-022-02198-6

Chin, Y., Choo, C. C., & Doshi, K. (2019). A Case of Subjective Cognitive Complaints in Older Adults: Anxiety, Stress, and Ageing in an Elderly Client. In *Clinical Psychology Casebook Across the Lifespan* (pp. 71-77). Springer.

Cholerton, B., Larson, E. B., Quinn, J. F., Zabetian, C. P., Mata, I. F., Keene, C. D., Flanagan, M., Crane, P. K., Grabowski, T. J., & Montine, K. S. (2016). Precision medicine: clarity for the complexity of dementia. *The American Journal of Pathology*, *186*(3), 500-506. https://doi.org/10.1016/j.ajpath.2015.12.001

Choo, C. C., Harris, K. M., Chew, P. K., & Ho, R. C. (2017). Does ethnicity matter in risk and protective factors for suicide attempts and suicide lethality? *PloS one*, *12*(4), e0175752. https://doi.org/10.1371/journal.pone.0175752

Christensen, H., Mackinnon, A., Jorm, A. F., Korten, A., Jacomb, P., Hofer, S. M., & Henderson, S. (2004). The Canberra Longitudinal Study: Design, aims, methodology, outcomes and recent empirical investigations. *Ageing Neuropsychology and Cognition*, *11*(2-3), 169-195. https://doi.org/10.1080/13825580490511053

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). *Rasch models in health*. Wiley Online Library.

Chuang, Y.-F., Hayden, K. M., Norton, M. C., Tschanz, J., Breitner, J. C., Welsh-Bohmer, K. A., & Zandi, P. P. (2010). Association between APOE ε4 allele and vascular dementia: the Cache County study. *Dementia and Geriatric Cognitive Disorders*, *29*(3), 248-253. https://doi.org/10.1159/000285166

Ciolek, C. H., & Lee, S. Y. (2019). Cognitive Issues in the Older Adult. In *Guccione's Geriatric Physical Therapy E-Book* (pp. 425-452). https://doi.org/10.1016/B978-0-323-60912-8.00019-1

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Publishing Co.

Cojan, Y., Saj, A., & Vuilleumier, P. (2021). Brain Substrates for Distinct Spatial Processing Components Contributing to Hemineglect in Humans. *Brain Sciences*, *11*(12), 1584. https://doi.org/10.3390/brainsci11121584

Contreras, A., Nieto, I., Valiente, C., Espinosa, R., & Vazquez, C. (2019). The study of psychopathology from the network analysis perspective: a systematic review. *Psychotherapy and Psychosomatics*, *88*(2), 71-83. https://doi.org/10.1159/000497425

Craik, F. I., Bialystok, E., & Freedman, M. (2010). Delaying the onset of Alzheimer disease: Bilingualism as a form of cognitive reserve. *Neurology*, *75*(19), 1726-1729. https://doi.org/10.1212/WNL.0b013e3181fc2a1c

Crimmins, E. M., Saito, Y., Kim, J. K., Zhang, Y. S., Sasson, I., & Hayward, M. D. (2018). Educational differences in the prevalence of dementia and life expectancy with dementia: changes from 2000 to 2010. *The Journals of Gerontology: Series B*, *73*(suppl_1), S20-S28. https://doi.org/10.1093/geronb/gbx135

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137-163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Crook, T. H., Feher, E. P., & Larrabee, G. J. (1992). Assessment of memory complaint in age-associated memory impairment: the MAC-Q. *International Psychogeriatrics*, *4*(2), 165-176. https://doi.org/10.1017/s1041610292000991

Crooks, V. C., Clark, L., Petitti, D. B., Chui, H., & Chiu, V. (2005). Validation of multi-stage telephone-based identification of cognitive impairment and dementia. *BMC Neurology*, *5*(1), 1-8. https://doi.org/10.1186/1471-2377-5-8

Crooks, V. C., Lubben, J., Petitti, D. B., Little, D., & Chiu, V. (2008). Social network, cognitive function, and dementia incidence among elderly women. *American Journal of Public Health*, *98*(7), 1221-1227. https://doi.org/10.2105/AJPH.2007.115923

Crooks, V. C., Petitti, D. B., Robins, S. B., & Buckwalter, J. G. (2006). Cognitive domains associated with performance on the Telephone Interview for Cognitive Status—Modified. *American Journal of Alzheimer's Disease & Other Dementias*, *21*(1), 45-53. https://doi.org/10.1177/153331750602100104

Crumley, J. J., Stetler, C. A., & Horhota, M. (2014). Examining the relationship between subjective and objective memory performance in older adults: a meta-analysis. *Psychology and Ageing*, *29*(2), 250-263. https://doi.org/10.1037/a0035908

Cullum, S., Mullin, K., Zeng, I., Yates, S., Payman, V., Fisher, M., & Cheung, G. (2018). Do community-dwelling Māori and Pacific peoples present with dementia at a younger age and at a later stage compared with NZ Europeans? *International Journal of Geriatric Psychiatry*, *33*(8), 1098-1104. https://doi.org/10.1002/gps.4898

De Craen, A. J. M., Gussekloo, J., Vrijsen, B., & Westendorp, R. G. J. (2005). Meta-analysis of nonsteroidal antiinflammatory drug use and risk of dementia. *American Journal of Epidemiology*, *161*(2), 114-120. https://doi.org/10.1093/aje/kwi029

de Jager, C. A., Budge, M. M., & Clarke, R. (2003). Utility of TICS-M for the assessment of cognitive function in older adults. *International Journal of Geriatric Psychiatry*, *18*(4), 318-324. https://doi.org/10.1002/gps.830

de La Torre, J. C. (2012). Cardiovascular risk factors promote brain hypoperfusion leading to cognitive decline and dementia. *Cardiovascular psychiatry neurology*, *2012*. https://doi.org/10.1155/2012/367516

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *The California verbal learning test - Adult version.* San Antonio, TX: The Psychological Corporation.

Dotson, V. M., Beydoun, M. A., & Zonderman, A. B. (2010). Recurrent depressive symptoms and the incidence of dementia and mild cognitive impairment. *Neurology*, *75*(1), 27-34. https://doi.org/10.1212/WNL.0b013e3181e62124

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193-242. https://doi.org/10.1037/h0044139

El-Hayeck, R., Baddoura, R., Wehbé, A., Bassil, N., Koussa, S., Abou Khaled, K., Richa, S., Khoury, R., Alameddine, A., & Sellal, F. (2019). An Arabic version of the mini-mental state examination for the Lebanese population: Reliability, validity, and normative data. *Journal of Alzheimer's Disease*, *71*(2), 525-540. https://doi.org/10.3233/JAD-181232

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*, 1-18. https://doi.org/10.18637/jss.v048.i04

Epstein, S. (1984). Trait theory as personality theory: Can a part be as great as the whole? *Psychological Inquiry, 5*, 120-122. https://doi.org/10.1207/s15327965pli0502_4

Farina, M. P., Hayward, M. D., Kim, J. K., & Crimmins, E. M. (2020). Racial and educational disparities in dementia and dementia-free life expectancy. *The Journals of Gerontology: Series B, 75*(7), e105-e112. https://doi.org/10.1093/geronb/gbz046

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/bf03193146

Feeney, J., Savva, G. M., O'Regan, C., King-Kallimanis, B., Cronin, H., & Kenny, R. A. (2016). Measurement error, reliability, and minimum detectable change in the Mini-Mental State Examination, Montreal Cognitive Assessment, and Color Trails Test among community living

middle-aged and older adults. *Journal of Alzheimer's Disease*, *53*(3), 1107-1114. https://doi.org/10.3233/JAD-160248

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Fisher, W. (1992). *Reliability statistics. Rasch Measurement Transactions*. Chicago, USA: MESA Press.

Flatt, J. D., Gilsanz, P., Quesenberry Jr, C. P., Albers, K. B., & Whitmer, R. A. (2018). Post-traumatic stress disorder and risk of dementia among members of a health care delivery system. *Alzheimer's & Dementia*, *14*(1), 28-34. https://doi.org/10.1016/j.jalz.2017.04.014

Fleminger, S., Oliver, D., Lovestone, S., Rabe-Hesketh, S., & Giora, A. (2003). Head injury as a risk factor for Alzheimer's disease: the evidence 10 years on; a partial replication. *Journal of Neurology, Neurosurgery, Psychiatry*, *74*(7), 857-862. https://doi.org/10.1136/jnnp.74.7.857

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189-198. https://doi.org/10.1016/0022-3956(75)90026-6

Fong, T. G., Fearing, M. A., Jones, R. N., Shi, P., Marcantonio, E. R., Rudolph, J. L., Yang, F. M., Kiely, D. K., & Inouye, S. K. (2009). Telephone interview for cognitive status: Creating a crosswalk with the Mini-Mental State Examination. *Alzheimer's & Dementia*, *5*(6), 492-497. https://doi.org/10.1016/j.jalz.2009.02.007

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, *45*(1), 30.

Franco, R., Li, S., Rodriguez-Rocha, H., Burns, M., & Panayiotidis, M. I. (2010). Molecular mechanisms of pesticide-induced neurotoxicity: Relevance to Parkinson's disease. *Chemico-biological Interactions*, *188*(2), 289-300. https://doi.org/10.1016/j.cbi.2010.06.003

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129-1164. https://doi.org/10.1002/spe.4380211102

Ganguli, M., Vander Bilt, J., Saxton, J. A., Shen, C., & Dodge, H. H. (2005). Alcohol consumption and cognitive function in late life: a longitudinal community study. *Neurology*, *65*(8), 1210-1217. https://doi.org/10.1212/01.wnl.0000180520.35181.24

Gao, S., Hendrie, H. C., Hall, K. S., & Hui, S. (1998). The relationships between age, sex, and the incidence of dementia and Alzheimer disease: a meta-analysis. *Archives of General Psychiatry*, *55*(9), 809-815. https://doi.org/10.1001/archpsyc.55.9.809

Goodglass, H., & Kaplan, E. (1983). *The Assessment of Aphasia and Related Disorders, 2nd edn*. Philadelphia, USA: Lea & Febiger.

Garcia, M. A., Downer, B., Chiu, C. T., Saenz, J. L., Ortiz, K., & Wong, R. (2021). Educational Benefits and Cognitive Health Life Expectancies: Racial/Ethnic, Nativity, and Gender Disparities. *Gerontologist*, *61*(3), 330-340. https://doi.org/10.1093/geront/gnaa112

Garcia, M. A., Saenz, J., Downer, B., & Wong, R. (2018). The role of education in the association between race/ethnicity/nativity, cognitive impairment, and dementia among older adults in the United States. *Demographic Research*, *38*, 155-168. https://doi.org/10.4054/DemRes.2018.38.6

Guo, Z., Cupples, L. A., Kurz, A., Auerbach, S. H., Volicer, L., Chui, H., Green, R. C., Sadovnick, A. D., Duara, R., DeCarli, C., & ...Farrer, L. A. (2000). Head injury and the risk of AD in the MIRAGE study. *Neurology*, *54*(6), 1316-1323. https://doi.org/10.1212/wnl.54.6.1316

Hagell, P., & Westergren, A. (2016). Sample Size and Statistical Conclusions from Tests of Fit to the Rasch Model According to the Rasch Unidimensional Measurement Model (Rumm) Program in Health Outcome Measurement. *Journal of Applied Measurement*, *17*(4), 416-431.

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences*, *36*(4), 955-968. https://doi.org/10.1016/S0191-8869(03)00164-8

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. (2007). The integrated trait-state model. *Journal of Research in Personality, 41,* 295–315. https://doi.org/10.1016/j.jrp.2006.04.003

Hamer, M., & Chida, Y. (2009). Physical activity and risk of neurodegenerative disease: a systematic review of prospective evidence. *Psychological Medicine*, *39*(1), 3-11. https://doi.org/10.1017/S0033291708003681

Harada, C. N., Love, M. C. N., & Triebel, K. L. (2013). Normal cognitive ageing. *Clinics in Geriatric Medicine*, *29*(4), 737-752. https://doi.org/10.1016/j.cger.2013.07.002

Harrison, J. K., Fearon, P., Noel-Storr, A. H., McShane, R., Stott, D. J., & Quinn, T. J. (2015). Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within a secondary care setting. *Cochrane Database of Systematic Reviews*(3). https://doi.org/10.1002/14651858.CD010772.pub2

Harrison, J. K., Stott, D. J., McShane, R., Noel-Storr, A. H., Swann-Price, R. S., & Quinn, T. J. (2016). Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the early diagnosis of dementia across a variety of healthcare settings. *Cochrane Database of Systematic Reviews*, *11*(11), CD011333. https://doi.org/10.1002/14651858.CD011333.pub2

Harvey, P. D. (2022). Domains of cognition and their assessment. *Dialogues in Clinical Neuroscience*, *21*(3), 227-237. https://doi.org/10.31887/DCNS.2019.21.3/pharvey

Haslbeck, J., & Waldorp, L. J. (2018). How well do network models predict observations? On the importance of predictability in network models. *Behavior Research Methods*, *50*(2), 853-861. https://doi.org/10.3758/s13428-017-0910-x

Heeren, A., & McNally, R. J. (2016). An integrative network approach to social anxiety disorder: The complex dynamic interplay among attentional bias for threat, attentional control, and symptoms. *Journal of Anxiety Disorders*, *42*, 95-104. https://doi.org/10.1016/j.janxdis.2016.06.009

Hernán, M. A., Alonso, A., & Logroscino, G. (2008). Cigarette smoking and dementia: Potential selection bias in the elderly. *Epidemiology*, 448-450. https://doi.org/10.1097/EDE.0b013e31816bbe14

Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy*, *23*(4), 290-301. https://doi.org/10.1016/j.bjpt.2018.12.006

Hildreth, K. L., & Church, S. (2015). Evaluation and management of the elderly patient presenting with cognitive complaints. *Medical Clinics*, *99*(2), 311-335. https://doi.org/10.1016/j.mcna.2014.11.006

Hobart, J., & Cano, S. (2009). *Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods*. Prepress Projects.

Hodge, D. R., & Gillespie, D. F. (2007). Phrase completion scales: a better measurement approach than Likert scales? *Journal of Social Service Research*, *33*(4), 1-12. https://doi.org/10.1300/J079v33n04_01

Hopkins, E. G., Leman, P. J., Cervin, M., Numbers, K., Brodaty, H., Kochan, N. A., Sachdev, P. S., & Medvedev, O. N. (2023). Network of mental activities, cognitive function and depression in older men and women. *Journal of Psychiatric Research*, *162*, 113-122. https://doi.org/10.1016/j.jpsychires.2023.05.030

Horton Jr, A. M., Slone, D. G., & Shapiro, S. (1987). Neuropsychometric correlates of the Mini-Mental State Examination: preliminary data. *Perceptual and Motor Skills*, *65*(1), 64-66. https://doi.org/10.2466/pms.1987.65.1.64

Hughes, T. F. (2010). Promotion of cognitive health through cognitive activity in the ageing population. *Ageing Health*, *6*(1), 111-121. https://doi.org/10.2217/ahe.09.89

Hugo, J., & Ganguli, M. (2014). Dementia and cognitive impairment: epidemiology, diagnosis, and treatment. *Clinics in Geriatric Medicine*, *30*(3), 421-442. https://doi.org/10.1016/j.cger.2014.04.001

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality and Quantity*, *34*(4), 331-351. https://doi.org/10.1023/A:1004782230065

Jamieson, S. (2004). Likert scales: How to (ab) use them? *Medical Education*, *38*(12), 1217-1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jansen, A. P. D., van Hout, H. P. J., Nijpels, G., van Marwijk, H. W. J., Gundy, C., de Vet, H. C. W., & Stalman, W. A. B. (2008). Self-reports on the IQCODE in older adults: a psychometric evaluation. *Journal of Geriatric Psychiatry and Neurology*, *21*(2), 83-92. https://doi.org/10.1177/0891988707311558

Jellinger, K. A. (2004). Head injury and dementia. *Current Opinion in Neurology*, *17*(6), 719-723. https://doi.org/10.1097/00019052-200412000-00012

Jessen, F., Amariglio, R. E., Van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., & ...Subjective Cognitive Decline Initiative (SCD-I) Working Group. (2014). A conceptual framework for

research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's & Dementia*, *10*(6), 844-852. https://doi.org/10.1016/j.jalz.2014.01.001

Jones, P. J., Heeren, A., & McNally, R. J. (2017). Commentary: A network theory of mental disorders. *Frontiers in Psychology*, *8*, 1305. https://doi.org/10.3389/fpsyg.2017.01305

Jonker, C., Geerlings, M. I., & Schmand, B. (2000). Are memory complaints predictive for dementia? A review of clinical and population-based studies. *International Journal of Geriatric Psychiatry*, *15*(11), 983-991. https://doi.org/10.1002/1099-1166(200011)15:11<983::aid-gps238>3.0.co;2-5

Jorm, A. F. (1994). A short form of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): development and cross-validation. *Psychological Medicine*, *24*(1), 145-153. https://doi.org/10.1017/s003329170002691x

Jorm, A. F. (2004). The Informant Questionnaire on cognitive decline in the elderly (IQCODE): a review. *International Psychogeriatrics*, *16*(3), 275-293. https://doi.org/10.1017/s1041610204000390

Jorm, A. F., Christensen, H., Korten, A. E., Jacomb, P. A., & Henderson, A. S. (2000). Informant ratings of cognitive decline in old age: validation against change on cognitive tests over 7 to 8 years. *Psychological medicine*, *30*(4), 981-985. https://doi.org/10.1017/s0033291799002299

Jorm, A. F., & Jolley, D. (1998). The incidence of dementia: a meta-analysis. *Neurology*, *51*(3), 728-733. https://doi.org/10.1212/wnl.51.3.728

Jorm, A. F., Scott, R., Cullen, J. S., & MacKinnon, A. J. (1991). Performance of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) as a screening test for dementia. *Psychological Medicine*, *21*(3), 785-790. https://doi.org/10.1017/s0033291700022418

Justin, B. N., Turek, M., & Hakim, A. M. (2013). Heart disease as a risk factor for dementia. *Clinical Epidemiology*, *5*, 135. https://doi.org/10.2147/CLEP.S30621

Kaplan, E., Goodglass, H., Weintraub, S., & Segal, O. (2001). *Boston naming test*. Lippincott Williams & Wilkins.

Karp, A., Paillard-Borg, S., Wang, H.-X., Silverstein, M., Winblad, B., & Fratiglioni, L. (2006). Mental, physical and social components in leisure activities equally contribute to decrease dementia risk. *Dementia Geriatric Cognitive Disorders*, *21*(2), 65-73. https://doi.org/10.1159/000089919

Kim, C., Wu, B., Tanaka, E., Watanabe, T., Watanabe, K., Chen, W., Ito, S., Okumura, R., Arai, T., & Anme, T. (2016). Association between a change in social interaction and dementia among elderly people. *International Journal of Gerontology*, *10*(2), 76-80. https://doi.org/10.1016/j.ijge.2016.03.006

Kim, H., Yun, C.-H., Thomas, R. J., Lee, S. H., Seo, H. S., Cho, E. R., Lee, S. K., Yoon, D. W., Suh, S., & Shin, C. (2013). Obstructive sleep apnea as a risk factor for cerebral white matter change in a middle-aged and older general population. *Sleep*, *36*(5), 709-715. https://doi.org/10.5665/sleep.2632

Kivipelto, M., Ngandu, T., Fratiglioni, L., Viitanen, M., Kåreholt, I., Winblad, B., Helkala, E.-L., Tuomilehto, J., Soininen, H., & Nissinen, A. (2005). Obesity and vascular risk factors at midlife and the risk of dementia and Alzheimer disease. *Archives of Neurology*, *62*(10), 1556-1560. https://doi.org/10.1001/archneur.62.10.1556

Knopman, D. S., Roberts, R. O., Geda, Y. E., Boeve, B. F., Pankratz, V. S., Cha, R. H., Tangalos, E. G., Ivnik, R. J., & Petersen, R. C. (2009). Association of prior stroke with cognitive function and cognitive impairment: a population-based study. *Archives of Neurology*, *66*(5), 614-619. https://doi.org/10.1001/archneurol.2009.30

Kramarow, E. A., & Tejada-Vera, B. (2019). System Dementia Mortality in the United States, 2000-2017. *National Vital Statistics Reports*, *68*(2), 1-29. https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_02-508.pdf

Lacruz, M. E., Emeny, R. T., Bickel, H., Linkohr, B., & Ladwig, K. H. (2013). Feasibility, internal consistency and covariates of TICS-m (telephone interview for cognitive status-modified) in a population-based sample: findings from the KORA-Age study. *International Journal of Geriatric Psychiatry*, *28*(9), 971-978. https://doi.org/10.1002/gps.3916

Lacy, M., Kaemmerer, T., & Czipri, S. (2015). Standardized Mini-Mental State Examination Scores and Verbal Memory Performance at a Memory Center:Implications for Cognitive Screening.

American Journal of Alzheimer's Disease & Other Dementias, *30*(2), 145-152. https://doi.org/10.1177/1533317514539378

Lambert, B. (2018). *A student's guide to Bayesian statistics*. Sage.

Langa, K. M., Larson, E. B., Crimmins, E. M., Faul, J. D., Levine, D. A., Kabeto, M. U., & Weir, D. R. (2017). A comparison of the prevalence of dementia in the United States in 2000 and 2012. *JAMA Internal Medicine*, *177*(1), 51-58. https://doi.org/10.1001/jamainternmed.2016.6807

Langa, K. M., & Levine, D. A. (2014). The diagnosis and management of mild cognitive impairment: a clinical review. *Jama*, *312*(23), 2551-2561.

Launer, L. J., Hoes, A. W., Ott, A., Hofman, A., Breteler, M. M. B., & Stricker, B. C. (1998). NSAIDs and incident Alzheimer's disease. The Rotterdam study. *Neurobiology of Ageing*, *19*(6), 607-611. https://doi.org/10.1016/s0197-4580(98)00096-7

Lenehan, M. E., Klekociuk, S. Z., & Summers, M. J. (2012). Absence of a relationship between subjective memory complaint and objective memory impairment in mild cognitive impairment (MCI): is it time to abandon subjective memory complaint as an MCI diagnostic criterion? *International Psychogeriatrics*, *24*(9), 1505-1514. https://doi.org/10.1017/S1041610212000695

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research*, *37*(4), 412-421. https://doi.org/10.1080/01488376.2011.580697

Leung, Y.-Y., Png, M.-E., Conaghan, P., & Tennant, A. (2014). A systematic literature review on the application of Rasch analysis in musculoskeletal disease—A special interest group report of OMERACT 11. *The Journal of Rheumatology*, *41*(1), 159-164. https://doi.org/10.3899/jrheum.130814

Leys, D. (2009). Epidemiology of post-stroke dementia. *Journal of the Neurological Sciences*, *283*(1), 275. https://doi.org/10.1016/j.jns.2009.02.136

Lezak, M. D., Howieson, D. B., Loring, D. W., & Fischer, J. S. (2004). *Neuropsychological assessment*. Oxford University Press, USA.

Linacre, J. M. (2000). Comparing “partial credit” and “rating scale” models. *Rasch Measurement Transactions*, *14*(3), 768.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106.

Linacre, J. M. (2004). Estimation methods for Rasch measures. In E. V. Smith Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 25-48). JAM Press.

Lipnicki, D. M., Sachdev, P. S., Crawford, J., Reppermund, S., Kochan, N. A., Trollor, J. N., Draper, B., Slavin, M. J., Kang, K., Lux, O., & ...Brodaty, H. (2013). Risk factors for late-life cognitive decline and variation with age and sex in the Sydney Memory and Ageing Study. *PloS one*, *8*(6), e65841. https://doi.org/10.1371/journal.pone.0065841

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company.

Lourida, I., Soni, M., Thompson-Coon, J., Purandare, N., Lang, I. A., Ukoumunne, O. C., & Llewellyn, D. J. (2013). Mediterranean diet, cognitive function, and dementia: a systematic review. *Epidemiology*, 479-489. https://doi.org/10.1097/EDE.0b013e3182944410

Lundgren-Nilsson, A., Jonsdottir, I. H., Ahlborg, G., & Tennant, A. (2013). Construct validity of the psychological general well being index (PGWBI) in a sample of patients undergoing treatment for stress-related exhaustion: a rasch analysis. *Health and Quality of Life Outcomes*, *11*(1), 2. https://doi.org/10.1186/1477-7525-11-2

Lundgren-Nilsson, A., & Tennant, A. (2011). Past and present issues in Rasch analysis: the functional independence measure (FIM™) revisited. *Journal of Rehabilitation Medicine*, *43*(10), 884-891. https://doi.org/10.2340/16501977-0871

Mahncke, H. W., Connor, B. B., Appelman, J., Ahsanuddin, O. N., Hardy, J. L., Wood, R. A., Joyce, N. M., Boniske, T., Atkins, S. M., & Merzenich, M. M. (2006). Memory enhancement in healthy older adults using a brain plasticity-based training program: a randomized, controlled study. *Proceedings of the National Academy of Sciences*, *103*(33), 12523-12528. https://doi.org/10.1073/pnas.0605194103

Mehta, K. M., & Yeo, G. W. (2017). Systematic review of dementia prevalence and incidence in United States race/ethnic populations. *Alzheimers Dement*, *13*(1), 72-83. https://doi.org/10.1016/j.jalz.2016.06.2360

Maloney, E. A., Sattizahn, J. R., & Beilock, S. L. (2014). Anxiety and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(4), 403-411. https://doi.org/10.1002/wcs.1299

Mantella, R. C., Butters, M. A., Dew, M. A., Mulsant, B. H., Begley, A. E., Tracey, B., Shear, M. K., Reynolds III, C. F., & Lenze, E. J. (2007). Cognitive impairment in late-life generalized anxiety disorder. *The American Journal of Geriatric Psychiatry*, *15*(8), 673-679. https://doi.org/10.1097/JGP.0b013e31803111f2

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

Mayeux, R., Ottman, R., Tang, M. X., Noboa-Bauza, L., Marder, K., Gurland, B., & Stern, Y. (1993). Genetic susceptibility and head injury as risk factors for Alzheimer's disease among community-dwelling elderly persons and their first-degree relatives. *Annals of Neurology*, *33*(5), 494-501. https://doi.org/10.1002/ana.410330513

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

Medvedev, O. N., Berk, M., Dean, O. M., Brown, E., Sandham, M. H., Dipnall, J. F., McNamara, R. K., Sumich, A., Krägeloh, C. U., & Narayanan, A. (2020). A novel way to quantify schizophrenia symptoms in clinical trials. *European Journal of Clinical Investigation*, e13398. https://doi.org/10.1111/eci.13398

Medvedev, O. N., & Krägeloh, C. U. (2022). Rasch Measurement Model. In *Handbook of Assessment in Mindfulness Research* (pp. 1-18). Springer.

Medvedev, O. N., Krägeloh, C. U., Hill, E. M., Billington, R., Siegert, R. J., Webster, C. S., Booth, R. J., & Henning, M. A. (2019). Rasch analysis of the Perceived Stress Scale: Transformation from an ordinal to a linear measure. *Journal of Health Psychology*, *24*(8), 1070-1081. https://doi.org/10.1177/1359105316689603

Medvedev, O. N., Krägeloh, C. U., Narayanan, A., & Siegert, R. J. (2017). Measuring mindfulness: applying generalizability theory to distinguish between state and trait. *Mindfulness*, *8*(4), 1036-1046. https://doi.org/10.1007/s12671-017-0679-0

Medvedev, O. N., Krägeloh, C. U., Titkova, E. A., & Siegert, R. J. (2020). Rasch analysis and ordinal-to-interval conversion tables for the Depression, Anxiety and Stress Scale. *Journal of Health Psychology*, 1359105318755261. https://doi.org/10.1177/1359105318755261

Medvedev, O. N., Pratscher, S. D., & Bettencourt, A. (2020). Psychometric Evaluation of the Interpersonal Mindfulness Scale Using Rasch Analysis. *Mindfulness*, *11*(8), 2007-2015. https://doi.org/10.1007/s12671-020-01415-5

Medvedev, O. N., Siegert, R. J., Kersten, P., & Krägeloh, C. U. (2017). Improving the precision of the Five Facet Mindfulness Questionnaire using a Rasch approach. *Mindfulness*, *8*(4), 995-1008. https://doi.org/10.1007/s12671-016-0676-8

Medvedev, O. N., Siegert, R. J., Mohamed, A. D., Shepherd, D., Landhuis, E., & Krägeloh, C. U. (2017). The Oxford Happiness Questionnaire: transformation from an ordinal to an interval measure using Rasch analysis. *Journal of Happiness Studies*, *18*(5), 1425-1443. https://doi.org/10.1007/s10902-016-9784-3

Medvedev, O. N., Titkova, E. A., Siegert, R. J., Hwang, Y. S., & Krageloh, C. U. (2018). Evaluating Short Versions of the Five Facet Mindfulness Questionnaire Using Rasch Analysis. *Mindfulness*, *9*(5), 1411-1422. https://doi.org/10.1007/s12671-017-0881-0

Medvedev, O. N., Turner-Stokes, L., Ashford, S., & Siegert, R. J. (2018). Rasch analysis of the UK Functional Assessment Measure in patients with complex disability after stroke. *Journal of Rehabilitation Medicine*, *50*(5), 420-428. https://doi.org/10.2340/16501977-2324

Meng, X., & D'arcy, C. (2012). Education and dementia in the context of the cognitive reserve hypothesis: a systematic review with meta-analyses and qualitative analyses. *PloS one*, *7*(6), e38268. https://doi.org/10.1371/journal.pone.0038268

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine Rehabilitation*, *70*(4), 308-312.

Merkin, A. G., Medvedev, O. N., Sachdev, P. S., Tippett, L., Krishnamurthi, R., Mahon, S., Kasabov, N., Parmar, P., Crawford, J., & Doborjeh, Z. G. (2020). New avenue for the geriatric depression scale: Rasch transformation enhances reliability of assessment. *Journal of Affective Disorders*, *264*, 7-14. https://doi.org/10.1016/j.jad.2019.11.100

Mitchell-Parker, K., Medvedev, O. N., Krägeloh, C. U., & Siegert, R. J. (2018). Rasch analysis of the Frost multidimensional perfectionism scale. *Australian Journal of Psychology*, *70*(3), 258-268. https://doi.org/10.1111/ajpy.12192

Mitchell, A., Beaumont, H., Ferguson, D., Yadegarfar, M., & Stubbs, B. (2014). Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: meta-analysis. *Acta Psychiatrica Scandinavica*, *130*(6), 439-451. https://doi.org/10.1111/acps.12336

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of clinical psychology*, *47*(6), 790-801. https://doi.org/10.1002/1097-4679(199111)47:6<790::aid-jclp2270470610>3.0.co;2-c

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8-14. https://doi.org/10.1177/0963721411429458

Mortimer, J. A., Alladi, S., Bak, T. H., Russ, T. C., Shailaja, M., & Duggirala, V. (2014). Bilingualism delays age at onset of dementia, independent of education and immigration status. *Neurology*, *82*(21), 1936. https://doi.org/10.1212/WNL.0000000000000400

Mukamal, K. J., Kuller, L. H., Fitzpatrick, A. L., Longstreth Jr, W. T., Mittleman, M. A., & Siscovick, D. S. (2003). Prospective study of alcohol consumption and risk of dementia in older adults. *Jama*, *289*(11), 1405-1413. https://doi.org/10.1001/jama.289.11.1405

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171-189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Nation, D. A., Edland, S. D., Bondi, M. W., Salmon, D. P., Delano-Wood, L., Peskind, E. R., Quinn, J. F., & Galasko, D. R. J. N. (2013). Pulse pressure is associated with Alzheimer biomarkers in cognitively normal older adults. *Neurology*, *81*(23), 2024-2027.

Norquist, J. M., Fitzpatrick, R., Dawson, J., & Jenkinson, C. (2004). Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Medical Care*, 125-136. https://doi.org/10.1097/01.mlr.0000103530.13056.88

Norris, D. R., Clark, M. S., & Shipley, S. (2016). The mental status examination. *American family physician*, *94*(8), 635-641.

Nosheny, R. L., Amariglio, R., Sikkes, S. A., Van Hulle, C., Bicalho, M. A. C., Dowling, N. M., Brucki, S. M. D., Ismail, Z., Kasuga, K., & Kuhn, E. (2022). The role of dyadic cognitive report and subjective cognitive decline in early ADRD clinical research and trials: Current knowledge, gaps, and recommendations. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *8*(1), e12357. https://doi.org/10.1002/trc2.12357

Numbers, K., Crawford, J. D., Kochan, N. A., Draper, B., Sachdev, P. S., & Brodaty, H. (2020). Participant and informant memory-specific cognitive complaints predict future decline and incident dementia: Findings from the Sydney Memory and Ageing Study. *PloS one*, *15*(5), e0232961. https://doi.org/10.1371/journal.pone.0232961

O'Brien, J., Jackson, J. W., Grodstein, F., Blacker, D., & Weuve, J. (2014). Postmenopausal hormone therapy is not associated with risk of all-cause dementia and Alzheimer's disease. *Epidemiologic Reviews*, *36*(1), 83-103. https://doi.org/10.1093/epirev/mxt008

O'Keeffe, E., Mukhtar, O., & T O'Keeffe, S. (2011). Orientation to time as a guide to the presence and severity of cognitive impairment in older hospital patients. *Journal of Neurology, Neurosurgery & Psychiatry*, *82*(5), 500-504. https://doi.org/10.1136/jnnp.2010.214817

O'Bryant, S. E., Humphreys, J. D., Smith, G. E., Ivnik, R. J., Graff-Radford, N. R., Petersen, R. C., & Lucas, J. A. (2008). Detecting dementia with the mini-mental state examination in highly educated individuals. *Archives of Neurology*, *65*(7), 963-967. https://doi.org/10.1001/archneur.65.7.963

Panza, F., Frisardi, V., Capurso, C., D'Introno, A., Colacicco, A. M., Imbimbo, B. P., Santamato, A., Vendemiale, G., Seripa, D., Pilotto, A., Capurso, A., & Solfrizzi, V. (2010). Late-life depression, mild cognitive impairment, and dementia: possible continuum? *The American*

Journal of Geriatric Psychiatry, *18*(2), 98-116. https://doi.org/10.1097/JGP.0b013e3181b0fa13

Park, M. H. (2017). Informant questionnaire on cognitive decline in the elderly (IQCODE) for classifying cognitive dysfunction as cognitively normal, mild cognitive impairment, and dementia. *International Psychogeriatrics*, *29*(9), 1461-1467. https://doi.org/10.1017/S1041610217000965

Paterson, J., Medvedev, O. N., Sumich, A., Tautolo, E.-S., Krägeloh, C. U., Sisk, R., McNamara, R. K., Berk, M., Narayanan, A., & Siegert, R. J. (2018). Distinguishing transient versus stable aspects of depression in New Zealand Pacific Island children using Generalizability Theory. *Journal of Affective Disorders*, *227*, 698-704. https://doi.org/10.1016/j.jad.2017.11.075

Patnode, C. D., Perdue, L. A., Rossom, R. C., Rushkin, M. C., Redmond, N., Thomas, R. G., & Lin, J. S. (2020). Screening for Cognitive Impairment in Older Adults: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *Jama*, *323*(8), 764-785. https://doi.org/10.1001/jama.2019.22258

Pendlebury, S. T., & Rothwell, P. M. (2009). Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis. *The Lancet Neurology*, *8*(11), 1006-1018.

Perroco, T. R., Damin, A. E., Frota, N. A., Silva, M.-N. M., Rossi, V., Nitrini, R., & Bottino, C. (2008). Short IQCODE as a screening tool for MCI and dementia: preliminary results. *Dementia & Neuropsychologia*, *2*(4), 300-304. https://doi.org/10.1590/S1980-57642009DN20400012

Peters, R., Booth, A., Rockwood, K., Peters, J., D'Este, C., & Anstey, K. J. (2019). Combining modifiable risk factors and risk of dementia: a systematic review and meta-analysis. *BMJ Open*, *9*(1), e022846. https://doi.org/10.1136/bmjopen-2018-022846

Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, *256*(3), 183-194. https://doi.org/10.1111/j.1365-2796.2004.01388.x

Petersen, R. C. (2016). Mild cognitive impairment. *Continuum: Lifelong Learning in Neurology*, *22*(2 Dementia), 404. https://doi.org/10.1212/CON.0000000000000313

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology, 56,* 303-308. https://doi.org/10.1001/archneur.56.3.303

Petersen, R. C., Thomas, R. G., Grundman, M., Bennett, D., Doody, R., Ferris, S., Galasko, D., Jin, S., Kaye, J., Levey, A., Pfeiffer, E., Sano, M., van Dyck, C. H., Thal, L. J., & Alzheimer's Disease Cooperative Study Group. (2005). Vitamin E and donepezil for the treatment of mild cognitive impairment. *New England Journal of Medicine*, *352*(23), 2379-2388. https://doi.org/10.1056/NEJMoa050151

Phung, T. K. T., Chaaya, M., Asmar, K., Atweh, S., Ghusn, H., Khoury, R. M., Prince, M., & Waldemar, G. (2015). Performance of the 16-item informant questionnaire on cognitive decline for the elderly (IQCODE) in an Arabic-speaking older population. *Dementia and Geriatric Cognitive Disorders*, *40*(5-6), 276-289. https://doi.org/10.1159/000437092

Piguet, O., Hornberger, M., Shelley, B. P., Kipps, C. M., & Hodges, J. R. (2009). Sensitivity of current criteria for the diagnosis of behavioral variant frontotemporal dementia. *Neurology, 72*(8), 732-737. https://doi.org/10.1212/01.wnl.0000343004.98599.45

Plassman, B. L., Havlik, R. J., Steffens, D. C., Helms, M. J., Newman, T. N., Drosdick, D., Phillips, C., Gau, B. A., Welsh–Bohmer, K. A., Burke, J. R., Guralnik, J. M., & Breitner, J. C. (2000). Documented head injury in early adulthood and risk of Alzheimer's disease and other dementias. *Neurology*, *55*(8), 1158-1166. https://doi.org/10.1212/wnl.55.8.1158

Plassman, B. L., Newman, T. T., Welsh, K. A., & Helms, M. (1994). Properties of the Telephone Interview for Cognitive Status: application in epidemiological and longitudinal studies. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*, *7*(3), 235–241.

Ponds, R. W. H. M., & Jolles, J. (1996). Memory complaints in elderly people: The role of memory abilities, metamemory, depression, and personality. *Educational Gerontology: An International Quarterly*, *22*(4), 341-357. https://doi.org/10.1080/0360127960220404

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. P. (2013). The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's, Dementia and Geriatric Cognitive Disorders*, *9*(1), 63-75. e62. https://doi.org/10.1016/j.jalz.2012.11.007

Qiu, C., Winblad, B., Marengoni, A., Klarin, I., Fastbom, J., & Fratiglioni, L. (2006). Heart failure and risk of dementia and Alzheimer disease: a population-based cohort study. *Archives of Internal Medicine, 166*(9), 1003-1008. https://doi.org/10.1001/archinte.166.9.1003

Quik, M., Bordia, T., & O'Leary, K. (2007). Nicotinic receptors as CNS targets for Parkinson's disease. *Biochemical Pharmacology, 74*(8), 1224-1234. https://doi.org/10.1016/j.bcp.2007.06.015

Quik, M., Perez, X. A., & Bordia, T. (2012). Nicotine as a potential neuroprotective agent for Parkinson's disease. *Movement Disorders, 27*(8), 947-957. https://doi.org/10.1002/mds.25028

Rait, G., Fletcher, A., Smeeth, L., Brayne, C., Stirling, S., Nunes, M., Breeze, E., Ng, E. S.-W., Bulpitt, C. J., & Jones, D. (2005). Prevalence of cognitive impairment: results from the MRC trial of assessment and management of older people in the community. *Age Ageing, 34*(3), 242-248. https://doi.org/10.1093/ageing/afi039

Ramanaiah, N. V., Franzen, M., & Schill, T. (1983). A psychometric study of the State-Trait Anxiety Inventory. *Journal of Personality Assessment, 47*(5), 531-535. https://doi.org/10.1207/s15327752jpa4705_14

Rami, L., Mollica, M. A., García-Sanchez, C., Saldaña, J., Sanchez, B., Sala, I., Valls-Pedret, C., Castellví, M., Olives, J., & Molinuevo, J. L. (2014). The subjective cognitive decline questionnaire (SCD-Q): a validation study. *Journal of Alzheimer's Disease, 41*(2), 453-466. https://doi.org/10.3233/JAD-132027

Ramirez, D., Wood, R. C., Becho, J., Owings, K., & Espino, D. V. (2010). Mini-Mental State Exam domains predict falls in an elderly population: Follow-up from the Hispanic Established Populations for Epidemiologic Studiesof the Elderly (H-EPESE) Study. *Ethnicity & Disease, 20*(1), 48-52.

Randolph, C., Tierney, M. C., Mohr, E., & Chase, T. N. (1998). The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary clinical validity. *Journal of Clinical and Experimental Neuropsychology, 20*(3), 310-319. https://doi.org/10.1076/jcen.20.3.310.823

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology.* Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, University of California.

Ravaglia, G., Forti, P., Maioli, F., Chiappelli, M., Montesi, F., Tumini, E., Mariani, E., Licastro, F., & Patterson, C. (2007). Blood inflammatory markers and risk of dementia: The Conselice Study of Brain Ageing. *Neurobiology of Ageing*, *28*(12), 1810-1820. https://doi.org/10.1016/j.neurobiolageing.2006.08.012

Reid, L. M., & MacLullich, A. M. J. (2006). Subjective memory complaints and cognitive impairment in older people. *Dementia and Geriatric Cognitive Disorders*, *22*(5-6), 471-485. https://doi.org/10.1159/000096295

Reid, M., Parkinson, L., Gibson, R., Schofield, P., D'Este, C., Attia, J., Tavener, M., & Byles, J. (2012). Memory complaint questionnaire performed poorly as screening tool: validation against psychometric tests and affective measures. *Journal of Clinical Epidemiology*, *65*(2), 199-205. https://doi.org/10.1016/j.jclinepi.2011.06.006

Reitan, R. M., & Wolfson, D. (1993). The Halstead–Reitan neuropsychological test battery. In: Theory and clinical interpretation. South Tucson: Neuropsychology Press.

Richard, E., Reitz, C., Honig, L. H., Schupf, N., Tang, M. X., Manly, J. J., Mayeux, R., Devanand, D., & Luchsinger, J. A. (2013). Late-life depression, mild cognitive impairment, and dementia. *JAMA Neurology*, *70*(3), 383-389. https://doi.org/10.1001/jamaneurol.2013.603

Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, *125*(6), 747. https ://doi.org/10.1037/abn0000181

Rocca, M. A., Riccitelli, G. C., Meani, A., Pagani, E., Del Sette, P., Martinelli, V., Comi, G., Falini, A., & Filippi, M. (2019). Cognitive reserve, cognition, and regional brain damage in MS: a 2-year longitudinal study. *Multiple Sclerosis Journal*, *25*(3), 372-381. https://doi.org/10.1177/1352458517750767

Roses, M., Allen D. (1996). Apolipoprotein E alleles as risk factors in Alzheimer's disease. *Annual Review of Medicine*, *47*(1), 387-400. https://doi.org/10.1146/annurev.med.47.1.387

Roth, G. A., Abate, D., Maulik, P. K., Sachdev, P. S., Wolfe, C. D. A., Wu, S. L., Wu, Y. C., Murray, C. J. L., & ...GBD Causes Death Collaborotors. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1736-1788. https://doi.org/10.1016/S0140-6736(18)32203-7

Ruitenberg, A., van Swieten, J. C., Witteman, J. C. M., Mehta, K. M., van Duijn, C. M., Hofman, A., & Breteler, M. M. B. (2002). Alcohol consumption and risk of dementia: the Rotterdam Study. *The Lancet Neurology*, *359*(9303), 281-286. https://doi.org/10.1016/S0140-6736(02)07493-7

Sachdev, P. S., Brodaty, H., Reppermund, S., Kochan, N. A., Trollor, J. N., Draper, B., Slavin, M. J., Crawford, J., Kang, K., Broe, G. A., Mather, K. A., Lux, O., & Memory and Ageing Study Team. (2010). The Sydney Memory and Ageing Study (MAS): methodology and baseline medical and neuropsychiatric characteristics of an elderly epidemiological non-demented cohort of Australians aged 70-90 years. *International Psychogeriatrics*, *22*(8), 1248. https://doi.org/10.1017/S1041610210001067

Sachdev, P. S., Lammel, A., Trollor, J. N., Lee, T., Wright, M. J., Ames, D., Wen, W., Martin, N. G., Brodaty, H., & Schofield, P. R. (2009). A comprehensive neuropsychiatric study of elderly twins: the Older Australian Twins Study. *Twin Research and Human Genetics*, *12*(6), 573-582. https://doi.org/10.1375/twin.12.6.573

Sandham, M. H., Medvedev, O. N., Hedgecock, E., Higginson, I. J., & Siegert, R. J. (2019). A Rasch analysis of the integrated palliative care outcome scale. *Journal of Pain and Symptom Management*, *57*(2), 290-296. https://doi.org/10.1016/j.jpainsymman.2018.11.019

Santangeli, P., Di Biase, L., Bai, R., Mohanty, S., Pump, A., Brantes, M. C., Horton, R., Burkhardt, J. D., Lakkireddy, D., & Reddy, Y. M. (2012). Atrial fibrillation and the risk of incident dementia: a meta-analysis. *Heart Rhythm*, *9*(11), 1761-1768. https://doi.org/10.1016/j.hrthm.2012.07.026

Schmidt, R., Schmidt, H., Curb, J. D., Masaki, K., White, L. R., & Launer, L. (2002). Early inflammation and dementia: a 25-year follow-up of the Honolulu-Asia Ageing Study. *Annals of Neurology*, *52*(2), 168-174. https://doi.org/10.1002/ana.10265

Schmitt, A. L., Livingston, R. B., Goette, W. F., & Galusha-Glasscock, J. M. (2016). Relationship between the Mini-Mental State Examination and the Repeatable Battery for the Assessment of Neuropsychological Status in patients referred for a dementia evaluation. *Perceptual and Motor Skills*, *123*(3), 606-623. https://doi.org/10.1177/0031512516667674

Seo, E. H., Lee, D. Y., Kim, S. G., Kim, K. W., Kim, D. H., jo Kim, B., Kim, M.-D., Kim, S. Y., Kim, Y. H., & Kim, J.-L. (2011). Validity of the telephone interview for cognitive status (TICS) and modified TICS (TICSm) for mild cognitive imparment (MCI) and dementia screening. *Archives of Gerontology and Geriatrics*, *52*(1), e26-e30. https://doi.org/10.1016/j.archger.2010.04.008

Shao, H., Breitner, J. C., Whitmer, R. A., Wang, J., Hayden, K., Wengreen, H., Corcoran, C., Tschanz, J., Norton, M., & Munger, R. (2012). Hormone therapy and Alzheimer disease dementia: new findings from the Cache County Study. *Neurology*, *79*(18), 1846-1852. https://doi.org/10.1212/WNL.0b013e318271f823

Shatil, E., Metzer, A., Horvitz, O., & Miller, A. (2010). Home-based personalized cognitive training in MS patients: a study of adherence and cognitive performance. *Neurorehabilitation*, *26*(2), 143-153. https://doi.org/10.3233/NRE-2010-0546

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). SAGE Publications, Inc.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922-932. https://doi.org/10.1037/0003-066X.44.6.922

Shumaker, S. A., Legault, C., Kuller, L., Rapp, S. R., Thal, L., Lane, D. S., Fillit, H., Stefanick, M. L., Hendrix, S. L., & Lewis, C. E. (2004). Conjugated equine estrogens and incidence of probable dementia and mild cognitive impairment in postmenopausal women: Women's Health Initiative Memory Study. *Jama*, *291*(24), 2947-2958. https://doi.org/10.1001/jama.291.24.2947

Shumaker, S. A., Legault, C., Rapp, S. R., Thal, L., Wallace, R. B., Ockene, J. K., Hendrix, S. L., Jones III, B. N., Assaf, A. R., & Jackson, R. D. (2003). Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *Jama*, *289*(20), 2651-2662. https://doi.org/10.1001/jama.289.20.2651

Sierra-Rio, A., Balasa, M., Olives, J., Antonell, A., Iranzo, A., Castellví, M., Bosch, B., Grau-Rivera, O., Fernandez-Villullas, G., & Rami, L. (2016). Cerebrospinal fluid biomarkers predict clinical evolution in patients with subjective cognitive decline and mild cognitive impairment. *Neurodegenerative Diseases*, *16*(1-2), 69-76.

Skoog, I., Börjesson-Hanson, A., Kern, S., Johansson, L., Falk, H., Sigström, R., & Östling, S. (2017). Decreasing prevalence of dementia in 85-year olds examined 22 years apart: the influence of education and stroke. *Scientific Reports*, *7*(1), 1-8. https://doi.org/10.1038/s41598-017-05022-8

Slavin, M. J., Sachdev, P. S., Kochan, N. A., Woolf, C., Crawford, J. D., Giskes, K., Reppermund, S., Trollor, J. N., Draper, B., & Delbaere, K. (2015). Predicting cognitive, functional, and diagnostic change over 4 years using baseline subjective cognitive complaints in the Sydney Memory and Ageing Study. *The American Journal of Geriatric Psychiatry*, *23*(9), 906-914. https://doi.org/10.1016/j.jagp.2014.09.001

Smith Jr, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205-231.

Spearman, C. (1927). The measurement of intelligence. *Nature*, *120*(3025), 577-578.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack Jr, C. R., Kaye, J., & Montine, T. J. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Ageing-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 280-292. https://doi.org/10.1016/j.jalz.2011.03.003

Spielberger, C. D. (1999). *Staxi-2: state-trait anger expression inventory-2; professional manual*. Psychological Assessment Resources.

Spielberger, C. D., Gorsuch, R. L., Lushene, R. E., Vagg, P. R., & Jacobs, G. A. (1970). *Manual for State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spreen, O. (1977). Neurosensory center comprehensive examination for aphasia. *Neuropsychological Laboratory*.

Srinivasan, R., Davidson, Y., Gibbons, L., Payton, A., Richardson, A. M. T., Varma, A., Julien, C., Stopford, C., Thompson, J., & Horan, M. A. (2006). The apolipoprotein E ε4 allele selectively increases the risk of frontotemporal lobar degeneration in males. *Journal of Neurology, Neurosurgery Psychiatry*, *77*(2), 154-158. https://doi.org/10.1136/jnnp.2005.063966

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. American chemical society.

Striepens, N., Scheef, L., Wind, A., Popp, J., Spottke, A., Cooper-Mahkorn, D., Suliman, H., Wagner, M., Schild, H. H., & Jessen, F. (2010). Volume loss of the medial temporal lobe structures in subjective memory impairment. *Dementia and Geriatric Cognitive Disorders*, *29*(1), 75-81.

Stucki, G., Daltroy, L., Katz, J. N., Johannesson, M., & Liang, M. H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *Journal of Clinical Epidemiology*, *49*(7), 711-717. https://doi.org/10.1016/0895-4356(96)00016-9

Sundström, A., Nilsson, L.-G., Cruts, M., Adolfsson, R., Van Broeckhoven, C., & Nyberg, L. (2007). Increased risk of dementia following mild head injury for carriers but not for non-carriers of the APOE ε4 allele. *International Psychogeriatrics*, *19*(1), 159-165. https://doi.org/10.1017/S1041610206003498

Swiger, K. J., Manalac, R. J., Blumenthal, R. S., Blaha, M. J., & Martin, S. S. (2013). Statins and cognition: a systematic review and meta-analysis of short-and long-term cognitive effects. Mayo clinic proceedings,

Swiss Society for Research in Education Working Group. (2006). *EDUG User Guide*. Neuchatel, Switzerland: IRDP

Szekely, C. A., Breitner, J. C., Fitzpatrick, A. L., Rea, T. D., Psaty, B. M., Kuller, L. H., & Zandi, P. P. (2008). NSAID use and dementia risk in the Cardiovascular Health Study*: Role of APOE and NSAID type. *Neurology*, *70*(1), 17-24. https://doi.org/10.1212/01.wnl.0000284596.95156.48

Tang, W. K., Wong, E., Chan, S. S., Chiu, H. F., Wong, K. S., Kwok, T. C., Mok, V., & Ungvari, G. S. (2004). The scoring scheme of the informant questionnaire on cognitive decline in the elderly

needs revision: results of rasch analysis. *Dementia and Geriatric Cognitive Disorders*, *18*(3-4), 250-256. https://doi.org/10.1159/000080024

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*(8), 1358-1362. https://doi.org/10.1002/art.23108

Tennant, A., Horton, M., & Pallant, J. (2011). Introductory rasch analysis: a workbook. *Leeds, UK: Department of Rehabilitation Medicine, University of Leeds*.

Thurstone, L. L. (1931). The Measurement of Social Attitude. *Journal of Abnormal & Social Psychology*, *26*(3), 249-269. https://doi.org/10.1037/h0070363

Tierney, M. C., Szalai, J. P., Snow, W. G., Fisher, R. H., & Dunn, E. (1997). Domain specificity of the subtests of the Mini-Mental State Examination. *Archives of Neurology*, *54*(6), 713-716. https://doi.org/10.1001/archneur.1997.00550180035009

Tilvis, R. S., Kähönen-Väre, M. H., Jolkkonen, J., Valvanne, J., Pitkala, K. H., & Strandberg, T. E. (2004). Predictors of cognitive decline and mortality of aged people over a 10-year period. *The journals of gerontology. Series A, Biological Sciences and Medical Sciences*, *59*(3), 268-274. https://doi.org/10.1093/gerona/59.3.m268

Tombaugh, T. N. (2004). Trail Making Test A and B: normative data stratified by age and education. *Archives of clinical neuropsychology*, *19*(2), 203-214.

Trollor, J. N., Smith, E., Agars, E., Kuan, S. A., Baune, B. T., Campbell, L., Samaras, K., Crawford, J., Lux, O., & Kochan, N. A. (2012). The association between systemic inflammation and cognitive performance in the elderly: the Sydney Memory and Ageing Study. *Age*, *34*(5), 1295-1308. https://doi.org/10.1007/s11357-011-9301-x

Truong, Q. C., Cervin, M., Choo, C. C., Numbers, K., Bentvelzen, A. C., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., & Kochan, N. A. (2023). Using network analysis to validate domains of the modified telephone interview for cognitive status. *European Journal of Clinical Investigation*, e14016. https://doi.org/10.1111/eci.14016

Truong, Q. C., Choo, C., Numbers, K., Bentvelzen, A., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N. A., & Medvedev, O. N. (2023). Enhancing precision of the Telephone

Interview for Cognitive Status-modified (TICS-M) using the Rasch Model. *Psychological Assessment*. https://doi.org/10.1037/pas0001233

Truong, Q. C., Choo, C., Numbers, K., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N. A., & Medvedev, O. N. (2021a). Applying generalizability theory to examine assessments of subjective cognitive complaints: whose reports should we rely on – participant versus informant? . *International Psychogeriatrics*, *34*(12), 1–11. https://doi.org/10.1017/s1041610221000363

Truong, Q. C., Choo, C., Numbers, K., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N. A., & Medvedev, O. N. (2021b). Enhancing precision of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE-16) using Rasch methodology. *International Psychogeriatrics*, 1-11. https://doi.org/10.1017/S1041610221002568

Truong, Q. C., Choo, C., Numbers, K., Merkin, A. G., Sachdev, P. S., Feigin, V. L., Brodaty, H., Kochan, N. A., & Medvedev, O. N. (2022). Clinical investigation of dynamic and enduring aspects of global cognition in aged population. *European Journal of Clinical Investigation*, *52*(2), e13681. https://doi.org/10.1111/eci.13681

Truong, Q. C., Krageloh, C. U., Siegert, R. J., Landon, J., & Medvedev, O. N. (2020). Applying Generalizability Theory to Differentiate Between Trait and State in the Five Facet Mindfulness Questionnaire (FFMQ). *Mindfulness*, *11*(4), 953-963. https://doi.org/10.1007/s12671-020-01324-7

Truong, Q. C., Numbers, K., Choo, C. C., Bentvelzen, A. C., Catts, V. S., Cervin, M., Jorm, A. F., Kochan, N. A., Brodaty, H., Sachdev, P., & Medvedev, O. N. (2023). Establishing conversion of the 16-item Informant Questionnaire on Cognitive Decline in the Elderly scores into interval-level data across multiple samples using Rasch methodology. *Psychogeriatrics*. https://doi.org/10.1111/psyg.12946

Tsuang, D., Leverenz, J. B., Lopez, O. L., Hamilton, R. L., Bennett, D. A., Schneider, J. A., Buchman, A. S., Larson, E. B., Crane, P. K., Kaye, J. A., & Zabetian, C. P. (2013). APOE ϵ4 increases risk for dementia in pure synucleinopathies. *JAMA neurology*, *70*(2), 223-228. https://doi.org/10.1001/jamaneurol.2013.600

US Army. (1944). *Army individual test battery: Manual of directions and scoring*. Washington, DC: War Department, Adjunct General's Office.

Van Den Berg, E., Ruis, C., Biessels, G. J., Kappelle, L. J., & Van Zandvoort, M. J. (2012). The Telephone Interview for Cognitive Status (Modified): relation with a comprehensive neuropsychological assessment. *Journal of Clinical and Experimental Neuropsychology*, *34*(6), 598-605. https://doi.org/10.1080/13803395.2012.667066

Verghese, J., LeValley, A., Derby, C., Kuslansky, G., Katz, M., Hall, C., Buschke, H., & Lipton, R. B. (2006). Leisure activities and the risk of amnestic mild cognitive impairment in the elderly. *Neurology*, *66*(6), 821-827. https://doi.org/10.1212/01.wnl.0000202520.68987.48

Verghese, J., Lipton, R. B., Katz, M. J., Hall, C. B., Derby, C. A., Kuslansky, G., Ambrose, A. F., Sliwinski, M., & Buschke, H. (2003). Leisure activities and the risk of dementia in the elderly. *New England Journal of Medicine*, *348*(25), 2508-2516. https://doi.org/10.1056/NEJMoa022252

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1. https://doi.org/10.1037/met0000107

Wang, H.-X., Karp, A., Winblad, B., & Fratiglioni, L. (2002). Late-life engagement in social and leisure activities is associated with a decreased risk of dementia: a longitudinal study from the Kungsholmen project. *American Journal of Epidemiology*, *155*(12), 1081-1087. https://doi.org/10.1093/aje/155.12.1081

Ward, A., Arrighi, H. M., Michels, S., & Cedarbaum, J. M. (2012). Mild cognitive impairment: disparity of incidence and prevalence estimates. *Alzheimer's & Dementia*, *8*(1), 14-21. https://doi.org/10.1016/j.jalz.2011.01.002

Wechsler, D. (1981a). *WAIS-R manual: Wechsler adult intelligence scale-revised*. Psychological Corporation.

Wechsler, D. (1981b). *Wechsler adult intelligence scale: WAIS-R manual*. Harcourt Brace Jovanovich [for] The Psychological Corporation.

Wechsler, D. (1997). *Wechsler memory scale (WMS-III)* (Vol. 14). Psychological corporation San Antonio, TX.

Wechsler, D. (2003). *WISC-IV: Administration and scoring manual*. Psychological Corporation.

Welsh, K. A., Breitner, J. C., & Magruder-Habib, K. M. (1993). Detection of dementia in the elderly using telephone screening of cognitive status. *Cognitive and Behavioral Neurology*, *6*(2), 103-110.

Weston, W. W. (1987). Screening for dementia in family practice. *Canadian Family Physician*, *33*, 2495.

Whitmer, R. A., Quesenberry, C. P., Zhou, J., & Yaffe, K. (2011). Timing of hormone therapy and dementia: the critical window theory revisited. *Annals of Neurology*, *69*(1), 163-169. https://doi.org/10.1002/ana.22239

Williams, D. R., & Mulder, J. (2020). BGGM: Bayesian Gaussian graphical models in R. *Journal of Open Source Software*, *5*(51), 2111. https://doi.org/10.31234/osf.io/t2cn7

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.

Wilson, R. S., & Bennett, D. A. (2005). Assessment of cognitive decline in old age with brief tests amenable to telephone administration. *Neuroepidemiology*, *25*(1), 19-25. https://doi.org/10.1159/000085309

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L. O., Nordberg, A., Bäckman, L., Albert, M., & Almkvist, O. (2004). Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, *256*(3), 240-246. https://doi.org/10.1111/j.1365-2796.2004.01380.x

Wong, W. B., Lin, V. W., Boudreau, D., & Devine, E. B. (2013). Statins in the prevention of dementia and Alzheimer's disease: a meta-analysis of observational studies and an assessment of confounding. *Pharmacoepidemiology and Drug Safety*, *22*(4), 345-358. https://doi.org/10.1002/pds.3381

World Health Organization. (2003). *ICD-10: International Statistical Classification of Diseases and Related Health Problems: tenth Revision, 2nd ed*. Geneva: World Health Organization.

World Health Organization. (2018). *ICD-11: International Statistical Classification of Diseases and Related Health Problems: eleventh Revision*. Geneva: World Health Organization.

World Health Organization. (2022). Mental Health. World Health Organization. https://www.who.int/en/news-room/fact-sheets/detail/mental-health-strengthening-our-response

Wortmann, M. (2011). P3-367: Global economic cost of dementia. *Alzheimer's & Dementia*, *7*, S633-S633.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Xu, J., Kochanek, K. D., Murphy, S. L., & Tejada-Vera, B. (2010). Deaths: final data for 2007 *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System 58*(19), 1-19.

Yaffe, K., Vittinghoff, E., Lindquist, K., Barnes, D., Covinsky, K. E., Neylan, T., Kluse, M., & Marmar, C. (2010). Posttraumatic stress disorder and risk of dementia among US veterans. *Archives of General Psychiatry*, *67*(6), 608-613. https://doi.org/10.1001/archgenpsychiatry.2010.61

Ye, Z. J., Zhang, Z., Zhang, X. Y., Tang, Y., Chen, P., Liang, M. Z., Sun, Z., & Yu, Y. L. (2020). State or trait? Measuring resilience by generalisability theory in breast cancer. *European Journal of Oncology Nursing*, *46*, 101727. https://doi.org/10.1016/j.ejon.2020.101727

Yin, Y.-W., Li, J.-C., Wang, J.-Z., Li, B.-H., Pi, Y., Yang, Q.-W., Fang, C.-Q., Gao, C.-Y., & Zhang, L.-L. (2012). Association between apolipoprotein E gene polymorphism and the risk of vascular dementia: a meta-analysis. *Neuroscience Letters*, *514*(1), 6-11. https://doi.org/10.1016/j.neulet.2012.02.031

Yip, A. G., Green, R. C., Huyck, M., Cupples, L. A., & Farrer, L. A. (2005). Nonsteroidal anti-inflammatory drug use and Alzheimer's disease risk: the MIRAGE Study. *BMC Geriatrics*, *5*(1), 1-6. https://doi.org/10.1186/1471-2318-5-2

Zandi, P. P., Carlson, M. C., Plassman, B. L., Welsh-Bohmer, K. A., Mayer, L. S., Steffens, D. C., Breitner, J. C., & Investigators, C. C. M. S. (2002). Hormone replacement therapy and incidence of Alzheimer disease in older women: the Cache County Study. *Jama*, 288(17), 2123-2129. https://doi.org/10.1001/jama.288.17.2123