ADVANCEMENT OF THE CHEDOKE ARM AND HAND ACTIVITY INVENTORY

APPLYING MEASUREMENT THEORIES TO THE ADVANCEMENT OF THE CHEDOKE ARM AND HAND ACTIVITY INVENTORY

By Xinyi Silvana Choo, B.ASc. (OT), OT Reg (Singapore)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree

Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2018) Hamilton, Ontario (Rehabilitation Science)

TITLE: Applying measurement theories to the advancement of the Chedoke Arm and Hand Activity Inventory

AUTHOR: Xinyi Silvana Choo, B.ASc. (OT) (University of Sydney)

SUPERVISOR: Dr. Julie Richardson, PhD, PT.

NUMBER OF PAGES: xxii, 281

Lay Abstract

Weakness in the arm/hand is a common problem after a stroke and can affect one's ability to carry out daily tasks. The function of the affected arm/hand should be measured to track improvements after treatment. The Chedoke Arm and Hand Activity Inventory (CAHAI) is a tool used to measure arm/hand function for persons with stroke. This thesis describes the development and testing of a Singapore version of the CAHAI. We conducted this research because when a tool is used outside of the country where it was developed, there is a need to test how well it works. This thesis also describes the testing of the CAHAI to find out if it behaves like a ruler. This research is important because when the CAHAI behaves like a ruler, we know that it measures arm/hand function in the same manner when it is used with different persons with stroke.

Abstract

Background: The use of outcome measures to evaluate upper extremity function after stroke is highly recommended in clinical practice and research. The Chedoke Arm and Hand Activity Inventory (CAHAI) is a recommended measure as it has strong psychometric properties and clinical utility. However, the measure has not been validated in Asia and there are also gaps in the knowledge about the psychometric properties of the CAHAI.

Aim & Objectives: This thesis is dedicated to the continued evaluation of the CAHAI with two main objectives: (1) to develop a Singapore version of the CAHAI, and (2) to re-evaluate the original CAHAI using modern test theories.

Method: We conducted a study to cross-culturally adapt the CAHAI and evaluated the psychometric properties in a stroke sample in Singapore. Two studies were conducted to re-evaluate the original CAHAI using modern test theories. In the first study, item response theory and Rasch measurement theory were used to evaluate the psychometric properties of the measure. Following which, both measurement theories were used to revise the CAHAI in the second study.

Results: Two test items were modified for the Singapore version of the CAHAI, and the measure had good inter-rater reliability (intra-class correlation coefficient = 0.95 – 0.97) and construct validity. The evaluation of the original CAHAI using modern test theories identified three main problems: (1) the scoring scale was not working as intended, (2) local dependency, and (3) the measure was not unidimensional. Revisions to the CAHAI included collapsing the 7-category scale to four categories, deleting two test items, and developing two new shortened versions.

Conclusion: The Singapore version of the CAHAI is a valid and culturally relevant outcome measure that can be used to evaluate post-stroke upper extremity function. The original CAHAI was refined into a new 11- and 5-item versions with a 4-category scale which clinicians may find easier to use.

Acknowledgements

To all the people who believed

How much I could achieve

And saw my potential when I never did

Nothing could have taken me this far

Know that it was because of all of you

Your enduring support and belief

Over the past years

Undoubtedly is the heart in all that I have accomplished

Supervisor, Dr. Julie Richardson

Supervisory Committee, Dr. Jocelyn Harris, Dr. Jackie Bosch, & Professor Paul Stratford

School of Rehabilitation Science Faculty Member, Dr. Ayse Kuspinar

Funding Agency & Employer, Singapore General Hospital

Singapore General Hospital, Ms Leila Nasron, Dr. Therma Cheung & Colleagues

My Dearest Family & Friends

Preface

This thesis is structured as a sandwich thesis. It consists of four manuscripts (Chapters 2 – 5) that are positioned between the introductory (Chapter 1) and concluding (Chapter 6) chapters of this thesis. Each manuscript is presented in the format according to submission requirements of the target peer-reviewed journal for publication. I, Xinyi Silvana Choo, am the first author for all of the included manuscripts.

Chapters 2 and 3

These two chapters present the original study on the development of the Singapore version of the Chedoke Arm and Hand Activity Inventory. The study was conducted between September 2016 and June 2017. Two manuscripts are presented in Chapters 2 and 3, entitled "*Cross-cultural adaptation and psychometric evaluation of the Singapore version of the Chedoke Arm and Hand Activity*" and "*Reliability and validity of the shortened Singapore versions of the Chedoke Arm and Hand Activity Inventory*" respectively.

I, Xinyi Silvana Choo, initiated and originated the research idea, developed the initial study design, obtained ethics approval, implemented the data collection, conducted the data analysis, and wrote the manuscripts. The co-authors on these papers include Dr. Jackie Bosh, Dr. Julie Richardson, Professor Paul Stratford, and Dr. Jocelyn Harris.

Dr. Jackie Bosh contributed to the data collection process (access to the original administration manual), the interpretation of findings, and provided critical review of earlier versions of the manuscripts and the final approval of the version to be published.

Dr. Julie Richardson contributed to the refinement of the study design, the interpretation of findings, and provided critical review of earlier versions of the manuscripts and the final approval of the version to be published.

Professor Paul Stratford contributed to the refinement of the study design, statistical analysis, the interpretation of findings, and provided critical review of earlier versions of the manuscripts and the final approval of the version to be published.

Dr. Jocelyn Harris contributed to the refinement of the study design, the interpretation of findings, supervised the data collection process, and provided critical review of earlier versions of the manuscripts and the final approval of the version to be published.

Chapters 4 and 5

These two chapters present the re-evaluation of the Chedoke Arm and Hand Activity Inventory using data from a previous validation study of the measure. The secondary data analysis for this thesis was conducted between May 2017 and April 2018. Two manuscripts are presented in Chapters 4 and 5, entitled "*Measurement theories in rehabilitation: Introduction to item response theory and Rasch measurement theory*" and

"*Revising the Chedoke Arm and Hand Activity Inventory using modern test theories*" respectively.

I, Xinyi Silvana Choo, defined the research questions, developed the initial statistical analysis plans, conducted the data analysis, and wrote the manuscripts. The co-authors on these papers include Professor Paul Stratford, Dr. Ayse Kuspinar, Dr. Julie Richardson, Dr. Jocelyn Harris, and Dr. Jackie Bosch.

Professor Paul Stratford contributed to the refinement of the research questions and study design, the statistical analyses, the interpretation of findings, and provided critical review of earlier versions of the manuscripts and the final approval of the version to be published.

Dr. Ayse Kuspinar contributed to the refinement of the research questions and study design, the statistical analyses, the interpretation of findings, and provided critical review of earlier versions of the manuscripts and the final approval of the version to be published.

Dr. Julie Richardson contributed to the refinement of the research questions, the interpretation of findings, the overall structure of the manuscripts, and provided critical review of earlier versions of the manuscripts and the final approval of the version to be published.

Dr. Jocelyn Harris contributed to the refinement of the research questions, and provided critical review of earlier versions of the manuscripts and the final approval of the version to be published.

Dr. Jackie Bosch contributed to the refinement of the research questions, provided critical review of earlier versions of the manuscripts and its relevance to rehabilitation professionals, and also provided the final approval of the version of the manuscripts to be published.

Table of Contents

List of Figures

Chapter One: Introduction

Chapter Four: Measurement theories in rehabilitation: Introduction to item response theory and Rasch measurement theory

Chapter 5: Revising the Chedoke Arm and Hand Activity Inventory using modern test theories

List of Tables

Chapter One: Introduction

Chapter Two: Cross-cultural adaptation and psychometric evaluation of the Singapore version of the Chedoke Arm and Hand Activity

Chapter Three: Reliability and validity of the shortened Singapore versions of the Chedoke Arm and Hand Activity Inventory

Chapter Four: Measurement theories in rehabilitation: Introduction to item response theory and Rasch measurement theory

Chapter 5: Revising the Chedoke Arm and Hand Activity Inventory using modern test theories

List of Abbreviations and Symbols

Abbreviations

ADL – activities of daily living

ANOVA – analysis of variance

ARAT – Action Research Arm Test

BADL – basic activities of daily living

CAHAI – Chedoke Arm and Hand Activity Inventory

CAHAI-SG – Singapore version of the Chedoke Arm and Hand Activity Inventory

CFI – comparative fit index

CI – confidence interval

CMSA – Chedoke-McMaster Stroke Assessment

COSMIN – Consensus-based standards for the selection of health status measurement instruments

CTT – classical test theory

DIF – differential item functioning

FMA-UE – Fugl-Meyer Assessment of Upper Extremity

IADL – instrumental activities of daily living

ICC – intraclass correlation coefficient

ICF – International Classification of Functioning, Disability, and Health

IIF – item information function

IRT – item response theory

OT – occupational therapist

RMSEA – root mean square error of approximation

RMT – Rasch measurement theory

ROC – receiver operating characteristic

RUMM – Rasch Unidimensional Measurement Model

SD – standard deviation

SEM – standard error of measurement

TCCA-OAO – translation and cross-cultural adaptation of objectively-assessed outcome measures

TLI – Tucker-Lewis index

UE – upper extremity

Symbols

θ – theta, used to represent the latent trait

a – item discrimination

b – item difficulty

df – degrees of freedom

r – statistical symbol representing the correlation coefficient

r_s – statistical symbol representing Spearman's rank correlation coefficient

p – probability, used to represent the probability threshold for significance

x^2 – chi-square

Chapter One: Introduction

Chapter One: Introduction

> The eye cannot say to the hand, "I have no need of thee."
>
> Blessed be the hand. Thrice blessed be the hands that work!
>
> – Helen Keller, *The World I Live In*, 1908

Background

The upper extremity (UE) works as a synchronized unit to execute complex manipulative actions unique to humans (Carr & Shepherd, 2010). The control and movement of the proximal segments transport the hand, which enables the hand to position, orientate, and/or manipulate objects to achieve a specific goal, such as drinking coffee, driving, or waving hello (Lang & Beebe, 2007). Losing the ability to use one's UE, a common consequence after a stroke, would, therefore, affect every aspect of daily living.

Measuring UE function is essential to stroke rehabilitation as it provides credible justification for intervention, facilitates clinical decision-making about appropriate interventions, and demonstrates the effectiveness of intervention and rehabilitation services (Beaton, Bombardier, Katz, & Wright, 2001; Fawcett, 2007; McDonnell, Hillier, & Esterman, 2013). This thesis is devoted to the thorough examination of the Chedoke Arm and Hand Activity Inventory (CAHAI), an outcome measure of post-stroke UE function. The first part of this thesis (Chapters 2 and 3) evaluates the CAHAI in a clinical population beyond the country where it was initially developed. The second part (Chapters 4 and 5) examines the CAHAI's psychometric properties using item response theory and Rasch measurement theory.

This introductory chapter will describe the effects of stroke on UE function and discuss the loss of UE function according to the International Classification of Functioning, Disability, and Health (World Health Organization, 2001). The loss of UE function from the perspectives of individuals with stroke will also be described. Next, recommended post-stroke UE outcome will be introduced. A framework will then be applied to justify how the CAHAI was selected among the recommended post-stroke UE outcome measures as the most appropriate post-stroke UE outcome measure. Thereafter, the overall thesis objectives, the structure of this thesis, and the manuscripts included will be discussed in detail.

Upper Extremity Function after a Stroke

Stroke is a cerebrovascular disease that can be defined as a non-traumatic abrupt development of neurological dysfunction caused by focal ischemia and/or collection of blood within the brain (Sacco et al., 2013). It occurs in approximately 297 per 100,000 individuals in Canada and 160 per 100,000 individuals in Singapore (National Registry of Diseases Office, 2018; Public Health Agency of Canada, 2017). Stroke incidence increases with age as 70% of stroke occurs in adults above 65 years of age (Kelly-Hayes, 2010). Stroke is a major health concern worldwide as it is the second and third most common cause of death and disability respectively (Feigin, Norrving, & Mensah, 2017). Although stroke mortality rates have declined, the number of individuals who survive a stroke and are living with the effects of a stroke have increased (Feigin et al., 2017). Among older adults aged 65 years and older, an estimated 5% of Canadians (about 270,000) and 9.3% of Singaporeans (about 31,000) are living with the effects of a stroke

(Krueger et al., 2015; Teh et al., 2018). The economic burden of stroke is substantial; stroke care accounted for an average of 3% of total healthcare expenditure and an average of 0.27% of gross domestic product across eight countries (Evers et al., 2004). Both the economic and health burden of stroke is also expected to continue to escalate in the next decade as stroke prevalence is expected to rise with the increasing age of the population worldwide (Feigin et al., 2017).

Common sequelae following a stroke include sensorimotor, cognitive, and speech impairments, such as weakness in the arm and leg, swallowing difficulties, memory deficits, and aphasia (Tatemichi et al., 1994; Vidović, Sinanović, Sabaskić, Haticić, & Brkić, 2011; Warlow et al., 2008). Post-stroke impairments lead to difficulties in performing basic self-care tasks, instrumental activities of daily living (e.g., food preparation and household chores), and social integration (Jørgensen et al., 1995; Lo et al., 2008; Sturm et al., 2002), and these difficulties may persist even five years after stroke (Gall, Dewey, Sturm, Macdonell, & Thrift, 2009; Patel et al., 2006).

Post-stroke Upper Extremity Deficits

UE deficits are one of the most common and persistent consequences of stroke (Connell, Lincoln, & Radford, 2008; Lai, Studenski, Duncan, & Perera, 2002; Lawrence et al., 2001; Wade & Hewer, 1987). The loss of UE function after a stroke can be described according to the International Classification of Functioning, Disability, and Health (ICF) (World Health Organization, 2001). The ICF broadly classifies human functioning into two main components: body functions and structures, and activity and participation. Negative changes to body functions and structures are regarded as

impairments, while difficulties with task execution and participation are viewed as activity limitations and participation restrictions respectively (World Health Organization, 2001)

Motor impairments. Motor impairments are one of the most common UE deficits after a stroke (Lawrence et al., 2001; Wade & Hewer, 1987). Stroke affects signal transmissions from the motor cortex to the spinal cord and causes delayed, or even a lack of, initiation or termination of muscle contractions (Raghavan, 2015). This decreased (paresis) or loss (plegia) of volitional control over motor units typically occurs on one side of the body (hemiparesis or hemiplegia) (Gemperline, Allen, Walk, & Rymer, 1995; Sathian et al., 2011; Young & Mayer, 1982). Hemiparesis is the most common post-stroke UE impairment, occurring in 70% to 80% of individuals with an acute stroke (Lawrence et al., 2001; Nakayama, Jørgensen, Raaschou, & Olsen, 1994). UE hemiparesis also persists beyond the acute phase of stroke; 3 out of 4 individuals with severe hemiparesis do not regain UE function three months after a stroke (Nakayama et al., 1994). Even at four years post-stroke, 50% of individuals with hemiparesis do not recover full functional use of their UE (Broeks, Lankhorst, Rumping, & Prevo, 1999).

The loss of fractionated movement is another common post-stroke UE motor impairment, where individuals experience reduced or loss of voluntary ability to activate selective muscles and move independent segments of the UE (Lang, Bland, Bailey, Schaefer, & Birkenmeier, 2013; Lang & Schieber, 2004). For example, during voluntary elbow flexion, a person with stroke may inadvertently abduct their shoulder and pronate their forearm simultaneously. Difficulties with fractionated movements are similar to the

abnormal synergistic movements or “associated reactions” that are present in the affected UE after a stroke (Twitchell, 1951).

Somatosensory impairments. Stroke can also affect the signal transmissions from the somatosensory cortex to the spinal cord and cause deficits in the perception of body senses, such as diminished touch perception, decreased proprioception, and loss of pain sensation (Connell et al., 2008; Roland, 1987). UE somatosensory impairments are present in up to 63% of individuals with an acute stroke, with decreased or loss of touch and proprioception as the most common impairments (Connell et al., 2008; Tyson, Hanley, Chillala, Selley, & Tallis, 2008). Similar to UE motor impairments, UE somatosensory impairments are also persistent. Approximately 33% of individuals continue to have impaired touch sensation even after two years post-stroke (Bowden, Lin, & McNulty, 2014).

It is important to also consider the interaction between somatosensory and motor functions. This is because the ability to purposefully use the UE requires both efferent and afferent connections between the UE and the central nervous system to be intact (Twitchell, 1954). Post-stroke somatosensory impairments produce unreliable and erratic feedback from the UE to the cortex, disrupting the feed-forward motor control (Frey et al., 2011). Consequently, UE somatosensory impairments may result in motor impairments.

Activity limitations. Complete synchronization of the shoulder, elbow, forearm, wrist, thumb, and fingers are needed for humans to interact and manipulate objects in

their environment with accuracy, precision, and efficiency (Carr & Shepherd, 2010; Lang & Beebe, 2007). UE motor and somatosensory impairments, which may occur in isolation or in combination, can therefore lead to activity limitations in all aspects of daily living (Lang et al., 2013). Individuals with stroke often experience difficulties carrying out basic and instrumental activities of daily living (BADL and IADL respectively), with 25% to 74% of individuals with stroke requiring the assistance of, or are dependent on caregivers (Miller et al., 2010). These individuals may continue to require assistance in ADLs even up to 10 years after the stroke onset (Blomgren et al., 2018; Walsh, Galvin, Loughnane, Macey, & Horgan, 2015; Wolfe et al., 2011).

Individuals with stroke may have difficulties performing activities with their affected UE. For example, an individual with post-stroke weakness in his hand and shoulder may be unable to hold a cup and have a drink; another individual with impaired proprioception may have difficulties adjusting her grip and may squeeze a paper cup too tightly and cause the contents to spill. Although post-stroke UE impairments typically occur in the affected/involved UE, impairments in the affected UE impacts the coordination of both UEs in daily activities (Waller & Whitall, 2008). Many daily activities require bilateral UE use, such as cutting a piece of steak, playing the violin, and lifting a barbell. The ability to use both UEs together is thus highly associated with independence in carrying out IADLs (Haaland et al., 2012).

Several studies have demonstrated the relationship between UE impairments and activity limitations. Grip strength was significantly correlated to performance in BADLs, where greater paresis in the affected UE was associated with more difficulties performing

daily self-care tasks (Harris & Eng, 2007; Kim, 2016; Sunderland, Tinson, Bradley, & Hewer, 1989). Greater UE motor impairments were also associated with more difficulties with IADLs, such as household chores and meal preparation (Desrosiers et al., 2003; Poole, Sadek, & Haaland, 2011; Sveen, Bautz-Holter, Sødring, Wyller, & Laake, 1999). Similarly, UE sensory impairments correlated significantly with performance in daily tasks; impaired proprioception and touch sensation were associated with decreased independence in both BADLs and IADLs (Tyson et al., 2008). Unsurprisingly, the combination of UE motor and somatosensory impairments were also significantly related to activity limitations, with greater impairments associated with decreased independence in ADLs (Meyer, Karttunen, Thijs, Feys, & Verheyden, 2014).

Impact of Post-Stroke Upper Extremity Deficits

The loss of UE function after a stroke can impact beyond the reduced ability to use the UE and activity limitations in daily living. The following are quotes from participants in three qualitative studies that investigated the experience of UE impairments from the perspectives of individuals with stroke:

> It is hideously slow and I have no idea where it is most of the time unless it's in direct sight, I often think I'd be better off amputating it.
>
> *Internet-based account, 34-year-old man, 6 years post-stroke, UK*
>
> (Poltawski et al., 2016, p. 948)

I had not eaten out since the stroke. I didn't know if I would make a fool of myself.

Internet-based account, woman, 1 year post-stroke

(Poltawski et al., 2016, p. 948)

Frustration comes from that too. Things that you could do so automatically before require so much concentration now. Just one slip and then the whole thing, you know you just have to start all over.

Participant, 45-year-old man, 2.2 years post-stroke

(Doyle et al., 2014, p. 996)

Having stuff hit the floor in the kitchen was incredibly depressing. I cannot tell how much my heart would sink each time something would hit the floor in the kitchen. It was more than just oh darn now I have got to clean this up.

Participant, 56-year-old man, 1.5 years post-stroke

(Doyle et al., 2014, p. 996)

It would almost be easier if the arms came back. You could sit in a wheelchair, at least you could do something. When the leg comes back the only thing you learn to do is walk. But the number of things you can do with an arm…

Participant in a focus group

(Barker & Brauer, 2005, p. 1217)

> It's a big deal to be able to use your arm again. I think most of the doctors think it's not. It's a big deal to be able to use your arm again psychologically as well physically.
>
> *Participant in a focus group*
>
> (Barker & Brauer, 2005, p. 1217)

These quotes reflect the realities of individuals living with post-stroke UE deficits, where their psychosocial well-being and quality of life are adversely affected. The loss of UE function and the concomitant loss of independence in ADLs can lead to depression (Rao, n.d.). Pooled data from a systematic review estimates depression to be present in 31% of individuals with stroke at any time up to five years post-stroke (Hackett & Pickles, 2014). Quality of life of individuals with stroke was also found to be consistently lower than matched-control healthy adults (Bays, 2001). Post-stroke UE deficits negatively impact on health-related quality of life, where more severe UE impairments and activity limitations are associated with poorer quality of life (Morris, van Wijck, Joice, & Donaghy, 2013; Nichols-Larsen, Clark, Zeringue, Greenspan, & Blanton, 2005; Wyller, Sveen, Sødring, Pettersen, & Bautz-Holter, 1997).

Evaluating Post-stroke Upper Extremity Function

For individuals living with post-stroke UE deficits, regaining UE function is an important rehabilitation goal. The first and important step in the stroke rehabilitation process is the evaluation of UE function. There are two main goals when evaluating post-stroke UE function: the first is to establish the present status of UE function, and the

second goal is to measure the changes in UE function. Through the use of standardized and valid outcome measures, the severity of UE impairments and the extent of activity limitations can be quantified (Santisteban et al., 2016). Understanding the present status of the affected UE facilitates clinical decision-making about appropriate interventions to optimize recovery (Lang et al., 2013; McDonnell et al., 2013; Wolf, Kwakkel, Bayley, McDonnell, & Upper Extremity Stroke Algorithm Working Group, 2016). For example, constraint-induced movement therapy is more suitable for individuals with some voluntary motor control throughout their UE (Wolf et al., 2016).

Understanding the present status of the affected UE can also provide prognostic information about recovery. The degree of UE impairments at the onset of stroke is a strong predictor of UE recovery (Coupar, Pollock, Rowe, Weir, & Langhorne, 2012). UE motor impairments can also predict ADL outcomes beyond three months after stroke, where individuals with a lesser degree of impairments are more likely to regain independence in ADLs (Veerbeek, Kwakkel, Wegen, Ket, & Heymans, 2011). Prognostic information is useful to clinicians to enhance care, such as planning discharge destination and providing patients with accurate information about possible outcomes (Feys et al., 2000).

Simpson & Eng (2013) described, “Optimizing or augmenting changes in recovery is core to the rehabilitation process following stroke” (p. 240). Evaluating the status of the affected UE over time measures the changes in UE function. Objective measurements of changes in UE function provide a credible and reliable basis for the continuation/termination of rehabilitation (College of Occupational Therapists, 2013;

Fawcett, 2007). Measuring changes in UE function is also imperative to demonstrate the effectiveness of interventions in clinical trials and rehabilitation programs/services (Beaton et al., 2001; Guyatt, Deyo, Charlson, Levine, & Mitchell, 1989).

Post-stroke Upper Extremity Outcome Measures

Clinical practice guidelines for stroke rehabilitation recommend using standardized, valid, and reliable outcome measures to evaluate UE function (Hebert et al., 2016; Intercollegiate Stroke Working Party, 2012; Miller et al., 2010). However, these guidelines often do not provide recommendations on which outcome measures to use for clinical practice. There are approximately 53 post-stroke UE outcomes measures available (Murphy, Resteghini, Feys, & Lamers, 2015; Santisteban et al., 2016), and selecting the appropriate measures for clinical practice may be challenging. The Fugl-Meyer Assessment and the Action Research Arm Test are two UE outcome measures commonly used in clinical practice (van Wijck, Pandyan, Johnson, & Barnes, 2001). They are also UE measures recommended for outcomes measures in trials studying post-stroke sensorimotor recovery (Kwakkel et al., 2017). However, it is essential to consider that the 'popularity' or frequency of use of an outcome measure may not be indicative of its quality. The psychometric properties and clinical utility of existing post-stroke UE outcome measures were recently evaluated in an overview of systematic reviews by Murphy et al (2015). Six outcome measures with high-quality evidence of its psychometric properties and clinical utility were recommended: ABILHAND, Action Research Arm Test, Box and Block Test, Fugl-Meyer Assessment of Upper Extremity, Chedoke Arm and Hand Activity Inventory, and Wolf Motor Function Test. This thesis

focuses on one of the six recommended post-stroke UE outcome measure, the Chedoke Arm and Hand Activity Inventory (CAHAI).

Chedoke Arm and Hand Activity Inventory

The CAHAI is a standardized outcome measure that uses daily tasks to evaluate the function of the affected UE after a stroke (Barreca, Stratford, Lambert, Masters, & Streiner, 2005). The CAHAI was selected as the focus of this thesis for two main reasons. First, the theoretical constructs of the CAHAI concur with the philosophies of occupational therapy. A philosophical base of occupational therapy is participation in meaningful activities in daily living is an important element of health (American Occupational Therapy Association, 2011). Similarly, one of the CAHAI's theoretical construct is including domains of everyday living considered as important to individuals with stroke (Barreca et al., 2005). And thus, the CAHAI's use of real-life daily tasks that are important to individuals with stroke resonated to me as an occupational therapist. Second, the CAHAI possesses several inherent qualities of a good post-stroke UE outcome measure. In the following sections, the qualities of the CAHAI as a good outcome measure will be described according to the instrument evaluation framework by Rudman & Hannah (1998). This framework describes the factors that clinicians should consider when selecting instruments that measure UE function, which are categorized into five categories: clinical utility, standardization, purpose, psychometric properties, and clients' perspective (Figure 1). For each factor, desired qualities of a UE outcome measure will be described first, followed by a detailed explanation of how the CAHAI fulfils these qualities.

Category 1: Clinical utility. Although the quality of psychometric properties is crucial when selecting outcome measures, the instrument evaluation framework prioritizes clinical utility for efficiency (Rudman & Hannah, 1998). If a measure is clinically useful, clinicians can then embark on the time-consuming process of evaluating the psychometric properties of the outcome measure of interest. Five elements should be examined to determine if an outcome measure can be used in a clinical setting: specificity, clinical applicability, acceptability to clients, availability, and time demands (Rudman & Hannah, 1998).

Clinical applicability. Clinical applicability refers to the usefulness of information obtained from the outcome measure (Rudman & Hannah, 1998). The usefulness of information is indicated by the measures' administration method, type of tasks, type of results (quantitative and/or qualitative), and interpretation of results (Rudman & Hannah, 1998). Specific to post-stroke UE outcome measures, the administration method and type of tasks are two key considerations. The administration method should not demand high levels of cognitive, speech and language abilities from the individual with stroke. This is because individuals with stroke often have a myriad of cognitive and speech impairments in addition to their UE impairments (Douiri, Rudd, & Wolfe, 2013; Vidović et al., 2011). Using daily tasks and objects to evaluate UE function is therefore advantageous as it is more intuitive and can be easily understood by individuals with post-stroke cognitive or language deficits (Barreca et al., 2004). It can also provide a more accurate evaluation of UE function as individuals with stroke have faster and more efficient UE movements when performing tasks with real-life objects compared to performing similar tasks in a

simulated condition (e.g., holding a telephone receiver versus an object of the same size, weight, and color) (Trombly & Wu, 1999; Wu, Trombly, Lin, & Tickle-Degnen, 1998). Both unilateral and bilateral tasks should also be included (Rudman & Hannah, 1998). Bilateral tasks provide valuable information as post-stroke impairments in the affected UE can affect the coordination of both UEs in bimanual activities (Waller & Whitall, 2008).

Specificity. Specificity is the intended clinical population of the outcome measure (Rudman & Hannah, 1998). Only outcome measures developed and/or validated in stroke populations should be used to evaluate post-stroke UE function. This is because the psychometric properties of a measure are contextual in all ways and as such, one clinical population may not translate to a different clinical population (Streiner, Norman, & Cairney, 2015).

Acceptability to clients. The acceptability of an outcome measure to clients may affect their participation in the evaluation process (Rudman & Hannah, 1998). UE outcome measures that are easy to understand, such as a clear description of its purpose, not time-consuming, and use familiar objects are more acceptable to individuals with stroke (Murphy et al., 2015; Rudman & Hannah, 1998).

Availability. Post-stroke UE outcome measures that are readily available would be more likely to be used in clinical practice. Cost, availability of validated translations of the outcome measure, accessibility of required materials, and whether the measure is available in the public domain are some factors that may affect availability. Outcome

measures with minimal costs, have readily available materials (e.g., prefabricated instruments), and are available in a variety of languages supports its global use in clinical practice (Murphy et al., 2015; Rudman & Hannah, 1998).

Time demands. Time is an important factor that influences the clinical utility of the outcome measure. Clinicians prefer a short administration time due to the time constraints in clinical practice (Duncan & Murray, 2012). Thus, outcome measures with fewer test items or have shortened versions available are favoured. Other considerations include the time and cost required for training assessors on the measure, and the time needed to score and interpret the results (Rudman & Hannah, 1998).

Clinical utility of the CAHAI. The CAHAI meets all the desired qualities for clinical utility described in the framework. First, the CAHAI is an outcome measure that was developed to measure post-stroke UE function and was validated in a stroke sample (*specificity*). Second, of the six recommended post-stroke UE outcome measures (Murphy et al., 2015), the CAHAI is the only outcome measure that uses daily tasks in all test items. Thus, the unique strength of the CAHAI is the ecological validity of the test items, where daily objects and tasks are used to evaluate the function of the affected UE (*clinical applicability*). The use of daily objects and tasks is also beneficial to individuals with stroke for the reasons described earlier (*acceptability to clients*). Third, the CAHAI requires minimal cost as the administration manual and training videos are available in the public domain at no charge (www.cahai.ca), and the materials required are inexpensive (*availability*). Lastly, there are also shorter versions of the CAHAI which

require less administration time and scoring any of the versions is the simple addition of individual item scores (*time demands*).

Category 2: Standardization. Standardized measures are published assessment tools with detailed descriptions of the procedures for administering, scoring, and interpreting results (Finch, Brooks, Stratford, & Mayo, 2002). Standardized measures also have scientific evidence of its psychometric properties, such as reliability and validity (Finch et al., 2002). These measures are preferred over non-standardized measures as they enable consistency in the evaluation of UE function after stroke. The CAHAI is a standardized outcome measure with the detailed administration and scoring procedures. The CAHAI's psychometric properties were also evaluated in three studies (discussed in the later section).

Category 3: Purpose. The purposes of evaluation can be broadly categorized into three categories: (1) to describe status at one moment in time, (2) to predict concurrent or prospective outcomes, and (3) to measure change over time (Kirshner & Guyatt, 1985; Rudman & Hannah, 1998). Post-stroke UE outcome measures can be categorized based on their purpose as discriminative, predictive, and evaluative respectively (Kirshner & Guyatt, 1985). A measure may serve more than one purpose and it is essential to clarify the purpose(s) of the measure so that its psychometric properties can be evaluated accordingly (Rudman & Hannah, 1998). It also allows clinicians to determine whether the outcome measure addresses the needs of the clinical setting (e.g., evaluates changes in UE function to demonstrate program effectiveness).

The purposes of the CAHAI as a descriptive and evaluative outcome measure are suited to my clinical setting, a tertiary hospital in Singapore. Outcome measures are used to evaluate the UE status of individuals with stroke to assist clinicians in making decisions about appropriate interventions. Outcome measures are also routinely used to demonstrate the effectiveness of the hospital's inpatient rehabilitation services.

Category 4: Psychometric properties. As mentioned previously, an outcome measure should have published evidence that supports its psychometric properties. There are two fundamental psychometric properties required for all outcome measures: reliability, and validity (Guyatt et al., 1989). Reliability is the extent to which measurements on an outcome measure obtained under different circumstances (e.g., different assessors or different occasions, between which there should be no change in the measured parameter) produce similar results (Streiner et al., 2015). Validity refers to the extent an instrument measures what it is intended to measure (Portney & Watkins, 2009). An additional psychometric property is required for evaluative measures – sensitivity to change, or responsiveness (Guyatt et al., 1989; Kirshner & Guyatt, 1985). The types of reliability, validity, and responsiveness and their definitions are summarized in the COSMIN (COnsensus-based Standards for the selection of health status Measurement INstruments) definitions of measurement properties, as shown in Table 1 (Mokkink et al., 2010). Clinicians need to critically examine the statistical results and quality of the studies that evaluated the psychometric properties of the outcome measure (Rudman & Hannah, 1998). Post-stroke UE outcome measures that demonstrate good reliability, validity, and responsiveness should be prioritized, as clinicians can have confidence that

measurements are accurate, consistent, and measures UE function (Portney & Watkins, 2009; Rudman & Hannah, 1998).

The CAHAI is one of the six recommended outcome measures to evaluate post-stroke UE function as it has high-quality evidence of its psychometric properties (Murphy et al., 2015). There are three published studies that evaluated the psychometric properties of the CAHAI: internal consistency, inter-rater reliability, test-retest reliability, construct validity with the Chedoke-McMaster Stroke Assessment (arm and hand components) and the Action Research Arm Test, and longitudinal validity (Barreca et al., 2004, 2005; Barreca, Stratford, Masters, Lambert, Griffiths, et al., 2006). Table 2 provides a summary of the results from these three validation studies.

Category 5: Clients' perspective. Although performance-based outcome measures evaluate UE impairments and activity limitations, they do not provide information about UE function from the perspectives of individuals with stroke in their customary settings (e.g., home environment) (Ashford, Slade, Malaprade, & Turner-Stokes, 2008). Individuals who scored highly on performance-based UE impairment measures (i.e., minimal or no UE impairments) reported difficulties with UE movements in daily activities on self-report outcome measures (J. C. Stewart & Cramer, 2013). It is therefore vital for clinicians to use performance-based measures in conjunction with self-report UE outcome measures. However, a challenge to using self-report measures for individuals with stroke are cognitive and speech impairments that may confound the evaluation process (J. C. Stewart & Cramer, 2013).

Although the CAHAI is not a self-report measure, it meets the standards of the other four categories of the instrument evaluation framework (Rudman & Hannah, 1998). Furthermore, as most of the rehabilitation services within my hospital are inpatient care, self-report measures may not be suitable as individuals with stroke do not have experiences with using their affected UE in their customary settings.

Improving the Chedoke Arm and Hand Activity Inventory

The presence of evidence about the psychometric properties of an outcome measure should not mark the end of the evaluation of the measure. Continued evaluation of an outcome measure beyond its initial phase of development and validation is valuable. The evaluation of an outcome measure beyond the culture, setting, or country where it was initially developed enables the diffusion of the measure. Psychometric properties of existing outcome measures can also be further assessed using more rigorous study designs and sophisticated statistical analysis. There may be significant advances in technology since the initial development and validation of an outcome measure. Complex statistical analysis previously not feasible can now be conducted for dated outcome measures using advanced computers and software. By generating in-depth knowledge of outcome measures, revisions or refinements to existing outcome measures can be facilitated, and in turn, improve the efficiency and quality of measurement.

The continued evaluation of the CAHAI was approached two ways in this thesis. First, the evaluation was taken beyond the country where the CAHAI was developed (Canada) and was translated, cross-culturally adapted and validated for the stroke population in Singapore. Second, the CAHAI was evaluated using a different

measurement theory than the one initially employed in the previous studies, and its psychometric properties were re-evaluated using item response theory and Rasch measurement theory.

Conceptualization of the Research

The CAHAI was an outcome measure that my colleagues and I identified as a suitable outcome measure to evaluate post-stroke UE function in our clinical setting. However, the major barrier to using it was the lack of cross-cultural adaptation and validation of the measure for the stroke population in Singapore. Thus, the development of the Singapore version of the CAHAI was a research project that I first embarked on when I was enrolled in the master's program. I later transferred to the doctoral program and my research plan was to conduct a second project to continue the psychometric evaluation of the Singapore version. However, this research plan changed after analyzing the data from the initial study.

During data analysis, I noticed a pattern in the scores on all test items; the scores of 1 and 7 (the minimum and maximum scores respectively) were most frequently obtained, whereas the scores between 2 and 5 had low scoring frequencies. Professor Stratford, who was part of the validation study of the CAHAI (Barreca, Stratford, Masters, Lambert, & Griffiths, 2006), also observed a similar pattern in the study's data. The observation of this pattern sparked my interest to examine the scoring scale and individual test items. This led to changes in my research plan to evaluate the original CAHAI using item response theory and Rasch measurement theory. As it would be the first of such evaluation of the measure, the original CAHAI, rather than the Singapore

version, was chosen. This was to avoid the uncertainty of whether the study findings were due to inherent issues within the measure or the cross-cultural adaptation process should the Singapore version be evaluated.

Thesis Objectives and Structure

Thesis Objective

There are two overall objectives of this thesis: (1) to develop a Singapore version of the CAHAI, and (2) to re-evaluate the psychometric properties and clinical utility of the original CAHAI using item response theory and Rasch measurement theory.

Thesis Structure and Overall Context

Objective 1. The Singapore version of the CAHAI was developed through a research project conducted between September 2016 and June 2017. Data were collected prospectively at a tertiary hospital in Singapore for the validation of the measure. Two original research articles were published in peer-reviewed journals and are presented in Chapters 2 and 3 of this thesis.

Objective 2. The re-evaluation of the CAHAI was completed using data from a previous validation study of the CAHAI (Barreca, Stratford, Masters, Lambert, & Griffiths, 2006). The study compared the longitudinal validity of the full 13-item and short 9-item versions of the CAHAI (CAHAI-13 and CAHAI-9 respectively) against the Action Research Arm Test. It also compared the cross-sectional and longitudinal validity of the CAHAI-13 with the CAHAI-9. A total of 105 participants with post-stroke UE impairments were recruited and assessed on the CAHAI and the Action Research Arm

Test at two time points (baseline and discharge/completion of rehabilitation program). Data from the study, which was provided as a de-identified computer file, was used for analysis in this thesis. Results of all analyses conducted using this dataset are presented in Chapters 4 and 5 of this thesis. These two chapters are prepared for submission to peer-reviewed journals.

Beyond Borders: Translation and Cross-cultural Adaptation of the CAHAI

The first thesis objective was to develop a Singapore version of the CAHAI. This was achieved through the translation and cross-adaptation of the original CAHAI. Cross-cultural adaptation of an outcome measure is a process that examines both language (translation) and cultural relevance of an outcome measure in a target setting (Beaton, Bombardier, Guillemin, & Ferraz, 2000). This process is required when the target setting differs considerably from the original population in terms of culture, language, and country (Geisinger, 1994). There are two key advantages to cross-culturally adapt an existing outcome measure instead of developing a new measure. Cross-cultural adaptation of an outcome measure requires relatively lesser cost and shorter time compared to developing and validating a new measure (Hambleton & Patsula, 1998). It also facilitates the use of the same outcome measure across different cultures and countries. Using the same outcome measure enables nation-wide and international research studies (Guillemin, 1995) to compare the effectiveness of stroke interventions across regions and countries.

The process of translation and cross-cultural adaptation of an outcome measure involves three stages: translation, adaptation, and validation (Epstein, Santo, &

Guillemin, 2015). There are several published guidelines detailing the necessary steps and processes required to translate and adapt outcome measures (e.g. Beaton et al., 2000; Guillemin, 1995; Hambleton & Patsula, 1998; Schuster, Hahn, & Ettlin, 2010; Wild et al., 2005). Using such guidelines better ensures equivalence between the original and adapted measure (Guillemin, 1995). For example, conceptual equivalence is achieved when the adapted tool measures the same theoretical constructs as the original tool (Flaherty et al., 1988; A. L. Stewart & Nápoles-Springer, 2000). Measurement equivalence is also essential, where the adapted tool has similar psychometric properties as the original tool (Herdman, Fox-Rushby, & Badia, 1998; A. L. Stewart & Nápoles-Springer, 2000).

The process for translation and adaption of the CAHAI for the stroke population in Singapore would be best conducted if based on processes specifically designed for performance-based measures. When translating and adapting self-report and performance-based measures, test items must be relevant to the target clinical population. However, for performance-based measures, there is an additional need to ensure those using the outcome measure understand its requirements (instructions, administration, scoring, and interpretation of results) (Schuster et al., 2010). There is only one guideline available for the translation and cross-cultural adaptation of performance-based outcome measures; it is the Translation and Cross-Cultural Adaptation of Objectively-Assessed Outcome measures (TCCA-OAO) procedure (Schuster et al., 2010). This procedure was used to guide the systematic translation and adaptation of the CAHAI for the stroke

population in Singapore. The TCCA-OAO procedure is comprised of eight steps, as follows (Schuster et al., 2010):

1. Forward translation and region-specific adaptations by two reviewers
2. Merging of forward translation and region-specific adaptations
3. Preparation of translated and adapted outcome measure
4. Backward translation
5. Review of the translated and culturally-adapted outcome measure
6. Further adaptation and proof-reading
7. Pre-testing of the outcome measure
8. Psychometric evaluation of the adapted measure

The TCCA-OAO procedure was developed based on the guidelines for the translation and cross-cultural adaptation of self-report measures by Beaton et al. (2000). This guideline recommends six steps to translate and adapt self-report measures: (1) translation, (2) synthesis, (3) back translation, (4) expert committee review, (5) pre-testing, and (6) submitting and appraising all written reports by the original developers of the measure (Beaton et al., 2000). Thus, the steps in the TCCA-OAO procedure are similar to the guidelines for the translation and adaptation of self-report measures with one key difference. The TCCA-OAO procedure does not include a review by an expert committee as reviewers in Step 1 of the procedure are health care professionals (i.e., informed users) (Schuster et al., 2010). In contrast, naïve reviewers who unaware of the concepts evaluated in the measure and have no medical or clinical background are

recommended in the translation and adaptation of self-report measures, and thus, a review by an expert committee is required to assess for equivalence (Beaton et al., 2000).

After an outcome measure is translated and culturally adapted, the psychometric properties need to be evaluated (Beaton et al., 2000). This is to provide evidence of psychometric properties of the adapted measure and to also evaluate the extent to which measurement equivalence with the original measure was achieved (Beaton et al., 2000; Herdman et al., 1998). The design of measurement studies to evaluate the psychometric properties of an adapted measure can be a challenge. Unlike intervention studies where randomized controlled trials are recognized as the 'gold standard' study design to evaluate effectiveness, there is no recognized 'gold standard' study design for measurement studies. Therefore, quality assessment tools (e.g., the COSMIN Risk of Bias Checklist (Mokkink et al., 2018)) can be used to guide the design of measurement studies as these tools describe the required standards for high-quality studies. By using a more rigorous study design, we can ensure better quality evidence of the adapted measure's psychometric properties.

Summary of the Relevant Manuscripts

Chapter 2: Cross-cultural adaptation and psychometric evaluation of the Singapore version of the Chedoke Arm and Hand Activity. The CAHAI was developed for evaluation of post-stroke UE function, with all psychometric testing completed on a Canadian stroke population. The cultural differences between Canada and Singapore raise questions about the validity of the measure if used with the stroke population in Singapore. For example, one item in the CAHAI is to dial 911; however, in

Singapore, the emergency number for the police is 999 and 995 for fire and ambulance services. Thus, a Singapore version of the CAHAI is necessary. This manuscript describes the cross-cultural adaptation of the CAHAI for individuals with stroke in Singapore and reports the validation results of the full 13-item Singapore version in an acute and subacute stroke sample. This manuscript was published in *Disability and Rehabilitation* in 2018.

Chapter 3: Reliability and validity of the shortened Singapore versions of the Chedoke Arm and Hand Activity Inventory.

The second manuscript describes the validation of three shortened Singapore versions of the CAHAI with seven, eight, and nine items respectively. Items in the shortened Singapore versions corresponded to the items included in the original shortened versions of the CAHAI. Inter-rater reliability and construct validity of all three shortened Singapore versions were estimated by conducting further analyses of the data collected in the validation of the full 13-item version (Chapter 2). Thus, there are overlaps in the method sections between the Chapters 2 and 3 of this thesis. This manuscript was published in the *International Journal of Rehabilitation Research* in December 2018.

Beyond Singularity: Using Item Response Theory and Rasch Measurement Theory to Evaluate the CAHAI

The second thesis objective was to re-evaluate the psychometric properties of the CAHAI using different measurement theories than the one initially employed in its initial development and validation. Measurement theories serve as a framework for the

development, validation, and refinement of outcome measures (Hambleton & Jones, 1993). Each theory describes the mathematical model(s) and assumptions about factors that influence the observed scores on a rating scale of an outcome measure (Hobart & Cano, 2009). A measurement theory is selected to guide the development and validation of outcome measures. In rehabilitation, classical test theory (CTT) is the prevalent measurement theory used (Hobart & Cano, 2009). CTT assumes that an observed score comprises of a true score and an error component related to the observed score (Lord & Novick, 1968). Though CTT is commonly used in rehabilitation, this theory has several weaknesses. The following section describes a major weakness of CTT.

"Observations are Always Ordinal; Measurements, However, Must be Interval[1]"

All measurement scales can be classified into four levels (types): nominal, ordinal, interval, and ratio (Figure 2) (Stevens, 1946). For each type of scale, there are specific rules to assign numbers to observations (Stevens, 1946). For example, numbers are used only as labels in nominal scales (Stevens, 1946), such as assigning the left UE as '1' and the right UE as '2'. In contrast, numbers are assigned according to a ranked order in ordinal scales (Stevens, 1946), such as the CAHAI's 7-category scoring scale. The numbers 1 to 7 were consecutively assigned to an ordered level of function in the affected UE (from total assistance to independence) (Chedoke Arm and Hand Activity Inventory, n.d.). For most outcome measures, including the CAHAI, it is common practice to sum scores on individual items to obtain a total score. However, item scores are discrete

[1] Title of article by Wright, B. D., & Linacre, J. M, 1989, *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–860.

(ordinal) in nature, whereas total scores are assumed to be continuous (interval) data (Rusch, Lowry, Mair, & Treiblmaier, 2017). CTT does not evaluate the interval scaling of the outcome measure (Streiner, 2010), and this means that it is undetermined whether the distances (intervals) between the categories in the 7-category scoring scale of the CAHAI are consistent (Wright & Linacre, 1989). Consequently, arithmetic calculations of the CAHAI scores, such as summing up item scores or subtraction of total scores, are questionable (Wright & Linacre, 1989).

Using Modern Test Theories to Re-Evaluate the CAHAI

Each measurement theory has strengths and weaknesses; outcome measures naturally inherit strengths and weaknesses associated with the theory employed during its initial development and validation. The methods used to develop and evaluate the psychometric properties of the CAHAI were consistent with CTT (Barreca et al., 2004; Barreca, Stratford, Masters, Lambert, Griffiths, et al., 2006; Barreca et al., 2005; Barreca, Stratford, Masters, Lambert, & Griffiths, 2006), and therefore, the CAHAI has several inherent weaknesses due to the limitations of CTT. The main limitations of CTT include sample dependency, equal standard error of measure assumption, and inability to evaluate the interval scaling of a measure (Streiner, 2010; Streiner et al., 2015; Weiss & Davison, 1981). These limitations are described in further detail in Chapter 4 of this thesis.

Item response theory and Rasch measurement theory (referred to as modern test theories in this thesis) were developed to address the limitations of CTT. These theories describe a system of mathematical models about the relationships between the latent trait of interest and the properties of the test items (Allen & Yen, 1979; de Ayala, 2009).

Advantages of modern test theories include the ability to determine the interval scaling of a measure, precise estimation of the standard error of measurement, and sample independence (de Ayala, 2009). Thus, by using modern test theories to re-evaluate the CAHAI, its inherent weaknesses can be addressed. This re-evaluation, using different measurement theories, provides additional evidence about its psychometric properties, and refinements to the CAHAI can be made accordingly to improve its accuracy, precision, and clinical utility. With any refinements to improve the CAHAI, the psychometric properties of the revised version(s) need to be evaluated.

Summary of the Relevant Manuscripts

Chapter 4: Measurement theories in health care: Introduction to item response theory and Rasch measurement theory. Although modern test theories provide numerous advantages over CTT, these theories are not as widely used in rehabilitation compared to CTT because they are complex and the available information on the theories is not accessible to health care professionals. The nature of modern test theories demand advanced levels of understanding of mathematical concepts and are therefore less accessible to individuals from a clinical background (Hobart & Cano, 2009). Modern test theories were also developed and predominantly used in measurement in education and psychology fields, and therefore, information on these theories are often not presented in a manner intended for health care professionals (Hobart & Cano, 2009).

My knowledge of item response theory and Rasch measurement theory was acquired through self-directed learning under the guidance of my committee member, Professor Stratford, and faculty member, Dr. Kuspinar respectively. It was an arduous,

and at times, exasperating learning journey as the available materials on these theories were not easy to understand from a clinical background without advanced mathematical knowledge. Materials were mostly written in a technical manner and often contain a lot of jargon, and the clinical implications of the results were not always explicitly explained. Thus, this manuscript reports the psychometric re-evaluation of the CAHAI using modern test theories in a manner that can be easily understood by health care professionals. It first provides an overview of these two theories in a non-technical style. The CAHAI is then used as an example instrument to demonstrate the step-by-step application of the theories. Lastly, the clinical implications of the results and the challenges of using modern test theories in health care are discussed in details.

Chapter 5: Revising the Chedoke Arm and Hand Activity Inventory using modern test theories. The evaluation of the original versions of the CAHAI using modern test theories (Chapter 4) identified several limitations to the measures. This manuscript describes the revisions to the CAHAI, using modern test theories, to address these limitations. Rasch measurement theory was first applied to evaluate the psychometric properties of an 11-item version of the CAHAI and to revise its 7-category scoring scale. Item response theory was then applied to the new 11-item CAHAI version with a 4-category scoring scale to develop a new shortened version. Finally, classical test theory was used to estimate the reliability and validity of the new shortened versions and compared to the CAHAI-7 advocated previously by (Barreca, Stratford, Masters, Lambert, Griffiths, et al., 2006).

Chapters 4 and 5 of this thesis are highly technical due to the nature of item response theory and Rasch measurement theory. However, the technique of the methods used to derive the results is not described in detail in both manuscripts. Instead, the findings are reported as they apply to clinical practice to ensure the information is accessible to clinicians at a conceptual level.

Conclusion

The overall theme of this thesis is the continued evaluation of an outcome measure beyond its initial phase of development and validation, first to adapt the use for the Singapore stroke population and then to use newer measurement theories to better refine the measure. The detailed examination of the CAHAI will increase what is known about this measure and improve the accuracy, precision, and credibility of the measurements of UE function after a stroke. Ultimately, we must improve our ability to facilitate individuals with stroke to regain the functional abilities of their affected UE; improving the CAHAI is an important first step towards that goal.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, Calif: Brooks/Cole Pub. Co.

American Occupational Therapy Association. (2011). The philosophical base of occupational therapy. *American Journal of Occupational Therapy*, *65*(Suppl.), S65. https://doi.org/10.5014/ajot.2011.65S65

Ashford, S., Slade, M., Malaprade, F., & Turner-Stokes, L. (2008). Evaluation of functional outcome measures for the hemiparetic upper limb: a systematic review. *Journal of Rehabilitation Medicine*, *40*(10), 787–795. https://doi.org/10.2340/16501977-0276

Barker, R. N., & Brauer, S. G. (2005). Upper limb recovery after stroke: the stroke survivors' perspective. *Disability and Rehabilitation*, *27*(20), 1213–1223. https://doi.org/10.1080/09638280500075717

Barreca, S. R., Gowland, C. K., Stratford, P., Huijbregts, M., Griffiths, J., Torresin, W., … Masters, L. (2004). Development of the Chedoke Arm and Hand Activity Inventory: theoretical constructs, item generation, and selection. *Topics in Stroke Rehabilitation*, *11*(4), 31–42.

Barreca, S. R., Stratford, P. W., Lambert, C. L., Masters, L. M., & Streiner, D. L. (2005). Test-retest reliability, validity, and sensitivity of the Chedoke Arm and Hand Activity Inventory: a new measure of upper-limb function for survivors of stroke. *Archives of Physical Medicine and Rehabilitation*, *86*(8), 1616–1622. https://doi.org/10.1016/j.apmr.2005.03.017

Barreca, S. R., Stratford, P. W., Masters, L. M., Lambert, C. L., & Griffiths, J. (2006). Comparing 2 versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. *Physical Therapy*, *86*(2), 245–253.

Barreca, S. R., Stratford, P. W., Masters, L. M., Lambert, C. L., Griffiths, J., & McBay, C. (2006). Validation of three shortened versions of the Chedoke Arm and Hand Activity Inventory. *Physiotherapy Canada*, *58*(2), 148–156. https://doi.org/10.3138/ptc.58.2.148

Bays, C. L. (2001). Quality of life of stroke survivors: a research synthesis. *The Journal of Neuroscience Nursing: Journal of the American Association of Neuroscience Nurses*, *33*(6), 310–316.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186–3191.

Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemiology*, *54*(12), 1204–1217. https://doi.org/10.1016/S0895-4356(01)00407-3

Blomgren, C., Jood, K., Jern, C., Holmegaard, L., Redfors, P., Blomstrand, C., & Claesson, L. (2018). Long-term performance of instrumental activities of daily living (IADL) in young and middle-aged stroke survivors: Results from SAHLSIS outcome. *Scandinavian Journal of Occupational Therapy*, *25*(2), 119–126. https://doi.org/10.1080/11038128.2017.1329343

Bowden, J. L., Lin, G. G., & McNulty, P. A. (2014). The prevalence and magnitude of impaired cutaneous sensation across the hand in the chronic period post-stroke. *PLOS ONE*, *9*(8), e104153. https://doi.org/10.1371/journal.pone.0104153

Broeks, J. ., Lankhorst, G. J., Rumping, K., & Prevo, A. J. H. (1999). The long-term outcome of arm function after stroke: results of a follow-up study. *Disability and Rehabilitation*, *21*(8), 357–364. https://doi.org/10.1080/096382899297459

Carr, J. H., & Shepherd, R. B. (2010). *Neurological rehabilitation: optimizing motor performance* (2nd ed). Edinburgh ; New York: Churchill Livingstone.

Chedoke Arm and Hand Activity Inventory (CAHAI). (n.d.). Chedoke Arm and Hand Activity Inventory administration guidelines version 2. Retrieved from www.cahai.ca

College of Occupational Therapists. (2013). Occupational therapists' use of standardized outcome measures. Retrieved June 1, 2016, from https://www.cot.co.uk/position-statements/occupational-therapists%E2%80%99-use-standardised-outcome-measures

Connell, L. A., Lincoln, N. B., & Radford, K. A. (2008). Somatosensory impairment after stroke: frequency of different deficits and their recovery. *Clinical Rehabilitation*, *22*(8), 758–767. https://doi.org/10.1177/0269215508090674

Coupar, F., Pollock, A., Rowe, P., Weir, C., & Langhorne, P. (2012). Predictors of upper limb recovery after stroke: a systematic review and meta-analysis. *Clinical Rehabilitation*, *26*(4), 291–313. https://doi.org/10.1177/0269215511420305

de Ayala, R. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Desrosiers, J., Malouin, F., Bourbonnais, D., Richards, C. L., Rochette, A., & Bravo, G. (2003). Arm and leg impairments and disabilities after stroke rehabilitation: relation to handicap. *Clinical Rehabilitation*, *17*(6), 666–673. https://doi.org/10.1191/0269215503cr662oa

Douiri, A., Rudd, A. G., & Wolfe, C. D. A. (2013). Prevalence of poststroke cognitive impairment: South London stroke register 1995–2010. *Stroke*, *44*(1), 138–145. https://doi.org/10.1161/STROKEAHA.112.670844

Doyle, S. D., Bennett, S., & Dudgeon, B. (2014). Upper limb post-stroke sensory impairments: the survivor's experience. *Disability and Rehabilitation*, *36*(12), 993–1000. https://doi.org/10.3109/09638288.2013.825649

Duncan, E. A., & Murray, J. (2012). The barriers and facilitators to routine outcome measurement by allied health professionals in practice: a systematic review. *BMC Health Services Research*, *12*, 96. https://doi.org/10.1186/1472-6963-12-96

Epstein, J., Santo, R. M., & Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology*, *68*(4), 435–441. https://doi.org/10.1016/j.jclinepi.2014.11.021

Evers, S. M. A. A., Struijs, J. N., Ament, A. J. H. A., Genugten, M. L. L. van, Jager, J. (Hans) C., & Bos, G. A. M. van den. (2004). International comparison of stroke

cost studies. *Stroke*, *35*(5), 1209–1215. https://doi.org/10.1161/01.STR.0000125860.48180.48

Fawcett, A. L. (2007). *Principles of assessment and outcome measurement for occupational therapists and physiotherapists: theory, skills and application*. Hoboken, NJ, USA: John Wiley & Sons.

Feigin, V. L., Norrving, B., & Mensah, G. A. (2017). Global burden of stroke. *Circulation Research*, *120*(3), 439–448. https://doi.org/10.1161/CIRCRESAHA.116.308413

Feys, H., De Weerdt, W., Nuyens, G., van de Winckel, A., Selz, B., & Kiekens, C. (2000). Predicting motor recovery of the upper limb after stroke rehabilitation: value of a clinical examination. *Physiotherapy Research International: The Journal for Researchers and Clinicians in Physical Therapy*, *5*(1), 1–18.

Finch, E., Brooks, D., Stratford, P. W., & Mayo, N. E. (2002). *Physical rehabilitation outcome measures: a guide to enhanced clinical decision making* (2nd ed). Ontario, Canada: B.C. Decker.

Flaherty, J. A., Gaviria, F. M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J. A., & Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *The Journal of Nervous and Mental Disease*, *176*(5), 257–263.

Frey, S. H., Fogassi, L., Grafton, S., Picard, N., Rothwell, J. C., Schweighofer, N., … Fitzpatrick, S. M. (2011). Neurological principles and rehabilitation of action disorders: computation, anatomy, and physiology (CAP) model.

Neurorehabilitation and Neural Repair, *25*(5 Suppl), 6S-20S. https://doi.org/10.1177/1545968311410940

Gall, S. L., Dewey, H. M., Sturm, J. W., Macdonell, R. A. L., & Thrift, A. G. (2009). Handicap 5 years after stroke in the North East Melbourne Stroke Incidence Study. *Cerebrovascular Diseases (Basel, Switzerland)*, *27*(2), 123–130. https://doi.org/10.1159/000177919

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304–312. https://doi.org/10.1037/1040-3590.6.4.304

Gemperline, J. J., Allen, S., Walk, D., & Rymer, W. Z. (1995). Characteristics of motor unit discharge in subjects with hemiparesis. *Muscle & Nerve*, *18*(10), 1101–1114. https://doi.org/10.1002/mus.880181006

Guillemin, F. (1995). Cross-cultural adaptation and validation of health status measures. *Scandinavian Journal of Rheumatology*, *24*(2), 61–63.

Guyatt, G. H., Deyo, R. A., Charlson, M., Levine, M. N., & Mitchell, A. (1989). Responsiveness and validity in health status measurement: A clarification. *Journal of Clinical Epidemiology*, *42*(5), 403–408. https://doi.org/10.1016/0895-4356(89)90128-5

Haaland, K. Y., Mutha, P. K., Rinehart, J. K., Daniels, M., Cushnyr, B., & Adair, J. C. (2012). Relationship between arm usage and instrumental activities of daily living

after unilateral stroke. *Archives of Physical Medicine and Rehabilitation*, *93*(11), 1957–1962. https://doi.org/10.1016/j.apmr.2012.05.011

Hackett, M. L., & Pickles, K. (2014). Part I: frequency of depression after stroke: an updated systematic review and meta-analysis of observational studies. *International Journal of Stroke*, *9*(8), 1017–1025. https://doi.org/10.1111/ijs.12357

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research*, *45*(1–3), 153–171.

Harris, J. E., & Eng, J. J. (2007). Paretic upper-limb strength best explains arm activity in people with stroke. *Physical Therapy*, *87*(1), 88–97. https://doi.org/10.2522/ptj.20060065

Hebert, D., Lindsay, M. P., McIntyre, A., Kirton, A., Rumney, P. G., Bagg, S., … Teasell, R. (2016). Canadian stroke best practice recommendations: Stroke rehabilitation practice guidelines, update 2015. *International Journal of Stroke*, *11*(4), 459–484. https://doi.org/10.1177/1747493016643553

Herdman, M., Fox-Rushby, J., & Badia, X. (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Quality of*

Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, *7*(4), 323–335.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment (Winchester, England)*, *13*(12), iii, ix–x, 1–177. https://doi.org/10.3310/hta13120

Intercollegiate Stroke Working Party. (2012). *National clinical guideline for stroke* (4th ed.). London: Royal College of Physicians. Retrieved from https://www.rcplondon.ac.uk/guidelines-policy/stroke-guidelines

Jørgensen, H. S., Nakayama, H., Raaschou, H. O., Vive-Larsen, J., Støier, M., & Olsen, T. S. (1995). Outcome and time course of recovery in stroke. Part I: Outcome. The Copenhagen Stroke Study. *Archives of Physical Medicine and Rehabilitation*, *76*(5), 399–405.

Kelly-Hayes, M. (2010). Influence of Age and Health Behaviors on Stroke Risk: Lessons from Longitudinal Studies. *Journal of the American Geriatrics Society*, *58*(Suppl 2), S325–S328. https://doi.org/10.1111/j.1532-5415.2010.02915.x

Kim, D. J. (2016). The effects of hand strength on upper extremity function and activities of daily living in stroke patients, with a focus on right hemiplegia. *Journal of Physical Therapy Science*, *28*(9), 2565–2567. https://doi.org/10.1589/jpts.28.2565

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases*, *38*(1), 27–36.

Krueger, H., Koot, J., Hall, R. E., O'Callaghan, C., Bayley, M., & Corbett, D. (2015). Prevalence of individuals experiencing the effects of stroke in Canada: trends and projections. *Stroke*, *46*(8), 2226–2231. https://doi.org/10.1161/STROKEAHA.115.009616

Kwakkel, G., Lannin, N. A., Borschmann, K., English, C., Ali, M., Churilov, L., … Bernhardt, J. (2017). Standardized measurement of sensorimotor recovery in stroke trials: Consensus-based core recommendations from the Stroke Recovery and Rehabilitation Roundtable. *International Journal of Stroke: Official Journal of the International Stroke Society*, *12*(5), 451–461. https://doi.org/10.1177/1747493017711813

Lai, S.-M., Studenski, S., Duncan, P. W., & Perera, S. (2002). Persisting consequences of stroke measured by the Stroke Impact Scale. *Stroke*, *33*(7), 1840–1844.

Lang, C. E., & Beebe, J. A. (2007). Relating movement control at 9 upper extremity segments to loss of hand function in people with chronic hemiparesis. *Neurorehabilitation and Neural Repair*, *21*(3), 279–291. https://doi.org/10.1177/1545968306296964

Lang, C. E., Bland, M. D., Bailey, R. R., Schaefer, S. Y., & Birkenmeier, R. L. (2013). Assessment of upper extremity impairment, function, and activity following stroke: Foundations for clinical decision making. *Journal of Hand Therapy : Official Journal of the American Society of Hand Therapists*, *26*(2), 104–115. https://doi.org/10.1016/j.jht.2012.06.005

Lang, C. E., & Schieber, M. H. (2004). Reduced muscle selectivity during individuated finger movements in humans after damage to the motor cortex or corticospinal tract. *Journal of Neurophysiology*, *91*(4), 1722–1733. https://doi.org/10.1152/jn.00805.2003

Lawrence, E. S., Coshall, C., Dundas, R., Stewart, J., Rudd, A. G., Howard, R., & Wolfe, C. D. (2001). Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. *Stroke*, *32*(6), 1279–1284.

Lo, R. S. K., Cheng, J. O. Y., Wong, E. M. C., Tang, W. K., Wong, L. K. S., Woo, J., & Kwok, T. (2008). Handicap and its determinants of change in stroke survivors: one-year follow-up study. *Stroke*, *39*(1), 148–153. https://doi.org/10.1161/STROKEAHA.107.491399

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass. ; Don Mills, Ont: Addison-Wesley Pub. Co.

McDonnell, M. N., Hillier, S. L., & Esterman, A. J. (2013). Standardizing the approach to evidence-based upper limb rehabilitation after stroke. *Topics in Stroke Rehabilitation*, *20*(5), 432–440. https://doi.org/10.1310/tsr2005-432

Meyer, S., Karttunen, A. H., Thijs, V., Feys, H., & Verheyden, G. (2014). How do somatosensory deficits in the arm and hand relate to upper limb impairment, activity, and participation problems after stroke? A systematic review. *Physical Therapy*, *94*(9), 1220–1231. https://doi.org/10.2522/ptj.20130271

Miller, E. L., Murray, L., Richards, L., Zorowitz, R. D., Bakas, T., Clark, P., … Council, on behalf of the A. H. A. C. on C. N. and the S. (2010). Comprehensive Overview

of Nursing and Interdisciplinary Rehabilitation Care of the Stroke Patient A Scientific Statement From the American Heart Association. *Stroke*, *41*(10), 2402–2448. https://doi.org/10.1161/STR.0b013e3181e7512b

Mokkink, L. B., Vet, H. C. W. de, Prinsen, C. a. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research*, *27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, Lidwine B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Morris, J. H., van Wijck, F., Joice, S., & Donaghy, M. (2013). Predicting health related quality of life 6 months after stroke: the role of anxiety and upper limb dysfunction. *Disability and Rehabilitation*, *35*(4), 291–299. https://doi.org/10.3109/09638288.2012.691942

Murphy, M. A., Resteghini, C., Feys, P., & Lamers, I. (2015). An overview of systematic reviews on upper extremity outcome measures after stroke. *BMC Neurology*, *15*(1), 292. https://doi.org/10.1186/s12883-015-0292-6

Nakayama, H., Jørgensen, H. S., Raaschou, H. O., & Olsen, T. S. (1994). Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. *Archives of Physical Medicine and Rehabilitation*, *75*(4), 394–398.

National Registry of Diseases Office. (2018). *Singapore stroke registry annual report 2015*. Singapore: Ministry of Health. Retrieved from https://www.nrdo.gov.sg/publications

Nichols-Larsen, D. S., Clark, P. C., Zeringue, A., Greenspan, A., & Blanton, S. (2005). Factors influencing stroke survivors' quality of life during subacute recovery. *Stroke; a Journal of Cerebral Circulation*, *36*(7), 1480–1484. https://doi.org/10.1161/01.STR.0000170706.13595.4f

Patel, M. D., Tilling, K., Lawrence, E., Rudd, A. G., Wolfe, C. D. A., & McKevitt, C. (2006). Relationships between long-term stroke disability, handicap and health-related quality of life. *Age and Ageing*, *35*(3), 273–279. https://doi.org/10.1093/ageing/afj074

Poltawski, L., Allison, R., Briscoe, S., Freeman, J., Kilbride, C., Neal, D., … Dean, S. (2016). Assessing the impact of upper limb disability following stroke: a qualitative enquiry using internet-based personal accounts of stroke survivors. *Disability and Rehabilitation*, *38*(10), 945–951. https://doi.org/10.3109/09638288.2015.1068383

Poole, J. L., Sadek, J., & Haaland, K. Y. (2011). Meal preparation abilities after left or right hemisphere stroke. *Archives of Physical Medicine and Rehabilitation*, *92*(4), 590–596. https://doi.org/10.1016/j.apmr.2010.11.021

Portney, L., & Watkins, M. (2009). *Foundations of clinical research: applications to practice* (3rd ed.). New Jersey, USA: Pearson Prentice Hall.

Public Health Agency of Canada. (2017, September 19). Stroke in Canada: Highlights from the Canadian Chronic Disease Surveillance System, 2017 [education and awareness]. Retrieved June 11, 2018, from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/stroke-canada-fact-sheet.html

Raghavan, P. (2015). Upper Limb Motor Impairment After Stroke. *Physical Medicine and Rehabilitation Clinics of North America*, *26*(4), 599–610. https://doi.org/10.1016/j.pmr.2015.06.008

Rao, V. (n.d.). Neuropsychiatry of stroke. Retrieved June 18, 2018, from https://www.hopkinsmedicine.org/gec/series/neuropsych_stroke.html

Roland, P. E. (1987). Somatosensory detection of microgeometry, macrogeometry and kinesthesia after localized lesions of the cerebral hemispheres in man. *Brain Research Reviews*, *12*(1), 43–94.

Rudman, D., & Hannah, S. (1998). An instrument evaluation framework: Description and application to assessments of hand function. *Journal of Hand Therapy*, *11*(4), 266–277. https://doi.org/10.1016/S0894-1130(98)80023-9

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information and Management*, *54*(2), 189–203. https://doi.org/10.1016/j.im.2016.06.005

Sacco, R. L., Kasner, S. E., Broderick, J. P., Caplan, L. R., Connors, J. J. B., Culebras, A., … Council on Nutrition, Physical Activity and Metabolism. (2013). An updated definition of stroke for the 21st century: a statement for healthcare

professionals from the American Heart Association/American Stroke Association. *Stroke*, *44*(7), 2064–2089. https://doi.org/10.1161/STR.0b013e318296aeca

Santisteban, L., Térémetz, M., Bleton, J.-P., Baron, J.-C., Maier, M. A., & Lindberg, P. G. (2016). Upper Limb Outcome Measures Used in Stroke Rehabilitation Studies: A Systematic Literature Review. *PLoS ONE*, *11*(5). https://doi.org/10.1371/journal.pone.0154792

Sathian, K., Buxbaum, L. J., Cohen, L. G., Krakauer, J. W., Lang, C. E., Corbetta, M., & Fitzpatrick, S. M. (2011). Neurological principles and rehabilitation of action disorders: common clinical deficits. *Neurorehabilitation and Neural Repair*, *25*(5 0), 21S-32S. https://doi.org/10.1177/1545968311410941

Schuster, C., Hahn, S., & Ettlin, T. (2010). Objectively-assessed outcome measures: a translation and cross-cultural adaptation procedure applied to the Chedoke McMaster Arm and Hand Activity Inventory (CAHAI). *BMC Medical Research Methodology*, *10*(1), 106. https://doi.org/10.1186/1471-2288-10-106

Simpson, L. A., & Eng, J. J. (2013). Functional recovery following stroke: capturing changes in upper-extremity function. *Neurorehabilitation and Neural Repair*, *27*(3), 240–250. https://doi.org/10.1177/1545968312461719

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

Stewart, A. L., & Nápoles-Springer, A. (2000). Health-related quality-of-life assessments in diverse population groups in the United States. *Medical Care*, *38*(9 Suppl), II102-124.

Stewart, J. C., & Cramer, S. C. (2013). Patient-reported measures provide unique insights into motor function after stroke. *Stroke*, *44*(4), 1111–1116. https://doi.org/10.1161/STROKEAHA.111.674671

Streiner, D. L. (2010). Measure for measure: new developments in measurement and item response theory. *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, *55*(3), 180–186. https://doi.org/10.1177/070674371005500310

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use* (Fifth edition). Oxford: Oxford University Press.

Sturm, J. W., Dewey, H. M., Donnan, G. A., Macdonell, R. A. L., McNeil, J. J., & Thrift, A. G. (2002). Handicap after stroke: how does it relate to disability, perception of recovery, and stroke subtype?: the north North East Melbourne Stroke Incidence Study (NEMESIS). *Stroke*, *33*(3), 762–768.

Sunderland, A., Tinson, D., Bradley, L., & Hewer, R. L. (1989). Arm function after stroke. An evaluation of grip strength as a measure of recovery and a prognostic indicator. *Journal of Neurology, Neurosurgery, and Psychiatry*, *52*(11), 1267–1272.

Sveen, U., Bautz-Holter, E., Sødring, K. M., Wyller, T. B., & Laake, K. (1999). Association between impairments, self-care ability and social activities 1 year after stroke. *Disability and Rehabilitation*, *21*(8), 372–377.

Tatemichi, T. K., Desmond, D. W., Stern, Y., Paik, M., Sano, M., & Bagiella, E. (1994). Cognitive impairment after stroke: frequency, patterns, and relationship to

functional abilities. *Journal of Neurology, Neurosurgery, and Psychiatry*, *57*(2), 202–207.

Teh, W. L., Abdin, E., Vaingankar, J. A., Seow, E., Sagayadevan, V., Shafie, S., … Subramaniam, M. (2018). Prevalence of stroke, risk factors, disability and care needs in older adults in Singapore: results from the WiSE study. *BMJ Open*, *8*(3), e020285. https://doi.org/10.1136/bmjopen-2017-020285

Trombly, C. A., & Wu, C. Y. (1999). Effect of rehabilitation tasks on organization of movement after stroke. *The American Journal of Occupational Therapy: Official Publication of the American Occupational Therapy Association*, *53*(4), 333–344.

Twitchell, T. E. (1951). The restoration of motor function following hemiplegia in man. *Brain: A Journal of Neurology*, *74*(4), 443–480.

Twitchell, T. E. (1954). Sensory factors in purposive movement. *Journal of Neurophysiology*, *17*(3), 239–252. https://doi.org/10.1152/jn.1954.17.3.239

Tyson, S. F., Hanley, M., Chillala, J., Selley, A. B., & Tallis, R. C. (2008). Sensory loss in hospital-admitted people with stroke: characteristics, associated factors, and relationship with function. *Neurorehabilitation and Neural Repair*, *22*(2), 166–172. https://doi.org/10.1177/1545968307305523

van Wijck, F. M., Pandyan, A. D., Johnson, G. R., & Barnes, M. P. (2001). Assessing motor deficits in neurological rehabilitation: patterns of instrument usage. *Neurorehabilitation and Neural Repair*, *15*(1), 23–30. https://doi.org/10.1177/154596830101500104

Veerbeek, J. M., Kwakkel, G., Wegen, E. E. H. van, Ket, J. C. F., & Heymans, M. W. (2011). Early prediction of outcome of activities of daily living after stroke: a systematic review. *Stroke*, *42*(5), 1482–1488. https://doi.org/10.1161/STROKEAHA.110.604090

Vidović, M., Sinanović, O., Sabaskić, L., Hatić, A., & Brkić, E. (2011). Incidence and types of speech disorders in stroke patients. *Acta Clinica Croatica*, *50*(4), 491–494.

Wade, D. T., & Hewer, R. L. (1987). Motor loss and swallowing difficulty after stroke: frequency, recovery, and prognosis. *Acta Neurologica Scandinavica*, *76*(1), 50–54.

Waller, S. M., & Whitall, J. (2008). Bilateral arm training: Why and who benefits? *NeuroRehabilitation*, *23*(1), 29–41.

Walsh, M. E., Galvin, R., Loughnane, C., Macey, C., & Horgan, N. F. (2015). Community re-integration and long-term need in the first five years after stroke: results from a national survey. *Disability and Rehabilitation*, *37*(20), 1834–1838. https://doi.org/10.3109/09638288.2014.981302

Warlow, C. P., Gijn, J. van, Dennis, M. S., Wardlaw, J. M., Bamford, J. M., Hankey, G. J., … Rothwell, P. (2008). *Stroke: Practical Management* (3rd edition). Malden, MA: Blackwell Publishing.

Weisburd, D., & Britt, C. (2014). Measurement: the basic building block of research. In *Statistics in criminal justice* (pp. 13–35). Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-9170-5_2

Weiss, D. J., & Davison, M. L. (1981). Test theory and method. *Annual Review of Psychology*, *32*, 629–658.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural Adaptation. *Value in Health*, *8*(2), 94–104. https://doi.org/10.1111/j.1524-4733.2005.04054.x

Wolf, S. L., Kwakkel, G., Bayley, M., McDonnell, M. N., & Upper Extremity Stroke Algorithm Working Group. (2016). Best practice for arm recovery post stroke: an international application. *Physiotherapy*, *102*(1), 1–4. https://doi.org/10.1016/j.physio.2015.08.007

Wolfe, C. D. A., Crichton, S. L., Heuschmann, P. U., McKevitt, C. J., Toschke, A. M., Grieve, A. P., & Rudd, A. G. (2011). Estimates of outcomes up to ten years after stroke: analysis from the prospective South London Stroke Register. *PLoS Medicine*, *8*(5), e1001033. https://doi.org/10.1371/journal.pmed.1001033

World Health Organization. (2001). *International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–860.

Wu, C., Trombly, C. A., Lin, K., & Tickle-Degnen, L. (1998). Effects of object affordances on reaching performance in persons with and without cerebrovascular

accident. *The American Journal of Occupational Therapy: Official Publication of the American Occupational Therapy Association*, *52*(6), 447–456.

Wyller, T. B., Sveen, U., Sødring, K. M., Pettersen, A. M., & Bautz-Holter, E. (1997). Subjective well-being one year after stroke. *Clinical Rehabilitation*, *11*(2), 139–145. https://doi.org/10.1177/026921559701100207

Young, J. L., & Mayer, R. F. (1982). Physiological alterations of motor units in hemiplegia. *Journal of the Neurological Sciences*, *54*(3), 401–412.

Tables

Table 1. COSMIN definition of measurement properties (Mokkink et al., 2010)

Term			Definition
Domain	Measurement property	Aspect of a measurement property	
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same HR-PROs (internal consistency); over time (test-retest); by different persons on the same occasion (interrater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true' differences between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which a HR-PRO measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured

Table 1 (continued).

Term			Definition
Domain	Measurement property	Aspect of a measurement property	
Validity (continued)	Content validity (continued)	Face validity	The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of an HR-PRO instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the HR-PRO instrument validly measures the construct to be measured
		Structural validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument

Table 1 (continued).

Term			Definition
Domain	Measurement property	Aspect of a measurement property	
Validity (continued)		Criterion validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of a ‘‘gold standard’’
Responsiveness			The ability of an HR-PRO instrument to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability			The degree to which one can assign qualitative meaning – that is, clinical or commonly understood connotations – to an instrument’s quantitative scores or change in scores.

HR-PRO: Health-related patient-reported outcomes

Table 2. Summary of psychometric properties of all versions of the CAHAI

	CAHAI-13	CAHAI-9	CAHAI-8	CAHAI-7
Reliability				
Test-retest, ICC	0.98 (0.96 – 0.99)[a]	0.97 (0.94)[b]	0.97 (0.95)[b]	0.96 (0.92)[b]
Inter-rater, ICC	0.98	0.98	0.98	0.98
Internal consistency, α	0.98	0.98	0.98	0.97
Validity				
Cross-sectional*				
Convergent				
CMSA (Arm and hand)	0.81 (0.66 – 0.90)[a]	0.84 (0.73)[b]	0.84 (0.73)[b]	0.85 (0.75)[b]
ARAT	0.93 (0.87 – 0.96)[a]	0.94 (0.90)[b]	0.95 (0.91)[b]	0.95 (0.91)[b]
Discriminant*				
CMSA (Pain)	0.47 (0.18 – 0.68)[a]			
Longitudinal, area under ROC curve	0.95 (0.89)[b]	0.94 (0.87)[b]	0.93 (0.86)[b]	0.97 (0.94)[b]

[a] 95% confidence interval; [b] lower one-sided 95% confidence limit

* Correlation with comparative measure at baseline assessment

CAHAI: Chedoke Arm and Hand Activity Inventory; ICC, intra-class correlation coefficient; CMSA, Chedoke-McMaster Stroke Assessment; ARAT, Action Research Arm Test; ROC, receiver operating characteristics

Note: Psychometric properties were summarized from three studies (Barreca et al., 2004, 2005, 2006)

Figures

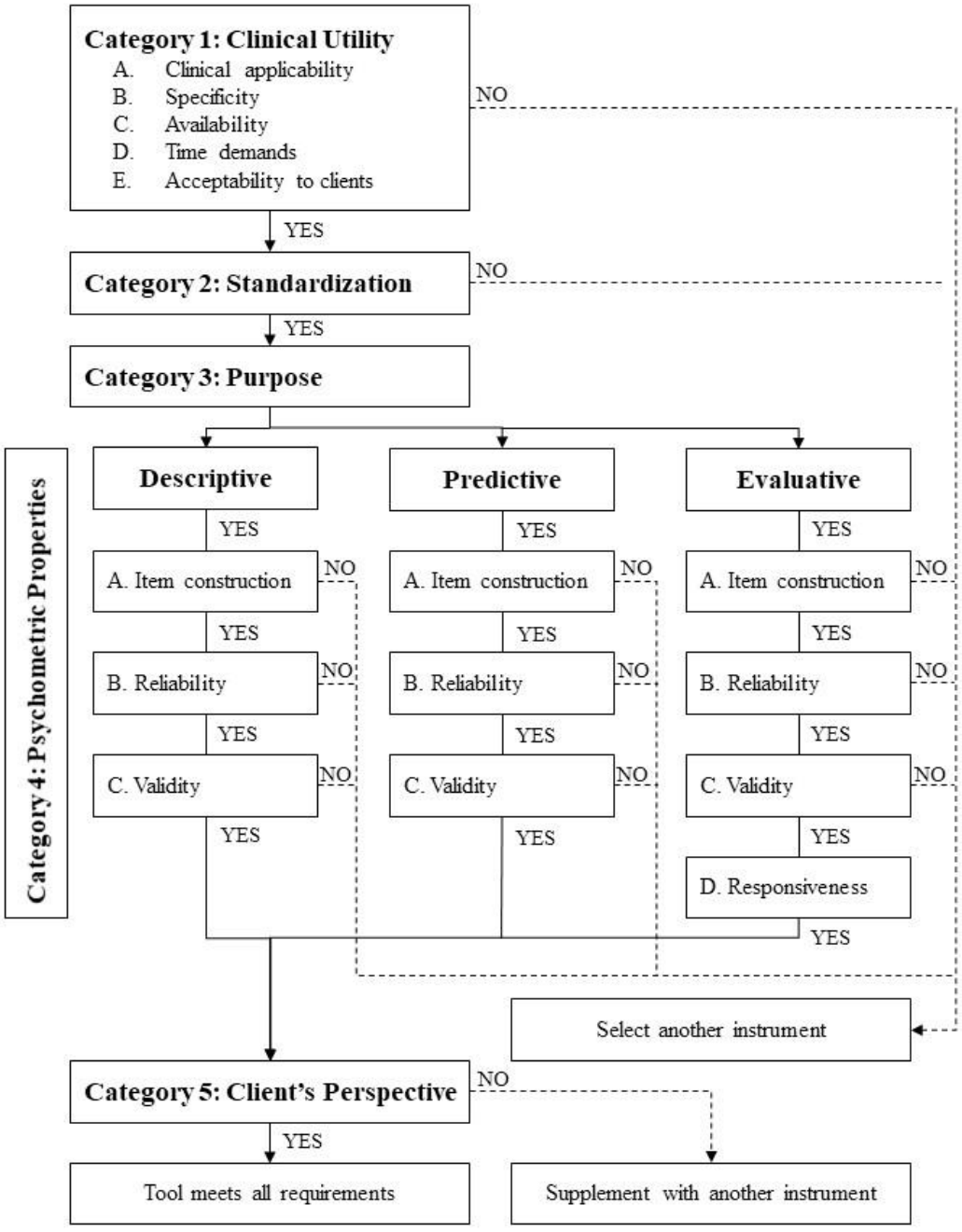


Figure 1. The instrument evaluation framework for selecting outcome measures of hand function (Adapted from Rudman & Hannah, 1998)

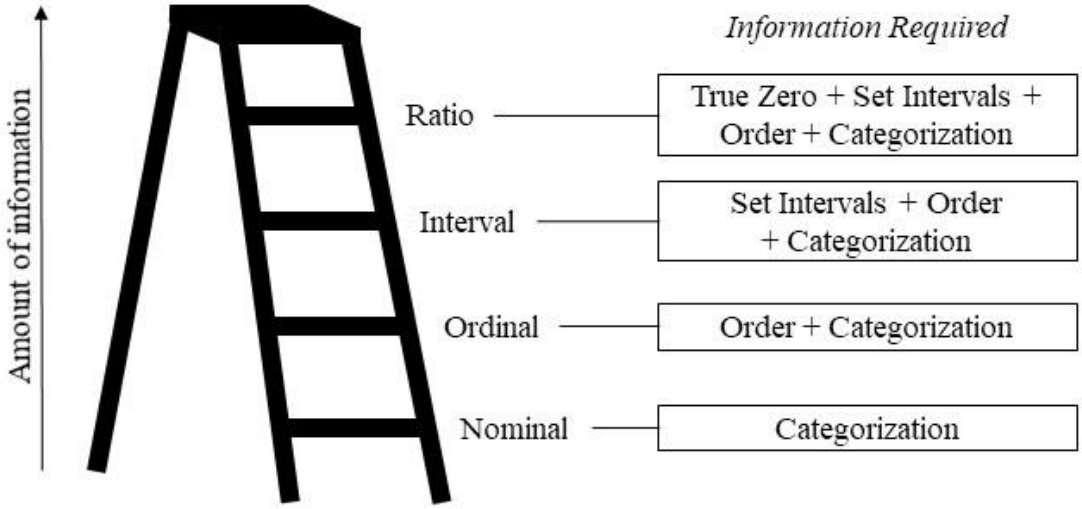


Figure 2. Levels of measurement conceptualized as a ladder (Adapted from Weisburd & Britt, 2014).

Chapter Two:

Cross-cultural adaptation and psychometric evaluation of the Singapore version of the Chedoke Arm and Hand Activity

Preface

In this chapter, the translation and cross-cultural adaptation of the Chedoke Arm and Hand Activity Inventory for the stroke population in Singapore is described. It also details the evaluation of the psychometric properties of the full 13-item Singapore version of the measure. The psychometric evaluation was conducted at a tertiary hospital in Singapore between December 2016 and May 2017. This is the first study to undertake the translation and cross-cultural adaptation of the Chedoke Arm and Hand Activity Inventory for an Asian stroke population. The Singapore version of the measure is also the first post-stroke upper extremity outcome measure that is cross-culturally adapted and validated for the stroke population in Singapore.

This manuscript was published in *Disability and Rehabilitation* and the complete citation is as follows:

> Choo, S. X., Bosch, J., Richardson, J., Stratford, P., & Harris, J. E. (2018). Cross-cultural adaptation and psychometric evaluation of the Singapore version of the Chedoke Arm and Hand Activity. *Disability and Rehabilitation*, 1–8. https://doi.org/10.1080/09638288.2018.1472817

Title Page

Cross-cultural adaptation and psychometric evaluation of the Singapore version of the Chedoke Arm and Hand Activity

Silvana X. Choo[a,b,*], Jackie Bosch[a], Julie Richardson[a], Paul Stratford[a], and Jocelyn E. Harris[a]

[a] *School of Rehabilitation Science, McMaster University, Hamilton, Ontario, Canada;*

[b]*Department of Occupational Therapy, Singapore General Hospital, Singapore.*

Correspondence concerning this article should be addressed to Silvana Choo, School of Rehabilitation Science, McMaster University, 1400 Main Street West, Institute of Applied Health Sciences, Room 308, Hamilton, Ontario L8S 2R8, Canada. Tel: +01-905-525-9140 ext. 26410. E-mail: silvana.choo.xinyi@sgh.com.sg.

Acknowledgements

We thank Elain Koh, Charmaine Magnus Kuan, and Cassandra Ng for their assistance in the cross-cultural adaptation process. We also thank Cheryl Sim, Weiqi Koh, and the Department of Occupational Therapy, Singapore General Hospital, for their contributions to the data collection and the study participants for volunteering their time to participate. The first author is supported by the Singapore General Hospital Scholarship Award.

Cross-cultural adaptation and psychometric evaluation of the Singapore version of the Chedoke Arm and Hand Activity

Abstract

Purpose: To develop a Singapore version of the Chedoke Arm and Hand Activity Inventory (CAHAI) and to estimate the construct validity and inter-rater reliability.

Materials and methods: The Translation and Cross-Cultural Adaptation of Objectively-Assessed Outcome measure procedure was used to systematically adapt the Chedoke Arm and Hand Activity Inventory. We recruited 56 adults admitted to an inpatient stroke facility to evaluate the psychometric properties of the Singapore version of the CAHAI. The Singapore version of the CAHAI, Fugl-Meyer Assessment of Upper Extremity (FMA-UE), and the Action Research Arm Test (ARAT) were administered to all participants. We used Spearman's rank correlation coefficients to estimate convergent and discriminative validity, and reliability was estimated using the intra-class correlation coefficient and standard error of measurement.

Results: Implementation of the Translation and Cross-Cultural Adaptation of Objectively-Assessed Outcome measure procedure resulted in the modification to two test items. The Singapore version of the CAHAI demonstrated convergent validity with the FMA-UE (r_s = 0.87; 95% CI: 0.76, 0.92) and ARAT (r_s = 0.80; 95% CI: 0.63, 0.9). Discriminative validity between the Singapore version of the CAHAI and FMA-UE pain subscale was r_s= 0.42 (95% CI: 0.22, 0.59). Reliability of the Singapore version of the CAHAI was 0.97 (95% CI: 0.94, 0.99) and standard error of measurement of 4.80 points (95% CI: 4.23, 5.55).

Conclusion: The Singapore version of the CAHAI demonstrated good validity and reliability, similar to the properties of the original CAHAI.

Keywords: cross-cultural adaptation; upper limb; outcome measure; measurement; stroke; psychometric evaluation

Introduction

Stroke is a major health concern worldwide and is a leading cause of disability in Singapore [1,2]. Upper extremity (UE) impairment is one of the most common deficits after a stroke; hemiparesis is present in more than 80% of individuals with stroke [3], with limited recovery where only 5% achieve full UE function [3–5]. Regaining UE function is an important rehabilitation goal for individuals with stroke [6] and an important first step in the rehabilitation process is the use of outcome measures. Outcome measures provide information on the extent of UE impairments and activity limitations [7] and enable therapists to evaluate change. This information facilitates clinical decisions to select appropriate interventions according to the severity of UE deficits [6].

There are several outcome measures available to assess post-stroke UE function [6,8]. A review of the literature evaluating psychometric properties and clinical utility of UE outcome measures identified a total of six outcome measures that demonstrated high levels of measurement quality and clinical utility [8]: one patient-reported measure (ABILHAND) and five performance-based measures (Fugl-Meyer Assessment of Upper Extremity, Action Research Arm Test, Box and Block Test, Chedoke Arm and Hand Activity Inventory, and Wolf Motor Function Test). The Fugl-Meyer Assessment of Upper Extremity and the Action Research Arm Test are also recommended outcome measures for intervention trials targeted at sensorimotor recovery after stroke [9].

Among all the recommended UE outcome measures, the Chedoke Arm and Hand Activity Inventory (CAHAI), which uses 13 real-life daily tasks to evaluate the function of the affected UE after a stroke [10], is the only measure that uses daily activities in all

test items. Using daily tasks and objects to evaluate the UE ensures a more accurate evaluation as UE movements are influenced by the meaning of objects and tasks to the individual [11,12]. Persons with stroke have more efficient, faster, and smoother UE movements when performing meaningful tasks with real-life objects (e.g. picking up a telephone receiver and dialing a number) compared to performing similar tasks in a simulated condition (e.g. picking up a stick of the same size, weight, and color as the telephone receiver) [11,12].

With regards to psychometric properties, the CAHAI also demonstrated greater sensitivity to change than the ARAT [13], which was a recommended outcome measure [9]. The CAHAI has strong psychometric properties when used with adults with sub-acute and chronic stroke: inter-rater reliability, intra-class correlation (ICC) = 0.98 [14]; test-retest reliability, ICC = 0.98 [14]; convergent cross-sectional validity, correlation with the Chedoke-McMaster Stroke Assessment (CMSA) arm and hand components and the ARAT of 0.81 and 0.93 respectively [14]; and discriminant cross-sectional validity, correlation with the CMSA pain scale of 0.47 [14]. In the same clinical population, the CAHAI has a longitudinal validity of an area under the receiver operating characteristics curve = 0.95, where the analysis was conducted to examine the CAHAI's ability to distinguish different amounts of change between groups of participants with acute and chronic stroke [14]. The CAHAI also has a high internal consistency of coefficient alpha = 0.98 [10].

Three shortened versions of the CAHAI containing seven, eight and nine items (CAHAI-7, CAHAI-8, CAHAI-9 respectively) were later developed and have

comparable psychometric properties to the original 13-item version [15]. The CAHAI has also been translated into five different languages (French, German, Hebrew, Italian and Portuguese) and is used clinically in Canada, Germany and Australia [16–18] but has not been validated in Asian stroke populations. The validity of the CAHAI in countries such as Singapore may be jeopardized due to cultural differences between Singapore and Canada (where the CAHAI was developed). The three major ethnic groups in Singapore are Chinese (74.3%), Malays (13.4%), and Indians (9.1%) [19], which differs from Canada. Hence, the daily tasks used in the CAHAI may not be culturally relevant to Singaporeans. For example, one item assesses the ability of the individual to use a knife or fork with their affected UE; in Singapore, a spoon and fork, chopsticks or bare hand are also used. Also, the CAHAI was developed in English; although English is an official language and the main language used in Singapore, approximately 17% of Singapore's population is illiterate in English [20] and English is also not frequently used by older adults [21]. There are also three other official languages in Singapore: Mandarin, Malay, and Tamil. Thus, cross-cultural adaptation of the CAHAI, a process that examines both language (translation) and cultural adaptation (cultural relevance) of the measure for the new population [22], is necessary to support the use of the measure in Singapore. The use of a systematic and validated approach in the cross-cultural adaptation process is recommended to ensure equivalence between the original and adapted version of the CAHAI [22].

The aim of our study was to develop a Singapore version of the CAHAI. The objectives were: (1) to cross-culturally adapt the CAHAI for the acute and subacute

stroke population in Singapore; and (2) to estimate the construct validity and inter-rater reliability of the Singapore version of the CAHAI.

Materials and methods

Our study followed the Translation and Cross-Cultural Adaption of Objectively-Assessed Outcome measures procedure [18]. This procedure consists of eight steps that systematically guide the translation and cultural adaptation process of an outcome measure (table 1). Ethics approval was obtained from the Singhealth Centralized Institutional Review Board (CIRB 2015/2915) and the Hamilton Integrated Research Ethics Board (HiREB 0758). The study protocol, participant information sheets, and all consent forms used in this study were approved by both ethics review boards.

Measure

The CAHAI consists of 13 items using real-life tasks to evaluate the functional ability of the affected UE post-stroke [10]. For each test item, standardized instructions (e.g. “Call 911 using both of your hands”) are provided to patients and the assessor also demonstrates each task at least once (twice if needed). The patient then performs each task and the assessor evaluates the performance of the affected UE on a 7-point scale, from total assistance (score = 1) to independence (score = 7). Scoring is based on the extent the affected UE demonstrates stabilization and manipulation abilities. Scores on each task are summed to obtain a total score that can range from 13 to 91, with higher scores indicating better UE function. The CAHAI is available without charge at: www.cahai.ca.

Translation and cross-cultural adaptation

The translation and cross-cultural adaptation of the CAHAI for the stroke population in Singapore followed *Step 1* to *Step 7* outlined below of the Translation and Cross-Cultural Adaption of Objectively-Assessed Outcome measures procedure [18].

Step 1: Forward translation and region-specific adaptations

The CAHAI administration manual (version 2) [23] was reviewed by two registered occupational therapists (OTs) from Singapore who were familiar with the CAHAI. Both reviewers are experienced in working with individuals with stroke (8 years and 10 years of clinical experience respectively) and also teach at the local occupational therapy education program. Each reviewer independently reviewed the manual and evaluated the cultural relevance of the tasks, equipment, material descriptions and scoring instructions to the local population and provided recommendations for necessary adaptations. Forward translation of the standard instructions (provided to patients) to Mandarin and Malay was completed by the reviewers.

Step 2: Merging the forward translations and region-specific adaptations

Recommendations from the first (R1) and second reviewer (R2) were synthesized to produce one common set of recommendations (R-12). This synthesis was achieved through group discussions between the two reviewers and the author of this paper [SC]. The synthesis process was documented by the author [SC], detailing how the consensus of recommendations was reached, issues addressed and how they were resolved.

Step 3: Preparation of the translated and adapted outcome measure

The set of recommendations from *Step 2* (R-12) formed the basis of the initial draft of the Singapore version of the CAHAI (CAHAI-SG). The CAHAI-SG administration manual was prepared by integrating the recommendations, and photographs in the original manual were also changed from a Caucasian male to an Asian male to enhance the cultural relevance of the manual.

Step 4: Backward translation

The standard instructions for patients in Mandarin and Malay were back-translated to English. Back-translation was completed independently by two local OTs who had no previous experience with the original CAHAI administration manual. The OTs had nine and seven years of clinical experience respectively.

Step 5: Review of the translated and culturally-adapted outcome measure

The initial draft of the CAHAI-SG administration manual and photographs of all equipment were sent to the original authors of the CAHAI for review. Documents from the synthesis process in *Step 2* were also made available to the original authors to provide information on the rationale for each adaptation to the CAHAI.

Step 6: Further adaptation and proof-reading

One of the original authors of the CAHAI reviewed the initial draft of the CAHAI-SG and determined that no further modifications were required. A final review of the CAHAI-SG administration manual was conducted by one of the reviewers from *Step 1* (quality control step).

Step 7: Pre-testing of the outcome measure

The CAHAI-SG was pre-tested by two registered OTs on four patients with unilateral UE deficits due to stroke at a tertiary hospital in Singapore. Participant eligibility criteria and clinical setting were identical to those in *Step 8* (described below). This pre-testing was intended to identify any emerging discrepancies regarding scoring and interpretation of instructions from the administration manual.

Psychometric evaluation of the CAHAI-SG (Step 8)

The final step (*Step 8*) of the Translation and Cross-Cultural Adaption of Objectively-Assessed Outcome measures procedure is the evaluation of the psychometric properties of the adapted measure on the intended population [18].

Study design and participants

We employed a cross-sectional study design. Study participants were recruited from the largest tertiary hospital in Singapore, where one of the two campuses of the National Neuroscience Institute is located. All consecutive patients admitted to the inpatient neurology and rehabilitation wards and referred to the Department of Occupational Therapy were screened for study eligibility. The inclusion criteria were: (1) $\geq$ 21 years, (2) diagnosis of unilateral hemispheric stroke, (3) acute or sub-acute phase of stroke (< 12 months), (4) ability to tolerate upright sitting in a chair for at least 30 minutes, and (5) able to follow simple two-step instructions. Patients with unstable medical conditions, pre-existing upper extremity impairments (prior to the stroke), or severe visual impairments (based on medical records) were excluded. Informed consent was obtained

for all participants.

Measures

Singapore version of the CAHAI (CAHAI-SG). The administration and scoring procedures of the CAHAI-SG follow the same procedures as the original CAHAI [23].

Fugl-Meyer Assessment of Upper Extremity (FMA-UE). The FMA-UE consists of four subscales (motor function, sensation, passive range of motion and joint pain) that evaluate post-stroke UE impairments [24]. Items are scored on a 3-point scale and the maximum total score for FMA-UE is 126, where a higher score indicates lesser impairments. The maximum sub-total scores for the motor function, sensation, passive range of motion, and joint pain subscales are 66, 12, 24, and 24 respectively. The inter-rater reliability of the FMA-UE ranged from ICC = 0.61 to 0.99 in acute and subacute stroke samples [25–27].

Action Research Arm Test (ARAT). The ARAT is a 19-item test that measures the impairment and activity level (e.g. pinch, grasp, and reach) of the affected UE post-stroke [28]. Each item is scored on a 4-point scale and the range of score on the ARAT is 0 to 57, with higher scores indicating better UE function. The inter-rater reliability of the ARAT in acute and subacute stroke sample was reported in two studies, with ICC = 0.92 and 0.99 respectively [29,30].

Raters

A total of eight registered OTs participated as raters in the evaluation of construct validity and inter-rater reliability of the CAHAI-SG. The mean years of clinical experience of the raters was 3.47 (standard deviation = 1.53). There were six female and two male raters. Most raters had minimal (administered < 5 assessments) or no prior experience administering the CAHAI (n = 4 and n = 3 respectively). Only one rater had experience (10-15 assessments) in administering the original CAHAI.

All raters were trained to administer and score the CAHAI-SG, FMA-UE, and ARAT through two training workshops, which comprised of lectures and demonstrations. To further ensure standardization of data collection, administration and scoring procedures of the FMA-UE and the ARAT described by Sullivan et al. [26] and Yozbatiran et al. [31] respectively were used.

Sample size calculations

Construct validity. Sample size was calculated based on an expected correlation of ≥ 0.8 between the total scores of the CAHAI-SG and the total scores of the FMA-UE and ARAT, and a lower one-sided 95% confidence interval width of 0.10 [32]. The estimated sample size was 54; however, as there were eight raters, the target sample size for construct validity was 56 (rounded to the nearest multiple of eight).

Inter-rater reliability. The number of participants required was calculated using an expected inter-rater reliability ICC of 0.90 with a lower one-sided 95% confidence interval width of 0.1. According to parameter estimation sample size calculation [33], 11

participants were required. However, as there were eight raters, the target sample size was 16 (rounded to the nearest multiple of eight).

Procedures

For each participant, a single rater administered the UE measures in a single session. The assessment session was conducted within two days after study recruitment, and the measures were administered in the following sequence: (1) FMA-UE, (2) CAHAI-SG, and (3) ARAT. A balanced incomplete block design was used to assign raters to participants [34]. There was a total of seven blocks, with eight participants in each block. Within each block, each rater conducted one assessment session and the order of the rater was based on convenience.

CAHAI-SG administration was video recorded for all participants. Sixteen video recordings from the 56 assessments completed were randomly selected, using sealed envelopes, to estimate inter-rater reliability. Each video was independently scored by all eight raters. To avoid an order effect in the scoring of the CAHAI-SG, an 8x8 Latin square design was used to determine the sequence of reviewing the 16 videos for each rater.

Statistical analysis

All statistical analyses were performed using STATA Version 14 [35]. Descriptive statistics were used to summarize participant characteristics, the central tendency, range and distribution of the scores on each outcome measure. Floor and ceiling effects for each measure was examined and were considered present if more than 15% of participants

achieved the lowest or highest possible score [36].

Construct validity. Construct validation of a measure involves the process of formulating theories about the relationships between attributes of interests and testing whether the measure provides results consistent with the theories [37]. In our study, it was theorized that an outcome measure designed to assess UE function should correlate highly with other similar measures of UE function (convergent validity), and correlate less with measures of pain (discriminant validity). Thus, we expected the CAHAI-SG total scores to correlate more highly with FMA-UE scores (total and motor-subscale) and ARAT total scores than with the FMA-UE joint pain subscale scores. Visual inspection and the Shapiro-Wilk test were used to assess normality of the data. Spearman's rank correlation coefficients (r_s) with 95% confidence intervals (CI) were computed for CAHAI-SG total scores with FMA-UE and ARAT total scores, and CAHAI-SG total scores with the FMA-UE motor and joint pain subscale scores.

To determine if the construct validity of the CAHAI-SG was similar to the CAHAI, their respective correlation coefficients with the ARAT were compared. Each correlation coefficient was converted to a z-score using Fisher's *r* to *z* transformation and compared using the following formula [38]:

$$z_{observed} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

A $z_{observed}$ score greater than 1.96 (two-tailed test of significance with alpha set at 0.05) was considered statistically significant.

Inter-rater reliability. A randomized analysis of variance (ANOVA) was computed that identified three sources of variances: participants, rater, and error. Using these sources of variance, relative reliability was estimated by the Type 2,1 ICC ($ICC_{2,1}$) [39]. The standard error of measurement (SEM), calculated by taking the square root of the mean-square error, with the 95% CI was used to estimate the absolute reliability [40,41].

To determine if the inter-rater reliability of the CAHAI-SG was similar to the CAHAI, their respective SEMs were compared using F-test. An F-score greater than 2.18 (two-tailed test of significance with alpha set at 0.05) was considered statistically significant.

Results

Translation and cross-cultural adaptation

Steps 1 – 3: Forward translation, region-specific adaptations, and preparations

Adaptations to the original CAHAI included modifications to two test items, changes to descriptions of equipment, improving the clarity of the manual, and the refinement and translation of instructions. The language of the original administration manual was also changed from Standard Canadian English to Standard Singapore English.

Adaptation of items. Two test items (‘open jar of coffee’ and ‘call 911’) were identified as not culturally relevant to the local population; the use of instant coffee in the form of sachets were more commonly used and the emergency numbers in Singapore are 995 (for fire and ambulance) and 999 (for police). These items were adapted to ‘open a jar of

peanut butter' and 'call 995' respectively in the CAHAI-SG. The item 'clean a pair of eyeglasses' was also modified to 'clean a pair of spectacles' in the CAHAI-SG as both reviewers identified eyeglasses as a term not commonly used in Singapore.

Changes to descriptions of equipment. The descriptions of seven pieces of equipment in the original manual were identified as not culturally relevant and changes to their descriptions were recommended. Two brand names described in the CAHAI (Dycem and Rubbermaid®) were changed in the CAHAI-SG (non-slip mat and Toyogo® respectively). *Pitcher*, *wash cloth* and *wash basin* were changed to more frequently used terms in Singapore (*jug*, *face towel*, and *plastic basin* respectively). Paper size dimensions were changed from 8.5" x 11" to the international standard A4 paper size [42]. The material of the poncho was changed from polar fleece to flannel, which is a material more readily available and suitable for the tropical climate in Singapore.

Improving clarity of the manual. Two footnotes were added to enhance the clarity of the CAHAI-SG administration manual. In the general section of the original administration manual, there was a sentence 'When attempting each task, always consider safety, especially for Stage 1 upper limb.' The reviewers reported that it was unclear what 'Stage 1' meant, and thus, a footnote 'Stage 1 of Brunnstrom stages of motor recovery (flaccid paralysis)' was added in the CAHAI-SG manual to clarify what 'Stage 1' referred to. A footnote detailing the website link to the instructions on making the poncho and the vest was added to the equipment section of the CAHAI-SG manual.

Refinement and translation of instructions. The standard instructions (in English) for patients of three CAHAI items were refined to enhance its clarity. Table 2 provides the detailed description and rationale for these changes. As English is the main language in Singapore, translation of the entire administration manual was not required. However, both reviewers recommended the standard instructions also be available in Mandarin and Malay as these languages are commonly spoken in Singapore [43]. Although Tamil is also an official language in Singapore, only 3.26% of the population speaks Tamil on a daily basis and 0.35% of Singaporeans are literate in only Tamil [43]. Thus, in consideration that the language is not commonly used, the instructions were not translated to Tamil.

The standard instructions for patients in all 13 items in the CAHAI-SG were translated into Mandarin and Malay (Bahasa Melayu) languages. For example, in the test item 'Call 995' of the CAHAI-SG, instructions given to a patient are available in three languages as follows:

- "Call 995 using both of your hands." (English)
- "用你的双手拨打 995。" (Mandarin)
- "Telefon 995 menggunakan kedua-dua tangan anda." (Malay)

Step 4: Backward translation

There were no major differences between the back-translations of the standard instructions from Mandarin and Malay and the instructions in English.

Steps 5 – 7: Review of the adapted measure, proof-reading, and pre-testing

No modifications to the initial draft of the CAHAI-SG manual were required after the review by one of the original authors of the CAHAI. In the pre-testing, no major discrepancies were raised by the two OTs who administered the CAHAI-SG. No further changes to the manual were made and the final version of the CAHAI-SG administration manual is available at: http://cahai.ca/manual.html.

Psychometric evaluation of the CAHAI-SG (Step 8)

A total of 56 participants completed the assessment session and data analysis was conducted on 55 participants. One participant with chronic stroke (> 12 months) was erroneously recruited and therefore excluded from data analysis. Table 3 describes the participant characteristics and table 4 provides a summary of the scores on all three outcome measures. Participants' mean age was 63 years (SD = 13 years) and the ethnicity of the sample was reflective of the ethnic composition of Singaporeans [19]. All participants had an ischaemic stroke (100%) and the mean days post-stroke was 11 days (SD = 14 days). Floor effects were found in three of the five subscales of the ARAT (grasp, grip, and pinch). Ceiling effects were present for all three measures, including each subscale of the FMA-UE and the ARAT.

Construct validity

The CAHAI-SG total scores correlated strongly with the FMA-UE total scores (r_s = 0.87, 95% CI: 0.76, 0.92) and the ARAT total scores (r_s = 0.80, 95% CI: 0.63, 0.9). The CAHAI-SG total scores also correlated strongly with the FMA-UE motor subscale scores

(r_s = 0.89, 95% CI: 0.78, 0.94). The correlation between the CAHAI-SG total scores and scores on the FMA-UE pain subscale was r_s= 0.42 (95% CI: 0.22, 0.59). These results support both the convergent and discriminant validity of the CAHAI-SG.

Comparing the correlation coefficients of the CAHAI-SG and CAHAI (r = 0.93, 95% CI: 0.87, 0.96) [14] with the ARAT, the CAHAI-SG had a significantly lower correlation coefficient than the CAHAI (z = 2.582, p = 0.010).

Inter-rater reliability

The SEM was 4.80 points (95% CI: 4.23, 5.55) and the $ICC_{2,1}$ of the CAHAI-SG was 0.97 (95% CI: 0.94, 0.99). Comparing the SEMs of the CAHAI-SG and the CAHAI (SEM = 2.8 CAHAI points, 95% CI: 2.3, 3.7) [14], the CAHAI-SG had significantly larger SEM than the CAHAI ($F(15,38)$ = 2.938, p = 0.004).

Discussion

Cross-cultural adaptation of a measure is required when the use of the measure on the intended new population differs considerably from the original population in terms of culture, language, and country [44]. Our study followed the Translation and Cross-Cultural Adaption of Objectively-Assessed Outcome measures procedure to culturally adapt the CAHAI and using a validated process ensured equivalence between the original and adapted versions of the measure [22]. The modifications in the CAHAI-SG and psychometric evaluation of the measure can be considered within four dimensions of equivalence: conceptual, semantic, operational, and measurement equivalence [45].

Conceptual equivalence focuses on the instrument measuring the same construct in each culture and items within the instrument represent the definition of the construct [45]. In the CAHAI-SG, the items 'opening a jar of coffee' and 'calling 911' in the CAHAI were modified to 'opening a jar of peanut butter' and 'calling 995' respectively; these modifications addressed the cultural relevance of the tasks to Singaporeans whilst maintaining the same functional demands on UE for successful task performance. Thus, conceptual equivalence was addressed with all 13 items in the CAHAI-SG evaluating the same UE functions as the original measure.

Semantic equivalence relates to technical features of language (e.g. grammar, syntax, and complexity) and the transfer of meaning of each item across languages [45,46]. Language modifications in the CAHAI-SG ensured the measure is understood by both its users (healthcare professionals) and individuals with stroke in Singapore. The CAHAI-SG administration manual uses the Standard Singapore English, which differed from the CAHAI manual which was written in Standard Canadian English. This led to several modifications to the original manual, such as changes to spelling (e.g. from 'stabilize' to 'stabilise') and nouns (e.g. from 'eyeglasses' to 'spectacles'). The standard instructions provided to patients in the CAHAI-SG are also available in languages commonly spoken in Singapore (English, Mandarin, and Malay). Translation methods of the standard instructions for each item of the CAHAI-SG from English to Mandarin and Malay languages included forward and backward translation, which ensured consistency in the meanings of the instructions.

Operational equivalence refers to the method of assessment, where the data collection method is acceptable for the target culture and does not affect the results differently [45,47]. The CAHAI-SG is a performance-based measure where the method of assessment is the task performance of each test item. Each test item in the original measure was carefully evaluated during the cross-cultural adaptation process to ensure that the test items in the Singapore version reflect common daily tasks familiar to Singaporeans. Thus, task performance of all items in the CAHAI-SG is regarded as acceptable within the context of Singapore. Furthermore, during the pre-testing and psychometric evaluation of the CAHAI-SG, no participant expressed unfamiliarity with any test item.

Measurement equivalence can be defined as the extent to which the adapted version of the measure has comparable psychometric properties as the original measure [45,46]. Overall, measurement equivalence of the CAHAI-SG was achieved as psychometric evaluation demonstrated comparable results to the CAHAI. In terms of construct validity, comparing the convergent validity of the measures, the CAHAI-SG correlated with FMA-UE total scores at $r_s = 0.87$ (95% CI: 0.76, 0.92), which was comparable to the correlation between the CAHAI and the CMSA arm and hand subscale ($r = 0.81$, 95% CI: 0.66, 0.90) [14]. Although different measures were used to estimate the convergent validity of the CAHAI-SG and the CAHAI, the similarities in the theoretical foundations of the FMA-UE and the CMSA supports the basis of comparison (both measures were developed using Brunnstrom stages of motor recovery [24,48]).

When comparing the correlation of the CAHAI and the CAHAI-SG with the ARAT, we found a significantly lower correlation on the CAHAI-SG (r_s = 0.80, 95% CI: 0.63, 0.9) compared to the CAHAI (r = 0.93, 95% CI: 0.87, 0.96) [14]. This significant lower correlation was unanticipated since all items in the CAHAI-SG maintained the same functional demands on the UE as the original measure. We attributed the lower correlation to different scoring procedures of the ARAT between studies. In the CAHAI study [14], the authors followed the original scoring procedures of the ARAT [49], thus, the speed of task completion on the CAHAI and the ARAT were evaluated similarly using subjective judgement (i.e. in a timely manner). In our study, participants were scored on the ARAT using the scoring procedures described by Yozbatiran et al. [31], where cut-off times were defined on the 4-point scale. While the use of cut-off times improved the distinction between the scores on the ARAT, it introduced the objective evaluation of speed. Speed was not objectively evaluated in the CAHAI-SG and the difference in how speed was evaluated (i.e., subjectively in the CAHAI-SG versus objectively in the ARAT) may account for the lower correlation between the CAHAI-SG and the ARAT.

Examining the discriminant validity, the CAHAI-SG correlated with the scores on the FMA-UE joint pain subscale at r_s = 0.42 (95% CI: 0.22, 0.59), which was similar to the correlation between the CAHAI with the CMSA shoulder pain subscale (r = 0.47, 95% CI: 0.18, 0.68) [14].

For reliability, the SEM of the CAHAI-SG was 4.80 points (95% CI: 4.23, 5.55), which was significantly higher than the CAHAI (SEM = 2.8 CAHAI points, 95% CI: 2.3,

3.7) [14]. The weaker absolute reliability of the CAHAI-SG may be due to differences in the training of raters. The raters in the study by Barreca et al. [14] were trained to 85% accuracy in the administration and scoring of the CAHAI; in contrast, our raters were not trained to a required level of accuracy of the CAHA-SG. This may have resulted in a greater error variance, and therefore contribute to the larger absolute reliability observed on the CAHAI-SG. However, the relative reliability of the CAHAI-SG was similar to the CAHAI, with ICC = 0.97 (95% CI: 0.94, 0.99) and ICC = 0.98 (95% CI: 0.87, 0.96) respectively.

Limitations

Outcome measures were consistently administered in the same order for all participants which may result in an order effect [50]. Scores on the ARAT (administered last) may be lower due to fatigue, which may influence the correlation between the CAHAI-SG and ARAT scores. The second limitation is the competency of the raters in our study. Although training was provided on the administration and scoring of all measures, there was no formal evaluation of the raters' competency. In contrast, the raters in the CAHAI validation study were trained to 85% accuracy in the administration and scoring of all measures [14]. However, despite the lack of formal competency evaluation in our study, the psychometric properties of the CAHAI-SG was comparable to the CAHAI. Another limitation is the use of video analysis to estimate inter-rater reliability. Although video recordings allowed the observation of exactly the same performance and was the most feasible method of data collection (i.e., minimal burden on participants) due to a large group of raters in our study, the observation and scoring of participants using video

recordings differs from clinical practice as the ratings by the raters were based on the handling procedures of one rater. The fourth limitation is the cross-sectional design of our study which restricted the types of validity and reliability evaluated. Future studies using a longitudinal design are therefore recommended to evaluate other psychometric properties of the CAHAI-SG (e.g. test-retest reliability, longitudinal validity, and responsiveness).

Conclusion

This is the first study to undertake a cross-cultural adaptation and validation the CAHAI for an Asian stroke population. The results of our study address the equivalence of the CAHAI-SG to the original version, as well as provide initial evidence in support of the measure's validity and reliability for adults with acute and subacute stroke. Clinicians and researchers now have access to a usable and culturally relevant outcome measure to evaluate UE function in a multicultural stroke population in Singapore.

Declaration of interest

The authors report no declarations of interest.

Implications for rehabilitation

- The Singapore version of the Chedoke Arm and Hand Activity Inventory demonstrates evidence of construct validity and inter-rater reliability.
- The Singapore version of the Chedoke Arm and Hand Activity Inventory can be used by clinicians and researchers to evaluate function in the affected upper extremity for persons with stroke in Singapore.

References

[1] Warlow C, Sudlow C, Dennis M, et al. Stroke. Lancet Lond. Engl. 2003;362:1211–1224.

[2] Epidemiology & Disease Control Division. Singapore burden of disease study 2010 [Internet]. Singapore: Ministry of Health, Singapore; 2014 [cited 2016 Jun 2]. Available from: https://www.moh.gov.sg/content/dam/moh_web/Publications/Reports/2014/Singapore%20Burden%20of%20Disease%20Study%202010%20Report_v3.pdf.

[3] Nakayama H, Jørgensen HS, Raaschou HO, et al. Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. Arch. Phys. Med. Rehabil. 1994;75:394–398.

[4] Kwakkel G, Kollen BJ, Wagenaar RC. Long term effects of intensity of upper and lower limb training after stroke: a randomised trial. J. Neurol. Neurosurg. Psychiatry. 2002;72:473–479.

[5] Sommerfeld DK, Eek EU-B, Svensson A-K, et al. Spasticity after stroke: its occurrence and association with motor impairments and activity limitations. Stroke J. Cereb. Circ. 2004;35:134–139.

[6] Lang CE, Bland MD, Bailey RR, et al. Assessment of upper extremity impairment, function, and activity following stroke: Foundations for clinical decision making. J. Hand Ther. Off. J. Am. Soc. Hand Ther. 2013;26:104–115.

[7] Duroz MT. Assessment of hand function. In: Duroz MT, editor. Hand Funct. Pract. Guide Assess. [Internet]. New York: Springer; 2014 [cited 2015 Feb 15]. p. 41–54. Available from: http://link.springer.com.libaccess.lib.mcmaster.ca/chapter/10.1007/978-1-4614-9449-2_3/fulltext.html.

[8] Murphy MA, Resteghini C, Feys P, et al. An overview of systematic reviews on upper extremity outcome measures after stroke. BMC Neurol. 2015;15:292.

[9] Kwakkel G, Lannin NA, Borschmann K, et al. Standardized measurement of sensorimotor recovery in stroke trials: Consensus-based core recommendations from the Stroke Recovery and Rehabilitation Roundtable. Int. J. Stroke. 2017;12:451–461.

[10] Barreca S, Gowland CK, Stratford P, et al. Development of the Chedoke Arm and Hand Activity Inventory: theoretical constructs, item generation, and selection. Top. Stroke Rehabil. 2004;11:31–42.

[11] Wu C, Trombly CA, Lin K, et al. Effects of object affordances on reaching performance in persons with and without cerebrovascular accident. Am. J. Occup. Ther. 1998;52:447–456.

[12] Trombly CA, Wu CY. Effect of rehabilitation tasks on organization of movement after stroke. Am. J. Occup. Ther. 1999;53:333–344.

[13] Barreca SR, Stratford PW, Masters LM, et al. Comparing 2 versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. Phys. Ther. 2006;86:245–253.

[14] Barreca SR, Stratford PW, Lambert CL, et al. Test-retest reliability, validity, and sensitivity of the Chedoke Arm and Hand Activity Inventory: a new measure of upper-limb function for survivors of stroke. Arch. Phys. Med. Rehabil. 2005;86:1616–1622.

[15] Barreca SR, Stratford PW, Masters LM, et al. Validation of three shortened versions of the Chedoke Arm and Hand Activity Inventory. Physiother. Can. 2006;58:148–156.

[16] Rowland TJ, Turpin M, Gustafsson L, et al. Chedoke Arm and Hand Activity Inventory-9 (CAHAI-9): perceived clinical utility within 14 days of stroke. Top. Stroke Rehabil. 2011;18:382–393.

[17] Gustafsson LA, Turpin MJ, Dorman CM. Clinical utility of the Chedoke Arm and Hand Activity Inventory for stroke rehabilitation. Can. J. Occup. Ther. 2010;77:167–173.

[18] Schuster C, Hahn S, Ettlin T. Objectively-assessed outcome measures: a translation and cross-cultural adaptation procedure applied to the Chedoke McMaster Arm and Hand Activity Inventory (CAHAI). BMC Med. Res. Methodol. 2010;10:106.

[19] National Population and Talent Division. Population in Brief 2016. Singapore: National Population and Talent Division, Prime Minister's Office; 2016.

[20] Singapore Department of Statistics. General Household Survey 2015 [Internet]. Singapore: Department of Statistics, Ministry of Trade & Industry, Republic of Singapore; 2016 [cited 2017 Jun 13]. Available from: http://www.singstat.gov.sg/publications/publications-and-papers/GHS/ghs2015content.

[21] Wong YM, Teo Z. The elderly in Singapore. Stat. Singap. Newsl. [Internet]. 2011 Sep [cited 2017 Jun 13]; Available from: http://www.singstat.gov.sg/publications/statistics-singapore-newsletter.

[22] Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine. 2000;25:3186–3191.

[23] Chedoke Arm and Hand Activity Inventory (CAHAI). Chedoke Arm and Hand Activity Inventory administration guidelines version 2 [Internet]. [cited 2015 Feb 27]. Available from: www.cahai.ca.

[24] Fugl-Meyer AR, Jääskö L, Leyman I, et al. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scand. J. Rehabil. Med. 1975;7:13–31.

[25] Lin J-H, Hsu M-J, Sheu C-F, et al. Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. Phys. Ther. 2009;89:840–850.

[26] Sullivan KJ, Tilson JK, Cen SY, et al. Fugl-Meyer assessment of sensorimotor function after stroke: standardized training procedure for clinical practice and clinical trials. Stroke J. Cereb. Circ. 2011;42:427–432.

[27] Sanford J, Moreland J, Swanson LR, et al. Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke. Phys. Ther. 1993;73:447–454.

[28] Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int. J. Rehabil. Res. 1981;4:483–492.

[29] Nijland R, van Wegen E, Verbunt J, et al. A comparison of two validated tests for upper limb function after stroke: The Wolf Motor Function Test and the Action Research Arm Test. J. Rehabil. Med. 2010;42:694–696.

[30] Hsueh I-P, Lee M-M, Hsieh C-L. The Action Research Arm Test: is it necessary for patients being tested to sit at a standardized table? Clin. Rehabil. 2002;16:382–388.

[31] Yozbatiran N, Der-Yeghiaian L, Cramer SC. A standardized approach to performing the action research arm test. Neurorehabil. Neural Repair. 2008;22:78–90.

[32] Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control. Clin. Trials. 1981;2:93–113.

[33] Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat. Med. 1998;17:101–110.

[34] Fleiss JL. Balanced incomplete block designs for inter-rater reliability studies. Appl. Psychol. Meas. 1981;5:105–112.

[35] StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP; 2015.

[36] Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 2007;60:34–42.

[37] Streiner DL, Norman GR. Health measurement scales : a practical guide to their development and use. 4th ed. New York: Oxford University Press; 2008.

[38] Cohen J, Cohen P, West SG, et al. Applied multiple regression/correlation analysis for the behavioral sciences. 2nd ed. Mahwah: Lawrence Erlbaum Associates; 2002.

[39] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 1979;86:420–428.

[40] Stratford PW. Getting more from the literature: estimating the standard error of measurement from reliability studies. Physiother. Can. 2004;56:027.

[41] Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Phys. Ther. 1997;77:745–750.

[42] International Organization for Standardization (ISO). Writing paper and certain classes of printed matter - Trimmed sizes - A and B series, and indication of machine direction. Switzerland: ISO copyright office; 2007. Report No.: ISO 216:2007. .

[43] Singapore Department of Statistics. Census of population 2010 statistical release 1: demographic characteristics, education, language and religion. Singapore: Department of Statistics, Ministry of Trade & Industry, Republic of Singapore; 2011.

[44] Geisinger KF. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychol. Assess. 1994;6:304–312.

[45] Stewart AL, Nápoles-Springer A. Health-related quality-of-life assessments in diverse population groups in the United States. Med. Care. 2000;38:II102-124.

[46] Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. Qual. Life Res. Int. J. Qual. Life Asp. Treat. Care Rehabil. 1998;7:323–335.

[47] Flaherty JA, Gaviria FM, Pathak D, et al. Developing instruments for cross-cultural psychiatric research. J. Nerv. Ment. Dis. 1988;176:257–263.

[48] Moreland J, Gowland C, Van Hullenaar S, et al. Theoretical basis of the Chedoke-McMaster Stroke Assessment. Physiother. Can. 1993;45:231–238.

[49] Carroll D. A quatitative test of upper extremity function. J. Chronic Dis. 1965;18:479–491.

[50] Portney L, Watkins M. Foundations of clinical research: applications to practice. 3rd ed. New Jersey, USA: Pearson Prentice Hall; 2009.

Tables

Table 1. Eight-step procedure for Translation and Cross-Cultural Adaptation of Objectively-Assessed Outcome measures [18, p. 108].

Step	Aim	Required personnel
1	To produce two independent forward translations and make necessary region-specific adaptations of the test manual including task and material descriptions, and scoring instructions into the target language.	The two informed translators are native speakers of the target language and profession. Translators are aware of the study objectives.
2	To merge the two forward translations from step 1 to form only one translation. To check comprehensiveness by a therapist/person of the target profession for consistency and adequate vocabulary.	The synthesis is done by another independent and informed person of the target profession or the project manager.
3	To review layout, grammar, and typography. This can be very time consuming but it is important to provide an error-free, professional document for all following steps.	This check is done by the project manager or another person not involved in the translation process but with expertise in the target profession.
4	To backward translate the merged version by an informed person to assure detection of inconsistencies or conceptual errors, and discrepancies.	The back translator should be bilingual or a native speaker of the source language and should have not seen the original before.
5	To review all translations, including all photo or video material showing the necessary test-specific material (e.g. wooden cubes, cups, clothes or zippers). The review should verify a consistent translation and adaptation process. If the translation process fails, a second forward and backward translation is recommended.	This review of all translations and created documents should be done by the original authors including the material that will be used.

Table 1 (continued).

Step	Aim	Required personnel
6	To adapt and re-check the merged forward translation based on the review comments and for grammatical, typographical or other errors, in particular, for consistency in the task and scoring descriptions, and client instructions (quality control step).	The project manager or one of the forward translators could do this check.
7	To pre-test the translated version with 2 to 4 patients including the comprehensiveness of the test manual, and the task and scoring descriptions. Emerging discrepancies of scoring or interpretation of results shall be discussed. Based on severity of required adaptations go back to steps 5 or 6.	Two professionals should test the pre-final version with patients.
8	To evaluate the quality factors of the trans-adapted OAO in patient studies. Depending on the OAO types of validity, reliability and responsiveness have to be determined.	This is the most time and human resources consuming part involving: patients, health professional of different disciplines, a project manager, assistants, and a statistician.

Table 2. Refinement of standard instructions read to patients (in English).

Item	Original CAHAI	CAHAI-SG	Reason for change
Do up five buttons	Do up five buttons using both of your hands, starting at the top.	Button up all five buttons using both of your hands, starting at the top.	'Do up' was not a commonly used term in Singapore
Dry back with towel	Dry your entire back with the towel using both of your hands.	Dry your back completely with the towel using both of your hands.	The phrasing of sentence not easily understood by Singaporeans
Put toothpaste on toothbrush	Put the toothpaste on the toothbrush using both of your hands.	Put some toothpaste on the toothbrush using both of your hands.	To avoid potential misunderstanding of placing the entire tube of toothpaste on the toothbrush

Table 3. Participant characteristics (*n* = 55).

Variables	Sample (%)
Total *n*	55
Gender	
Male	27 (49.1)
Female	28 (50.9)
Age, in years	
Mean (SD)	62.6 (13.2)
Min, Max	23, 83
Ethnicity	
Chinese	45 (81.8)
Malay	6 (10.9)
Indian	3 (5.5)
Others	1 (1.8)
Days since stroke	
Mean (SD)	11.1 (14.3)
Min, Max	0, 94
Type of stroke	
Ischemic	55 (100%)
Affected upper limb	
Right	24 (43.6)
Left	31 (56.4)

SD, standard deviation; min, minimum; max, maximum

Table 4. Distribution of scores on the outcome measures (n = 55).

Measure	Mean (SD)	Min, max	Median (25^{th}, 75^{th} percentile)	Floor / ceiling effect (%)
FMA-UE				
Motor	53.3 (18.2)	0, 66	62 (49, 65)	1.8 / 21.8
Sensation	11.4 (1.3)	6, 12	12 (11, 12)	0 / 74.5
PROM	23.8 (0.7)	20, 24	24 (24, 24)	0 / 85.5
Pain	23.7 (0.9)	20, 24	24 (24, 24)	0 / 85.5
Total	112.2 (19.0)	60, 126	122 (107, 125)	0 / 18.1
CAHAI-SG				
Total	65.4 (28.1)	13, 91	78 (45, 87)	12.7 / 20
ARAT				
Grasp	13.9 (6.9)	0, 18	18 (12, 18)	16.4 / 67.3
Grip	8.8 (4.6)	0, 12	11 (7, 12)	18.1 / 49.1
Pinch	13.0 (7.2)	0, 18	18 (8, 18)	20 / 56.4
Gross motor	7.5 (2.8)	0, 9	9 (6, 9)	7.3 / 72.7
Total	43.1 (20.8)	0, 57	55 (36, 57)	7.3 / 41.8

SD, standard deviation; FMA-UE, Fugl-Meyer Assessment of Upper Extremity; PROM, passive range of motion; CAHAI-SG, Singapore version of the Chedoke Arm and Hand Activity Inventory; ARAT, Action Research Arm Test

Chapter Three:

Reliability and validity of the shortened Singapore versions of the Chedoke Arm and Hand Activity Inventory

Preface

This chapter describes the evaluation of the psychometric properties of three shortened Singapore versions of the Chedoke Arm and Hand Activity Inventory. The inter-rater reliability and the construct validity of the all three shortened versions were estimated by conducting further analyses of the data from the validation of the full 13-item version (Chapter 2). The validation of the shortened versions supports the use of the measure in clinical practice where time constraints are a barrier to the routine use of outcome measures.

This manuscript was published in the *International Journal of Rehabilitation Research* and the copyright holder of this published scholarly work is Wolters Kluwer. The electronic version of the published work (Accepted Manuscript) is presented in this thesis chapter with the following acknowledgement:

This is a non-final version of an article published in final form in:

> Choo, S. X., Bosch, J., Richardson, J., Stratford, P., & Harris, J. E. (2018). Reliability and validity of the shortened Singapore versions of the Chedoke Arm and Hand Activity Inventory. *International Journal of Rehabilitation Research, 41*(4), 297. https://doi.org/10.1097/MRR.0000000000000318

Title Page

Reliability and validity of the shortened Singapore versions of the Chedoke Arm and Hand Activity Inventory

Silvana X. Choo[a,b*], Jackie Bosch[a], Julie Richardson[a], Paul Stratford[a], and Jocelyn E. Harris[a]

[a]*School of Rehabilitation Science, McMaster University, Hamilton, Ontario, Canada;*
[b]*Department of Occupational Therapy, Singapore General Hospital, Singapore*

Short title: Reliability and validity of CAHAI-SG

Corresponding author

Silvana X. Choo

School of Rehabilitation Science, McMaster University, 1400 Main Street West, Institute of Applied Health Sciences, Room 308, Hamilton, Ontario L8S 2R8, Canada. Tel: +01-905-525-9140 ext. 26410. E-mail: silvana.choo.xinyi@sgh.com.sg.

Acknowledgements

We thank the Department of Occupational Therapy, Singapore General Hospital, for assisting with participant screening and assessments, and all participants for volunteering their time. S.X.C. is supported by the Singapore General Hospital Scholarship Award.

Conflicts of interest

There are no conflicts of interest.

Funding

None declared.

Abstract

Background: Upper limb deficits are common sequelae after a stroke and negatively impact daily living and quality of life. The use of outcome measures to evaluate upper limb function is essential to assess sensorimotor recovery and to determine the effectiveness of rehabilitation.

Objective: To estimate the construct validity and inter-rater reliability of three shortened versions of the Singapore version of the Chedoke Arm and Hand Activity Inventory (CAHAI-SG) comprising 7, 8, and 9 test items.

Methods: The sample consisted of 55 inpatients with acute/ subacute stroke to whom the CAHAI-SG, Fugl-Meyer Assessment of Upper Extremity (FMA-UE) and the Action Research Arm Test (ARAT) were administered. To estimate convergent and discriminative construct validity, Spearman's rank correlation coefficient and 95% confidence intervals were computed for CAHAI-SG scores with FMA-UE and ARAT scores. Reliability was estimated using intra-class correlation coefficient (ICC) (relative reliability) and the standard error of measurement (SEM) (absolute reliability).

Results: Convergent validity with the FMA-UE was 0.79, 0.80, and 0.81 for 7-, 8-, and 9-item versions of the CAHAI-SG respectively, and 0.81 with the ARAT for all shortened versions. Discriminative validity with the FMA-UE pain subscale was between 0.37 and 0.38. The absolute reliability was 3.09, 3.65, and 3.98, and relative reliability was 0.96, 0.95, and 0.96 for the 7-, 8-, and 9-item versions respectively.

Conclusion: All shortened versions of the CAHAI-SG demonstrated similar psychometric properties to the full (13 item) version, meaning clinicians may use these shorter versions that require less time to administer and score.

Keywords: upper limb; outcome measure; measurement; stroke; psychometric evaluation

Manuscript Text

Reliability and validity of the shortened Singapore versions of the Chedoke Arm and Hand Activity Inventory

Introduction

Stroke is one of the leading causes of disability worldwide and is among the top three primary causes of disability among older adults in Singapore (Epidemiology & Disease Control Division, 2014; Feigin *et al.*, 2017). Stroke prevalence in Singapore is estimated at 7.6% (Teh *et al.*, 2018), with a disease burden of approximately 804 disability-adjusted life-years per 100,000 individuals (Venketasubramanian *et al.*, 2017). One of the most common deficit following stroke is upper extremity (UE) impairments (Lawrence *et al.*, 2001). More than 80% of persons with stroke experience hemiparesis (Nakayama *et al.*, 1994), and among them, only 5% to 34% regain full functional use of the UE (Wade, 1989; Nakayama *et al*., 1994; Kwakkel *et al*., 2002; Nijland *et al*., 2010b). UE impairments are associated with decreased potential to achieve independence in daily activities, restricted social participation, and reduced quality of life (Wolfe, 2000; Nichols-Larsen *et al.*, 2005). Measurement of UE function is thus critical in rehabilitation to enable effective care – the evaluation process determines the extent of UE impairments and activity limitations, and clinical decisions can be made to provide appropriate interventions according to the severity of the UE deficits (Lang *et al.*, 2013; Duruoz, 2014).

There are currently several stroke-specific outcome measures available to assess UE function (Lang *et al.*, 2013; Murphy *et al.*, 2015). The Fugl-Meyer Assessment of Upper Extremity (FMA-UE) (Fugl-Meyer *et al.*, 1975) and the Action Research Arm Test (ARAT) (Lyle, 1981) are widely used in clinical practice (van Wijck *et al.*, 2001). The FMA-UE is often used to describe and evaluate post-stroke UE impairments and requires about 20 minutes to administer (Velstra *et al.*, 2011; Murphy *et al.*, 2015). In contrast, the ARAT evaluates the manual ability (i.e., activity level) of the affected UE after stroke and the administration time is approximately 10 minutes (Velstra *et al.*, 2011; Murphy *et al.*, 2015). Both the FMA-UE and the ARAT are also recommended performance-based outcome measures for intervention trials targeted at sensorimotor recovery after stroke (Kwakkel *et al.*, 2017). The Chedoke Arm and Hand Activity Inventory (CAHAI) is another measure among the recommended performance-based outcome measures that has demonstrated high levels of measurement quality and clinical utility (Murphy *et al.*, 2015).

The CAHAI consists of 13 test items that use real-life daily tasks to evaluate function in the affected UE after a stroke (Barreca *et al.*, 2004). There are four versions of the measure: the original 13-item version and three shortened versions (7, 8, and 9 items) (Barreca *et al.*, 2006). All four versions of the CAHAI have strong psychometric properties: test-retest reliability of intraclass-correlation coefficient (ICC) ranging from 0.96 to 0.98 (Barreca *et al.*, 2005, 2006); inter-rater reliability of ICC = 0.98 (Barreca *et al.*, 2005); convergent cross-sectional validity of correlations ranging from 0.81 to 0.87 with the Chedoke-McMaster Stroke Assessment (arm and hand components) and 0.93 to

0.95 with the ARAT(Barreca *et al.*, 2005); and longitudinal validity with area under the receiver operating characteristics curve of 0.93 to 0.97 (Barreca *et al.*, 2005).

Aside from strong psychometric properties, the CAHAI also possesses several merits that support its use in clinical practice. Firstly, all items in the CAHAI use real-life daily tasks to evaluate UE function. The use of daily tasks and objects facilitates a more accurate evaluation of UE function as UE movements are influenced by the meaning objects and tasks have for the individual (Wu *et al.*, 1998; Trombly and Wu, 1999). The CAHAI's task-related evaluation approach is also beneficial for persons with post-stroke cognitive or communication deficits, where the use of familiar everyday tasks is more intuitive and demand less advanced cognitive or communication skills (Barreca *et al.*, 2004). Secondly, the administration manual of the CAHAI is available at no charge at its website (www.cahai.com) and the materials required are inexpensive daily objects. The low cost of the CAHAI is advantageous as the excessive cost of an outcome measure is a barrier to its use in clinical practice (Duncan and Murray, 2012). Thirdly, the availability of shortened versions of the CAHAI addresses clinicians' time concerns to administer an outcome measure (Barreca *et al.*, 2006). The full 13-item version takes approximately 25 minutes to administer, while the shortened 7-item version requires half the time (approximately 12 minutes) (Barreca *et al.*, 2006). Short administration time are beneficial as a lengthy time to complete an outcome measure is recognized as a barrier to its use in clinical practice (Duncan and Murray, 2012).

Considering the high level of measurement quality and clinical utility, the CAHAI was recently cross-culturally adapted for the Asian stroke population (Choo *et al.*, 2018).

The Singapore version of the CAHAI (CAHAI-SG) was developed using a systematic cross-cultural adaptation approach and the psychometric properties of the full (13-item) version demonstrated good inter-rater reliability and construct validity in an acute and subacute stroke sample (Choo *et al.*, 2018). While the psychometric properties of the full version of the CAHAI-SG was comparable to the original CAHAI (Choo *et al.*, 2018), the reliability and validity of its shortened versions are currently unknown. There is a need to examine the psychometric properties of the shortened versions as their brief administration time is a facilitator to support the routine use of the CAHAI-SG in clinical practice. The primary aim of this study was to estimate the construct validity and inter-rater reliability of three shortened versions of the CAHAI-SG in an acute and subacute stroke sample in Singapore. The secondary aim was to determine if there was a difference in the validity and reliability between the Singapore and original shortened versions.

Methods

Ethics

Ethics approval was obtained from the Singhealth Centralized Institutional Review Board (CIRB 2015/2915) and the Hamilton Integrated Research Ethics Board (HiREB 0758) prior to administration of any study procedures. Consent was obtained for all participants.

Study design, participants, and raters

Data collected from the validation study of the 13-item CAHAI-SG were used for analysis (Choo *et al.*, 2018). In summary, 55 participants were recruited from the

inpatient neurology and rehabilitation wards at a tertiary hospital in Singapore. The inclusion criteria were: age ≥ 21 years, diagnosis of unilateral hemispheric stroke, acute or subacute phase of stroke (< 12 months), able to sit upright in a chair for at least 30 minutes, and able to follow simple two-step instructions. Individuals with unstable medical conditions, pre-existing UE impairments (prior to the stroke), or severe visual impairments were excluded.

Each participant was assessed once on the CAHAI-SG, the FMA-UE, and the ARAT to estimate construct validity. The administration of the CAHAI-SG on all participants was video recorded and 16 video recordings were randomly selected to estimate inter-rater reliability. Eight registered occupational therapists participated as raters in the evaluation of the construct validity and inter-rater reliability of the CAHAI-SG. Their mean years of clinical experience was 3.5 years (standard deviation, *SD* = 1.5 years). All raters were trained to administer and score all outcome measures through two training workshops.

Measures

Singapore version of the Chedoke Arm and Hand Activity Inventory

The CAHAI-SG uses real-life daily tasks to evaluate function of the affected UE after a stroke (Choo *et al.*, 2018). Task performance on each item is rated on a 7-point scale, from total assistance (score = 1) to independence (score = 7). A total score is obtained by summing scores on each task, with higher scores indicating better UE function. Similar to the original CAHAI, there are three shortened versions of the CAHAI-SG with seven,

eight, and nine items (CAHAI-SG-7, CAHAI-SG-8, and CAHAI-SG-9 respectively). The items included in each version of the CAHAI-SG can be found in Appendix 1. The range of possible scores on the CAHAI-SG-7, CAHAI-SG-8, and CAHAI-SG-9 are 7 to 49, 8 to 56, and 9 to 63 respectively. The full version of the CAHAI-SG demonstrated good inter-rater reliability (ICC = 0.97, 95% CI: 0.94, 0.99), convergent cross-section validity (FMA-UE, rs = 0.87, 95% CI: 0.76, 0.92; ARAT, rs = 0.80, 95% CI: 0.63, 0.90), and discriminative cross-sectional validity (FMA-UE pain subscale, rs = 0.42, 95% CI: 0.22, 0.59) (Choo *et al.*, 2018).

Fugl-Meyer Assessment of Upper Extremity

The FMA-UE is a measure of post-stroke UE impairment (Fugl-Meyer *et al.*, 1975). It consists of four subscales (motor function, sensation, passive range of motion and joint pain) and items are scored on a 3-point scale. The maximum subtotal scores for the motor function, sensation, passive range of motion, and joint pain subscales are 66, 12, 24, and 24 respectively. The range of total scores on the FMA-UE is 0 to 126, with higher scores indicating lesser impairments. The inter-rater reliability of the FMA-UE in acute and subacute stroke samples ranged from ICC = 0.61 to 0.99 (Sanford *et al.*, 1993; Lin *et al.*, 2009; Sullivan *et al.*, 2011).

Action Research Arm Test

The ARAT evaluates the impairments and activity limitations (e.g. pinch, grasp, and reach) of the affected UE post-stroke (Lyle, 1981). There are 19 test items and each item is scored on a 4-point scale. The validation study of the 13-item CAHAI-SG (Choo *et al.*,

2018) followed the scoring procedures described by Yozbatiran *et al.* (2008), where time limits were defined for each score category (e.g., < 5s for a score of 3) instead of qualitative descriptors (e.g., takes abnormally long) in the original scoring. The range of scores on the ARAT is 0 to 57, with higher scores indicating better UE function. Its inter-rater reliability in acute and subacute stroke samples were ICC = 0.92 and 0.99 respectively (Hsueh *et al.*, 2002; Nijland *et al.*, 2010a).

Procedure

Participants' scores on all outcome measures were used in this study. Total scores for each shortened versions of the CAHAI-SG were calculated by summing the first seven, eight and nine items. For the comparison measures, the FMA-UE total, motor sub-scale and pain sub-scale scores, and the total ARAT scores were used.

Data analysis

All statistical analysis was performed using STATA Version 14 (StataCorp, College Station, Texas, USA). Participant characteristics and scores on all outcome measures were summarized as medians with 1st and 3rd quartiles. The following analyses were computed for each shortened version of the CAHAI-SG.

Construct validity

Spearman's rank correlation coefficients with 95% CIs were used to estimate construct validity. 95% CIs were used instead of *p*-values as the primary aim of the study addressed parameter estimation rather than hypothesis testing. For convergent cross-sectional

validity, we examined the extent to which the shortened versions of the CAHAI-SG were closely related to similar measures of UE function (Streiner and Norman, 2008). Thus, the correlation between CAHAI-SG total scores and FMA-UE (total and motor sub-scale) and ARAT (total) scores were computed. Discriminant cross-sectional validity was estimated with the correlation between CAHAI-SG total scores and FMA-UE joint pain subscale scores. We expected lower correlations between the shortened versions of the CAHAI-SG and the FMA-UE joint pain subscale as they assessed dissimilar outcomes (Streiner and Norman, 2008).

To determine if the construct validity of the shortened versions of the CAHAI-SG were similar to the shortened versions of the CAHAI as reported by (Barreca *et al.*, 2006), their respective correlation coefficients with the ARAT were compared. Correlation coefficients with the ARAT of all shortened versions (both CAHAI-SG and CAHAI) were converted to a z-score using Fisher's *r* to *z* transformation and head-to-head comparisons (e.g., CAHAI-SG-7 versus CAHAI-7) were performed using the following formula (Cohen *et al.*, 2002):

$$z_{observed} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

A $z_{observed}$ score greater than 1.96 (two-tailed test of significance with alpha set at 0.05) was considered statistically significant.

Inter-rater reliability

Randomized analysis of variance (ANOVA) was computed with the CAHAI-SG score as the dependent variable, and the factors were participants (16 levels) and raters (8 levels). The sources of variances (participants, rater, and error) identified in the ANOVA analysis were then used to estimate relative reliability using Shrout and Fleiss Type 2,1 ICC (Shrout and Fleiss, 1979) with 95% CI. The SEM, which was calculated by taking the square root of the mean-square error, and its associated 95% CI was used to estimate absolute reliability (Stratford and Goldsmith, 1997; Stratford, 2004).

To determine if the inter-rater reliability of the shortened versions of the CAHAI-SG were similar to the shortened versions of the CAHAI, a head-to-head comparison of their respective SEMs were conducted using an *F*-test. A *p*-value of < 0.05 was used to indicate a significant difference between the absolute reliability of the CAHAI-SG and the CAHAI.

Results

Participant characteristics

Participant characteristics are summarized in Table 1. The mean age was 63 years (*SD* = 13 years) and 51% (n = 28) were female. All participants had an ischemic stroke and the median days post-stroke was 7 days (1^{st}, 3^{rd} quartile: 4, 14 days). The median total scores on the CAHAI-SG-7, CAHAI-SG-8, and CAHAI-SG-9 were 46, 53 and 60 respectively. Table 2 summarizes the scores on all three outcome measures.

Construct validity

Estimation of construct validity

The construct validity of the CAHAI-SG shortened versions are presented in Table 3. For convergent cross-sectional validity, the correlation between the CAHAI-SG-7, CAHAI-SG-8, and CAHAI-SG-9 scores with the FMA-UE total scores and motor subscale scores were 0.79, 0.8, and 0.8 respectively, and the correlation with the ARAT was 0.81 for all three measures. In contrast, the discriminative cross-sectional validity of the CAHAI-SG-7, CAHAI-SG-8, and CAHAI-SG-9 were 0.37, 0.38, and 0.37 respectively. The correlations between the CAHAI-SG shortened versions and the FMA-UE joint pain sub-scale were low as anticipated since dissimilar outcomes were assessed.

Comparison between CAHAI-SG and CAHAI

The correlation coefficients for all shortened versions of the CAHAI-SG with the ARAT were smaller than those of the CAHAI (0.81 versus 0.95, 0.95, 0.94 for the 7-, 8-, 9-item versions respectively). The differences between the correlation coefficients of the Singapore and original shortened versions were statistically significant (CAHAI-SG-7: $z = 3.25$, $p = 0.001$, CAHAI-SG-8: $z = 3.25$, $p = 0.001$, CAHAI-SG-9: $z = 2.82$, $p = 0.005$).

Inter-rater reliability

Estimation of inter-rater reliability

Table 3 summarizes the inter-rater reliability of all three shortened versions of the CAHAI-SG. The absolute reliability of the CAHAI-SG-7, CAHAI-SG-8, and CAHAI-SG-9 was SEM = 3.09, 3.65, and 3.98 CAHAI points respectively; their relative

reliability coefficients were ICC = 0.96, 0.95, and 0.96 respectively.

Comparison between CAHAI-SG and CAHAI

The SEMs of all shortened versions of the CAHAI-SG (3.09, 3.65, and 3.98 for the 7-, 8-, 9-item versions respectively) were larger than those of the CAHAI (2.32, 2.26, and 2.57 for the 7-, 8-, 9-item versions respectively). However, only the SEMs of the CAHAI-SG-8 ($F(15, 38) = 2.59$, $p = 0.009$) and CAHAI-SG-9 ($F(15, 38) = 2.40$, $p = 0.015$) were significantly larger.

Discussion

This study examined the construct validity and inter-rater reliability of the three shortened versions of the CAHAI-SG in an acute and subacute stroke sample in Singapore. Overall, all three shortened versions demonstrated similar validity and reliability as the full version of the CAHAI-SG, and the study findings provide psychometric evidence to support their clinical use. The three shortened versions of the CAHAI-SG also had similar validity and relative reliability, while the CAHAI-SG-7 demonstrated better absolute reliability (lowest SEM). The SEM, which is expressed in the same unit as the CAHAI-SG scores, quantifies the precision of scores on the CAHAI-SG (Streiner and Norman, 2008). Thus, a lower SEM value is desired as it indicates a smaller error in the observed score of a measure.

Comparing the psychometric properties of the shortened versions of the CAHAI-SG with the shortened versions of the CAHAI, the CAHAI-SG versions demonstrated weaker construct validity and inter-rater reliability. However, this may be due to the

study procedures employed in the validation study of the CAHAI-SG by Choo *et al.* (2018). For cross-sectional convergent validity, the correlations between the CAHAI-SG versions and ARAT total scores were significantly lower than the CAHAI versions. These significant differences may be attributed to the different scoring procedures that the ARAT employed. The validation study of the shortened versions of the CAHAI by Barreca *et al.* (2006) followed the original scoring procedures of the ARAT (Carroll, 1965), and thus, the speed of task completion were evaluated similarly using subjective judgment (i.e., in a timely manner) in both the CAHAI and the ARAT. In contrast, this study followed scoring procedures of the ARAT described by Yozbatiran *et al.* (2008), where cut-off times were defined on the 4-point scoring scale. Although cut-off times improved the distinction between the score categories on the ARAT scoring scale, it introduced the objective evaluation of speed. The difference in how speed was evaluated in all versions of the CAHAI-SG (subjectively) and the ARAT (objectively) may explain the lower correlation between the CAHAI-SG shortened versions and the ARAT, as compared to the CAHAI shortened versions.

The relative reliability of the CAHAI-SG shortened versions (ICC = 0.95 to 0.96) were similar to the CAHAI shorted versions (ICC = 0.98). However, the SEMs of the CAHAI-SG-8 and CAHAI-SG-9 were significantly larger than the SEMs of the CAHAI-8 and CAHAI-9. The weaker absolute reliability of the CAHAI-SG shortened versions may be due to the difference in the training of the raters. In the validation study of the shortened versions of the CAHAI (Barreca *et al.*, 2006), the raters were trained to 85% accuracy in the administration and scoring of the CAHAI. In comparison, although raters

in this study were trained in the administration and scoring of the CAHAI-SG, the raters were not trained to criterion. This may have contributed to a greater error variance associated with raters in the Singapore version, and therefore, resulted in the weaker absolute reliability observed on the CAHAI-SG shortened versions. Thus, it is important to consider including a formal evaluation component in future training workshops on the administration and scoring of the CAHAI-SG in clinical practice.

Study limitations

To estimate inter-rater reliability, video analysis was used with the intention of minimizing participant burden due to a large number of raters in this study. However, the use of video recordings differs from clinical practice where scoring is based on the assessor's hands-on administration of the CAHAI-SG. The observation of the exact same performance on the CAHAI-SG through video recordings may have produced a lower rater variance, and consequently increased the ICC values. Another study limitation is the consistent order in which the outcome measures were administered for all participants. This may have resulted in an order effect (Portney and Watkins, 2009), where scores on the ARAT (administered last) may be lower due to participants' fatigue. This could possibly influence the correlation between the CAHAI-SG and ARAT scores. Lastly, as mentioned earlier, all raters in our study were trained to administer and score the outcome measures, but there was no formal evaluation of their competency as we did not train them to criterion. This may have lowered the absolute reliability of the CAHAI-SG shortened versions.

Conclusion

This study provides evidence in support of the validity and reliability of the shortened versions of the CAHAI-SG for adults with acute and subacute stroke in Singapore. With comparable psychometric properties to the full version, the shortened versions of the CAHAI-SG may be suited for use in clinical programs or research studies where administration time of an outcome measure may be a concern. The choice between the three shortened versions is, however, less straightforward due to the marginal differences in their psychometric properties. It is also important to consider the items in each shortened version and select the version that includes test items which assess the targeted UE function of interest.

References

Barreca S, Gowland CK, Stratford P, Huijbregts M, Griffiths J, Torresin W, *et al*. (2004). Development of the Chedoke Arm and Hand Activity Inventory: theoretical constructs, item generation, and selection. *Top Stroke Rehabil* **11**, 31–42.

Barreca SR, Stratford PW, Lambert CL, Masters, LM, Streiner DL (2005). Test-retest reliability, validity, and sensitivity of the Chedoke Arm and Hand Activity Inventory: a new measure of upper-limb function for survivors of stroke. *Arch Phys Med Rehabil* **86**, 1616–1622.

Barreca SR, Stratford PW, Masters LM, Lambert CL, Griffiths J, McBay C (2006). Validation of three shortened versions of the Chedoke Arm and Hand Activity Inventory. *Physiotherapy Canada* **58**, 148–156.

Carroll D (1965). A quatitative test of upper extremity function. *J Chronic Dis* **18**, 479–491.

Choo SX, Bosch J, Richardson J, Stratford P, Harris JE (2018). Cross-cultural adaptation and psychometric evaluation of the Singapore version of the Chedoke Arm and Hand Activity. *Disabil Rehabil*. Epub ahead of print. https://www.tandfonline.com/doi/abs/10.1080/09638288.2018.1472817 Accessed 22 May 2018.

Cohen J, Cohen P, West SG, Aiken LS, (2002). Applied multiple regression/correlation analysis for the behavioral sciences, 2nd ed. Lawrence Erlbaum Associates, Mahwah.

Duncan EA and Murray J (2012). The barriers and facilitators to routine outcome measurement by allied health professionals in practice: a systematic review. *BMC Health Serv Res* **12**, 96.

Duruoz MT (2014). Assessment of hand function. In MT Duruoz (ed), *Hand function: a practical guide to assessment*. Springer: New York, pp; 41–54.

Epidemiology & Disease Control Division (2014). Singapore burden of disease study 2010. Ministry of Health, Singapore, Singapore.

Feigin VL, Norrving B, Mensah GA (2017). Global burden of stroke. *Circ Res* **120**, 439–448.

Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S (1975). The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scand J Rehabil Med* **7**, 13–31.

Hsueh IP, Lee MM, Hsieh CL (2002). The Action Research Arm Test: is it necessary for patients being tested to sit at a standardized table? *Clin Rehabil* **16**, 382–388.

Kwakkel G, Kollen BJ, Wagenaar RC (2002). Long term effects of intensity of upper and lower limb training after stroke: a randomised trial. *J Neurol Neurosurg Psychiatry* **72**, 473–479.

Kwakkel G, Lannin NA, Borschmann K, English C, Ali M, Churilov L *et al*. (2017). Standardized measurement of sensorimotor recovery in stroke trials: consensus-based core recommendations from the Stroke Recovery and Rehabilitation Roundtable. *Int J Stroke* **12**, 451–461.

Lang CE, Bland MD, Bailey RR, Schaefer SY, Birkenmeier RL (2013). Assessment of upper extremity impairment, function, and activity following stroke: foundations for clinical decision making. *J Hand Ther* **26**, 104–115.

Lawrence ES, Coshall C, Dundas R, Stewart J, Rudd AG, Howard R, Wolfe CD (2001). Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. *Stroke* **32**, 1279–1284.

Lin JH, Hsu MJ, Sheu CF, Wu TS, Lin RT, Chen CH, Hsieh, CL (2009). Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. *Phys Ther* **89**, 840–850.

Lyle RC (1981). A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *Int J Rehabil Res* **4**, 483–492.

Murphy MA, Resteghini C, Feys P, Lamers I (2015). An overview of systematic reviews on upper extremity outcome measures after stroke. *BMC Neurol* **15**, 292.

Nakayama H, Jørgensen HS, Raaschou HO, Olsen TS (1994). Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. *Arch Phys Med Rehabil* **75**, 394–398.

Nichols-Larsen DS, Clark PC, Zeringue A, Greenspan A, Blanton S (2005). Factors influencing stroke survivors' quality of life during subacute recovery. *Stroke* **36**, 1480–1484.

Nijland R, van Wegen E, Verbunt J, van Wijk R, van Kordelaar J, Kwakkel G (2010a). A comparison of two validated tests for upper limb function after stroke: the Wolf

Motor Function Test and the Action Research Arm Test. *J Rehabil Med* **42**, 694–696.

Nijland R, van Wegen E, Harmeling-van der Wel, B, Kwakkel G, EPOS Investigators (2010b). Presence of finger extension and shoulder abduction within 72 hours after stroke predicts functional recovery: early prediction of functional outcome after stroke: the EPOS cohort study. S*troke* **41**, 745–750.

Portney L, Watkins M (2009). *Foundations of clinical research: applications to practice*. 3rd ed. New Jersey, USA: Pearson Prentice Hall.

Sanford J, Moreland J, Swanson LR, Stratford PW, Gowland C (1993). Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke. *Phys Ther* **73**, 447–454.

Shrout PE, Fleiss, JL (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* **86**, 420–428.

Stratford PW (2004). Getting more from the literature: estimating the standard error of measurement from reliability studies. *Physiother Can* **56**, 027.

Stratford PW and Goldsmith CH (1997). Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther* **77**, 745–750.

Streiner DL and Norman GR (2008). *Health measurement scales: a practical guide to their development and use*. 4th ed. Oxford, New York: Oxford University Press.

Sullivan KJ, Tilson JK, Cen SY, Rose DK, Hershberg J, Correa A *et al*. (2011). Fugl-Meyer assessment of sensorimotor function after stroke: standardized training procedure for clinical practice and clinical trials. *Stroke* **42**, 427–432.

Teh WL, Abdin E, Vaingankar JA, Seow E, Sagayadevan V, Shafie S *et al*. (2018). Prevalence of stroke, risk factors, disability and care needs in older adults in Singapore: results from the WiSE study. *BMJ Open* **8**, e020285.

Trombly CA and Wu CY (1999). Effect of rehabilitation tasks on organization of movement after stroke. *Am J Occup Ther* **53**, 333–344.

van Wijck FM, Pandyan AD, Johnson GR, Barnes MP (2001). Assessing motor deficits in neurological rehabilitation: patterns of instrument usage. *Neurorehabil Neural Repair* **15**, 23–30.

Velstra I-M, Ballert CS, Cieza A (2011). A systematic literature review of outcome measures for upper extremity function using the international classification of functioning, disability, and health as reference. *PM R* **3**, 846–860.

Venketasubramanian N, Yoon BW, Pandian J, Navarro JC (2017). Stroke epidemiology in South, East, and South-East Asia: a review. *J Stroke* **19**, 286–294.

Wade DT (1989). Measuring upper limb impairment and disability after stroke. *Int Disabil Stud* **11**, 89–92.

Wolfe CD (2000). The impact of stroke. *Br Med Bull* **56**, 275–286.

Wu C, Trombly CA, Lin K, Tickle-Degnen L (1998). Effects of object affordances on reaching performance in persons with and without cerebrovascular accident. *Am J Occup* Ther **52**, 447–456.

Yozbatiran N, Der-Yeghiaian L, Cramer SC (2008). A standardized approach to performing the action research arm test. *Neurorehabil Neural Repair* **22**, 78–90.

Appendix

Appendix 1

Items included in each Singapore versions of the Chedoke Arm and Hand Activity Inventory (CAHAI-SG)

CAHAI-SG-7

(1) Open jar of peanut butter

(2) Call 995

(3) Draw a line with a ruler

(4) Pour a glass of water

(5) Wring out face towel

(6) Button five buttons

(7) Dry back with towel

CAHAI-SG-8

(8) Put toothpaste on toothbrush

CAHAI-SG-9

(9) Cut medium-resistance putty

CAHAI-SG (full version)

(10) Zip up the zipper

(11) Clean a pair of spectacles

(12) Place box on table

(13) Carry bag up the stairs

Tables

Table 1. Participant characteristics (n = 55).

Variables	Sample (%)
Total *n*	55
Sex	
Female	28 (50.9)
Age, in years	
Mean (SD)	63(13.2)
Min, Max	23, 83
Race	
Chinese	45 (81.8)
Malay	6 (10.9)
Indian	3 (5.5)
Others	1 (1.8)
Days since stroke	
Median (1^{st}, 3^{rd} quartile)	7 (4, 14)
Min, Max	0, 94
Type of stroke	
Ischemic	55 (100%)
Affected upper extremity	
Right	24 (43.6)

SD, standard deviation; min, minimum; max, maximum

Table 2. Summary of scores on all outcome measures (n = 55).

Measure	Mean (SD)	Median (1^{st}, 3^{rd} quartile)	Min, max
CAHAI-SG			
7-item version	37.8 (15.6)	46 (27, 49)	7, 49
8-item version	43.2 (17.8)	53 (32, 56)	8, 56
9-item version	48.2 (20.1)	60 (35, 63)	9, 63
FMA-UE			
Motor	53.3 (18.2)	62 (49, 65)	0, 66
Sensation	11.4 (1.3)	12 (11, 12)	6, 12
PROM	23.8 (0.7)	24 (24, 24)	20, 24
Pain	23.7 (0.9)	24 (24, 24)	20, 24
Total	112.2 (19.0)	122 (105, 125)	60, 126
ARAT			
Total	43.1 (20.8)	55 (36, 57)	0, 57

SD, standard deviation; FMA-UE, Fugl-Meyer Assessment of Upper Extremity; PROM, passive range of motion; CAHAI-SG, Singapore version of the Chedoke Arm and Hand Activity Inventory; ARAT, Action Research Arm Test

Table 3. Construct validity and inter-rater reliability of three shortened versions of the CAHAI-SG.

	Construct validity				Inter-rater Reliability	
	FMA-UE, r_s (95% CI)	FMA-UE motor subscale, r_s (95% CI)	ARAT, r_s (95% CI)	FMA-UE pain subscale, r_s (95% CI)	Relative reliability, ICC (95% CI)	Absolute reliability, SEM (95% CI)
CAHAI-SG-7	0.79	0.79	0.81	0.37	0.96	3.09
	(0.63, 0.89)	(0.63, 0.90)	(0.64, 0.91)	(0.10, 0.57)	(0.91, 0.98)	(2.73, 3.58)
CAHAI-SG-8	0.80	0.80	0.81	0.38	0.95	3.65
	(0.63, 0.89)	(0.64, 0.90)	(0.64, 0.91)	(0.10, 0.58)	(0.90, 0.98)	(3.22, 4.22)
CAHAI-SG-9	0.81	0.81	0.81	0.37	0.96	3.98
	(0.65, 0.90)	(0.65, 0.91)	(0.65, 0.91)	(0.09, 0.57)	(0.91, 0.98)	(3.51, 4.60)
CAHAI-SG*	0.89	0.89	0.80	0.41	0.97	4.80
	(0.80, 0.94)	(0.78, 0.94)	(0.63, 0.90)	(0.20, 0.58)	(0.94, 0.99)	(4.23, 5.55)

CAHAI-SG: Singapore version of the Chedoke Arm and Hand Activity Inventory; FMA-UE: Fugl-Meyer Assessment of Upper Extremity; ARAT: Action Research Arm Test; r_s: Spearman's rank correlation coefficient; CI: confidence interval; ICC: intra-class correlation coefficient; SEM: standard error of measurement.

*Construct validity and inter-rater reliability of the full version of CAHAI-SG are presented for reference.

Chapter Four:

Measurement theories in health care: Introduction to item response theory and Rasch measurement theory

Preface

In this chapter, the original versions of the Chedoke Arm and Hand Activity Inventory are evaluated using two measurement theories, item response theory and Rasch measurement theory. This manuscript describes a brief overview of both measurement theories and the Chedoke Arm and Hand Activity Inventory is used as an example instrument to demonstrate the application of the theories. Data from a previous validation study of the measure was used to conduct the analysis. This manuscript is intended to increase the awareness of item response theory and Rasch measurement theory among rehabilitation professionals and introduce both theories at a conceptual level. Hence, technical aspects of underlying complex mathematical concepts and computations are not described in detail, and results are reported as they apply to clinical practice.

This manuscript was submitted to *PLOS One* on November 27, 2018.

Title Page

Measurement theories in health care: Introduction to item response theory and Rasch measurement theory

Silvana X. Choo[1,2*], Paul Stratford[1], Julie Richardson[1], Jocelyn E. Harris[1], Jackie Bosch[1], and Ayse Kuspinar[1]

[1] School of Rehabilitation Science, McMaster University, Hamilton, Ontario, Canada

[2] Department of Occupational Therapy, Singapore General Hospital, Singapore

*Corresponding author

silvana.choo.xinyi@sgh.com.sg

Abstract

The development and validation of outcome measures are traditionally guided by classical test theory. However, modern test theories, specifically item response theory and Rasch measurement theory, are alternative measurement theories that address the limitations of classical test theory. Modern test theories are increasingly applied in health research and thus, information on these theories are needed to be presented in a manner that can be easily understood by health care professionals. This paper provides an overview of item response theory and Rasch measurement theory and how each theory can be applied to evaluate outcome measures. The basic concepts, models, and assumptions of each theory are introduced. To demonstrate the typical application of the theories, we used data from a previous validation study of the Chedoke Arm and Hand Activity Inventory (CAHAI). The sample consisted of 105 participants with post-stroke upper extremity deficits who were assessed on the CAHAI at admission and at the completion of their rehabilitation program. The working example will enable health care professionals to better understand the use of modern test theories to evaluate outcome measures.

Key words: outcome measures; rehabilitation; psychometrics

Measurement theories in healthcare: Introduction to item response theory and Rasch measurement theory

Introduction

Measurement is highly valued in health research as the quantification of clinical observations enables the demonstration of treatment/program effectiveness, justification for interventions, and facilitates clinical decision-making [1–3]. Measurement, the assignment of numbers to observations according to a set of rules to represent their quantity or magnitude [4,5], can be achieved through the use of outcome measures. To ensure outcome measures are accurate and credible, their construction, scoring, validation, and refinement follow rigorous methods that are traditionally guided by classical test theory. Classical test theory has been the prevalent measurement theory used to guide the development and evaluation of outcome measures [6]. However, the limitations of outcome measures developed using classical test theory are well established [7].

Consider this typical clinical scenario. John and William recently had a stroke and are receiving rehabilitation at an inpatient rehabilitation facility. Their occupational therapist uses a standardized tool, the 7-item version of the Chedoke Arm and Hand Activity Inventory [8] (CAHAI-7) to evaluate the functional abilities of their affected upper extremity. John and William each performed seven functional tasks (e.g., opening a jar of coffee and calling 911) and their performance on each task is scored on a 7-point ordinal scale, from total assistance (score = 1) to independence (score = 7). Scores on each task are summed and both John and William obtained a total score of 28 (Table 1).

Although they have the same total score, can their occupational therapist conclude that John and William have similar upper extremity function?

One objection that might be raised is the demand of upper extremity function for successful task completion differs between tasks. For example, doing up five buttons requires greater hand dexterity than strength compared to pouring a glass of water, which demands greater strength than dexterity. Accordingly, John and Williams' scores on each item suggest that William has better upper extremity function, demonstrating both strength and dexterity albeit needing minimal assistance in all tasks. In contrast, John appears to have only regained strength in his affected upper extremity, requiring maximal assistance in tasks that demand hand dexterity. How did this problem arise despite the CAHAI having established psychometric properties of reliability, validity, and internal consistency [9–11]?

This hypothetical clinical scenario demonstrates one of the several limitations of classical test theory. Although item response theory (IRT) and Rasch measurement theory (RMT), were developed in the 1960s to address the limitations of classical test theory, both theories have only been increasingly applied in the recent decade [12,13]. Notably, the National Institutes of Health funded a 5-year multisite collaborative research (Patient-Reported Outcome Measurement Information System) to standardize patient-reported outcomes guided by IRT [14,15]. With both IRT and RMT gaining momentum and prevalence in health research, accessible information for health care professionals on their underlying complex mathematical concepts and how these theories can be applied is necessary.

This article is intended to provide health care professionals with an overview of modern measurement theories (IRT and RMT) and demonstrate the evaluation of psychometric properties of outcome measures using these theories. First, we present an overview of classical test theory and highlight its limitations. Then, we describe both IRT and RMT and use an example instrument to illustrate how common analyses within each theory is conducted. Last, we compare the results from IRT and RMT analyses and discuss the strengths and challenges of using modern measurement theories. As the goal of this article is to introduce IRT and RMT at a conceptual level to health care professionals, technical details of the underlying complex mathematical concepts and computations are not described. Readers who are interested in the technical details of the theories may refer to the references mentioned in the article.

What is measurement theory?

Measurement is the assignment of numbers to observations or a characteristic of an object to represent their quantity or magnitude [4,5]. Measurement theory uses mathematical model(s) to describe the factors that influence scores obtained from a rating scale (e.g., an outcome measure) and also defines the underlying assumptions of its mathematical concepts [6]. Measurement theories provide a framework for the development, evaluation of quality and refinement of outcome measures to improve the accuracy, relevance, and meaningfulness of measurements [16].

Classical test theory

Classical test theory (CTT) has been the prevalent measurement theory in the development and evaluation of outcome measures in clinical research [6], and concepts

such as reliability and validity are familiar to health care professionals. CTT, also known as true score theory, states that an observed score (O) is composed of a true score (T) and the error associated with the observed score (E), or O = T + E [17]. The theory assumes an individual has a true score that would be obtained with an error-free measurement. However, as measurement tools are not perfect, observed scores for an individual may differ from their true ability as their scores include an error component [18]. Thus, when the development and refinement of outcome measures are guided by CTT, one main objective is to reduce measurement error [19]. CTT also focuses on test-level information only, meaning its mathematical models describe the association between total scores to true scores rather than scores on individual items to the true scores [16]. This implies that all test items in an outcome measure need to be administered or completed to obtain a valid and reliable score [20], and difficulties with score interpretation arise with missing information/data on a single test item.

The summation of item scores to obtain a total score on an outcome measure, like the CAHAI, is common practice. However, for outcome measures developed using CTT, there are two major problems in doing so. First, the total scores are assumed to be continuous (interval) data even though scores on each item are discrete (ordinal) in nature [21]. Yet, interval scaling of an outcome measure is not evaluated in CTT [22]. Using the analogy of a ruler, a ruler's measurement scale is interval, where the distance between each interval is equally spaced throughout the ruler. Thus, to measure the length of an object, we can simply add the intervals and accordingly, mathematical calculations of length are justifiable. For example, 7cm – 2cm = 5cm, or a length of 4cm is twice that of

2cm. In the CAHAI, consecutive numbers (1 to 7) were assigned to the qualitative descriptions of total assistance, maximal assistance, moderate assistance, minimal assistance, supervision, modified independence, and independence respectively [8]. This assignment of numbers to qualitative descriptors says nothing about the distances between the categories in the 7-point scale nor that all items in the CAHAI employ the scoring scale consistently [23]. Consequently, it is unknown whether the 'distance' between a score of 1 (total assistance) and 2 (maximal assistance) is the same as the 'distance' between a score of 6 (modified independence) and 7 (independence). Likewise, we also do not know that a score of 4 (minimal assistance) is twice that of a score of 2 (maximal assistance). The second problem with summing item scores is the assumption that all test items contribute equally to the total score [19]. However, the contributions of individual test items are not evaluated within CTT [19].

Two other problems associated with CTT are sample dependency and the assumption of equal standard error of measurement [7,19]. Estimates of reliability and validity of an outcome measure only apply to the specific sample of participants on which the data was collected (i.e., sample dependent) [7]. Thus, a measure's psychometric properties can be reasonably applied to individuals who share similar characteristics as the sample; otherwise, there is a need to re-establish its psychometric properties when individuals have different characteristics from the sample (e.g. different diagnosis, different severity of the condition) [7,19]. The assumption of equal standard error of measurement in CTT is that the error associated with the observed score is consistent along the scale [19]. Consequently, a single estimate of the standard error of

measurement is calculated for an outcome measure. However, if the observed scores follow a normal distribution, where there are more individuals in the middle than the upper and lower score ranges, the standard error of measurement should differ accordingly (i.e., the smallest error in the middle range and increases at the extreme scores) [19]. Thus, multiple estimates of the standard error of measurement may instead be more appropriate.

Modern test theories – Item response theory and Rasch measurement theory

The limitations of classical test theory led to the development of two measurement theories, item response theory (IRT) and Rasch measurement theory (RMT) (we refer to both theories as modern test theories in this article). Modern test theories are a system of mathematical models that describe how the latent trait of individuals and test item properties are predictors of observed scores/responses on test items [24,25]. The subsequent sections describe IRT and RMT. First, three key terms common to both IRT and RMT are defined. Following which, the mathematical models and assumptions in each theory are briefly explained. The underlying mathematical concepts within both theories are not the focus of this article, and thus, readers are recommended to refer to textbooks on IRT and RMT for more details [16,24,26].

Definitions of key terms

Latent trait. Latent trait (θ) refers to the underlying ability or characteristic of a person that is purported to be evaluated by the outcome measure [27]. They are not directly observable but rather, are ascertained based on a person's presentation or task

performance [20]. For example, upper extremity function is the latent trait measured in the CAHAI that is observed through the performance of the affected upper extremity in daily tasks.

Scale properties. The logit scale is analogous to the centimeter/inches units on a ruler, where log odds units are used to measure the latent trait levels [20,28]. It approximates an equal-interval scale and the scale typically ranges from – 3 logits to 3 logits [20]. IRT and RMT transform ordinal scores to the logit scale, and therefore, data from outcome measures evaluated using these theories can provide evidence of the measures interval-level measurement characteristics [28].

Item parameters. The relationship between the latent trait and the outcome measure is expressed as mathematical models in IRT and RMT [20]. In the mathematical models, item parameters are used to explain the relationship between individuals' levels of the latent trait and how they would respond/score on each item in the outcome measure [20]. Two important parameters are the item difficulty and the item discrimination parameters. Item difficulty (*b*) represents the location of a test item on the logit scale where the probability of a response is 0.50 (e.g., 50% probability of responding 'yes' in a dichotomous 'yes/no' scale) or the probability of obtaining a score in two adjacent score categories is equal (e.g., the probability of scoring 2 or 3 on the CAHAI 7-point scale is equal at 50%) [24,29]. Test items with higher logit values can be interpreted as more 'difficult' because individuals require higher levels of the trait to endorse the item or obtain a score [20]. Item discrimination (*a*) represents the degree to which a test item is able to discriminate among individuals with varying latent trait levels [29]. In a test item

that has a better discriminating ability (i.e., higher item discrimination value), individuals with different amounts of the latent trait will respond differently to the test item [20].

Item response theory

Item response theory uses a system of mathematical models to explain the relationship between the test item, the latent trait, and the probability of individuals' responses on the item [20,24]. IRT models can be broadly categorized based on the scoring/response options of the outcome measure – dichotomous (e.g., yes/no, and able/unable) or polytomous (i.e., > 2 response options). The following paragraphs provide a brief description of unidimensional (i.e., measuring a single latent trait) IRT models.

Dichotomous models

For outcome measures with dichotomous scoring/response options, there are three IRT models: one-, two-, and three-parameter logistic models [29]. The one-parameter logistic model is the simplest model, where item difficulty parameters are estimated for each test item in an outcome measure. The item discrimination parameter is held constant and does not vary across test items in the one-parameter logistic model [24]. Two-parameter logistic models estimate both the item difficulty and item discrimination parameters for each test item. The item discrimination parameter is no longer held constant and is allowed to vary across test items in the two-parameter logistic model [24]. The three-parameter logistic model not only estimates both the item difficulty and item discrimination parameters for each test item, but also a third parameter, the pseudo-

guessing parameter [24]. The pseudo-guessing parameter accounts for circumstances in which guessing is a factor in the performance on an outcome measure [29]. In health care, this pseudo-guessing parameter may be relevant in educational situations, such as guessing the right answer on a competency test. All three logistic models form a series of hierarchy models such that the simpler models are nested within the more complex models, meaning the three-parameter model can be mathematically simplified to the two-parameter model, and the two-parameter model can be simplified to the one-parameter model [24].

Polytomous models

Polytomous models can be categorized into two groups, based on whether the scoring/response options are nominal (e.g., categorical responses of 'yes', 'no', or 'maybe') or ordinal (e.g., ordered response on a Likert scale). The nominal response model and the multiple-choice model can be applied to outcome measures with nominal polytomous response options [24]. For ordered polytomous response/scoring options, the polytomous models are extensions of the dichotomous models. The partial credit model [30] and the rating scale model [31] are extensions of the one-parameter logistic polytomous models. For two-parameter logistic polytomous models, there are the generalized partial credit model [32] and the graded response model [33]. Polytomous models examine the probability of obtaining a particular score category (or response to an option). However, the approach in how these score/response probabilities are regarded differ between polytomous models. For example, the generalized partial credit model can provide the probabilities of obtaining a particular score category on the CAHAI's 7-point

scale for a test item [24]. In contrast, the graded response model provides the cumulative probabilities of obtaining various subsets of score categories [24], such as the probability of scoring 2 or higher in a test item. The detailed discussion of the differences between polytomous models is beyond the scope of this article.

Assumptions in IRT

There are two key assumptions in IRT, unidimensionality and local independence [29]. Unidimensionality means that all test items in an outcome measure evaluate a single latent trait [29]. However, this assumption may not always be fully met for outcome measures used in clinical research as the outcome (latent trait) may be comprised of several factors. For instance, cognitive outcome measures may include test items that evaluate different cognitive functions, such as attention, orientation, and memory. In such outcome measures, the unidimensionality assumption can be sufficiently met if it is demonstrated the test items measure a single dominant factor [29]. Local independence means that scores/responses on one test item are independent of the scores/responses to any other test items (assuming the individual's level of the latent trait is held constant) [24,29]. This implies that individuals' scores/responses should be determined only by their performance on each test item and not by how they scored/responded to other test items [24,34].

Rasch measurement theory

Rasch measurement theory, similar to IRT, uses mathematical models (the Rasch model [35] and its extensions) to describe the relationship between the test item, latent

trait, and the probability of individuals' responses on the item [24]. The primary mathematical model in RMT is the Rasch model [35]. The Rasch model for outcome measures with a dichotomous scoring/response options is mathematically equivalent to the one-parameter logistic model in IRT with the exception of the estimation of the item discrimination parameter [24,35]. In the Rasch model, the item discrimination parameter is assumed to have a value of 1.0; in contrast, the item discrimination parameter in the one-parameter logistic IRT model is a constant value but need not have a value equal to 1.0 [24]. There are two polytomous Rasch models, the partial credit model and the rating scale model [26], which are also found in IRT.

Assumptions in RMT

The key assumptions underlying RMT are identical to IRT: unidimensionality and local independence [26]. Within RMT, there is also a requirement (rather than an assumption) of invariance [34]. The Rasch model was developed based on the criterion of invariance [36], meaning that measurements from an instrument are independent of what is being measured [24]. For example, the CAHAI-7 should measure upper extremity function in individuals with stroke regardless of whether the individual is a male or female.

IRT and RMT – How are they different?

The similarities in the mathematical models and underlying assumptions within IRT and RMT bring about confusion and debate between these two theories. There are two perspectives on these two measurement theories: in one, IRT and RMT share the

same paradigm and the Rasch models are regarded as special cases of IRT models; in the other, IRT and RMT are distinct and have different paradigms [36,37]. In this article, we take the position that IRT and RMT are different paradigms. The key differences between the theories stem from a philosophical perspective: in IRT (experimental paradigm), a model is selected to fit the observed data from an outcome measure; in RMT (measurement paradigm), the observed data is instead assessed if it fits the Rasch model [24,36]. Readers can refer to Andrich's articles [36,37] for in-depth details about the philosophical differences between IRT and RMT.

In the following sections, we use the CAHAI as an example instrument to demonstrate how IRT and RMT are applied to evaluate the psychometric properties of an outcome measure.

Methods

Sample

We used data from a previous validation study of the CAHAI (n = 105) [11]. Participants were recruited from four rehabilitation facilities in Hamilton, Ontario, Canada, and the inclusion criteria were first-ever stroke and a combined score on the Chedoke-McMaster Stroke Assessment [38] arm and hand subscales of either ≤ 5 or between 7 and 11. These scores indicate that some participants had greater upper extremity impairment (flaccid paralysis or some movements within synergistic patterns), while others had full range of synergistic movements [38]. The median days post-stroke was 38 days (1^{st}, 3^{rd} quartile: 27, 80 days). Participants were assessed on the CAHAI by their treating therapist (either an occupational therapist or physical therapist) at two time

points: at baseline (admission/ initial visit) and at discharge (completion of their rehabilitation program). The time between baseline and discharge assessments ranged from 2 to 6 weeks (median = 30 days; 1st, 3rd quartiles: 21, 42 days) [11].

Measure

The CAHAI evaluates the function of the affected upper extremity after a stroke using real-life daily tasks [8]. There are four versions of the CAHAI (Table 2): the original 13-item (CAHAI-13), and three shortened versions with 9, 8, and 7 items (CAHAI-9, CAHAI-8, and CAHAI-7 respectively). Performance of the affected upper extremity on each test item is rated on a 7-point scale, from total assistance (score = 1) to independence (score = 7). A total score is obtained by summing scores on each item, with higher scores indicating better function in the affected upper extremity. The range of scores for the CAHAI-13, CAHAI-9, CAHAI-8, and CAHAI-7 are 13 to 91, 9 to 63, 8 to 56, and 7 to 49 respectively.

The psychometric properties of all four versions of the CAHAI have been evaluated using CTT methods when it was developed. The psychometric properties evaluated include: inter-rater reliability of intraclass-correlation coefficient (ICC) = 0.98 [9]; test-retest reliability of ICC = 0.96 to 0.98 [9,10]; convergent cross-sectional validity of correlation = 0.81 to 0.87 with the Chedoke-McMaster Stroke Assessment (arm and hand components) and correlation = 0.93 to 0.95 with the Action Research Arm Test [9]; and longitudinal validity (ability to detect different amounts of change in upper extremity function between individuals with acute and chronic stroke) of an area under the receiver operating characteristics curve of 0.93 to 0.97 [9].

Procedure

Permission to use the data was obtained from the corresponding author of the previous validation study [11]. All data were provided in a de-identified format consistent with the original data collection method. For this study, all 105 participants were included, and participants' CAHAI scores at either baseline or discharge were used. This was intended to optimize the frequency of scores for each score category in the 7-category scale in the dataset.

Data analysis

Participant characteristics and total scores on each version of the CAHAI were summarized using descriptive statistics. For each item of the CAHAI, the frequency of scores in each score category was calculated as percentages.

Item response theory

In IRT analysis, a model is selected and the fit of this model with the observed scores is evaluated [39]. These were the steps taken in our analysis:

1. Preliminary selection of potential IRT models
2. Fitting and comparison of potential IRT models, and selection of the best fit model
3. Estimation of item difficulty and discriminating parameters for the model of interest
4. Evaluation of model assumptions

This analysis was first conducted for the full version of the CAHAI (13 items), and the same analysis was then repeated for each shortened version of the CAHAI (7, 8, and 9 items). All IRT analyses were conducted using STATA version 14 [40].

Preliminary selection of potential IRT models. The first and most important consideration in the preliminary selection of IRT models is the outcome measure's scoring/response options. As presented earlier, different IRT models exist for dichotomous and polytomous scoring/response options and there are IRT models that include the plausibility of guessing responses. The IRT models selected should be consistent with the outcome measure's scoring/response options. Two other factors should also be considered in the preliminary selection of IRT models: one, whether the number of score/response categories are the same for all items in the instrument, and two, assumptions about the item discrimination parameter (either held constant or allowed to vary). In our analysis, all three factors were considered in the preliminary selection of IRT models.

Fitting and comparison of potential IRT models. After the preliminary selection process, potential IRT models are compared to identify the model that best fit with the data. Presently, there is no consensus on how best to evaluate the fit of IRT models to the observed data [39]. However, there are several suggested methods including the comparison of nested models, checking the extent to which model assumptions are met, determining if expected model features of invariance were achieved, and evaluating the accuracy of model predictions [29]. In our analysis, we compared nested models and assessed the extent to which model assumptions (unidimensionality and local

independence) were met. The analyses conducted to assess model assumptions are described later.

The comparison of nested models evaluates the relative model fit, where the goal is to select the simplest IRT model from a set of potential models that best fit with the observed data (i.e., adequately explains the observed data) [41]. When the models are nested, the likelihood ratio test, which approximately follows a chi-square distribution, can be used to compare models [41]. There are situations where IRT models are not nested and the use of the likelihood ratio test for comparison is not appropriate [42,43]. However, the comparison of non-nested IRT models is beyond the scope of this article. In our analysis, potential IRT models were fitted and compared using the likelihood ratio test to whether it was appropriate to reduce a more complex model (generalized partial credit model) to a simpler model (partial credit or rating scale model) [39,41]. A significant result ($p < 0.05$) indicated that the reduction to a simpler model decreased the model fit [41]. This meant that the complex model was preferred as it had a better fit with the observed data (i.e., could better explain the observed scores on the CAHAI items) [41].

Estimating item difficulty and discriminating parameters. The characteristics of the best fit model determine the item parameters that need to be estimated. For example, for one-parameter logistic models, item difficulty parameters (*b*) are estimated for each item and a single item discriminating parameter (*a*) is estimated (since all items are assumed to have equal discriminating abilities); in comparison, for two-parameter logistic models, item difficulty and discriminating parameters are estimated for each item [24]. In

our analysis, after the best fit model was selected, both item difficulty and discriminating parameters were estimated according to the model's characteristics.

In addition, the item parameters of each CAHAI item were examined graphically using category characteristic curves and item information functions. Category characteristic curves plot the probability of obtaining a specific score category (*y*-axis) against the latent trait, θ (*x*-axis) [44]. Accordingly, each CAHAI item has 7 category characteristic curves. The points at which the curves of adjacent score categories intersect indicate the transition points from one score category to the next [44]. Item information functions plot the item information (*y*-axis) against the latent trait (*x*-axis) [44], displaying the contribution of each test item along the latent trait continuum [29]. Item information functions also illustrate the variation of an item's precision along the latent trait continuum and are comparable to the reliability of measurement [39,45]. Items with greater discriminating ability contribute more information (indicated by a taller peak of the curve) and are associated with greater measurement precision and smaller error variance [39,45]. In our analysis, item information functions of each CAHAI item were plotted to assess the precision of each item and the test information function (the sum of item information functions [44]) was used to examine the precision of the CAHAI in its entirety.

Evaluation of model assumptions. The two assumptions in IRT, unidimensionality and local independence, must be evaluated. In our analysis, confirmatory factor analysis (one-factor model) was used to assess these assumptions. The evidence for unidimensionality was a good fit with the one-factor model, based on four different

criteria: chi-square (x^2) statistic, root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI).[46] In general, a nonsignificant x^2, lower RMSEA values, and higher CFI and TLI values indicate a better fit.[46,47] We used a nonsignificant x^2, RMSEA $\leq$ 0.06, and CFI and TLI $\geq$ 0.95 as the criteria for a good fit with the one-factor model. [46,47] The assumption of local independence was assessed by examining the residual correlation matrix of the confirmatory factor analysis. Residual correlations > 0.20 was used as the criterion to indicate local dependency between pairs of items (i.e., violation of local independence assumption).[48]

Rasch measurement theory

All Rasch analyses were performed using the Rasch Unidimensional Measurement Model software, RUMM2030 [49]. Briefly, Rasch analysis examines the extent to which participants' scores (observed scores) on each CAHAI item agrees with the predicted scores from the Rasch model [36]. The extent of the agreement between observed scores and the Rasch model should not be regarded as an absolute agreement [50]. Rather, this agreement is a relative one, which is determined through the interpretation of the results of a range of procedures [50]. In our analysis, five categories of analyses were performed: fit, targeting, dependency, reliability, and stability. This analysis was first conducted for the full version of the CAHAI (13 items), and the same analysis was then repeated for each shortened versions of the CAHAI (7, 8, and 9 items).

Fit – Do the items work together? Individual test items should work together (i.e., fit) so that it is appropriate to sum scores on each item (ordinal scores) to obtain a total

score (continuous score) [51]. The extent to which the CAHAI items fit together provides evidence of its validity to measure upper extremity function in individuals with stroke [51]. We assessed the fit of the CAHAI items with the ordering of item scoring categories (ordering of thresholds [52]), two statistical tests (fit residual and x^2 statistics), and one graphical indicator (item characteristics curve) [53].

The CAHAI score categories represent an ordered continuum of upper extremity function, where a higher score represents greater upper extremity function. This means that as an individual's upper extremity function improves, their scores on each CAHAI item should progress sequentially. Evaluating the ordering of item scoring categories determines whether the CAHAI 7-category scale works as intended. Category probability curves of each CAHAI item were used to assess the ordering of the scoring categories. Category probability curves (also known as category characteristic curves) plot the probabilities of obtaining a specific score category in an item [54]. Accordingly, each CAHAI item has 7 category characteristic curves. The points at which the curves of adjacent score categories intersect indicate the thresholds (transition points) from one score category to the next.[44] If the CAHAI's 7-category scale works as intended, the thresholds of adjacent score categories should be ordered (i.e., 1-2 < 2-3 < 3-4 < 4-5 < 5-6 < 6-7). Disordered thresholds indicate the current 7-category scale is not working as intended.

Two statistical tests were conducted to evaluate the fit of the CAHAI with the Rasch model. Firstly, residuals were used to summarize individual person- and item- fit statistics. Residuals are the deviations of observed scores on each CAHAI item from their

expected values using Rasch analysis [55]. Residuals between $\pm$ 2.5 were considered an adequate fit with the Rasch model [56]. Secondly, x^2 statistics were used to summarize item-trait (item-by-ability level) interaction for each CAHAI item, and a significant x^2 indicated the item does not fit the Rasch model [57]. Bonferroni adjustment was applied to adjust the level of significance to account for multiple hypothesis testing [58].

Graphical representations of CAHAI items' fit with the Rasch model were presented as item characteristic curves. An item characteristic curve describes the relationship between the trait (upper extremity function) and the probability of obtaining a particular score category in an item, and items with a good fit will show plots lying on the curve [24,56]. The slope of an item characteristic curve provides information about the item's discrimination abilities, where a steeper slope indicates better discrimination [59].

Targeting – Does the range of upper extremity function measured by the CAHAI match the range of upper extremity function in the sample? Scale-to-sample targeting refers to the match between the range of upper extremity function measured by the CAHAI and participants' range of upper extremity function. Targeting was examined with the person-item threshold distribution, which plots the histograms of two relative distributions of person locations (i.e., participants' upper extremity function) and item locations (i.e., item difficulty) [55]. A greater similarity between the distributions represents better targeting.

The location of CAHAI items (based on its difficulty level) and its associated standard error, as well as the spread of item locations, were also examined [53]. Item locations can be understood using the analogy of a ruler. When measuring with a ruler, objects have longer (or shorter) length; similarly, CAHAI items that are more difficult (or easier) requires greater (or lesser) upper extremity function to complete, which indicates an individual having greater (or lesser) upper extremity function. Test items should essentially be located at different points on this ruler (the CAHAI) such that it measures a continuum of upper extremity function, from less to more. The locations of items in each CAHAI version were estimated and the mean item locations were set at 0. This arbitrary constraint imposed on the mean item location was required to estimate the relative locations of the items [60].

The mean item and person location scores were then compared. A mean person location of 0 indicated the CAHAI is a well-targeted (not too easy or difficult) measure [56]. A mean person location of > 0 indicated that participants had greater upper extremity function than the task difficulty of the CAHAI (i.e., the CAHAI was 'too easy'). Conversely, a mean person location of < 0 indicated that participants had lesser upper extremity function than the tasks demanded, suggesting the CAHAI may be a 'difficult' measure.

Dependency – Do scores on one CAHAI item bias scores on the other items? As mentioned earlier, an underlying assumption of RMT is local independence where scores on each item should be independent of scores on the other items. High levels of

dependency can artificially inflate the CAHAI's reliability [55]. The assumption of local independence was assessed by examining the residual correlations and the general rule of thumb of not exceeding $\pm$ 0.3 was used as an acceptable degree of correlation [55].

Reliability – To what extent are the CAHAI scores not associated with random error? The Person Separation Index was used as the reliability index, defined as the ratio of true-score variance to observed variance [59]. Higher values indicate that the CAHAI has greater reliability.

Stability – Does each CAHAI item perform similarly across different groups of individuals? Using the earlier analogy of the CAHAI as a measurement ruler of upper extremity function, this ruler should perform similarly across various groups of individuals, such as different age groups and stroke chronicity. This property is also known as invariance [61]. The stability of the CAHAI was assessed across age, sex, and factors associated with the recovery of upper extremity function after stroke (stroke chronicity [62,63], baseline upper extremity impairment [64], and unilateral spatial neglect [65]). CAHAI items that displayed differential item functioning (DIF) indicated the item was unstable [24].

There are two types of DIF, uniform and non-uniform. In items displaying uniform DIF, there is a consistent systematic difference in scores between groups of individuals across the range of the attribute measured (i.e., the latent trait continuum) [56]. For example, males may consistently obtain a higher score than females in a test item. In contrast, items that display non-uniform DIF indicate that there are varying

differences between groups of individuals across the latent trait continuum [56]. For example, males may obtain a higher score than females when $\theta \geq 0.5$ whereas females may obtain a higher score than males when $\theta < 0.5$. Analysis of variance was conducted for each CAHAI item where scores across each level of the person factor (e.g. sex) were compared across different levels of the latent trait (also known as class intervals) [60]. A significant main effect for the person factor indicated the presence of uniform DIF while a significant interaction effect (person factor x class interval) indicated non-uniform DIF [56]. The level of significance was adjusted with Bonferroni correction to account for multiple hypothesis testing [58].

Evaluation of assumptions. The assumption of local independence and the requirement of invariance were assessed using residual correlations and DIF respectively (as described earlier). The assumption of unidimensionality was assessed using principal component analysis of the residuals [66]. Residuals of the first factor were correlated with the items and then classified into two subsets of items (positively and negatively correlated items) [67]. These two subsets were subsequently used to generate person estimates, and the differences between the estimates for each person was assessed using the independent t-test [67]. If the percentage of persons with significant differences between the estimates exceeded 5% [68], the CAHAI measure was deemed as unidimensional.

Results

Participant characteristics

Participant characteristics and scores on each version of the CAHAI are summarized in Table 3. Participants' median age was 72 years (1^{st}, 3^{rd} quartile: 62, 78) and 51% (n = 54) were males. The median days post-stroke was 38 days (1^{st}, 3^{rd} quartile: 27, 80) and 82% (n = 78) had an ischaemic stroke. The median total scores on the 13-, 9-, 8-, and 7-item versions of the CAHAI are 38, 29, 25, and 22 respectively. Fig 1 shows the frequency of scores in each score category for each CAHAI item.

Item response theory

Preliminary selection of potential IRT models

All items in the CAHAI use the same 7-category (ordered) scoring scale and guessing of scores was not possible. No assumptions about the item discriminating parameters were made, meaning it could either be similar between items or allowed to vary across items. This assumption was based on our belief that each CAHAI item had a similar (or different) ability to discriminate between individuals with greater or lesser upper extremity function as each item demands a similar (or varying) extent of upper extremity function for successful task completion. For example, item 6 (do up five buttons) and item 11 (clean a pair of eyeglasses) require greater hand dexterity than grip strength for task completion; in comparison, item 4 (pour a glass of water) and item 12 (place container on table) require greater grip and upper extremity strength than hand dexterity. Thus, our preliminary selection of IRT models included one-parameter (partial

credit model and rating scale model) and two-parameter (generalized partial credit model) logistic polytomous models. The graded response model was not considered as the partial credit and rating scale models were not nested within it (this paper focuses on the comparison of nested IRT models).

Fitting and comparison of potential IRT models

The generalized partial credit model was first fitted to the data and compared against the nested models (partial credit and rating scale models). Likelihood ratio test results showed that the generalized partial credit model fitted significantly better than the rating scale model for all versions of the CAHAI (CAHAI-13: $x^2(72) = 225.53$, $p < 0.001$; CAHAI-9: $x^2(48) = 142.33$, $p < 0.001$; CAHAI-8: $x^2(42) = 142.10$, $p < 0.001$; CAHAI-7: $x^2(36) = 137.81$, $p < 0.001$).

The generalized partial credit model fitted significantly better than the partial credit model for only the CAHAI-13 (CAHAI-13: $x^2(12) = 51.49$, $p < 0.001$). For all shortened versions of the CAHAI, the likelihood ratio tests between the generalized partial credit model and the partial credit model were non-significant. This meant that choosing the simpler (i.e., more parsimonious) model did not decrease the model fit with the observed data. Thus, the partial credit model was selected as the best fit model for the shortened versions of the CAHAI.

Estimating item difficulty and discriminating parameters

With the generalized partial credit model selected as the best fit model for the CAHAI-13, item discriminating (a) and difficulty parameters (b_{1-6}) were estimated for each test item. Table 4 summarizes the parameter estimates and their standard errors for the CAHAI-13. Item 13 (carry a bag up the stairs) had the lowest item discriminating ability (a = 0.84) while Item 4 (pour a glass of water had the highest item discriminating ability (a = 3.72). Fig 2 shows an example of the graphical representation of the item parameters using category characteristic curves. The points at which curves of adjacent score categories intersect indicate the item difficult parameters [44]. For instance, the location where curves of score categories 1 and 2 intersect is 0.05 logits. This means that an individual with a level of upper extremity function of 0.05 logits has a 50/50 chance of obtaining a score of 1 or 2 on CAHAI Item 5 (wring out washcloth).

Fig 3 shows the item information functions and test information function of the CAHAI-13. There were varying heights of the peaks in the item information functions, which were consistent with the items' discriminating ability. Some item information functions were non-unimodal (e.g. Item 5, wring out washcloth). Non-unimodal item information functions were expected since each score category contributes to its own information and its maximum information may peak at different locations along the latent trait continuum [44]. The locations of the peaks for the CAHAI-13 were generally located between $\theta = \pm 1$ logit. The test information function peaked at about $\theta = 0$, with the width of the curve approximately between $\theta = -2$ logits to 2 logits.

As the partial credit model was selected as the best fit model for the shortened versions of the CAHAI, 6 item difficulty parameters (b_{1-6}) were estimated for each test

item and a single discriminating parameter (*a*) was estimated for each CAHAI shortened version. The item discriminating ability of the CAHAI-9, CAHAI-8, and CAHAI-7 were 2.55 (0.26), 2.50 (0.27), and 2.76 (0.33) respectively. This indicates that the discriminating abilities of the three shortened versions were similar. The item information functions of the shortened versions of the CAHAI were similar to the CAHAI-13, displaying peaks (some non-unimodal) with varying heights and at different locations along the continuum. Details for the shortened versions of the CAHAI are described in the supplementary files (S1 – S3 Figs).

Evaluation of model assumptions

Results from the confirmatory factor analysis revealed that all versions of the CAHAI did not meet IRT assumption of unidimensionality. None of the fit indices of the CAHAI-13 met the predefined fit criteria (CAHAI-13, $x^2(65) = 243.02$ ($p < 0.001$), RMSEA = 0.16, CFI = 0.90, TLI = 0.88). For the shortened versions of the CAHAI, there were mixed fit with the one-factor model, which suggests they were also not unidimensional. All shortened versions had a significant x^2 and RMSEA > 0.06 (CAHAI-9, $x^2(27) = 78.94$, $p < 0.0001$, RMSEA = 0.14; CAHAI-8, $x^2(20) = 65.22$, $p < 0.0001$, RMSEA = 0.15; CAHAI-7, $x^2(14) = 41.66$, $p = 0.0001$, RMSEA = 0.14). However, they generally met the criteria of CFI and TLI $\geq$ 0.95 (CAHAI-9, CFI = 0.96, TLI = 0.94; CAHAI-8, CFI = 0.96, TLI = 0.94; CAHAI-7, CFI = 0.97, TLI = 0.95).

For local independence, residual correlation from the confirmatory factor analysis ranged from -0.41 to 1.67. For each version of the CAHAI, there were pairs of items with

residual correlation > 0.20, indicating the violation of IRT assumption of local independence. In the CAHAI-13, nine pairs of items (item pairs 1-5, 2-6, 3-8, 4-7, 5-1, 5-9, 5-11, 7-12, 7-13, and 12-13) had residual correlations between 0.21 to 1.67. The CAHAI-9, CAHAI-8, and CAHAI-7 had four (item pairs 1-5, 3-8, 4-7, and 5-9), three (item pairs 1-5, 3-8, and 4-7), and two (item pairs 1-5 and 4-7) pairs of items respectively that did not meet the predetermined criterion, with residuals ranging between 0.22 to 0.32.

Rasch measurement theory

Fit – Do the items work together?

Ordering of item scoring categories (ordering of item thresholds). In the CAHAI-13, all items had disordered thresholds. This was a similar finding in all the shortened versions of the CAHAI, with the exception of Item 6 (do up five buttons). For example, Fig 4 shows the category probability curves of Item 5 (wring out washcloth) and Item 6 (do up five buttons) in the CAHAI-7. The intersections (thresholds) of adjacent curves in Item 5 were disordered (1-2 < 2-3 < 3-4 < 6-7 < 4-5 < 5-6), indicating that lesser upper extremity function is required to obtain the highest score of 7 than lower score categories of 4 to 6 (Fig 4A). In comparison, the category probability curve of Item 6 (Fig 4B) shows ordered thresholds (i.e., 1-2 < 2-3 < 3-4 < 4-5 < 5-6 < 6-7). Although Item 6 had ordered thresholds, the distances between the thresholds were very close (0.007 to 0.252 logits), which suggests that clinicians still have some difficulty discriminating between the various score categories.

Fit-residual and x^2 statistics. All items in each CAHAI version had fit-residuals between ± 2.5. The residual mean and standard deviations of the CAHAI-13, -9, -8, and -7 were -0.14 (1.13), -0.10 (0.74), -0.19 (0.84), and -0.16 (0.78) respectively. For x^2 values, all items in the shortened versions of the CAHAI were non-significant, indicating a good fit with the Rasch model. However, for CAHAI-13, items 12 (place container on table) and 13 (carry bag up the stairs) had significant misfit with the Rasch model ($x^2(2) = 26.0$, $p < 0.0001$ and $x^2(2) = 64.3$, $p < 0.0001$ respectively) (Table 5).

Item characteristic curves. The item characteristic curves supported the statistical results of the CAHAI items' fit to the Rasch model. For example, Fig 5 shows the item characteristic curves for two CAHAI-13 items with the best (Item 7, dry back with a towel) and worst (Item 13, carry bag up the stairs) fit-statistic results. The plots (observed scores) in Fig 5A lie closely on the S-shaped curve (predicted scores), indicating a fit with the Rasch model. However, in Fig 5B, the plots lie distant from the curve which indicates a poor fit with the Rasch model. This is consistent with this item's significant x^2 value ($x^2(2) = 64.3$, $p < 0.0001$).

Targeting – Does the range of upper extremity function measured by the CAHAI match the range of UE function in the sample?

Item locations. Table 5 shows the item locations and its associated error for CAHAI-13. The location of items ranged from -0.68 to -0.80 logits, with item 1 (open jar of coffee) and item 4 (pour a glass of water) as the easiest and most difficult test items respectively. The mean location was 0 logit with a standard deviation (*SD*) of 0.50.

Details for the shortened versions of the CAHAI are described in the supplementary files (S1-S3 Tables).

Comparison between item and person locations. The mean person locations and standard deviations for the CAHAI-13, -9, -8, and -7 were -1.07 (2.13), -1.08 (2.53), -1.03 (2.36), and -1.10 (2.48). As the mean person locations were < 0, this suggests that the CAHAI may be a 'difficult' measure as participants had lower function in their affected upper extremity than the average difficulty of the CAHAI items.

Person-item threshold distribution. Fig 6 shows the person-item threshold distribution of the CAHAI-13. The spread of item locations between -3 logits and +2 logits indicates the CAHAI-13 measures a reasonable continuum of upper extremity function. However, there was an uneven spread along the continuum, with most items clustering around the location of 0.5 logits. There was a mismatch between the distribution of item and person locations, particularly at the lower end of the continuum (locations between -5 logits and -3 logits). This suggests inadequate scale-to-sample targeting where the CAHAI is unable to measure function in individuals with minimal upper extremity function (e.g., flaccidity). There were similar findings in the shortened versions of the CAHAI. The person-item threshold distributions of the shortened versions of the CAHAI are available in the supplementary files (S1 – S3 Figs).

Dependency – Do scores on one CAHAI item bias scores on the other items?

In all versions of the CAHAI, there were pairs of items with residual correlation slightly exceeding $\pm$0.30, indicating a minor violation of the local independence

assumption. Dependency was present in three pairs of items in the CAHAI-13; item-pairs 3-5, 9-12, and 12-13 had residual correlations of -0.32, -0.32, and 0.32 respectively. In both the CAHAI-9 and CAHAI-8, the same four pairs of items exceeded the predetermined criterion (item-pairs 2-7, 3-5, 8-1, and 8-6). The residual correlations of these item-pairs in the CAHAI-9 were -0.34 (item pair 2-7), -0.39 (item pair 3-5), -0.30 (item pair 8-1) and -0.34 (item pair 8-6); in the CAHAI-8 were -0.4 (item pair 2-7), -0.39 (item pair 3-5) -0.33 (item pair 8-1) and -0.34 (item pair 8-6). For the CAHAI-7, four pairs of items had residual correlation > 0.30 (item-pair 1-3, 3-5, 7-2, and 7-6).

Reliability – To what extent are the CAHAI scores not associated with random error?

The Person-Separation Index for the CAHAI-13, -9, -8, and -7 were 0.94, 0.93, 0.91, and 0.89 respectively. This suggests that all versions of the CAHAI demonstrated good reliability.

Stability – Does each CAHAI item perform similarly across different groups of individuals?

With the exception of CAHAI-7, all other CAHAI versions (CAHAI-8, CAHAI-9, and CAHAI-13) displayed non-uniform DIF, indicating the measures are unstable across different groups of individuals. For the CAHAI-13, non-uniform DIF was present for stroke type (infarct or hemorrhagic) in item 8 (put toothpaste on toothbrush) and for upper extremity impairment (mild-moderate or severe impairment) in item 3 (draw a line with a ruler), item 8 (put toothpaste on toothbrush) and item 10 (zip up the zipper). Similarly, non-uniform DIF was present for stroke type in item 8 and for upper extremity

impairment in items 1 and 3 for the CAHAI-9. For the CAHAI-8, there was non-uniform DIF in item 7 (dry back with a towel) across different age groups (40 – 59 years, 60 – 69 years, 70 – 79 years, and ≥ 80 years).

Evaluation of assumptions

In summary, there was the violation of the assumption of local independence in all versions of the CAHAI and invariance was found in the CAHAI-8, CAHAI-9, and CAHAI-13. For the assumption of unidimensionality, the CAHAI-8, CAHAI-9, and CAHAI-13 did not meet the criteria of < 5% of persons with significant differences between the estimates (6.7%, 6.7%, and 20.0% respectively). Only the CAHAI-7 was found to be unidimensional (0.95%).

Discussion

The evaluation of outcome measures guided by modern test theories can provide evidence of whether the measure assumes interval-level measurement characteristics [28], and describe the relationship between individuals' performance on each test item and the underlying latent trait assessed [29]. In our example instrument, results from both IRT and Rasch analyses indicate that the CAHAI does not have equal-interval scale properties and is not a unidimensional outcome measure. This means that the summation of scores on individual items is not appropriate and differences in the raw total scores need to be interpreted with caution since mathematical calculations are not justified. Although we took the position that IRT and RMT have different underlying paradigms, we identified two key similarities in our results. First, both analyses indicated that the

CAHAI's 7-category scoring scale was not working as intended (described below). Second, there was an overall violation of the assumptions of unidimensionality and local independence in both analyses. The following sections discuss these findings and provide recommendations of how the identified issues with the CAHAI may be resolved.

Similarities in findings between IRT and Rasch analyses

The 7-category scoring scale of the CAHAI not working as intended

The 7-category scale was intended to represent an ordered continuum of upper extremity function, where higher scores represent better upper extremity function. From Rasch analysis, there were disordered thresholds in all versions of the CAHAI. This indicates that the 7-category scoring scale was not working as intended as the transition between the score categories (from 1 to 7) did not correspond to a progressive level of upper extremity function measured. Although the order of scoring categories was not formally evaluated in IRT, estimates of the item difficulty parameter supported the Rasch analysis results. For example, recalling that item difficulty represents the location where the probability of obtaining a score in two adjacent score categories is equal [24,29], the item difficulty parameters of adjacent score categories of 1-2, 2-3, 3-4, 4-5, 5-6, 6-7 for Item 5 (wring out washcloth; Table 4) were 0.05, -0.37, 0.23, 1.10, 0.37, and 0.55 respectively. This means that better upper extremity function is required to obtain a score of 4 or 5 (θ = 1.10 logits) than the higher score categories of 6 and 7 (θ = 0.37 and 0.55 logits), which was not how the 7-category scoring scale was intended.

The CAHAI's scoring scale was based on the same 7-category scoring scale used in the Functional Independence Measure [69,70]. Studies have found disordered thresholds in the Functional Independence Measure's scale [71], and thus, it was not surprising that the CAHAI's scoring scale was not working as intended. Disordered thresholds may be attributed to too many score categories and assessors may have difficulties discriminating between them [56,72]. Another possible reason for disordered thresholds are problems with the labeling of the score categories (e.g., may be confusing or vague) [56,72]. One method of resolving disordered thresholds is to collapse score categories [73]. The collapsing from seven to five or four categories in the scoring scale of the Functional Independence Measure resolved the disordered thresholds [74–76]. Thus, it is possible that similar collapsing of score categories in the CAHAI's scale may resolve the disordered thresholds. A follow-up study employing Rasch analysis to revise the CAHAI is therefore needed to create a revised scale that works as intended.

Violation of underlying assumptions

Unidimensionality. Overall, the CAHAI did not present as a unidimensional outcome measure, meaning that it does not only measure upper extremity function. Examining the 13 test items, we postulated that two items, Item 12 (place container on table) and Item 13 (carry bag up the stairs), affected the dimensionality of the CAHAI. These two test items measure beyond upper extremity function as the ability to stand and/or climb a flight of stairs are needed for successful task completion. This notion is further supported by the International Classification of Functioning, Disability and Health (ICF)[77] categories linked to the CAHAI [78]. Two ICF categories related to body

positions (d4105, bending) and moving around (d4551, climbing) were linked to items 12 and 13 of the CAHAI [77], which suggests that the CAHAI items do not measure a single latent trait of upper extremity function. Thus, it is important to consider the theoretical and conceptual constructs of the test items during the development and refinement of outcome measures guided by IRT or RMT.

Although the shortened versions of the CAHAI do not contain items 12 and 13, they did not consistently demonstrate unidimensionality. This may be because the CAHAI measures different aspects of upper extremity function, such as strength, dexterity, reaching, and coordination. Outcome measures that are comprised of items that measure various aspects of a latent trait can better capture the complexity of the trait but at the expense of unidimensionality [79]. For example, health-related quality of life measures may include multi-components, such as physical and mental well-being, and social relationships. By evaluating specific domains, quality of life can be measured more holistically. However, this affects the dimensionality of the measure and it may have difficulties meeting IRT and RMT assumption of unidimensionality.

Local independence Local dependency among item pairs was present in all versions of the CAHAI, possibly for two reasons. First, the assumptions of unidimensionality and local independence are related, and local dependency may be present in outcome measures that are not unidimensional [29]. Second, as the development of the CAHAI was guided by CTT, the test items selected had moderate correlations with each other to ensure internal consistency [8]. However, the inter-item correlations may cause local dependency when the CAHAI is evaluated using modern

test theories. The appropriateness of local dependency depends on the purpose of the outcome measure [80]. In outcome measures that contain sets of items where responses on later items are based on earlier items (e.g., presence/absence of pain and how it affects upper extremity function), local dependency may be expected. However, in the case of the CAHAI, local dependency suggests that there may be items that are redundant [80]. For example, local dependence was present between Item 3 (draw a line with ruler) and Item 5 (wring out washcloth) for all versions of the CAHAI, suggesting one of the items may be redundant.

Revisions to the CAHAI is needed

The problems in the CAHAI's scoring scale and the violation of underlying assumptions of both IRT and RMT indicate the needed to revise the measure. Revisions to the CAHAI can be guided by modern test theories. For example, RMT may be employed to guide the revision of the scoring scale to ensure it works as intended. IRT may also be employed to guide the selection of items in the revised CAHAI to better ensure unidimensionality.

Reason for similarities in findings

The similarities in the results from both IRT and Rasch analyses in the example instrument may be due to our IRT model choices in the IRT analysis. The scope of this article focused only on nested IRT models, and the generalized partial credit model and the partial credit models were the models that best fitted our data. As the partial credit model is mathematically an extension of the Rasch model [81], this may account for the

similarities in the findings. It is important that readers are aware of the possibility of fitting other IRT models, such as the graded response model, which may produce different findings. The fitting of the graded response model, a popular model choice for polytomous items, may produce different results due to the unique characteristics of the model. This model describes the probability of obtaining a particular score category or higher versus lower score categories (e.g., the probability of scoring ≥ 4 versus scoring < 4), and consequently, the item difficulty parameters (ordering of item thresholds in RMT) will be sequentially ordered [24]. Thus, if the graded response model was fitted in our IRT analysis, the disordered thresholds in the 7-category scoring scale in the CAHAI would not be present and scaling issues with the scoring scale would not be identified.

Challenges with applying modern test theories

Sample size requirements

One of the biggest challenges in applying modern test theories in clinical research is the large sample sizes needed for stable estimations, compared to using CTT. In our example instrument, 105 participants may be considered small for IRT analysis (as a two-parameter logistic polytomous model was fitted for the CAHAI-13) and reasonable for Rasch analysis [24]. Our data was not uniformly distributed across all seven score categories, with some items having less than 10 observations in a score category. This may have influenced the accuracy of estimations in both IRT and Rasch analyses.

There are various sample size recommendations for IRT and RMT, however, they should not be regarded as rigid rules as there are several considerations when determining

sample sizes [24]. Factors to consider include the characteristics of the outcome measure (e.g., instrument length, number of scoring/response options), data characteristics (e.g., number of missing data), the distribution of the sample (e.g., frequency of responses in each score/response category), the estimation techniques and procedures (e.g., number of item parameters), and the ancillary techniques that will be used (e.g., factor analysis for the evaluation of unidimensionality) [24]. In general, the number of participants needed increases with the number of item parameters to be estimated, and thus, larger sample sizes are required when polytomous models are used (compared to dichotomous models) and IRT analysis require larger sample sizes in comparison to Rasch analysis. For dichotomous models, a minimum of 30 participants is recommended for the Rasch model [82], and approximately 100 and 250 to 500 participants are recommended for the one- and two-parameter logistic models respectively [24]. For polytomous models, a minimum sample size of 50 and at least 10 observations per category is recommended for polytomous Rasch models [82]. There are several recommendations for polytomous IRT models, such as using the 2:1 ratio of person to item parameters (e.g., 13 CAHAI items x 6 transition location parameters x 2 = 156 participants) for the partial credit model and a sample size of 500 for two-parameter logistic polytomous models [24]. A uniform distribution of observations across the scoring/response categories in the sample is also desired [24]. Overall, the need for larger sample sizes can affect the cost, complexity, and feasibility of conducting measurement studies guided by modern test theories.

Practical considerations

Statistical software. Both IRT and Rasch analyses often require statistical software designed for the purpose of such analyses, and it may be necessary for health care professionals to devote time to learn how to use a new statistical software. Examples of software packages available include RUMM2030 [49] and Winsteps [83] for Rasch analysis, and PARSCALE [84], Mplus [85], and BILOG [86] for IRT analysis. The capabilities of each software vary (e.g., model specifications available, estimation procedures, number of items, and number of participants) and thus, health care professionals need to be aware of the strengths and weaknesses of the software when purchasing and using it.

Accessibility of information. This article provides health care professionals with information about modern test theories in a non-technical manner; however, it is important to note that what we have presented only provides an overview of these theories. The abstruse nature of modern test theories demands the understanding of complex mathematical models and advanced knowledge of statistics. However, existing information about the theories is often not written in a manner that is easy for clinicians to understand [6]. Health care professionals may thus find it challenging to gain an in-depth understanding of IRT and RMT and to apply these theories proficiently.

Conclusion

The goal of this article was to provide rehabilitation professionals with an accessible overview of measurement theories and how to apply IRT and RMT in the

evaluation of outcome measures. The choice between IRT and RMT is dependent on the researcher and it is imperative that researchers provide a clear rationale for their choice of measurement theory. Researchers need to articulate their philosophical stance on modern test theories and use terminologies consistent with their philosophical stance. Research articles intended for the rehabilitation professional audiences should also endeavor to provide detailed and non-technical explanations to help clinicians understand the findings. Regardless of the theory chosen, the overall goal of both IRT and RMT is to improve the quality of outcome measures to ensure accurate and valid measurements of outcomes in rehabilitation, and consequently, improve the quality of patient care.

Acknowledgements

We thank Susan Barreca for permitting the use of the data from her study published in 2006. The first author is supported by the Singapore General Hospital scholarship award.

References

[1] Deyo RA, Carter WB. Strategies for improving and expanding the application of health status measures in clinical settings. A researcher-developer viewpoint. Med Care 1992;30:MS176-186; discussion MS196-209.

[2] College of Occupational Therapists. Occupational therapists' use of standardized outcome measures 2013. https://www.cot.co.uk/position-statements/occupational-therapists%E2%80%99-use-standardised-outcome-measures (accessed June 1, 2016).

[3] Fawcett AL. Principles of assessment and outcome measurement for occupational therapists and physiotherapists: theory, skills and application. Hoboken, NJ, USA: John Wiley & Sons.; 2007.

[4] Díez J. A hundred years of numbers. An historical introduction to measurement theory 1887–1990: Part I: The formation period. Two lines of research: Axiomatics and real morphisms, scales and invariance. Stud Hist Philos Sci Part A 1997;28:167–85. doi:10.1016/S0039-3681(96)00014-3.

[5] Stevens SS. On the theory of scales of measurement. Science 1946;103:677–80.

[6] Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol Assess Winch Engl 2009;13:iii, ix–x, 1–177. doi:10.3310/hta13120.

[7] Weiss DJ, Davison ML. Test theory and method. Annu Rev Psychol 1981;32:629–58.

[8] Barreca SR, Gowland CK, Stratford P, Huijbregts M, Griffiths J, Torresin W, et al. Development of the Chedoke Arm and Hand Activity Inventory: theoretical constructs, item generation, and selection. Top Stroke Rehabil 2004;11:31–42.

[9] Barreca SR, Stratford PW, Lambert CL, Masters LM, Streiner DL. Test-retest reliability, validity, and sensitivity of the Chedoke Arm and Hand Activity Inventory: a new measure of upper-limb function for survivors of stroke. Arch Phys Med Rehabil 2005;86:1616–22. doi:10.1016/j.apmr.2005.03.017.

[10] Barreca SR, Stratford PW, Masters LM, Lambert CL, Griffiths J, McBay C. Validation of three shortened versions of the Chedoke Arm and Hand Activity Inventory. Physiother Can 2006;58:148–56. doi:10.3138/ptc.58.2.148.

[11] Barreca SR, Stratford PW, Masters LM, Lambert CL, Griffiths J. Comparing 2 versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. Phys Ther 2006;86:245–53.

[12] Tesio L, Simone A, Bernardinello M. Rehabilitation and outcome measurement: where is Rasch analysis-going? Eur Medicophysica 2007;43:417–26.

[13] Belvedere SL, de Morton NA. Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. J Clin Epidemiol 2010;63:1287–97. doi:10.1016/j.jclinepi.2010.02.012.

[14] Ader D. Developing the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care n.d.;45:S1–2. doi:10.1097/01.mlr.0000260537.45076.74.

[15] Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45:S3–11. doi:10.1097/01.mlr.0000258615.42478.55.

[16] Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educ Meas Issues Pract 1993;12:38–47. doi:10.1111/j.1745-3992.1993.tb00543.x.

[17] Lord FM, Novick MR. Statistical theories of mental test scores. Reading, Mass. ; Don Mills, Ont: Addison-Wesley Pub. Co; 1968.

[18] Magno C. Demonstrating the difference between Classical test theory and item response theory using derived test data. Int J Educ Psychol Assess 2009;1:1–11.

[19] Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. Fifth edition. Oxford: Oxford University Press; 2015.

[20] Baylor C, Hula W, Donovan NJ, Doyle PJ, Kendall D, Yorkston K. An introduction to item response theory and Rasch models for speech-language pathologists. Am J Speech Lang Pathol 2011;20:243–59. doi:10.1044/1058-0360(2011/10-0079).

[21] Rusch T, Lowry PB, Mair P, Treiblmaier H. Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. Inf Manage 2017;54:189–203. doi:https://doi.org/10.1016/j.im.2016.06.005.

[22] Streiner DL. Measure for measure: new developments in measurement and item response theory. Can J Psychiatry Rev Can Psychiatr 2010;55:180–6. doi:10.1177/070674371005500310.

[23] Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989;70:857–60.

[24] de Ayala R. The theory and practice of item response theory. New York: Guilford Press; 2009.

[25] Allen MJ, Yen WM. Introduction to measurement theory. Monterey, Calif: Brooks/Cole Pub. Co; 1979.

[26] Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Second. Psychology Press; 2013.

[27] Reid CA, Kolakowsky-Hayner SA, Lewis AN, Armstrong AJ. Modern psychometric methodology: applications of item response theory. Rehabil Couns Bull 2007;50:177–88. doi:10.1177/00343552070500030501.

[28] Cook KF, Monahan PO, McHorney CA. Delicate balance between theory and practice: health status assessment and item response theory. Med Care 2003;41:571–4. doi:10.1097/01.MLR.0000064780.30399.A4.

[29] Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, Calif: Sage Publications; 1991.

[30] Masters GN. A rasch model for partial credit scoring. Psychometrika 1982;47:149–74. doi:10.1007/BF02296272.

[31] Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73. doi:10.1007/BF02293814.

[32] Muraki E. A Generalized Partial Credit Model: Application of an EM Algorithm. Appl Psychol Meas 1992;16:159–76. doi:10.1177/014662169201600206.

[33] Samejima F. Estimation of latent ability using a response pattern of graded scores. ETS Res Bull Ser 2014;1968:i–169. doi:10.1002/j.2333-8504.1968.tb00153.x.

[34] Ball S, Vickery J, Hobart J, Wright D, Green C, Shearer J, et al. Rasch measurement theory analysis of multiple sclerosis rating scale data. NIHR Journals Library; 2015.

[35] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research; 1960.

[36] Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res 2011;11:571–85. doi:10.1586/erp.11.59.

[37] Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004;42:I7-16. doi:10.1097/01.mlr.0000103528.48582.7c.

[38] Gowland C, Van Hullenaar S, Torresin W, et al. Chedoke-McMaster Stroke Assessment: development, validation and administration manual. Hamilton, Ontario: Chedoke-McMaster Hospitals and McMaster University; 1995.

[39] Nguyen TH, Han H-R, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. The Patient 2014;7:23–35. doi:10.1007/s40271-013-0041-0.

[40] StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP; 2015.

[41] Brown C, Templin J, Cohen A. Comparing the two- and three-parameter logistic models via likelihood ratio tests: a commonly misunderstood problem. Appl Psychol Meas 2015;39:335–48. doi:10.1177/0146621614563326.

[42] Kang T, Cohen A. IRT model selection methods for dichotomous items. Appl Psychol Meas 2007;31:331–58. doi:10.1177/0146621606292213.

[43] Kang T, Cohen A, Sung H-J. Model selection indices for polytomous items. Appl Psychol Meas 2009;33:499–518. doi:10.1177/0146621608327800.

[44] StataCorp. Stata 14 Base Reference Manual. College Station, TX: Stata Press; 2015.

[45] Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38:II28–42.

[46] Hill CD, Edwards MC, Thissen D, Langer MM, Wirth RJ, Burwinkle TM, et al. Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. Med Care 2007;45:S39-47. doi:10.1097/01.mlr.0000259879.05499.eb.

[47] Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Model Multidiscip J 1999;6:1–55. doi:10.1080/10705519909540118.

[48] Morizot J, Ainsworth AT, Reise, Steven P. Toward modern psychometrics: application of item response theory models in personality research. In: Robins RW, Fraley RC, Krueger RF, editors. Handb. Res. Methods Personal. Psychol., New York: Guilford Press; 2007, p. 407–23.

[49] RUMM Laboratory. RUMM 2030. Perth, Australia: RUMM Laboratory; 1998.

[50] Hagell P, Westergren A. 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. J Appl Meas 2016;17:416–31.

[51] Cano SJ, Mayhew A, Glanzman AM, Krosschell KJ, Swoboda KJ, Main M, et al. Rasch analysis of clinical outcome measures in spinal muscular atrophy. Muscle Nerve 2014;49:422–30. doi:10.1002/mus.23937.

[52] Hagquist C, Andrich D. Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. Personal Individ Differ 2004;36:955–68.

[53] Wright BD, Masters GN. Rating scale analysis: Rasch measurement. 1 edition. Chicago: MESA; 1982.

[54] Linacre JM. Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). Rasch Meas Trans 1999;13:675.

[55] Hobart J, Cano S, Posner H, Selnes O, Stern Y, Thomas R, et al. Putting the Alzheimer's cognitive test to the test II: Rasch Measurement Theory. Alzheimer's Dement 2013;9:S10–S20.

[56] Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007;46:1–18.

[57] Linacre JM. RUMM2020 item-trait chi-square and Winsteps DIF size. Rasch Meas Trans 2007;21:1096.

[58] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.

[59] Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther 2014;36:648–62. doi:10.1016/j.clinthera.2014.04.006.

[60] RUMM Laboratory. Interpreting RUMM2030. RUMM Laboratory Pty Ltd; 2013.

[61] Boone WJ. Rasch analysis for instrument development: why, when, and how? CBE Life Sci Educ 2016;15. doi:10.1187/cbe.16-04-0148.

[62] Kwakkel G, Kollen B, Twisk J. Impact of time on improvement of outcome after stroke. Stroke 2006;37:2348–53. doi:10.1161/01.STR.0000238594.91938.1e.

[63] Houwink A, Nijland RH, Geurts AC, Kwakkel G. Functional recovery of the paretic upper limb after stroke: who regains hand capacity? Arch Phys Med Rehabil 2013;94:839–44. doi:10.1016/j.apmr.2012.11.031.

[64] Coupar F, Pollock A, Rowe P, Weir C, Langhorne P. Predictors of upper limb recovery after stroke: a systematic review and meta-analysis. Clin Rehabil 2012;26:291–313. doi:10.1177/0269215511420305.

[65] Nijboer TCW, Kollen BJ, Kwakkel G. The impact of recovery of visuo-spatial neglect on motor recovery of the upper paretic limb after stroke. PloS One 2014;9:e100584. doi:10.1371/journal.pone.0100584.

[66] Brentani E, Golia S. Unidimensionality in the Rasch model: how to detect and interpret. Statistica 2007;67:253–61. doi:10.6092/issn.1973-2201/3508.

[67] Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3:205–31.

[68] Tennant A, Conaghan PG. 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 Rheum 2007;57:1358–62. doi:10.1002/art.23108.

[69] Chedoke Arm and Hand Activity Inventory (CAHAI). Chedoke Arm and Hand Activity Inventory administration guidelines version 2 n.d.

[70] Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. Adv Clin Rehabil 1987;1:6–18.

[71] Nilsson ÅL, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM^{TM}) revisited. J Rehabil Med 2011;43:884–91. doi:10.2340/16501977-0871.

[72] Cano SJ, Barrett LE, Zajicek JP, Hobart JC. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. Mult Scler Houndmills Basingstoke Engl 2011;17:214–22. doi:10.1177/1352458510385269.

[73] Wright BD, Linacre JM. Combining and splitting categories. Rasch Meas Trans 1992;6:233–5.

[74] Nilsson AL, Sunnerhagen KS, Grimby G. Scoring alternatives for FIM in neurological disorders applying Rasch analysis. Acta Neurol Scand 2005;111:264–73. doi:10.1111/j.1600-0404.2005.00404.x.

[75] Gosman-Hedström G, Blomstrand C. Evaluation of a 5-level functional independence measure in a longitudinal study of elderly stroke survivors. Disabil Rehabil 2004;26:410–8. doi:10.1080/09638280410001662978.

[76] Grimby G, Gudjonsson G, Rodhe M, Sunnerhagen KS, Sundh V, Ostensson ML. The functional independence measure in Sweden: experience for outcome measurement in rehabilitation medicine. Scand J Rehabil Med 1996;28:51–62.

[77] World Health Organization. International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organization; 2001.

[78] Choo SX, JN Ng C, Fayed N, Harris JE. International Classification of Functioning, Disability and Health Framework: Bridging adapted outcome measures. Int J Ther Rehabil 2017;24:494–500. doi:10.12968/ijtr.2017.24.11.494.

[79] Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas 2008;9:200–15.

[80] Chen W-H, Thissen D. Local dependence indexes for item pairs using item response theory. J Educ Behav Stat 1997;22:265–89. doi:10.3102/10769986022003265.

[81] Andrich D. The Rasch model explained. In: Maclean DR, Watanabe R, Baker R, Boediono D, Cheng PYC, Duncan DW, et al., editors. Appl. Rasch Meas. Book Ex., Springer Netherlands; 2005, p. 27–59. doi:10.1007/1-4020-3076-2_3.

[82] Linacre JM. Sample size and item calibration or person measure stability. Rasch Meas Trans 1994;7:328.

[83] Linacre JM. Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com; 2017.

[84] Muraki E, Bock D. PARSCALE. Chicago, IL: Scientific Software International, Inc; 2003.

[85] Muthén LK, Muthén BO. Mplus. Los Angeles, California: Muthén & Muthén; 2018.

[86] Zimowski M, Muraki E, Mislevy R, Bock D. BILOG-MG. Lincolnwood, IL: Scientific Software International, Inc; 2003.

Tables

Table 1. Summary data of the hypothetical case scenario using the 7-item Chedoke Arm and Hand Activity Inventory.

Item	John	William
	Score	Score
1. Open jar of coffee	6	4
2. Call 911	2	4
3. Draw a line with ruler	2	4
4. Pour a glass of water	5	4
5. Wring out washcloth	6	4
6. Do up five buttons	2	4
7. Dry back with towel	5	4
Total score	28	28

Table 2. Test items in each version of the Chedoke Arm and Hand Activity Inventory (CAHAI).

CAHAI-7
1. Open jar of coffee
2. Call 911
3. Draw a line with ruler
4. Pour a glass of water
5. Wring out washcloth
6. Do up five buttons
7. Dry back with towel
CAHAI-8
8. Put toothpaste on toothbrush
CAHAI-9
9. Cut medium-resistance putty
CAHAI-13
10. Zip up a zipper
11. Clean a pair of eyeglasses
12. Place container on table
13. Carry bag up the stairs

Table 3. Participant characteristics (*n* = 105).

Variables	Sample
Total *n*	105
Sex (%)	
Male	54 (51.4)
Female	51 (48.6)
Age, in years	
Median (1st, 3rd quartile)	72 (62, 78)
Min, Max	44, 92
Days since stroke	
Median (1st, 3rd quartile)	38 (27, 80)
Min, Max	3, 342
Type of stroke (%)	
Ischemic	78 (82.1)
Hemorrhagic	17 (17.9)
Unilateral spatial neglect (%)	
Absent	54 (51.9)
Present, client able to compensate	28 (26.9)
Present, client unable to compensate	22 (21.2)
Affected upper limb (%)	
Right	46 (43.8)
Left	57 (54.3)
Bilateral	2 (1.9)
Baseline upper limb impairment (%)	
Mild-moderate (CMSA score 7 – 11)	54 (51.4)
Severe (CMSA score ≤ 5)	51 (48.6)
CAHAI scores	
CAHAI-13	
Median (1st, 3rd quartile)	38 (16, 65)
Min, Max	13, 91
CAHAI-9	
Median (1st, 3rd quartile)	29 (11, 48)
Min, Max	9, 63
CAHAI-8	
Median (1st, 3rd quartile)	25 (10, 43)
Min, Max	8, 56
CAHAI-7	
Median (1st, 3rd quartile)	22 (8, 37)
Min, Max	7, 49

SD, standard deviation; min, minimum; max, maximum; CMSA, Chedoke-McMaster Stroke Assessment (arm and hand component); CAHAI, Chedoke Arm and Hand Activity Inventory

Table 4. Item response theory analysis: Generalized partial credit model item parameter estimates and standard errors for the CAHAI-13 (n = 105).

Item	*a*	b_1	b_2	b_3	b_4	b_5	b_6
1. Open jar of coffee	2.70 (0.59)	-0.62 (0.19)	-0.12 (0.20)	0.30 (0.21)	0.41 (0.21)	0.67 (0.21)	0.38 (0.21)
2. Call 911	2.87 (0.60)	-0.30 (0.18)	0.16 (0.18)	0.53 (0.19)	0.87 (0.25)	0.53 (0.25)	0.52 (0.20)
3. Draw a line with ruler	1.99 (0.42)	-0.07 (0.26)	-0.14 (0.28)	0.26 (0.26)	0.38 (0.25)	1.17 (0.34)	-0.11 (0.35)
4. Pour a glass of water	3.72 (0.88)	0.60 (0.20)	0.19 (0.19)	0.81 (0.17)	0.78 (0.18)	0.81 (0.18)	1.62 (0.25)
5. Wring out washcloth	2.35 (0.48)	0.05 (0.27)	-0.37 (0.27)	0.23 (0.20)	1.10 (0.29)	0.37 (0.28)	0.55 (0.22)
6. Do up five buttons	2.95 (0.67)	0.28 (0.19)	0.40 (0.20)	0.50 (0.21)	0.58 (0.21)	0.87 (0.22)	0.57 (0.21)
7. Dry back with towel	2.93 (0.65)	0.27 (0.18)	0.38 (0.18)	0.77 (0.19)	1.07 (0.27)	1.01 (0.34)	0.31 (0.32)
8. Put toothpaste on toothbrush	1.97 (0.39)	-0.73 (0.21)	0.26 (0.24)	0.04 (0.24)	1.28 (0.31)	0.78 (0.34)	0.31 (0.31)
9. Cut medium-resistance putty	2.43 (0.50)	-1.02 (0.19)	0.24 (0.21)	0.23 (0.21)	0.88 (0.24)	0.66 (0.25)	0.47 (0.23)
10. Zip up a zipper	1.81 (0.35)	-1.11 (0.22)	0.55 (0.26)	0.06 (0.26)	1.07 (0.25)	1.55 (0.41)	0.40 (0.41)
11. Clean a pair of eyeglasses	2.15 (0.42)	-0.08 (0.22)	0.05 (0.22)	0.67 (0.22)	0.96 (0.27)	0.77 (0.28)	0.74 (0.26)

Table 4 (Continued).

Item	*a*	b_1	b_2	b_3	b_4	b_5	b_6
12. Place container on table	0.93 (0.22)	2.04 (0.64)	1.06 (0.68)	-0.15 (0.70)	1.50 (0.69)	0.46 (0.71)	0.31 (0.58)
13. Carry bag up the stairs	0.84 (0.19)	3.02 (0.95)	0.59 (0.92)	-1.20 (0.88)	1.64 (0.63)	0.54 (0.66)	0.91 (0.60)

a, item discrimination parameter: a larger value indicates that a test item has greater ability to discriminate between individuals with lesser or more upper extremity function.[29]
$b_{1\text{-}6}$, item difficulty parameters (also known as threshold parameter): locations of thresholds where the probability of obtaining a score in two adjacent score categories is 0.5.[24,29] For example, *b1* represents the location where the probability of obtaining a score of 1 or 2 is 0.5.

Table 5. Rasch analysis: CAHAI-13 item locations, fit-residual and chi-square statistics ordered by item location.

Item	Location	SE	Fit residual	*df*	x^2	*P*-value
1. Open jar of coffee	-0.684	0.098	0.042	2	2.32	.314
9. Cut medium resistance putty	-0.524	0.094	0.003	2	0.58	.750
3. Draw a line with a ruler	-0.514	0.096	0.615	2	0.97	.616
8. Put toothpaste on toothbrush	-0.354	0.095	0.111	2	0.13	.936
5. Wring out washcloth	-0.332	0.095	-0.647	2	3.78	.151
2. Call 911	-0.202	0.090	-1.311	2	1.00	.608
10. Zip up the zipper	-0.152	0.102	0.488	2	1.85	.397
11. Clean a pair of eyeglasses	0.073	0.094	-0.349	2	1.04	.596
6. Do up five buttons	0.202	0.087	-0.861	2	1.14	.567
7. Dry back with towel	0.362	0.088	-2.004	2	2.79	.248
12. Place container on table	0.652	0.084	1.670	2	25.97	< .0001*
13. Carry bag up the stairs	0.674	0.086	1.835	2	64.33	< .0001*
4. Pour a glass of water	0.798	0.095	-1.364	2	3.01	.222

SE: standard error; df: degrees of freedom

* significant p-value (Bonferroni adjusted) of $p < .00077$

Figures

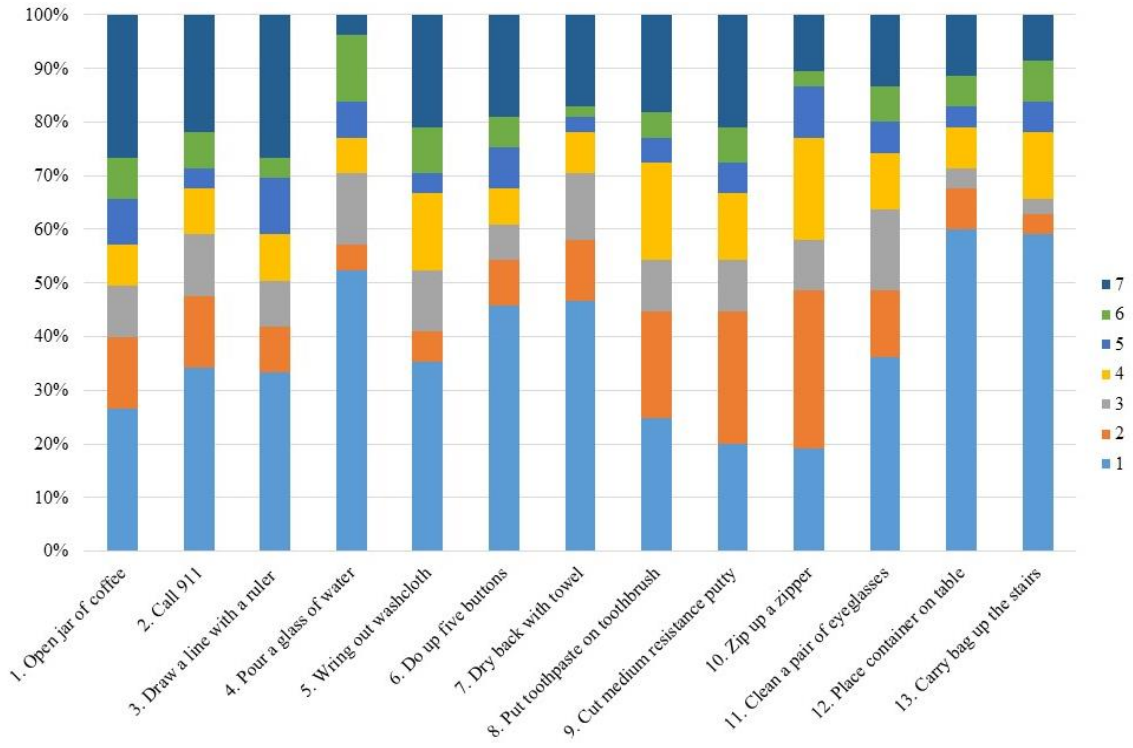


Fig 1. Frequency of scores of the 7-category scoring scale for each item in the CAHAI.

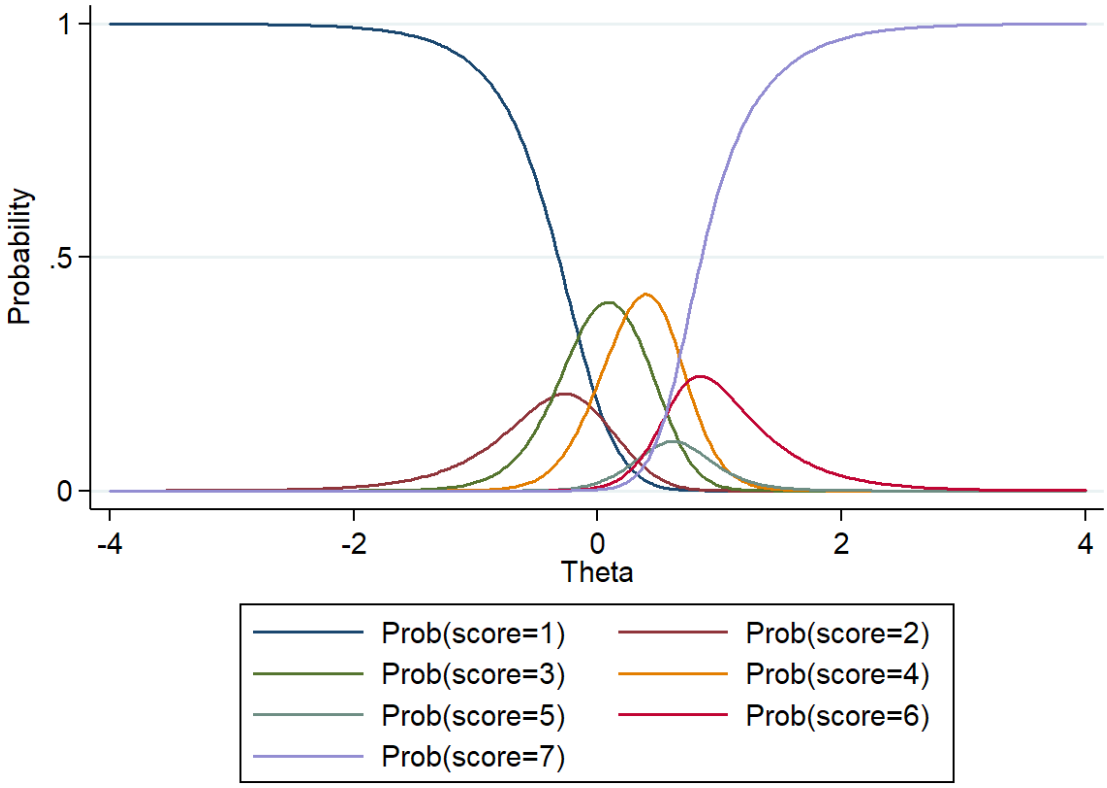


Fig 2. Category characteristic curves of item 5 (wring out washcloth) in CAHAI-13. The individual curves plot the probability of obtaining a particular score category along the latent trait (θ) continuum. The transition locations (points at which adjacent curves intersect) indicate the item difficulty parameters.

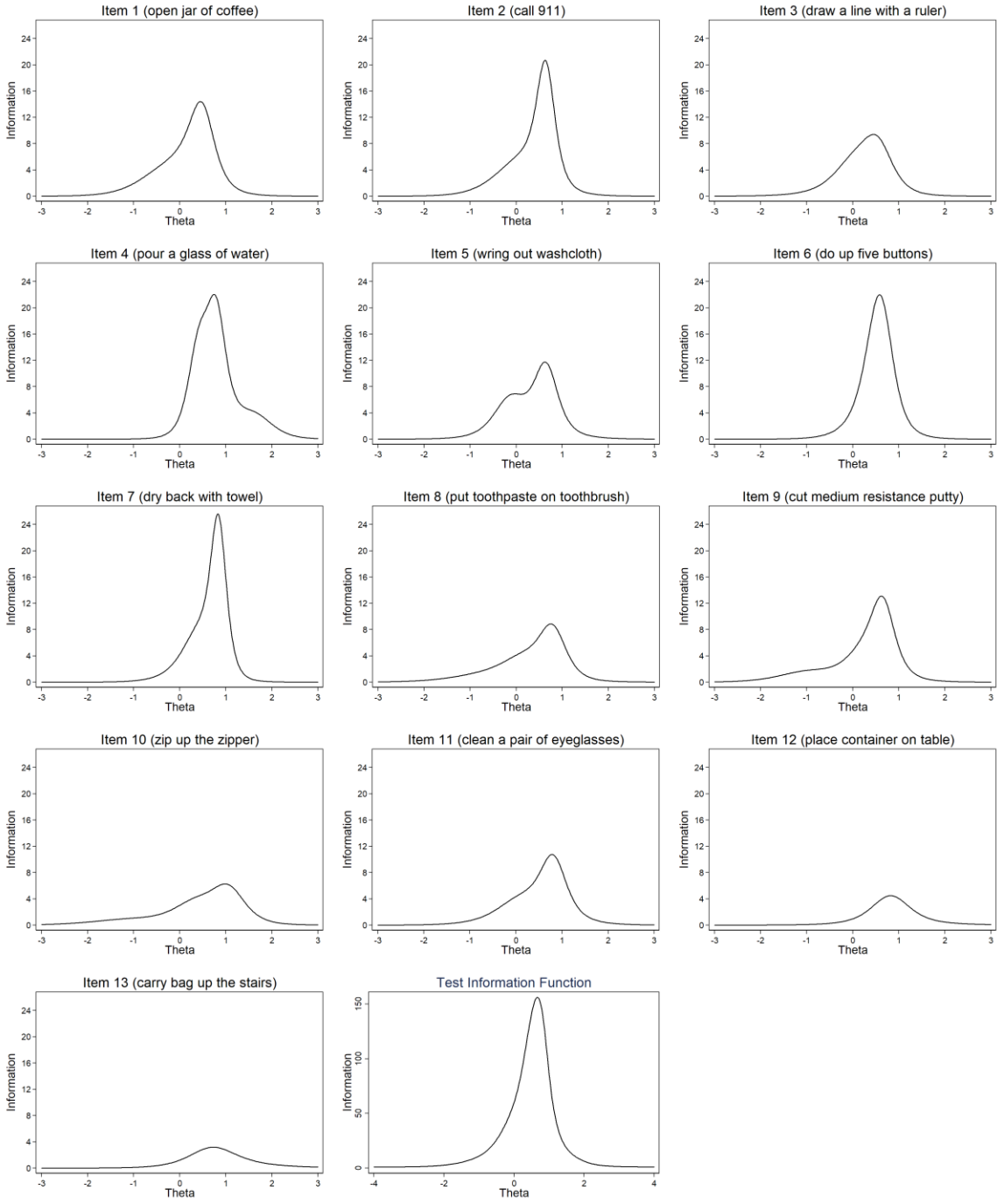


Fig 3. Item information functions and test information function of the CAHAI-13. A taller peak indicates greater item discriminating ability and a non-unimodal peak suggests inconsistent performance of the item across the latent trait continuum. The test information function is the sum of all item information functions.

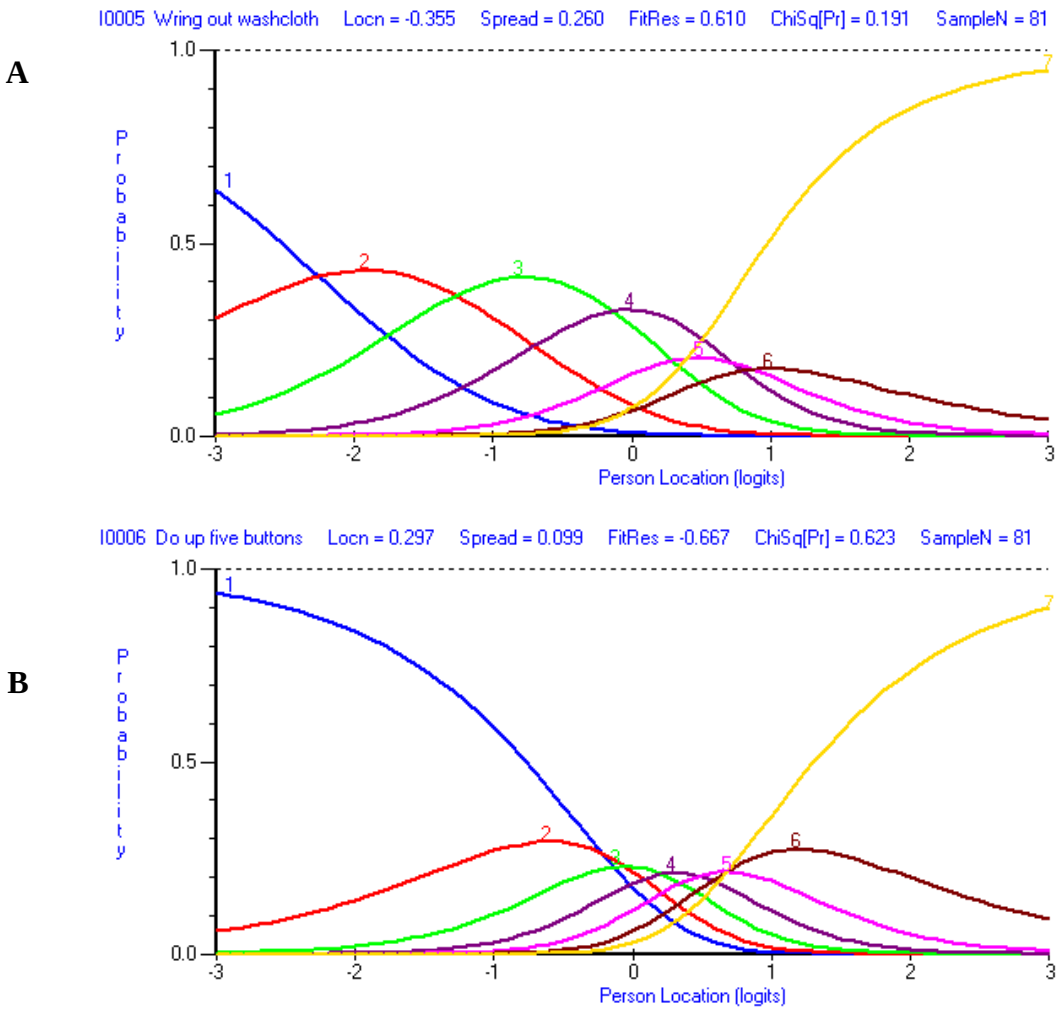
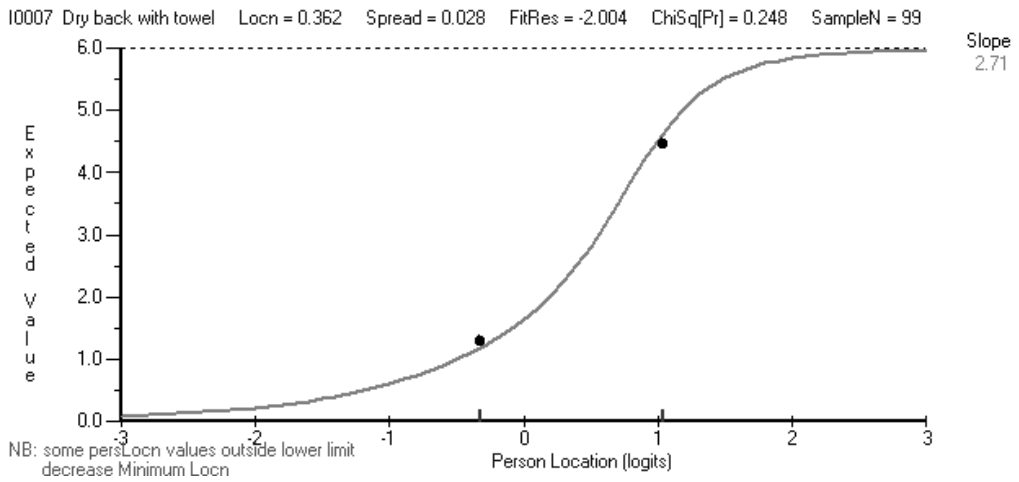


Fig 4. Example of two category probability curves in CAHAI-7. Each curve plots the probability of obtaining a score category on the 7-category scale. The points at which curves of adjacent score categories intersect (i.e., thresholds) indicate equal probability of obtaining a score in the two adjacent score categories. (A) The category probability curve of item 5 (wring out washcloth) shows that the intersections (thresholds) of adjacent curves were disordered, with threshold parameters (locations of the 1-2, 2-3, 3-4, 4-5, 5-6, and 6-7 thresholds) of -1.90, -0.91, 0.22, 1.08, 1.24, and 0.28 respectively. (B) The category probability curve of item 6 (do up five buttons) shows ordered thresholds, with the locations of the 1-2, 2-3, 3-4, 4-5, 5-6, and 6-7 thresholds at -0.51, -0.35, -0.09, -0.18, 0.37, and 0.39 respectively.

A



B

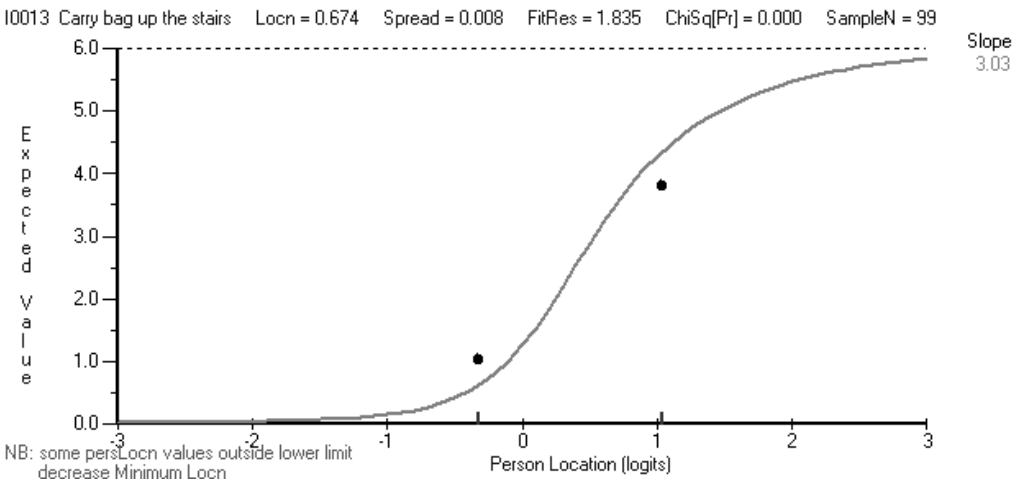


Fig 5. Item characteristic curves of two test items in the CAHAI-13. (A) Item characteristic curve for item 7 (dry back with a towel) shows a good fit with the Rasch model where the plots (observed scores) lie closely on the curve (predicted scores using the Rasch model). (B) Item characteristic curve for item 13 (carry bag up the stairs) shows misfit with the Rasch model as the plots are distant from the curve.

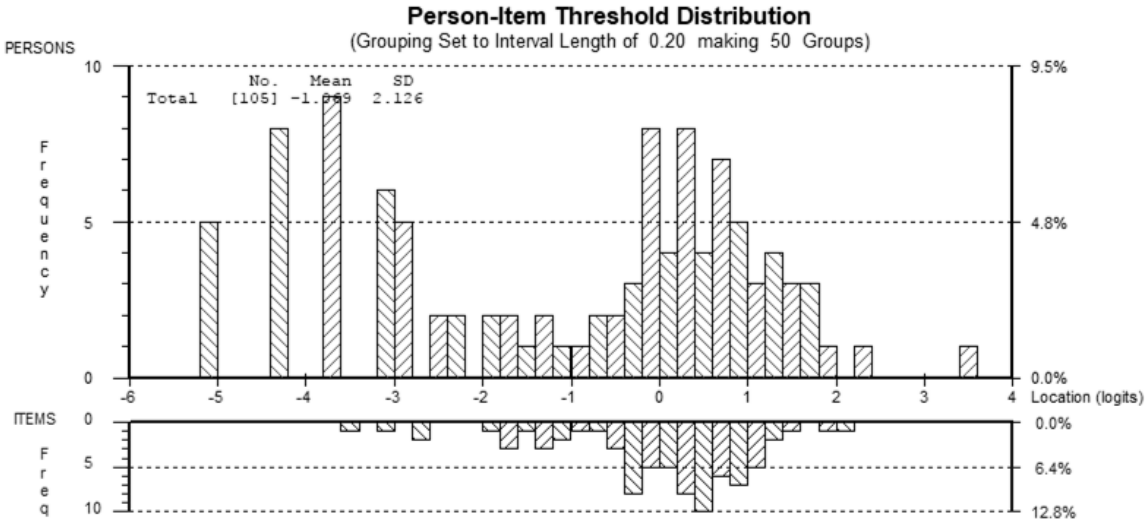


Fig 6. Person-item threshold distribution of the CAHAI-13. This figure shows the relative distributions of item and person locations. The mismatch between the range of upper extremity function measured by the CAHAI and the participants' range of upper extremity function suggests inadequate scale-to-sample targeting.

Supplementary File 1

Results for CAHAI-9

Table 1. Item response theory analysis: Partial credit model item parameter estimates and standard errors for the CAHAI-9 (n = 105).

Item	b_1	b_2	b_3	b_4	b_5	b_6
Item discriminating parameter (a) = 2.55 (0.26)						
1. Open jar of coffee	-0.57 (0.18)	-0.11 (0.21)	0.26 (0.22)	0.36 (0.22)	0.63 (0.22)	0.34 (0.19)
2. Call 911	-0.23 (0.18)	0.13 (0.19)	0.49 (0.20)	0.86 (0.26)	0.48 (0.27)	0.46 (0.20)
3. Draw a line with ruler	-0.18 (0.19)	-0.11 (0.22)	0.24 (0.21)	0.36 (0.20)	1.00 (0.26)	0.09 (0.24)
4. Pour a glass of water	0.78 (0.23)	-0.02 (0.23)	0.86 (0.21)	0.75 (0.24)	0.72 (0.22)	1.78 (0.30)
5. Wring out washcloth	0.03 (0.22)	-0.34 (0.24)	0.20 (0.19)	1.01 (0.25)	0.36 (0.26)	0.57 (0.19)
6. Do up five buttons	0.34 (0.19)	0.35 (0.23)	0.45 (0.23)	0.54 (0.23)	0.85 (0.24)	0.50 (0.21)
7. Dry back with towel	0.30 (0.18)	0.32 (0.19)	0.75 (0.21)	1.09 (0.29)	1.00 (0.38)	0.19 (0.32)
8. Put toothpaste on toothbrush	-0.73 (0.18)	0.20 (0.19)	0.09 (0.19)	1.09 (0.23)	0.75 (0.27)	0.47 (0.23)
9. Cut medium-resistance putty	-0.98 (0.19)	0.22 (0.19)	0.20 (0.20)	0.82 (0.22)	0.62 (0.24)	0.48 (0.20)

a, item discrimination parameter; $b_{1\text{-}6}$, item difficulty parameters

Table 2. Rasch analysis: CAHAI-9 item locations, fit-residual and chi-square statistics ordered by item location.

Item	Location	SE	Fit residual	*df*	x^2	*p*-value
1. Open jar of coffee	-0.643	0.106	0.444	2	3.81	.149
9. Cut medium resistance putty	-0.461	0.103	0.260	2	1.16	.560
3. Draw a line with a ruler	-0.424	0.105	0.613	2	0.26	.879
8. Put toothpaste on toothbrush	-0.251	0.104	1.084	2	0.24	.885
5. Wring out washcloth	-0.235	0.104	-0.046	2	2.84	.242
2. Call 911	-0.083	0.099	-0.917	2	0.22	.894
6. Do up five buttons	0.403	0.096	-0.591	2	0.31	.855
7. Dry back with a towel	0.589	0.096	-0.761	2	2.76	.252
4. Pour a glass of water	1.105	0.104	-0.941	2	1.92	.383

SE: standard error; df: degrees of freedom
Note: Significant *p*-value (Bonferroni adjusted) set at $p < .0011$

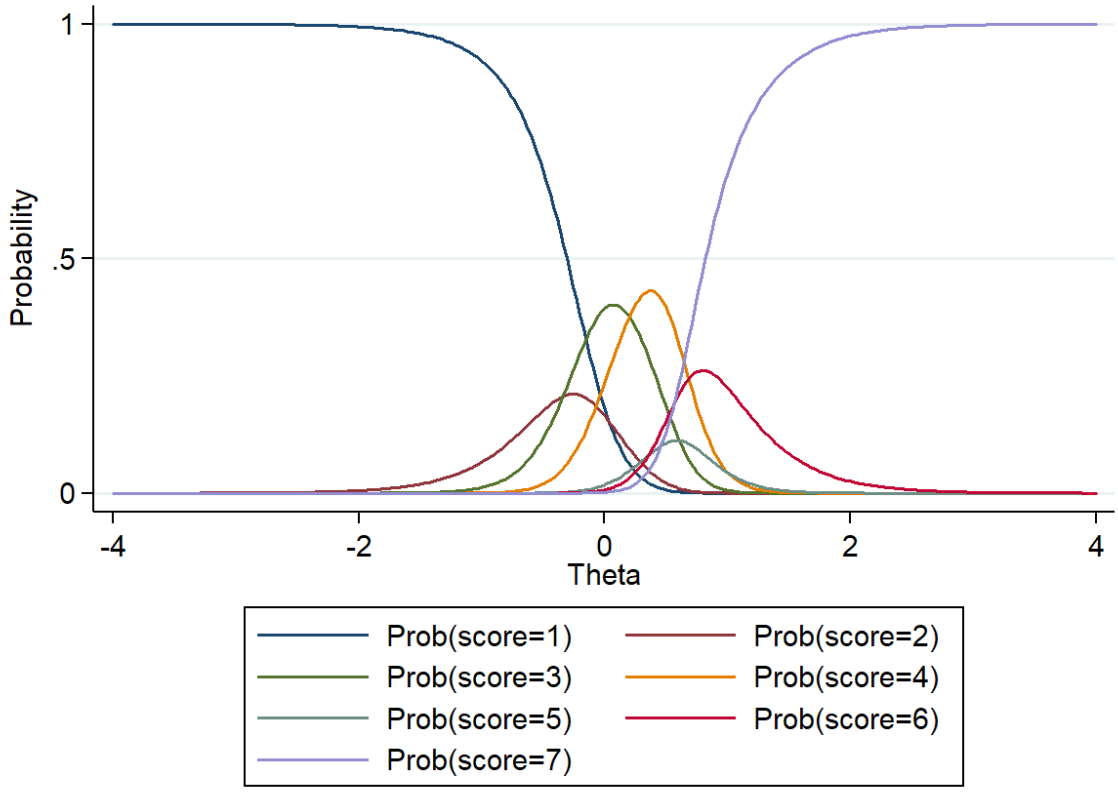


Figure 1. Category characteristic curves of item 5 (wring out washcloth) in CAHAI-9. The individual curves indicate the probability of obtaining the score category.

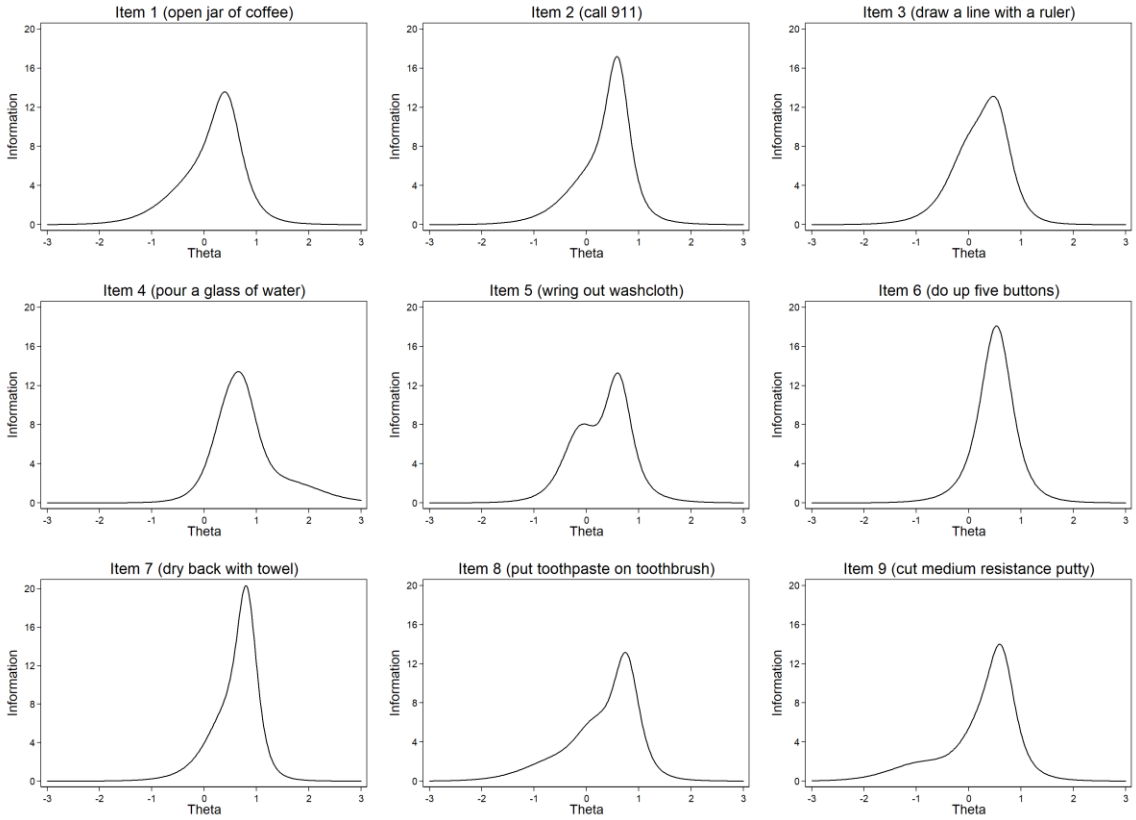


Fig 2. Item information functions of the CAHAI-9. A taller peak indicates more information provided by the item, and a non-unimodal peak suggests inconsistent performance of the item across the latent trait continuum.

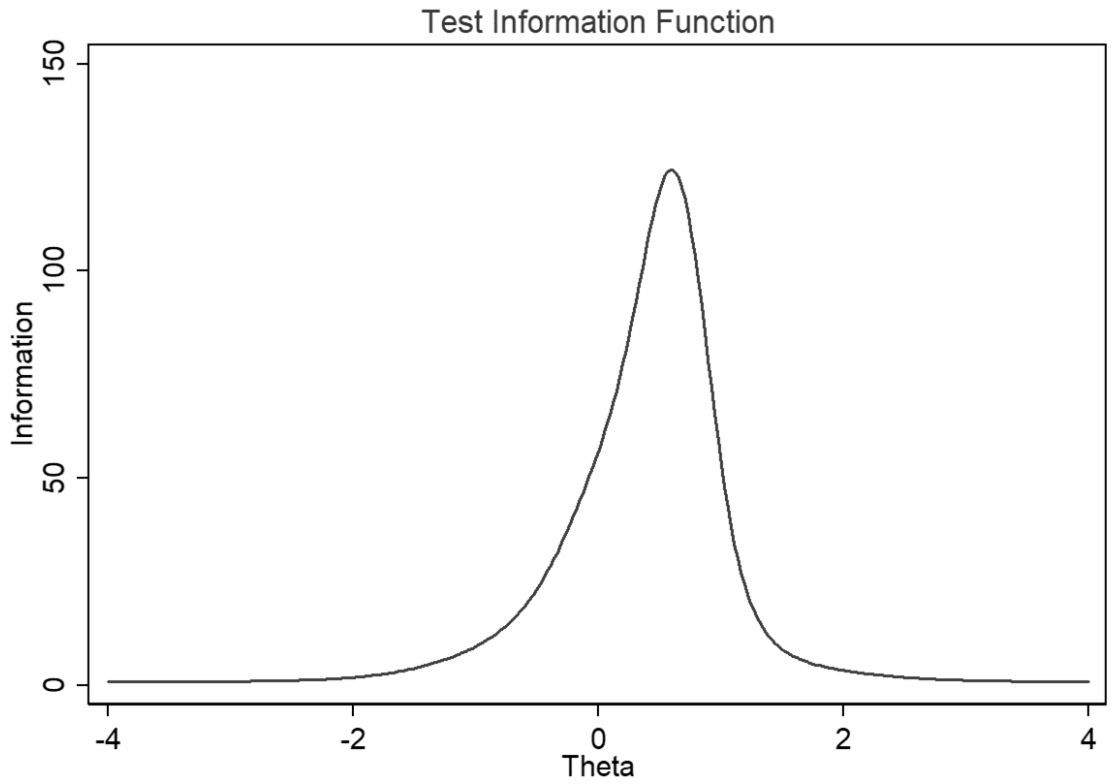


Fig 3. Test information function of the CAHAI-9. The test information function is the sum of all item information functions.

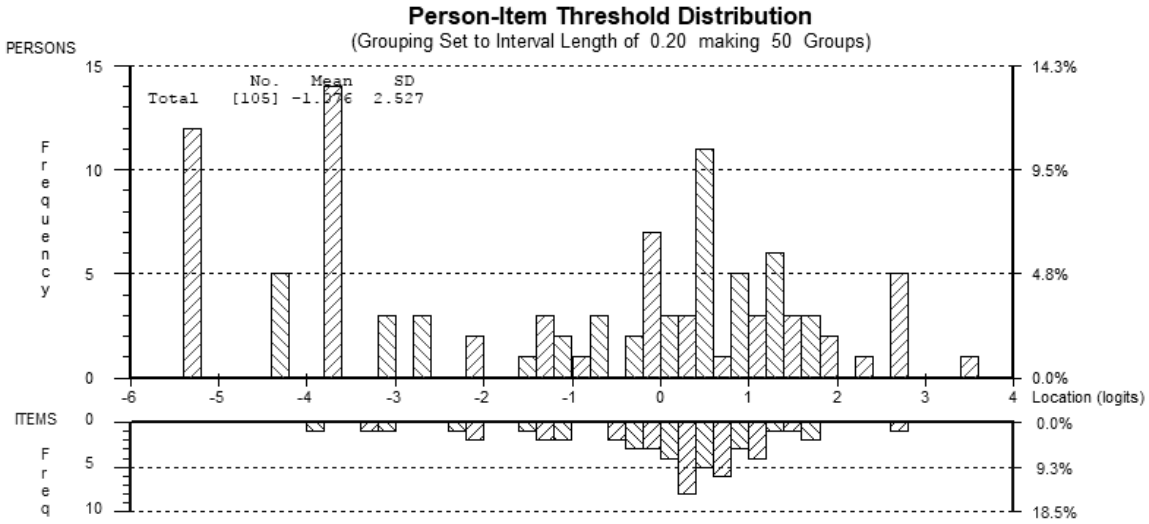


Fig 4. Person-item threshold distribution of the CAHAI-9. This figure shows the relative distributions of item and person locations. The mismatch between the range of upper extremity function measured by the CAHAI and the participants' range of upper extremity function suggests inadequate scale-to-sample targeting.

Supplementary File 2

Results for CAHAI-8

Table 1. Item response theory analysis: Partial credit model item parameter estimates and standard errors for the CAHAI-8 (n = 105).

Item	b_1	b_2	b_3	b_4	b_5	b_6
Item discriminating parameter (a) = 2.50 (0.27)						
1. Open jar of coffee	-0.55 (0.18)	-0.10 (0.21)	0.24 (0.22)	0.33 (0.22)	0.62 (0.22)	0.33 (0.19)
2. Call 911	-0.22 (0.18)	0.11 (0.20)	0.47 (0.20)	0.85 (0.27)	0.47 (0.27)	0.45 (0.20)
3. Draw a line with ruler	-0.16 (0.19)	-0.12 (0.23)	0.22 (0.22)	0.34 (0.21)	1.00 (0.26)	0.07 (0.24)
4. Pour a glass of water	0.79 (0.23)	-0.05 (0.24)	0.85 (0.22)	0.75 (0.24)	0.73 (0.22)	1.80 (0.31)
5. Wring out washcloth	0.06 (0.22)	-0.35 (0.24)	0.18 (0.19)	1.00 (0.26)	0.34 (0.26)	0.57 (0.19)
6. Do up five buttons	0.34 (0.19)	0.32 (0.23)	0.43 (0.24)	0.52 (0.23)	0.85 (0.24)	0.50 (0.22)
7. Dry back with towel	0.31 (0.18)	0.29 (0.19)	0.74 (0.21)	1.09 (0.30)	1.02 (0.39)	0.19 (0.32)
8. Put toothpaste on toothbrush	-0.72 (0.18)	0.20 (0.19)	0.07 (0.19)	1.08 (0.24)	0.75 (0.28)	0.47 (0.23)

a, item discrimination parameter; $b_{1\text{-}6}$, item difficulty parameters

Table 2. Rasch analysis: CAHAI-8 item locations, fit-residual and chi-square statistics ordered by item location.

Item	Location	SE	Fit residual	*df*	x^2	*p*-value
1. Open jar of coffee	-0.689	0.105	0.195	2	2.01	.366
3. Draw a line with a ruler	-0.463	0.104	0.748	2	0.06	.971
8. Put toothpaste on toothbrush	-0.298	0.105	1.067	2	0.50	.787
5. Wring out washcloth	-0.290	0.103	0.082	2	2.77	.251
2. Call 911	-0.139	0.099	-1.160	2	0.77	.680
6. Do up five buttons	0.324	0.095	-0.562	2	0.15	.930
7. Dry back with a towel	0.529	0.096	-0.840	2	0.67	.715
4. Pour a glass of water	1.025	0.104	-1.061	2	2.66	.265

SE: standard error; df: degrees of freedom

Note: Significant *p*-value (Bonferroni adjusted) set at $p < .0013$

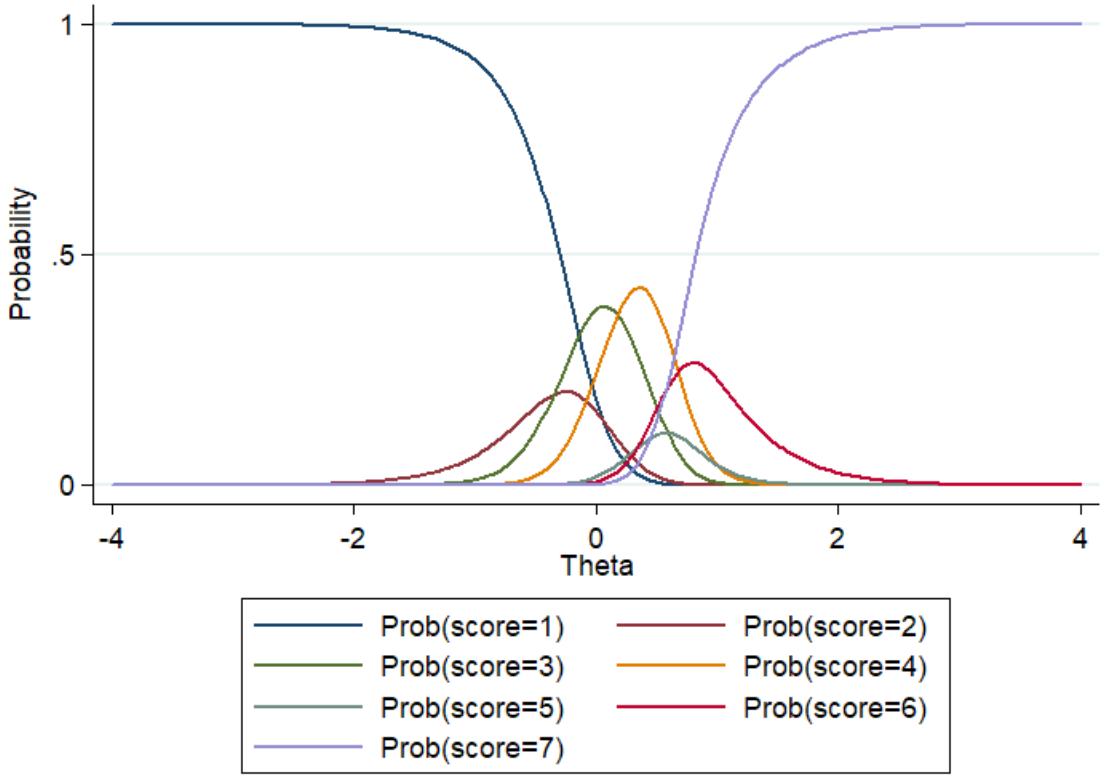


Fig 1. Category characteristic curves of item 5 (wring out washcloth) in CAHAI-8. The individual curves indicate the probability of obtaining the score category.

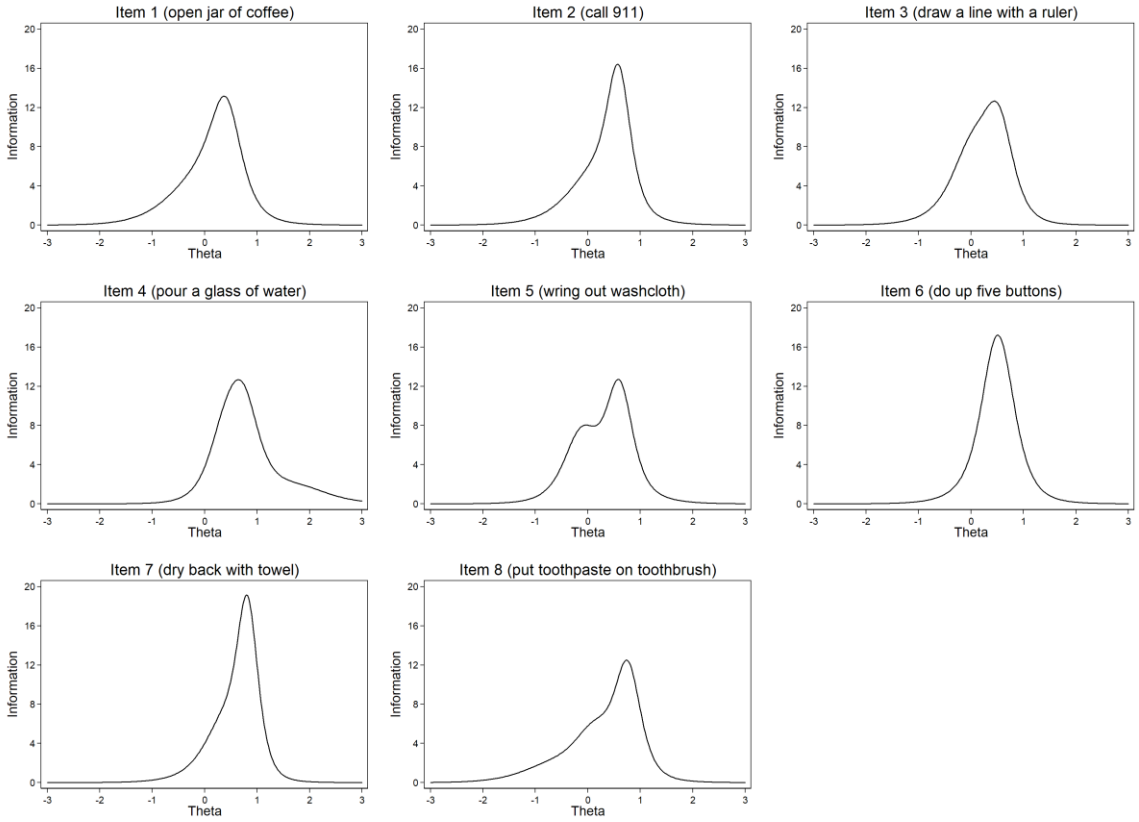


Fig 2. Item information functions of the CAHAI-8. A taller peak indicates more information provided by the item, and a non-unimodal peak suggests inconsistent performance of the item across the latent trait continuum.

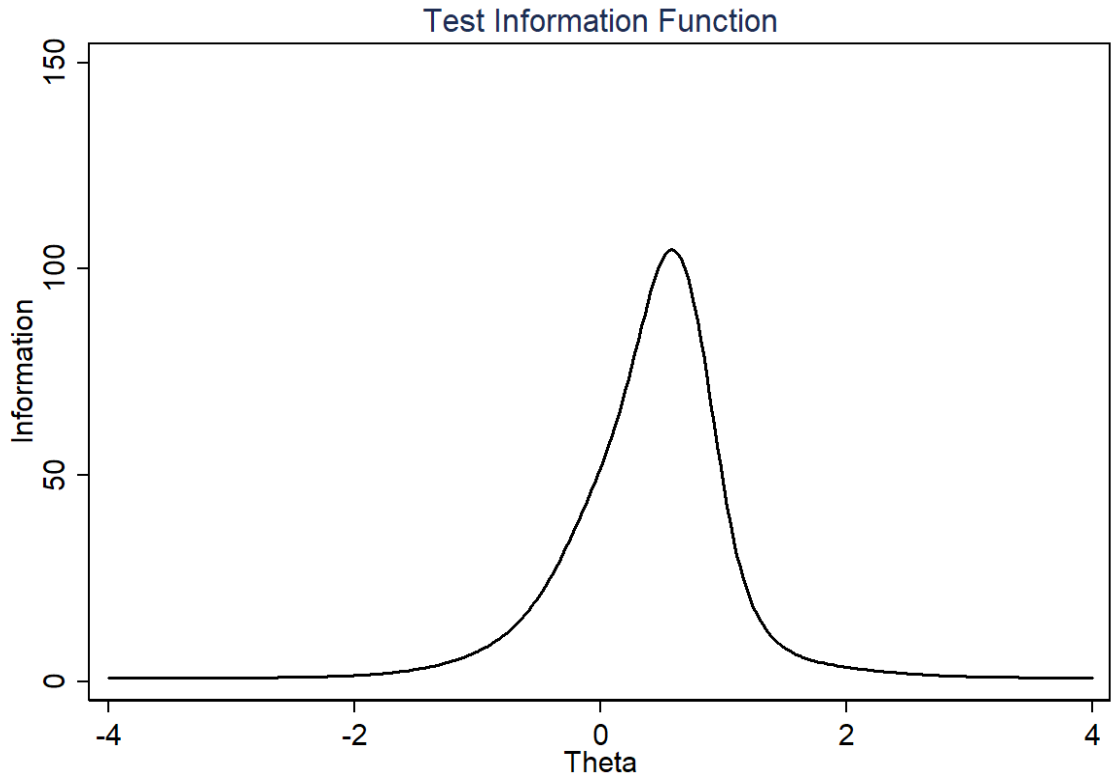


Fig 3. Test information function of the CAHAI-8. The test information function is the sum of all item information functions.

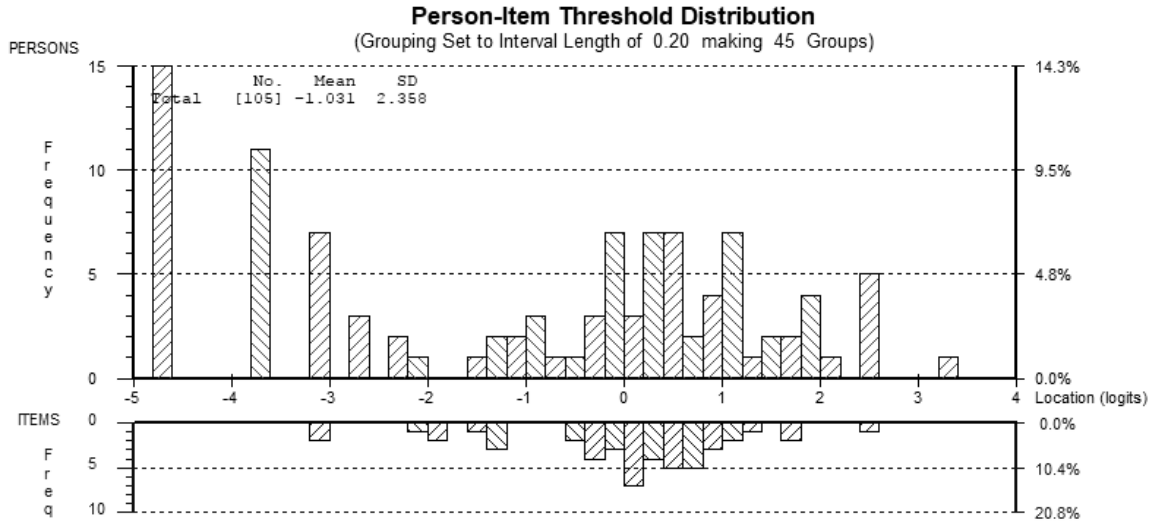


Fig 4. Person-item threshold distribution of the CAHAI-8. This figure shows the relative distributions of item and person locations. The mismatch between the range of upper extremity function measured by the CAHAI and the participants' range of upper extremity function suggests inadequate scale-to-sample targeting.

Supplementary File 3

Results for CAHAI-7

Table 1. Item response theory analysis: Partial credit model item parameter estimates and standard errors for the CAHAI-7 (n = 105).

Item	b_1	b_2	b_3	b_4	b_5	b_6
Item discriminating parameter (a) = 2.76 (0.33)						
1. Open jar of coffee	-0.54 (0.17)	-0.08 (0.19)	0.25 (0.20)	0.34 (0.20)	0.59 (0.20)	0.36 (0.18)
2. Call 911	-0.21 (0.16)	0.13 (0.18)	0.46 (0.19)	0.80 (0.24)	0.46 (0.25)	0.48 (0.19)
3. Draw a line with ruler	-0.16 (0.18)	-0.09 (0.21)	0.23 (0.20)	0.34 (0.19)	0.95 (0.24)	0.12 (0.22)
4. Pour a glass of water	0.73 (0.21)	-0.01 (0.22)	0.81 (0.20)	0.73 (0.22)	0.73 (0.20)	1.79 (0.30)
5. Wring out washcloth	0.04 (0.20)	-0.30 (0.22)	0.19 (0.17)	0.95 (0.24)	0.35 (0.24)	0.58 (0.18)
6. Do up five buttons	0.32 (0.18)	0.33 (0.21)	0.43 (0.22)	0.51 (0.21)	0.81 (0.22)	0.52 (0.20)
7. Dry back with towel	0.29 (0.16)	0.30 (0.18)	0.71 (0.19)	1.02 (0.27)	0.97 (0.35)	0.26 (0.30)

a, item discrimination parameter; $b_{1\text{-}6}$, item difficulty parameters

Table 2. Rasch analysis: CAHAI-7 item locations, fit-residual and chi-square statistics ordered by item location.

Item	Location	SE	Fit residual	*df*	x^2	*p*-value
1. Open jar of coffee	-0.792	0.108	0.137	2	2.06	.358
3. Draw a line with a ruler	-0.505	0.105	1.105	2	0.86	.651
5. Wring out washcloth	-0.355	0.105	0.610	2	3.31	.191
2. Call 911	-0.193	0.100	-0.730	2	1.56	.458
6. Do up five buttons	0.297	0.096	-0.667	2	0.95	.623
7. Dry back with a towel	0.504	0.098	-0.599	2	1.36	.506
4. Pour a glass of water	1.044	0.106	-0.943	2	3.94	.140

SE: standard error; df: degrees of freedom

Note: Significant *p*-value (Bonferroni adjusted) set at $p < 0.0014$

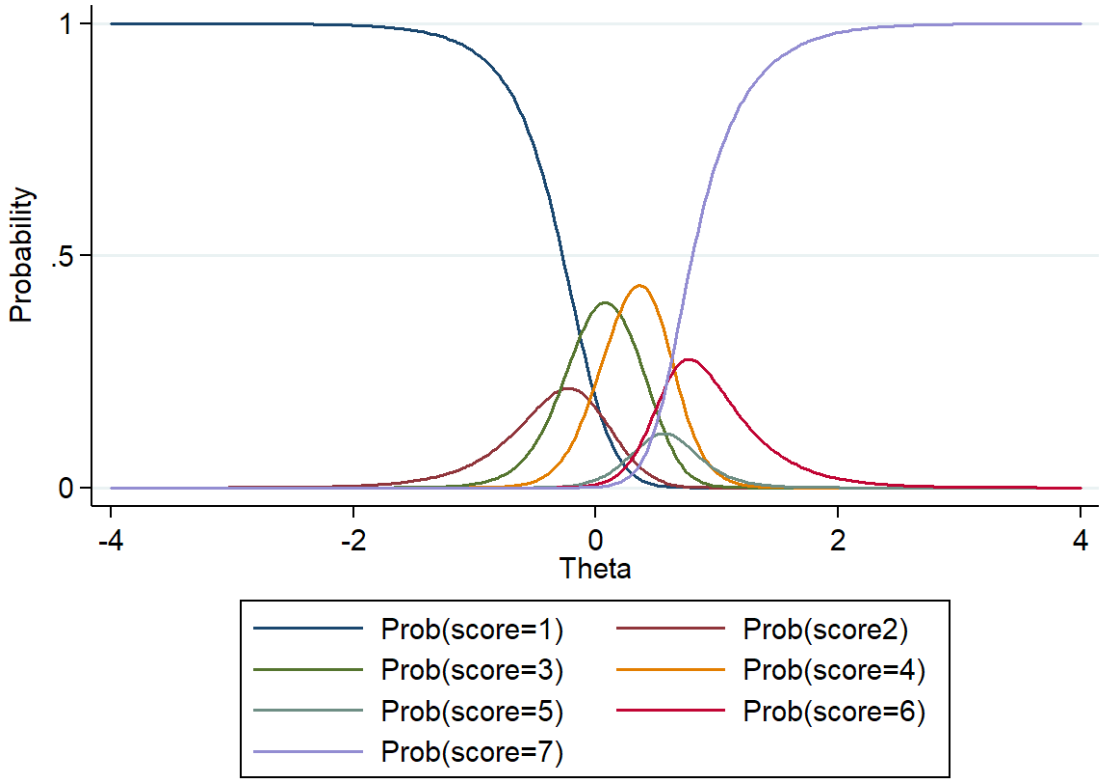


Fig 1. Category characteristic curves of item 5 (wring out washcloth) in CAHAI-7. The individual curves indicate the probability of obtaining the score category.

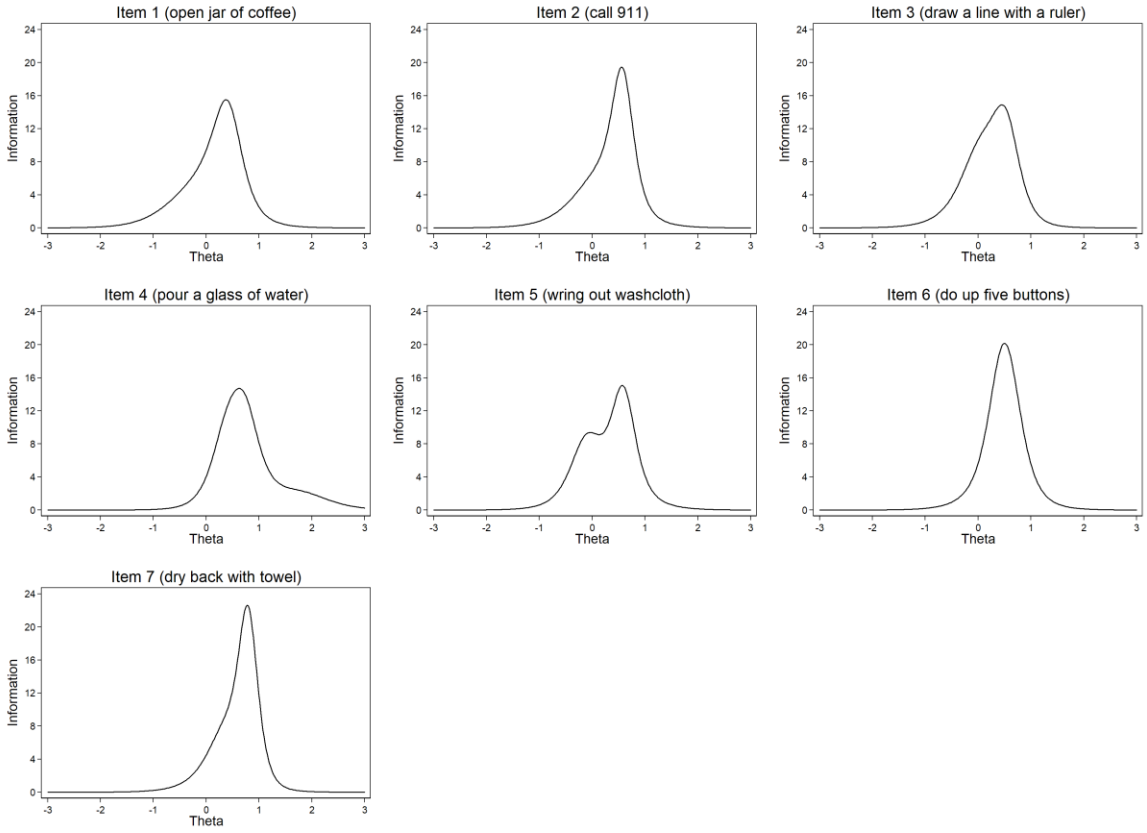


Fig 2. Item information functions of the CAHAI-7. A taller peak indicates more information provided by the item, and a non-unimodal peak suggests inconsistent performance of the item across the latent trait continuum.

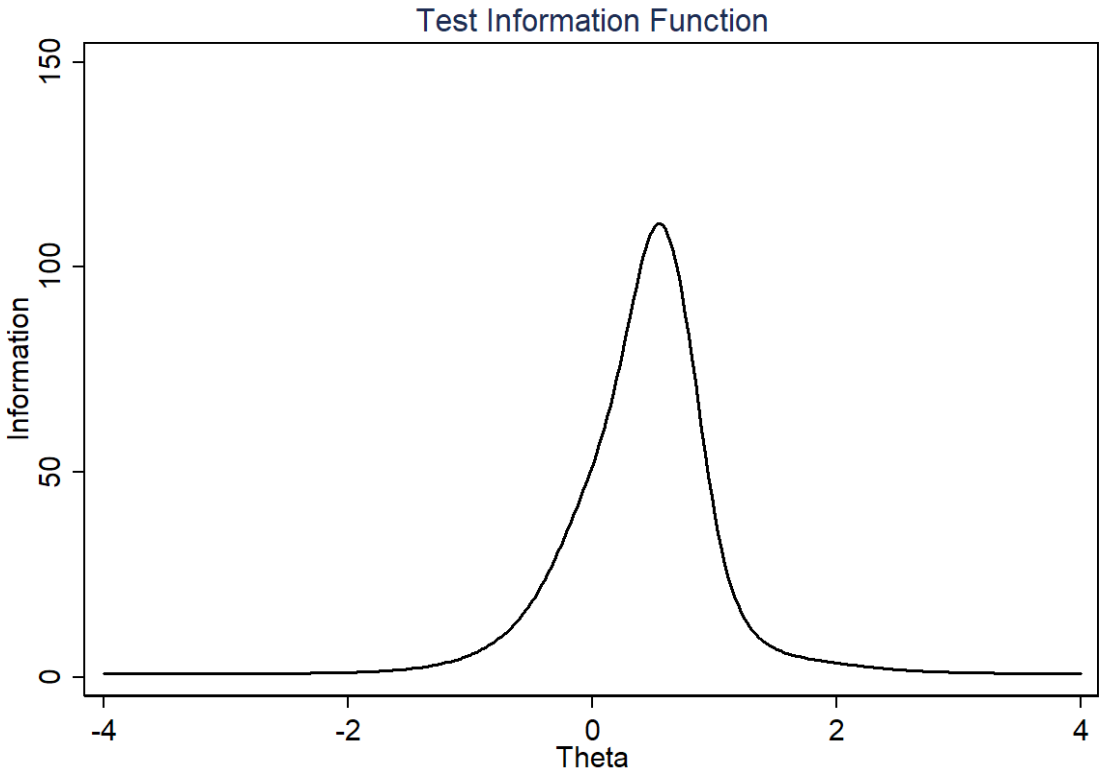


Fig 3. Test information function of the CAHAI-7. The test information function is the sum of all item information functions.

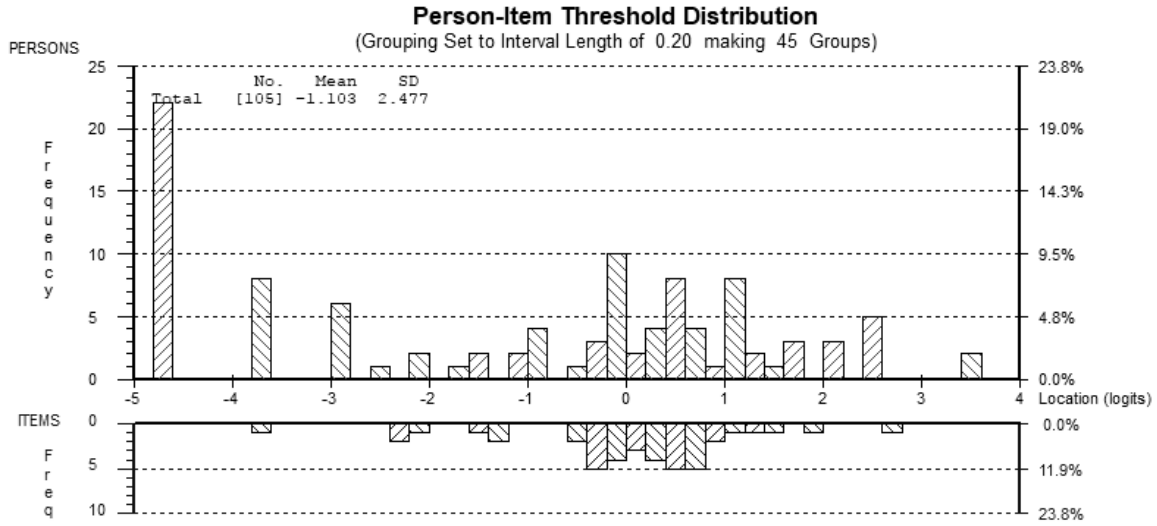


Fig 4. Person-item threshold distribution of the CAHAI-7. This figure shows the relative distributions of item and person locations. The mismatch between the range of upper extremity function measured by the CAHAI and the participants' range of upper extremity function suggests inadequate scale-to-sample targeting.

Chapter Five:

Revising the Chedoke Arm and Hand Activity Inventory using modern test theories

Preface

Following the findings in Chapter 4 of this thesis, the Chedoke Arm and Hand Activity Inventory was revised using item response theory and Rasch measurement theory. This chapter describes the revision process which used data from a previous validation study of the measure. The revised versions of the Chedoke Arm and Hand Activity Inventory had improved psychometric properties and clinical utility. Thus, this chapter demonstrates the importance of continued evaluation of outcome measures beyond its initial development and validation phase to improve the quality of existing measures. This is also the first study to demonstrate the novel use of both item response theory and Rasch measurement theory to revise an existing outcome measure.

This manuscript was prepared according to the submission guidelines of the *Journal of Clinical Epidemiology*.

Title Page

Revising the Chedoke Arm and Hand Activity Inventory using modern test theories

Silvana X. Choo,[1,2] Ayse Kuspinar,[1] Julie Richardson,[1] Jackie Bosch,[1] Jocelyn E. Harris,[1] and Paul Stratford[1]

[1]School of Rehabilitation Science, McMaster University, Hamilton, Ontario, Canada;

[2]Department of Occupational Therapy, Singapore General Hospital, Singapore

Corresponding Author

Silvana Choo

Department of Occupational Therapy, Singapore General Hospital, Outram Road, Singapore 169608. Telephone: +65 6326 5325. Email: silvana.choo.xinyi@sgh.com.sg

Declarations of Interest

None.

Acknowledgments

We appreciate the generosity and thank Susan Barreca in allowing the use of the data from her study published in 2006. The first author is supported by the Singapore General Hospital scholarship award.

Revising the Chedoke Arm and Hand Activity Inventory using modern test theories

Abstract

Objective: Building on findings from previous evaluation studies of the Chedoke Arm and Hand Activity Inventory (CAHAI), this study aimed to use item response theory and Rasch analyses to revise the measure.

Study design and setting: Secondary analysis was conducted using data from 105 participants in a previous validation study. Rasch analysis was used to evaluate an 11-item version and to revise the 7-category scoring scale. We then used item response theory to guide item selection for a new short version. The reliability and validity of the new shortened versions were compared to an original shortened version (CAHAI-7).

Results: Collapsing the 7-category scale to four categories resolved the disordering and improved the psychometric properties of the 11-item CAHAI. Two new shortened versions containing five and seven items were developed. The test-retest reliability (ICC = 0.96), construct validity (r = 0.87 – 0.90), and longitudinal validity (area under receiver operating characteristic curve = 0.75) of the new 5-item version was not inferior to the CAHAI-7.

Conclusion: This study demonstrated the novel use of both item response theory and Rasch analyses to revise outcome measures. The original CAHAI was refined into a new 11- and 5- item versions with a 4-category scoring scale.

Keywords: stroke; upper limb; evaluation; measurement; item response theory; Rasch

1. Introduction

Stroke is one of the leading causes of disability worldwide.[1] In Canada, stroke prevalence is estimated at 1.15% and about 405,000 Canadians are living with the effects of stroke.[2] A common effect of this condition is upper extremity (UE) deficits; approximately 70 – 80% of individuals experience UE hemiparesis in the acute phase of stroke.[3,4] Post-stroke UE deficits also remain beyond the initial stroke onset. About 1 in 4 individuals with severe hemiparesis regain full UE function 3 months post-stroke,[3] and 50% of individuals with UE hemiparesis do not have functional use of their UE 4 years post-stroke.[5] Post-stroke UE impairments reduce one's ability to perform routine self-care tasks and result in a loss of independence in daily living.[6,7] Consequently, the emotional and psychological well-being and quality of life of individuals with stroke are affected.[8,9] Therefore, assessment of UE function is critical for evaluating the severity of UE deficits to facilitate clinical decisions about appropriate interventions and to determine the prognosis of UE recovery.[10]

There is currently a range of outcome measures available to assess post-stroke UE function,[11] such as the widely used Fugl-Meyer Assessment and the Action Research Arm Test (ARAT),[12] and the Chedoke Arm and Hand Activity Inventory (CAHAI). The CAHAI is an outcome measure that uses real-life daily tasks to assess function in the affected UE after a stroke.[13] It is a recommended outcome measure to assess post-stroke UE function as it demonstrates high levels of clinical utility and measurement quality.[11] The use of real-life daily tasks, low cost, and brief administration time are qualities the CAHAI possesses that supports its clinical utility. A task-related evaluation approach is

suitable for individuals with post-stroke cognitive or communication impairments as familiar daily tasks are intuitive and demand less advanced cognitive or communication skills.[13] The low cost of the CAHAI administration is attributed to inexpensive materials required and the administration manual and training video are available online (www.cahai.com) at no charge. The CAHAI has a full 13 item version (CAHAI-13), and also three shortened versions with seven, eight, and nine items (CAHAI-7, CAHAI-8, and CAHAI-9 respectively). The shorter CAHAI versions are faster to administer (e.g., the 7-item version requires about 12 minutes compared to the 13-item version that takes about 25 minutes) and are beneficial to clinicians and researchers when limited time is an issue.

In addition to clinical utility, a measure must have adequate psychometric properties. The psychometric properties of all CAHAI versions were evaluated using classical test theory.[14,15] All versions of the CAHAI have good test-retest reliability (intraclass-correlation coefficient (ICC) 0.96 to 0.98),[14,15] inter-rater reliability (ICC = 0.98),[14] and internal consistency of coefficient alpha = 0.97 to 0.98.[13,15] All versions of the CAHAI also have high convergent validity with the Chedoke-McMaster Stroke Assessment (CMSA) arm and hand subscales (r = 0.81 to 0.87) and with the ARAT (r = 0.93 to 0.95),[14] and longitudinal validity with area under the receiver operating characteristics curve of 0.93 to 0.97.[14] Newer measurement theories have expanded the considerations for whether an instrument is performing as expected. To address the considerations from newer measurement theories, the psychometric properties of all CAHAI versions were evaluated using item response theory and Rasch analyses.[16]

The Rasch analysis of the CAHAI identified several issues.[16] First, 2 of the 13 items did not fit the Rasch model (Item 12, place container on table; and Item 13, carry bag up the stairs). Second, there were disordered thresholds for almost all items, indicating the CAHAI's scoring scale was not working as intended. Third, differential item functioning was present in the CAHAI-13, CAHAI-9, and CAHAI-8, which indicates the measures do not perform consistently when used across different groups of individuals (e.g., infarct versus hemorrhagic stroke types). Fourth, there was violation of the assumptions of unidimensionality and local independence, which was a similar finding in the item response theory analysis of the CAHAI. The study results indicated that the CAHAI does not assume interval-level measurement characteristics, and so, summing individual item scores to obtain a total score is inappropriate. Therefore, there is a need to revise the CAHAI, including its scoring scale, to improve its psychometric properties by ensuring unidimensionality and interval scaling.

There is also a need to re-examine the current shortened versions of the CAHAI as the item reduction process used only a qualitative approach.[15] Four experienced clinicians were asked to select nine items from the full 13-item version by ranking items according to three criteria related to content. Subsets of these nine items were also used to develop the 7- and 8-item versions. Methodological guidelines to shorten outcome measures recommend both qualitative content analysis and quantitative psychometric analysis.[17] While the approach used preserved the content of the CAHAI,[15] there is also a need to employ quantitative psychometric analysis in the reduction process.[17] For example, using item response theory analysis allows precise evaluation of each test item

and provides objective statistical information that can be used to select items for the shortened versions.[17]

This study builds on findings from the previous study that evaluated the CAHAI using item response theory (IRT) and Rasch analysis.[16] The aim was to use IRT and Rasch analysis to revise the CAHAI. The objectives were as follows:

(1) To evaluate the psychometric properties of an 11-item version of the CAHAI using Rasch analysis;
(2) To revise the current 7-category scoring scale of the CAHAI using Rasch analysis;
(3) (a) To develop a shortened version of the CAHAI using quantitative psychometric analysis (i.e., IRT) and (b) compare its test-retest reliability, cross-sectional validity, and longitudinal validity to one of the shortened versions advocated previously by Barreca et al.[15]

Both IRT and Rasch analyses were used to address the limitations of the other's analysis approach. When revising existing outcome measures, Rasch analysis can be applied to revise the measure's scoring scale to ensure it is working as intended;[18] however, there are no guidelines for the application of Rasch analysis in the item selection process when shortening outcome measures. In contrast, IRT analysis has been applied as quantitative approach to methodically shorten outcome measure.[17] Thus, both IRT and Rasch analyses were used in complementary in this study.

2. Methods

2.1 Study design and sample

This is a secondary analysis of data from a previous validation study of the CAHAI.[14,19] Briefly, the study recruited 105 adults with first-ever stroke and a combined score on the Chedoke-McMaster Stroke Assessment[20] (CMSA) arm and hand subscales of either $\leq$ 5 or between 7 and 11. These CMSA scores meant that participants had either severe (flaccid paralysis or some synergistic movements) or mild to moderate (full range of movements within synergistic patterns) UE impairments.[20] All participants were assessed on the CAHAI, CMSA, and the Action Research Arm Test[21] (ARAT) at baseline (admission/ initial visit to rehabilitation program) and at discharge. In addition, the CAHAI was administered to 39 of the 105 participants within 36 hours from baseline assessment in order to estimate inter-rater reliability.[14]

2.2 Measures

2.2.1 Chedoke Arm and Hand Activity Inventory

The CAHAI is comprised of test items that use daily tasks to assess the function of the affected UE after a stroke.[13] The test items included in each version of the CAHAI are presented in Table 1. Task performance of the affected UE is scored on a 7-category scale, from total assistance (score = 1) to independence (score = 7). The range of total scores for the CAHAI-13, -9, -8, and -7 are 13 to 91, 9 to 63, 8 to 56, and 7 to 49 respectively, with higher scores indicating better UE function. The psychometric properties of all CAHAI versions were described earlier.

2.2.2. Chedoke-McMaster Stroke Assessment

The CMSA was used as a comparative measure of UE function in the evaluation of construct validity of the CAHAI.[14,15] It has two parts that measure post-stroke physical impairments and activity limitations.[20] The physical impairment inventory of the CMSA includes six subscales (stage of recovery of the arm, hand, leg, and foot; shoulder pain; and postural control), and only the arm and hand subscales were used in the study. Each subscale is scored on a 7-point scale that corresponds to Brunnstrom stages of motor recovery (1, flaccid paralysis; 2, presence of spasticity; 3, marked spasticity and synergistic movements may be elicited voluntarily; 4, spasticity reduces; 5, diminished spasticity with voluntary movements outside of synergistic patterns; 6, near normal movement patterns and coordination; and 7, normal movement). The minimum total score of the CMSA arm and hand subscales is 2 and the maximum score is 14. The inter-rater reliability of the CMSA arm and hand subscales are ICC = 0.88 and 0.93 respectively, and their concurrent validity with the Fugl-Meyer Assessment (upper extremity) is 0.95.[20]

2.2.3 Action Research Arm Test

The ARAT was also a comparative measure of UE function used in the construct validation of the CAHAI.[14,15] It measures impairment and activity levels of the affected UE after a stroke.[20] The ARAT consists of 19 test items categorized into four subscales (grasp, grip, pinch, and gross movement), and each item is scored on a 4-point scale where higher scores indicate better UE function. The range of total scores is 0 to 57. The

psychometric properties of the ARAT in stroke samples include: inter-rater reliability of ICC ranging from 0.92 – 0.99[21–23]; test-retest reliability, ICC = 0.99[22]; and concurrent validity of correlation coefficient of 0.86[24] with the Wolf-Motor Function Test and 0.87[25] with the Fugl-Meyer Assessment (upper extremity motor scale).

2.3 Procedure

Permission for data use was obtained from the corresponding author of the previous CAHAI validation study,[19] and all data were provided in a de-identified format. In this study, we included all 105 participants and created two datasets. The first dataset (D1) consisted of participants' demographic data and CAHAI scores at one measurement point, either baseline or discharge. This approach was used to optimize the frequency of scores in each score category of the CAHAI's 7-point scale for item-level analysis. The second dataset (D2) was the full data from the original study, which contained demographic data and scores on all outcome measures at baseline and discharge.

2.4 Statistical analysis

2.4.1 Participant characteristics

Descriptive statistics were used to summarize participant characteristics. For each dataset, scores on each outcome measure were summarized as medians and 1st and 3rd quartiles. In the first dataset (D1), the frequencies of scores in each score category of the CAHAI were summarized as percentages.

2.4.2 Evaluation of the CAHAI using Rasch analysis

All Rasch analyses were conducted using the Rasch Unidimensional Measurement Model software, RUMM2030,[27] and only the first dataset (D1) was used.

2.4.2.1 Rasch analysis of the 11-item version of the CAHAI. The first study objective was to evaluate the psychometric properties of an 11-item version of the CAHAI. We first deleted two items (item 12, place container on table; item 13, carry bag up the stairs) from the full 13-item CAHAI to create a new 11-item version. Our basis for deletion was the quantitative and qualitative evidence about the CAHAI-13. First, Rasch analysis of the CAHAI-13 identified these two items as not fitting with the Rasch model.[16] Second, these two items do not appear to measure a single latent trait of UE function as they were linked to two categories of the International Classification of Functioning, Disability and Health[28] related to body positions (d4105, bending) and moving around (d4551, climbing).[29] Clinicians have also expressed that the ability to stand is a preceding criterion for the assessment of UE function in these two items (Choo et al., unpublished data, 2015).

Five categories of analyses were conducted to evaluate the psychometric properties of the 11-item version of the CAHAI: fit, targeting, dependency, reliability, and stability. Fit with the Rasch model was assessed with the ordering of item scoring categories (ordering of item thresholds),[30] two statistical tests (fit residual and x^2 statistics), and one graphical indicator (item characteristic curves).[31] Ordering of item scoring categories were evaluated using the category probability curves of each CAHAI

item, where the order of the thresholds were examined. Thresholds are points at which the category probability curves of adjacent score categories intersect.[32,33] At these intersections, the probability of obtaining a score in two adjacent score categories is equal (e.g. probability of scoring 1 or 2 on a CAHAI test item is equal at 0.5).[32,33] A sequential order of the thresholds (i.e., 1-2 < 2-3 < 3-4 < 4-5 < 5-6 < 6-7) indicated the 7-category scoring scale was working as intended while disordered thresholds indicated the scoring scale was not working as intended. Fit residuals larger than $\pm$ 2.5 or a significant x^2 indicated that the item did not fit the Rasch model.[34,35]

Targeting refers to the match between the range of UE function evaluated in the CAHAI and participants' range of UE function. To assess targeting, histograms of the relative distributions of item locations (i.e., item difficulty) and person locations (i.e., participants' level of UE function) were graphically examined using the person-item threshold distribution.[36] A greater similarity between the distributions indicated better targeting. Dependency is related to the underlying assumption of local independence in Rasch analysis, where scores on each item should be independent of scores on other items.[37] Residual correlations exceeding $\pm$ 0.3 indicated violation of local independence assumption.[36] Reliability was determined using the Person Separation Index, which is the ratio of true-score variance to observed variance.[38] Stability, or the invariance property of the CAHAI,[39] was assessed across age, sex, and factors associated with the recovery of UE function after stroke (stroke chronicity,[40,41] baseline UE impairment,[42] and unilateral spatial neglect[43]). Items that displayed differential item functioning (DIF) indicated instability.[32]

2.4.2.2 Revision of the 7-category scoring scale. The second study objective was to revise the CAHAI's 7-category scoring scale. Following the psychometric evaluation of the 11-item version of the CAHAI using Rasch analysis, score categories were collapsed where disordered thresholds were found. Different scoring structures of the 11-item version were explored, where the original scale was collapsed into five, four, and three categories. For each new scoring structure, the 11-item CAHAI was re-evaluated using the five categories of Rasch analysis described earlier. The best scoring structure was determined as meeting all our predetermined criteria of fit with the Rasch model, targeting, local independence, reliability, and stability.

2.4.3 Development of new shortened versions of the CAHAI using IRT

All analyses in the following sections were conducted using STATA version 14.[44]

2.4.3.1 Fitting of IRT model to data. The next study objective was to develop a shortened version of the CAHAI using a statistical approach. The starting point of our IRT analysis was the 11-item version of the CAHAI with a revised 4-category scoring scale (following results from the Rasch analyses described above). In our preliminary selection of potential IRT models, we considered only nested unidimensional polytomous models. The generalized partial credit model[45] was fitted to the first dataset (D1) and compared against two nested models, the partial credit model[46] and rating scale model.[47] Relative model fit was compared using the likelihood ratio test. A significant result ($p < 0.05$) indicated that the generalized partial credit model was preferred as the simpler model (partial credit or rating scale model) decreased the model fit.[48,49]

2.4.3.2 Item properties. After identifying the best fit model, it was re-fitted to the same dataset (D1) and the item difficulty and discriminating parameters were estimated. Item properties were also graphically examined using the category characteristic curves, item characteristic curves, and item information functions (IIFs).

2.4.3.3 Selection of items for the new shortened versions. The IIFs were used to guide the item reduction process to develop the new shortened CAHAI version. IIFs display the contribution of each item along the latent trait continuum by plotting the item's information (y-axis) against the latent trait (x-axis).[33,50] Items with greater discriminating ability contribute more information, as indicated by a taller peak of the curve, and have smaller error variance (i.e., greater measurement precision).[49,51]

The goal of the item reduction process was to retain/select as few items as possible while maximizing item information and minimizing content overlap.[52] Two approaches were used. In the first approach ('backward elimination'), the IIFs of all 11 items were plotted and items with similar plots were identified. When two or more IIFs were identified as similar, we examined their content (i.e., the specific aspects of UE function assessed) and only retained the item that had minimal content overlap with other items. In the second approach ('forward selection'), starting with a blank plot, items with the highest information levels were selected and added to the plot. Thereafter, items that assessed UE function at lowest and highest θ values (i.e., lowest and highest item difficulty) were added next. Finally, IIFs of all added items were examined to identify

gaps along the latent trait continuum, and items with IIFs that covered the identified gaps were added.

2.4.3.4 Preliminary validation of the new shortened versions. IRT analysis was conducted for each new shortened version of the CAHAI using the first dataset (D1). The steps taken in our analysis are described accordingly in sections *2.4.3.1* and *2.4.3.2*. To assess the underlying IRT assumptions of unidimensionality and local independence, confirmatory factor analysis (one-factor model) was conducted for each new shortened version. Unidimensionality was indicated by the fit with the one-factor model and the following criteria were used: a non-significant x^2 statistic, root mean square error of approximation (RMSEA) ≤ 0.06, comparative fit index (CFI) ≥ 0.95, and Tucker-Lewis index (TLI) ≥ 0.95.[53,54] Local independence assumption was assessed by examining the residual correlation matrix. Residual correlations > 0.20 indicated the presence of local dependency between pairs of items.[55]

2.4.4 Comparison of the new CAHAI versions with the previous versions

The final study objective was to compare the test-retest reliability, cross-sectional validity, and longitudinal validity of the new shortened CAHAI versions to one of the shortened versions advocated previously by Barreca et al.[15] We hypothesized that the new shortened CAHAI versions were non-inferior to the CAHAI-7. The second dataset (D2) was used to estimate the reliability and validity of the new shortened CAHAI versions and to conduct comparative analyses with the CAHAI-7. As the psychometric properties

of the new CAHAI versions and the CAHAI-7 were derived from the same sample, sample dependency was accounted for in all comparative analyses between the measures.

2.4.4.1 Test-retest reliability. For each new shortened CAHAI version, 3-way analysis of variance was first computed with three factors: participants, raters, and occasion. The sources of variances identified were then used to estimate relative reliability with ICCs.[56] Absolute reliability was estimated with the standard error of measurement (SEM), calculated by taking the square root of the mean-square error.[57,58] The differences between the ICCs of the new shortened versions and the CAHAI-7 were calculated, and the 95% CIs of the differences were computed using bootstrapping with 1000 replacements.[59] Our predefined non-inferiority margin was 0.1.

2.4.4.2 Cross-sectional validity. Convergent validity was estimated with the correlation between the scores on the new short CAHAI versions and the CMSA (arm and hand) and ARAT scores. The correlations between the new short CAHAI versions scores and CMSA shoulder pain subscale scores were used to estimate discriminant validity. We expected higher correlations between the new CAHAI versions and the CMSA (arm and hand subscales) and the ARAT, compared to the correlations with the CMSA pain subscale. This is because measures that assess similar outcomes (i.e., UE function) should correlate highly, while measures that assess dissimilar outcomes (i.e., UE function and pain) should have low correlations.[60] All correlation coefficients of the new short CAHAI versions were compared with the respective correlation coefficients of the

CAHAI-7 using the method proposed by Meng et al.[61] of comparing correlation coefficients from dependent sample.

2.4.4.3 Longitudinal validity. Receiver operating characteristic (ROC) curves were computed for each new short CAHAI version to estimate longitudinal validity, which is the ability of an instrument to measure change.[62] Pairwise comparison of the area under the ROC curves between the new CAHAI short versions and the CAHAI-7 were conducted using Hanley and McNeil's method of comparing ROC curves from the same sample.[63]

3. Results

3.1 Participant characteristics

Participant characteristics are described in Table 2. The median age was 72 years (1^{st}, 3^{rd} quartile: 62, 78) and 51% (n = 54) were males. Most participants (82%, n = 78) had an ischemic stroke and the median time since stroke was 38 days (1^{st}, 3^{rd} quartile: 27, 80). Table 3 summarizes the scores on the outcome measures for each dataset. For the first dataset (D1), the frequency of the score categories in each item of the CAHAI-13 are shown in Figure 1. Score categories 1 (total assistance) and 7 (complete independence) were most frequently used by therapists.

3.2 Evaluation of the CAHAI using Rasch analysis

3.2.1 Rasch analysis of the 11-item version of the CAHAI

All 11 items had disordered thresholds, implying the 7-category scoring scale was not working as intended. For example, Figure 2A shows the category probability curves

of item 5 (wring out washcloth). The thresholds were disordered (1-2 < 2-3 < 3-4 < 6-7 < 4-5 < 5-6) indicating lesser UE function is required to obtain the highest score of 7 than lower score categories of 4 to 6. The fit residuals of all items were between $\pm$ 2.5 and no items had a significant x^2 values, indicating that all items fitted with the Rasch model (Table 4). The item characteristic curves supported the statistical results and Figure 3A shows an example of an item characteristic curve for item 4, pour a glass of water. The plots (observed scores) lie close to the curve (predicted scores), again indicating fit with the Rasch model.

The mean person location was -1.096 logits (*SD*: 2.467) and Figure 4A shows the person-item threshold distribution. In general, the relative distributions of the item and person locations were similar between $\theta = \pm 2$ logits, indicating adequate targeting of the 11-item CAHAI. This means between $\theta = \pm 2$ logits, the range of UE function measured by the CAHAI items matched participants' range of UE function. However, there were no test items that evaluated individuals with minimal function in the affected UE (i.e., $\theta < -4$ logits).

Two pairs of items had residual correlations that slightly exceeded $\pm$ 0.30 (item-pairs 3-5, and 2-7 had residual correlation of -0.358 and -0.330 respectively), violating the assumption of local independence. For reliability, the Person-Separation Index of the 11-item CAHAI was 0.94. In terms of stability, items 2, 3, 4, 8, and 9 exhibited non-uniform DIF across baseline UE impairment levels (mild-moderate or severe impairment).

3.2.2 Revision of the 7-category scoring scale

As all items in the 11-item CAHAI had disordered thresholds, the 7-category scoring scale was revised by collapsing into five, four, and three categories. No 5-category scoring scale resolved the disordered thresholds. Only one 4-category scoring scale resulted in ordered thresholds for all 11 items. In this 4-category scale, category 1, total assistance in the original scale was kept as the new category 1. The original score category 2, maximal assistance, and category 3, moderate assistance, were combined into the new category 2. Original categories 4-6 were combined, forming new category 3, and the original category 7, complete independence, was rescored into the new category 4.

Six 3-category scoring scales resulted in ordered threshold for all 11 items. In the best 3-category scale, the original category 1, total assistance was kept as the new category 1. The original categories 2-6, moderate assistance to modified independence were combined to form the new category 2, and the original category 7, complete independence, was rescored into the new category 3.

Between the 4- and 3-category scoring scales, we selected the former scoring structure as it was the only scoring scale that met all our predetermined criteria of fit with the Rasch model, targeting, local independence, reliability, and stability. All items in the 11-item CAHAI version with a 4-category scale had ordered score categories, indicating that the scoring scale was working as intended. For example, Figure 2B shows the category probability curves of item 5 (wring out washcloth) where the thresholds were ordered. All items also had a good fit with the Rasch model as no items had a significant

x^2 value and the fit residuals were between $\pm$ 2.5 (Table 4). Figure 3B shows the item characteristic curve of item 4, pour a glass of water; there was a flattening of the slope with the revision from the 7- to 4-category scale. The collapsing of score categories to a 4-category scoring scale resulted in changes to the targeting of the 11-item CAHAI. Figure 4B shows the person-item threshold distribution, where items are located between $\theta = \pm 5$ logits. In contrast, the locations of items ranged from θ = -4 logits to +2.5 logits when the 7-category scoring scale was used (Figure 4A). This indicates that with a 4-category scoring scale, the test items measure a relatively wider range of UE function, although not consistently across the latent trait continuum. For example, there are no test items located between θ = -3 to -1 logits, indicating that there areas along the UE function continuum that is not measured by the CAHAI items. The location of items (i.e., item difficulty parameters) changed because the probabilities of obtaining scores in each score category that were used to compute the locations changed when the score categories were collapsed. There was good reliability with a Person-Separation Index of 0.95. DIF was not present in any items in the 11-item CAHAI with a revised 4-category scoring scale.

3.3 Development of the new shortened versions of the CAHAI using IRT

3.3.1 Fitting of IRT model and item properties

The likelihood ratio test between the generalized partial credit model and the rating scale model was significant ($x^2(30) = 126.62$, $p < 0.001$), but was non-significant with the partial credit model. Thus, the partial credit model was identified as the best fit model and Table 5 shows the item discrimination and difficulty parameters. Figure 5 shows the IIFs of all 11 items. Most IIFs were non-unimodal, indicating the item's

maximum information peaked at different locations along the latent trait continuum. There were overlaps between IIFs, with some items having similar IIFs (i.e., width, height, or shape of peaks). For example, the IIFs of items 1 and 8 overlapped, and the IIFs of items 2 and 7 had similar width and peak heights.

3.3.2 Selection of items for the new shortened versions

Two new shortened CAHAI versions containing seven and five items were developed through the item reduction process. From the first approach ('backward elimination'), we identified similar/overlapping IIFs between items 1 and 8, items 2 and 7, and items 4, 5 and 9. Based on the specific UE functions assessed in these items,[29] we removed items 2, 5, 8 and 9 due to content overlap with other items. Thus, a new 7-item CAHAI was developed, which contained item 1 (open jar of coffee), item 3 (draw a line with ruler), item 4 (pour a glass of water), item 6 (do up five buttons), item 7 (dry back with towel), item 10 (zip up the zipper), and item 11 (clean a pair of eyeglasses).

Using the second approach ('forward selection'), item 7 (dry back with towel) was selected first as it had the highest information level. Next, item 10 (zip up the zipper) and item 4 (pour a glass of water) with the lowest and highest item difficulty values respectively were then selected. Lastly, item 3 (draw a line with ruler) and item 11 (clean a pair of eyeglasses) were selected to address the gaps along the latent trait continuum. Accordingly, a new 5-item CAHAI was developed that contained these five items.

3.3.3 Preliminary validation of the new shortened versions

For both new 7-item and 5-item CAHAIs, the likelihood ratio tests between the generalized partial credit model and the rating scale model were significant ($x^2(18) = 110.04$, $p < 0.001$, $x^2(12) = 76.12$, $p < 0.001$ respectively), but were non-significant with the partial credit model. The partial credit model was thus the best fit model for both new versions of the CAHAI and Table 5 shows their item parameter estimates. The item discrimination parameter for the new 7-item and 5-item CAHAI are 3.71 logits and 3.67 logits respectively. Figure 6 shows the item and test information functions (sum of all IIFs)[50] for both new versions of the CAHAI. The peak heights of the test information functions of both the 7-item and 5-item CAHAI were similar.

Confirmatory factor analysis indicated that the new 7-item CAHAI violated the unidimensional assumption as it did not meet all predefined fit criteria. The new 7-item CAHAI had a significant x^2 ($p = 0.03$) and RMSEA of 0.09, but met the criteria for CFI and TLI ≥ 0.95 (CFI = 0.99, TLI = 0.98). The new 7-item CAHAI met the local independence assumption as no residual correlations were > 0.20. In contrast, the new 5-item CAHAI had a non-significant x^2 statistics, RMSEA of 0.050, CFI = 0.99, TFL = 0.99, and no residual correlations > 0.20. These indicated that the new 5-item CAHAI met both assumptions of unidimensionality and local independence.

3.4 Comparison of the new CAHAI versions with the CAHAI-7

The reliability and validity of the new 7-item and 5-item CAHAI were compared with the CAHAI-7 advocated previously by Barreca et al.[15] with the revised 4-category scoring scale. The psychometric properties of these three versions of the CAHAI are summarized in Table 6.

3.4.1 Test-retest reliability

As the rater variance was zero, the 3-way analysis of variance was reduced to 2-way with two factors (participants and occasion). The test-retest reliability of the new 7-item and 5-item CAHAI were similar with ICC = 0.96, and their SEMs were 1.20 and 0.83 CAHAI points respectively (Table 6). The differences between ICCs of the CAHAI-7 and the new 7-item and 5-item versions were 0.004 (95% CI: -0.01, 0.02) and 0.004 (95% CI: -0.02, 0.03) respectively. As both 95% CIs fall within our predefined inferiority margin of 0.1, the test-retest reliability of the new CAHAI versions were not inferior to the CAHAI-7.

3.4.2 Cross-sectional validity

The convergent validity of the new 7-item CAHAI was 0.91 with the ARAT and 0.88 with the CMSA arm and hand subscales, and its discriminant validity with the CMSA pain subscale was 0.57 (Table 6). Similarly, the convergent validity of the new 5-item CAHAI with the ARAT and CMSA arm and hand subscales was 0.90 and 0.87 respectively, and its discriminant validity was 0.57 with the CMSA pain subscale. The higher correlations with the ARAT and CMSA arm and hand subscales and lower correlations with the CMSA pain subscale support the construct validity of the new CAHAI versions.

There were no significant differences between correlations of the new CAHAI versions and the CAHAI-7 with the CMSA subscales (arm and hand, and pain). This implies that the construct validity of the new CAHAI versions with the CMSA were not

inferior to the CAHAI-7. However, the correlation of the CAHAI-7 with the ARAT was significantly higher than the correlations between the 7-item (p = 0.02, 1-sided) and 5-item (p = 0.03, 1-sided) CAHAI with the ARAT.

3.4.3 Longitudinal validity

The areas under the ROC curves are as follows: CAHAI-7, 0.81 (95% CI: 0.73, 0.90); new 7-item CAHAI, 0.78 (95% CI: 68, 0.87); and new 5-item CAHAI, 0.75 (95% CI: 0.65, 0.85). The differences between the areas under the ROC curves of the CAHAI-7 and both new CAHAI versions were not statistically significant. This indicates that the sensitivity to change of the new CAHAI versions were not inferior to the CAHAI-7.

4. Discussion

Building on findings from a previous study on the psychometric properties of the CAHAI using modern test theories (IRT and Rasch measurement theory), this study aimed to revise the CAHAI using these theories. The use of modern test theories allows a more detailed evaluation of the psychometric properties of an outcome measure, particularly at the item-level. This facilitates the identification of potential issues within the measure and appropriate revisions can be made to improve it. Three new versions of the measure were developed in this study: new full 11-item version, and two new shortened versions comprising of five and seven items. The scoring scale of the CAHAI was also revised from a 7-category to a 4-category scale.

Using Rasch analysis, the 11-item CAHAI with a 4-category scoring scale can be used as the revised full version of the CAHAI. Deleting two items in the CAHAI-13 that

did not fit the Rasch model improved the overall psychometric properties of the full version of the measure. With the two misfit items deleted, all items in the 11-item CAHAI demonstrated good fit to the Rasch model. There was a slight improvement in the targeting of the measure, where the items now evaluate a wider range of UE function, particularly at the previously identified gap between θ = -3 to -5 logits. This means the 11-item CAHAI is able to evaluate individuals with lesser UE function when compared to the CAHAI-13. Although local dependency was still present in the 11-item CAHAI, there was a reduction from 3 to 2 pairs of items with residual correlations exceeding the predefined criterion. Stability of the measure also improved, where non-uniform DIF is only present for baseline UE impairment levels in the 11-item CAHAI and no longer present for individuals with different stroke types. The reliability of the 11-item CAHAI was also identical to the CAHAI-13, with a Person-Separation Index of 0.94.

The revision of the scoring scale by collapsing the 7-category scale to a 4-category scale further improved the psychometric properties of the 11-item CAHAI. In particular, the problems with disordered thresholds and items exhibiting non-uniform DIF were fully resolved. The revision of the scoring scale to a 4-category scale was similar to previous studies that revised the scoring scale of the Functional Independence Measure.[64–66] The CAHAI's original scoring scale was based on the same 7-category scoring scale of the Functional Independence Measure.[67,68] Disordered thresholds were also identified in the 7-category scale of the Functional Independence Measure, and the collapsing from seven to five or four categories resolved the problem of disordered thresholds.[64–66]

Following the Rasch analysis of the 11-item CAHAI, IRT analysis was applied in the item reduction process to develop new shortened CAHAI versions. The use of a quantitative psychometric analysis in this study was intended as a follow-up to the qualitative approach originally used to shorten the CAHAI. The application of qualitative and quantitative approaches as two separate steps to shorten a measure is not uncommon. In a review of the current methodology used to shorten outcome measures,[17] 58% of the cases combined both qualitative and quantitative approaches as two distinct steps to shorten an outcome measure.

There were similar results in the items included in the shortened version of the CAHAI using a quantitative psychometric analysis and a qualitative content analysis. Most of the items in the new 5-item and 7-item CAHAI (three and five items respectively) were the same items included in the CAHAI-7. This suggests that the new shortened CAHAI versions retained the content of the measure as intended by the original expert panel. Furthermore, the new shortened CAHAI versions also maintained similar reliability and validity as the original shortened CAHAI version; the formal comparison of psychometric properties between the new shortened versions with the CAHAI-7 did not yield statistically significant differences (i.e., not inferior). In terms of clinical utility, the further reduction of administration time with the 5-item CAHAI is beneficial to clinicians, researchers, and individuals with stroke.

4.1 Implication of study findings

Results of this study suggest that the original full 13-item version of the CAHAI can be replaced by the 11-item version with the 4-category scoring scale. When

administration time is of concern, the new 5-item CAHAI with the 4-category scale can be used instead. The new 5-item CAHAI is recommended as it has the strongest psychometric properties among all the shortened versions of the CAHAI when evaluated using both classical and modern test theories. With the revisions to the CAHAI, there is also a need to revise the administration manual and both clinicians and researchers need to familiarize themselves with the new scoring scale.

4.2 Study limitations

Our sample size (n = 105) was sufficient for both Rasch and IRT analysis; a minimum of 50 participants is recommended for polytomous models in Rasch analysis,[69] and 66 participants for IRT analysis using the 2:1 ratio of person to item parameters when fitting the partial credit model.[32] However, the distribution of observations across the score categories in our sample was not uniform and there were less than 10 observations in score categories of some items. This may have affected the stability of our estimations and future studies is recommended to validate our study findings in a larger sample with a sufficient and uniform distribution of observations across each score category.[32,69] The second limitation was the use of data derived from the same sample of the original validation study of the CAHAI in our study. Our study conducted the calibration and validation of the new CAHAI versions using the same sample and the study results may have limited generalizability. Thus, there is a need for further psychometric evaluation of the new CAHAI versions in a different sample of individuals with stroke to ensure the validity of our findings.

5. Conclusion

This study demonstrated the novel use of both IRT and Rasch analyses in the revision of outcome measures. We used Rasch analysis to help us revise the CAHAI's scoring scale, and employed IRT analysis to guide the item reduction process to develop new shortened versions of the measure. The new versions of the CAHAI developed through these processes not only demonstrated interval scaling properties, they also had good reliability and cross-sectional and longitudinal validity. Our findings support the notion that the refinement of outcome measures beyond its initial development and validation phase is essential to improving the measurement quality of existing measures.

What is New?

Key findings

- The new 5-item and 11-item versions of the CAHAI have interval-level scaling characteristics.

What this adds to what was known?

- Previously identified issues with disordered score categories, and violation of unidimensionality, local independence and invariance assumptions of the CAHAI were resolved with the new 11-item version and a 4-category scoring scale.
- The original short versions were developed using qualitative analysis and this study employed quantitative psychometric analysis to develop a new short CAHAI.

What is the implication and what should change now?

- Individual items scores in the CAHAI can be summed to create a total score
- The revised full (11-item) and short (5-item) CAHAI with a 4-category scoring scale may be used.

References

1. Feigin VL, Norrving B, Mensah GA. Global burden of stroke. *Circ Res*. 2017;120(3):439-448. doi:10.1161/CIRCRESAHA.116.308413

2. Krueger H, Koot J, Hall RE, O'Callaghan C, Bayley M, Corbett D. Prevalence of individuals experiencing the effects of stroke in Canada: trends and projections. *Stroke*. 2015;46(8):2226-2231. doi:10.1161/STROKEAHA.115.009616

3. Nakayama H, Jørgensen HS, Raaschou HO, Olsen TS. Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. *Arch Phys Med Rehabil*. 1994;75(4):394-398.

4. Lawrence ES, Coshall C, Dundas R, et al. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. *Stroke*. 2001;32(6):1279-1284.

5. Broeks J., Lankhorst GJ, Rumping K, Prevo AJH. The long-term outcome of arm function after stroke: results of a follow-up study. *Disabil Rehabil*. 1999;21(8):357-364. doi:10.1080/096382899297459

6. Lai S-M, Studenski S, Duncan PW, Perera S. Persisting consequences of stroke measured by the Stroke Impact Scale. *Stroke*. 2002;33(7):1840-1844.

7. Sveen U, Bautz-Holter E, Sødring KM, Wyller TB, Laake K. Association between impairments, self-care ability and social activities 1 year after stroke. *Disabil Rehabil*. 1999;21(8):372-377.

8. Wyller TB, Sveen U, Sødring KM, Pettersen AM, Bautz-Holter E. Subjective well-being one year after stroke. *Clin Rehabil*. 1997;11(2):139-145. doi:10.1177/026921559701100207

9. Nichols-Larsen DS, Clark PC, Zeringue A, Greenspan A, Blanton S. Factors influencing stroke survivors' quality of life during subacute recovery. *Stroke J Cereb Circ*. 2005;36(7):1480-1484. doi:10.1161/01.STR.0000170706.13595.4f

10. Lang CE, Bland MD, Bailey RR, Schaefer SY, Birkenmeier RL. Assessment of upper extremity impairment, function, and activity following stroke: Foundations for clinical decision making. *J Hand Ther Off J Am Soc Hand Ther*. 2013;26(2):104-115. doi:10.1016/j.jht.2012.06.005

11. Murphy MA, Resteghini C, Feys P, Lamers I. An overview of systematic reviews on upper extremity outcome measures after stroke. *BMC Neurol*. 2015;15(1):292. doi:10.1186/s12883-015-0292-6

12. van Wijck FM, Pandyan AD, Johnson GR, Barnes MP. Assessing motor deficits in neurological rehabilitation: patterns of instrument usage. *Neurorehabil Neural Repair*. 2001;15(1):23-30. doi:10.1177/154596830101500104

13. Barreca SR, Gowland CK, Stratford P, et al. Development of the Chedoke Arm and Hand Activity Inventory: theoretical constructs, item generation, and selection. *Top Stroke Rehabil*. 2004;11(4):31-42.

14. Barreca SR, Stratford PW, Lambert CL, Masters LM, Streiner DL. Test-retest reliability, validity, and sensitivity of the Chedoke Arm and Hand Activity Inventory: a new measure of upper-limb function for survivors of stroke. *Arch Phys Med Rehabil*. 2005;86(8):1616-1622. doi:10.1016/j.apmr.2005.03.017

15. Barreca SR, Stratford PW, Masters LM, Lambert CL, Griffiths J, McBay C. Validation of three shortened versions of the Chedoke Arm and Hand Activity Inventory. *Physiother Can*. 2006;58(2):148-156. doi:10.3138/ptc.58.2.148

16. Choo SX, Stratford P, Richardson J, Harris JE, Bosch J, Kuspinar A. Measurement theories in rehabilitation: Introduction to item response theory and Rasch measurement theory. [Doctoral thesis]. Hamilton, ON: McMaster University; 2018.

17. Goetz C, Coste J, Lemetayer F, et al. Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. *J Clin Epidemiol*. 2013;66(7):710-718. doi:10.1016/j.jclinepi.2012.12.015

18. Tennant A, Conaghan PG. 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 Rheum*. 2007;57(8):1358-1362. doi:10.1002/art.23108

19. Barreca SR, Stratford PW, Masters LM, Lambert CL, Griffiths J. Comparing 2 versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. *Phys Ther*. 2006;86(2):245-253.

20. Gowland C, Van Hullenaar S, Torresin W, et al. *Chedoke-McMaster Stroke Assessment: Development, Validation and Administration Manual*. Hamilton, Ontario: Chedoke-McMaster Hospitals and McMaster University; 1995.

21. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *Int J Rehabil Res Int Z Für Rehabil Rev Int Rech Réadapt*. 1981;4(4):483-492.

22. Hsieh CL, Hsueh IP, Chiang FM, Lin PH. Inter-rater reliability and validity of the action research arm test in stroke patients. *Age Ageing*. 1998;27(2):107-113.

23. Hsueh I-P, Lee M-M, Hsieh C-L. The Action Research Arm Test: is it necessary for patients being tested to sit at a standardized table? *Clin Rehabil*. 2002;16(4):382-388.

24. Nijland R, van Wegen E, Verbunt J, van Wijk R, van Kordelaar J, Kwakkel G. A comparison of two validated tests for upper limb function after stroke: The Wolf Motor Function Test and the Action Research Arm Test. *J Rehabil Med*. 2010;42(7):694-696. doi:10.2340/16501977-0560

25. Nijboer TCW, Ten Brink AF, Kouwenhoven M, Visser-Meily JMA. Functional assessment of region-specific neglect: are there differential behavioural consequences of peripersonal versus extrapersonal neglect? *Behav Neurol*. 2014;2014:526407. doi:10.1155/2014/526407

26. Rabadi MH, Rabadi FM. Comparison of the action research arm test and the Fugl-Meyer assessment as measures of upper-extremity motor weakness after stroke. *Arch Phys Med Rehabil*. 2006;87(7):962-966. doi:10.1016/j.apmr.2006.02.036

27. RUMM Laboratory. *RUMM 2030*. Perth, Australia: RUMM Laboratory; 1998.

28. World Health Organization. *International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization; 2001.

29. Choo SX, JN Ng C, Fayed N, Harris JE. International Classification of Functioning, Disability and Health Framework: Bridging adapted outcome measures. *Int J Ther Rehabil*. 2017;24(11):494-500. doi:10.12968/ijtr.2017.24.11.494

30. Hagquist C, Andrich D. Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personal Individ Differ*. 2004;36(4):955-968.

31. Wright BD, Masters GN. *Rating Scale Analysis: Rasch Measurement*. 1 edition. Chicago: MESA; 1982.

32. de Ayala R. *The Theory and Practice of Item Response Theory*. New York: Guilford Press; 2009.

33. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of Item Response Theory*. Newbury Park, Calif: Sage Publications; 1991.

34. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol*. 2007;46(Pt 1):1-18.

35. Linacre JM. RUMM2020 item-trait chi-square and Winsteps DIF size. *Rasch Meas Trans*. 2007;21(1):1096.

36. Hobart J, Cano S, Posner H, et al. Putting the Alzheimer's cognitive test to the test II: Rasch Measurement Theory. *Alzheimer's Dement*. 2013;9:S10–S20.

37. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Second. Psychology Press; 2013.

38. Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther*. 2014;36(5):648-662. doi:10.1016/j.clinthera.2014.04.006

39. Boone WJ. Rasch analysis for instrument development: why, when, and how? *CBE Life Sci Educ*. 2016;15(4). doi:10.1187/cbe.16-04-0148

40. Kwakkel G, Kollen B, Twisk J. Impact of time on improvement of outcome after stroke. *Stroke*. 2006;37(9):2348-2353. doi:10.1161/01.STR.0000238594.91938.1e

41. Houwink A, Nijland RH, Geurts AC, Kwakkel G. Functional recovery of the paretic upper limb after stroke: who regains hand capacity? *Arch Phys Med Rehabil*. 2013;94(5):839-844. doi:10.1016/j.apmr.2012.11.031

42. Coupar F, Pollock A, Rowe P, Weir C, Langhorne P. Predictors of upper limb recovery after stroke: a systematic review and meta-analysis. *Clin Rehabil*. 2012;26(4):291-313. doi:10.1177/0269215511420305

43. Nijboer TCW, Kollen BJ, Kwakkel G. The impact of recovery of visuo-spatial neglect on motor recovery of the upper paretic limb after stroke. *PloS One*. 2014;9(6):e100584. doi:10.1371/journal.pone.0100584

44. StataCorp. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP; 2015.

45. Muraki E. A Generalized Partial Credit Model: Application of an EM Algorithm. *Appl Psychol Meas*. 1992;16(2):159-176. doi:10.1177/014662169201600206

46. Masters GN. A rasch model for partial credit scoring. *Psychometrika*. 1982;47(2):149-174. doi:10.1007/BF02296272

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

48. Brown C, Templin J, Cohen A. Comparing the two- and three-parameter logistic models via likelihood ratio tests: a commonly misunderstood problem. *Appl Psychol Meas*. 2015;39(5):335-348. doi:10.1177/0146621614563326

49. Nguyen TH, Han H-R, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. *The Patient*. 2014;7(1):23-35. doi:10.1007/s40271-013-0041-0

50. StataCorp. *Stata 14 Base Reference Manual*. College Station, TX: Stata Press; 2015.

51. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care*. 2000;38(9 Suppl):II28-II42.

52. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res Int J Qual Life Asp Treat Care Rehabil*. 2007;16 Suppl 1:5-18. doi:10.1007/s11136-007-9198-0

53. Hill CD, Edwards MC, Thissen D, et al. Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. *Med Care*. 2007;45(5 Suppl 1):S39-47. doi:10.1097/01.mlr.0000259879.05499.eb

54. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Model Multidiscip J*. 1999;6(1):1-55. doi:10.1080/10705519909540118

55. Morizot J, Ainsworth AT, Reise, Steven P. Toward modern psychometrics: application of item response theory models in personality research. In: Robins RW, Fraley RC, Krueger RF, eds. *Handbook of Research Methods in Personality Psychology*. New York: Guilford Press; 2007:407-423.

56. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.

57. Stratford PW. Getting more from the literature: estimating the standard error of measurement from reliability studies. *Physiother Can*. 2004;56(01):027. doi:10.2310/6640.2004.15377

58. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther*. 1997;77(7):745-750.

59. Mooney CZ, Duval RD. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. California, USA: SAGE Publications Ltd; 1993.

60. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Fifth edition. Oxford: Oxford University Press; 2015.

61. Meng X, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. *Psychol Bull*. 1992;111(1):172-175. doi:10.1037/0033-2909.111.1.172

62. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care*. 2000;38(9 Suppl):II84-90.

63. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839-843. doi:10.1148/radiology.148.3.6878708

64. Nilsson AL, Sunnerhagen KS, Grimby G. Scoring alternatives for FIM in neurological disorders applying Rasch analysis. *Acta Neurol Scand*. 2005;111(4):264-273. doi:10.1111/j.1600-0404.2005.00404.x

65. Gosman-Hedström G, Blomstrand C. Evaluation of a 5-level functional independence measure in a longitudinal study of elderly stroke survivors. *Disabil Rehabil*. 2004;26(7):410-418. doi:10.1080/09638280410001662978

66. Grimby G, Gudjonsson G, Rodhe M, Sunnerhagen KS, Sundh V, Ostensson ML. The functional independence measure in Sweden: experience for outcome measurement in rehabilitation medicine. *Scand J Rehabil Med*. 1996;28(2):51-62.

67. Chedoke Arm and Hand Activity Inventory (CAHAI). Chedoke Arm and Hand Activity Inventory administration guidelines version 2. www.cahai.ca. Accessed February 27, 2015.

68. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil*. 1987;1:6-18.

69. Linacre JM. Sample size and item calibration or person measure stability. *Rasch Meas Trans*. 1994;7(4):328.

Tables

Table 1. Test items in the original versions of the Chedoke Arm and Hand Activity Inventory (CAHAI).

CAHAI-7

1. Open jar of coffee
2. Call 911
3. Draw a line with ruler
4. Pour a glass of water
5. Wring out washcloth
6. Do up five buttons
7. Dry back with towel

CAHAI-8

8. Put toothpaste on toothbrush

CAHAI-9

9. Cut medium-resistance putty

CAHAI-13

10. Zip up a zipper
11. Clean a pair of eyeglasses
12. Place container on table
13. Carry bag up the stairs

Table 2. Participant characteristics (*n* = 105).

Characteristic	**Sample (%)**
Total *n*	105
Sex	
Male	54 (51.4)
Female	51 (48.6)
Age, in years	
Median (1st, 3rd quartile)	72 (62, 78)
Min, Max	44, 92
Days since stroke	
Median (1st, 3rd quartile)	38 (27, 80)
Min, Max	3, 342
Type of stroke	
Ischemic	78 (82.1)
Hemorrhagic	17 (17.9)
Unilateral spatial neglect	
Absent	54 (51.9)
Present, client able to compensate	28 (26.9)
Present, client unable to compensate	22 (21.2)
Affected upper limb	
Right	46 (43.8)
Left	57 (54.3)
Bilateral	2 (1.9)
Baseline upper limb impairment	
Mild-moderate (CMSA score 7 – 11)	54 (51.4)
Severe (CMSA score ≤ 5)	51 (48.6)

SD, standard deviation; min, minimum; max, maximum; CMSA, Chedoke-McMaster Stroke Assessment (arm and hand component)

Table 3. Summary of scores on all outcome measures in both datasets.

Variables	Dataset 1 (D1)* (*n* = 105)		Dataset 2 (D2)			
			Baseline (*n* = 105)		Discharge (*n* = 100)	
	Median (1st, 3rd quartile)	Min, Max	Median (1st, 3rd quartile)	Min, Max	Median (1st, 3rd quartile)	Min, Max
CAHAI						
CAHAI-13	38 (16, 65)	13, 91	32 (14, 60)	13, 87	50 (16, 72)	13, 91
CAHAI-9	29 (12, 46)	9, 63	23 (10, 41)	9, 62	33 (11, 54)	9, 63
CAHAI-8	26 (10, 42)	8, 56	20 (9, 35)	8, 55	30 (9, 47)	8, 56
CAHAI-7	22 (9, 37)	7, 49	17 (7, 31)	7, 48	26 (8, 41)	7,49
CMSA						
Arm and hand	-	-	7 (4, 9)	2, 11	8 (4, 10)	2, 12
Shoulder pain	-	-	6 (5, 7)	2, 7	6 (5, 7)	2, 7
ARAT						
Total	-	-	18 (0, 43)	0, 57	23 (0, 52)	0, 57

*Data was derived using participants' CAHAI scores at one measurement point (either at baseline or discharge)

CMSA, Chedoke-McMaster Stroke Assessment; CAHAI, Chedoke Arm and Hand Activity Inventory; ARAT, Action Research Arm Test

Table 4. Rasch analysis results of two versions of the 11-item CAHAI.

Item	7-category scoring scale						4-category scoring scale					
	Location	**SE**	**Fit residual**	***df***	x^2	***P*-value**	**Location**	**SE**	**Fit residual**	***df***	x^2	***P*-value**
1. Open jar of coffee	-0.652	0.105	0.659	2	3.12	.210	-1.114	0.195	-0.737	2	0.99	.611
2. Call 911	-0.101	0.097	-1.053	2	0.65	.721	-0.240	0.192	-0.146	2	0.56	.757
3. Draw a line with a ruler	-0.456	0.104	0.686	2	1.14	.566	-0.752	0.198	0.010	2	1.34	.512
4. Pour a glass of water	1.060	0.102	-1.024	2	1.81	.404	2.198	0.213	-0.712	2	0.92	.633
5. Wring out washcloth	-0.251	0.103	-0.317	2	4.22	.121	-0.347	0.199	0.527	2	6.10	.047
6. Do up five buttons	0.378	0.094	-0.602	2	0.17	.921	0.602	0.196	-0.919	2	1.32	.516
7. Dry back with towel	0.567	0.093	-1.014	2	1.50	.472	0.967	0.192	-0.736	2	1.32	.517
8. Put toothpaste on toothbrush	-0.270	0.102	-1.313	2	1.02	.600	-0.575	0.199	0.625	2	1.78	.410
9. Cut medium resistance putty	-0.460	0.101	0.129	2	0.51	.777	-0.914	0.197	-0.763	2	0.45	.798
10. Zip up the zipper	-0.031	0.108	1.050	2	2.30	.317	-0.226	0.206	1.415	2	1.79	.410
11. Clean a pair of eyeglasses	0.213	0.100	-0.160	2	0.81	.669	0.399	0.199	0.084	2	2.65	.266

SE: standard error; df: degrees of freedom

Table 5. Item response theory analysis of three CAHAI version with 4-category scoring scale (n = 105).

	11-item CAHAI			New 7-item CAHAI			New 5-item CAHAI		
Item	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
1. Open jar of coffee	-0.73 (0.15)	0.05 (0.15)	0.72 (0.15)	-0.71 (0.15)	0.02 (0.15)	0.70 (0.15)	-	-	-
2. Call 911	-0.47 (0.15)	0.38 (0.15)	0.84 (0.15)	-	-	-	-	-	-
3. Draw a line with ruler	-0.47 (0.15)	0.06 (0.15)	0.72 (0.15)	-0.46 (0.15)	0.03 (0.15)	0.70 (0.15)	-0.45 (0.15)	0.04 (0.15)	0.70 (0.15)
4. Pour a glass of water	0.22 (0.14)	0.60 (0.15)	1.85 (0.25)	0.19 (0.14)	0.58 (0.15)	1.88 (0.26)	0.21 (0.15)	0.58 (0.15)	1.86 (0.26)
5. Wring out washcloth	-0.40 (0.14)	0.10 (0.15)	0.92 (0.15)	-	-	-	-	-	-
6. Do up five buttons	0.01 (0.15)	0.37 (0.15)	0.95 (0.16)	-0.02 (0.15)	0.34 (0.15)	0.93 (0.16)	-	-	-
7. Dry back with towel	-0.02 (0.14)	0.75 (0.16)	0.94 (0.17)	-0.04 (0.14)	0.72 (0.16)	0.92 (0.17)	-0.02 (0.14)	0.73 (0.16)	0.90 (0.17)
8. Put toothpaste on toothbrush	-0.81 (0.15)	0.20 (0.14)	1.02 (0.16)	-	-	-	-	-	-
9. Cut medium resistance putty	-0.99 (0.16)	0.22 (0.14)	0.91 (0.15)	-	-	-	-	-	-
10. Zip up the zipper	-1.03 (0.16)	0.31 (0.14)	1.36 (0.18)	-1.00 (0.16)	0.29 (1.4)	1.35 (0.19)	-1.00 (0.16)	0.30 (0.14)	1.33 (0.19)
11. Clean a pair of eyeglasses	-0.41 (0.14)	0.49 (0.14)	1.19 (0.17)	-0.41 (0.14)	0.46 (0.14)	1.17 (0.17)	-0.39 (0.14)	0.47 (0.14)	1.16 (0.17)
Item discriminating (*a*)	3.60 (0.33)			3.71 (0.39)			3.67 (0.43)		

b, item difficulty parameters*;* values in brackets are standard errors.

Table 6. Psychometric properties of the new CAHAI versions (7-item and 5-item) and the CAHAI-7 with the revised 4-category scoring scale (n = 105).

	Construct validity			**Test-retest reliability**	
	ARAT	***CMSA***			
	r_s (95% CI)	**Arm and hand, r_s (95% CI)**	**Pain, r_s (95% CI)**	**ICC (95% CI)**	**SEM (95% CI)**
CAHAI-7*	0.92 (0.90, 0.95)	0.88 (0.84, 0.92)	0.59 (0.45, 0.72)	0.96 (0.92, 0.98)	1.28 (1.05, 1.66)
New 7-item CAHAI	0.91 (0.88, 0.94)	0.88 (0.84, 0.92)	0.57 (0.43, 0.71)	0.96 (0.93, 0.98)	1.20 (0.98, 1.55)
New 5-item CAHAI	0.90 (0.87, 0.93)	0.87 (0.82, 0.92)	0.57 (0.43, 0.71)	0.96 (0.93, 0.99)	0.83 (0.68, 1.08)

ARAT: Action Research Arm Test; CMSA, Chedoke-McMaster Stroke Assessment; r_s: Spearman's rank correlation coefficient; ICC: intra-class correlation coefficient; SEM: standard error of measurement.

*Original CAHAI-7 with revised 4-category scoring scale

Figures

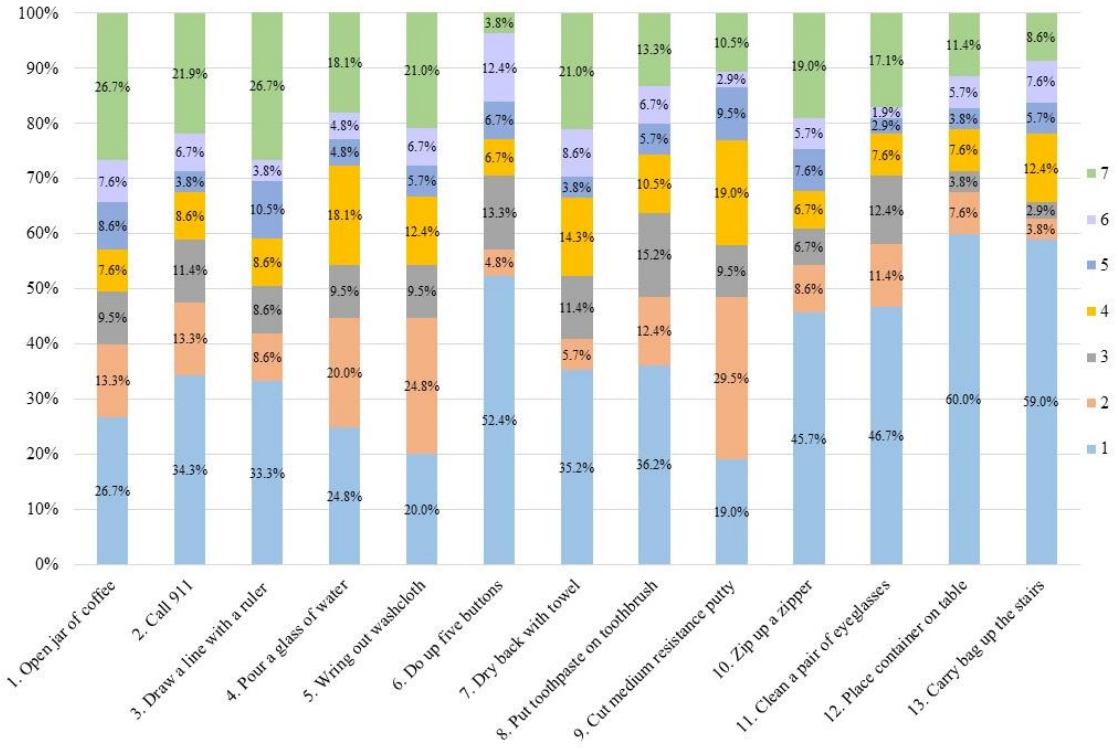
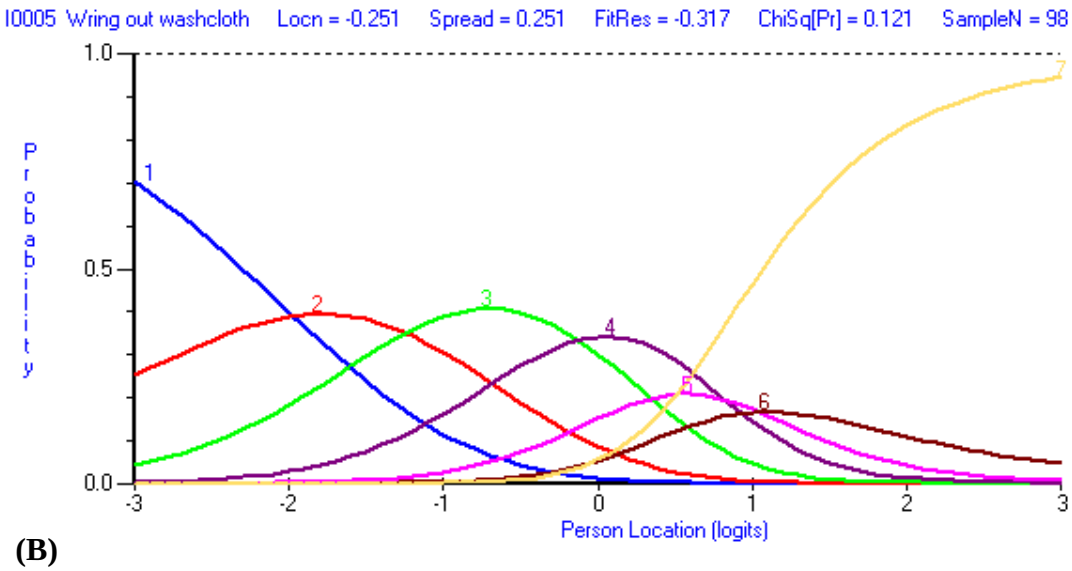


Figure 1. Frequency of score categories for each CAHAI item in the first dataset (D1).

(A)



(B)

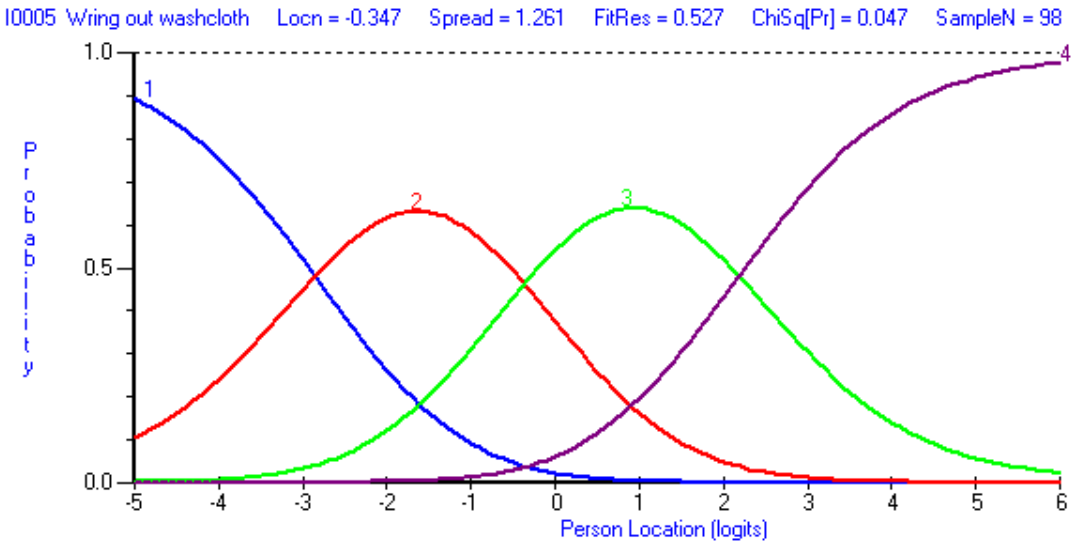
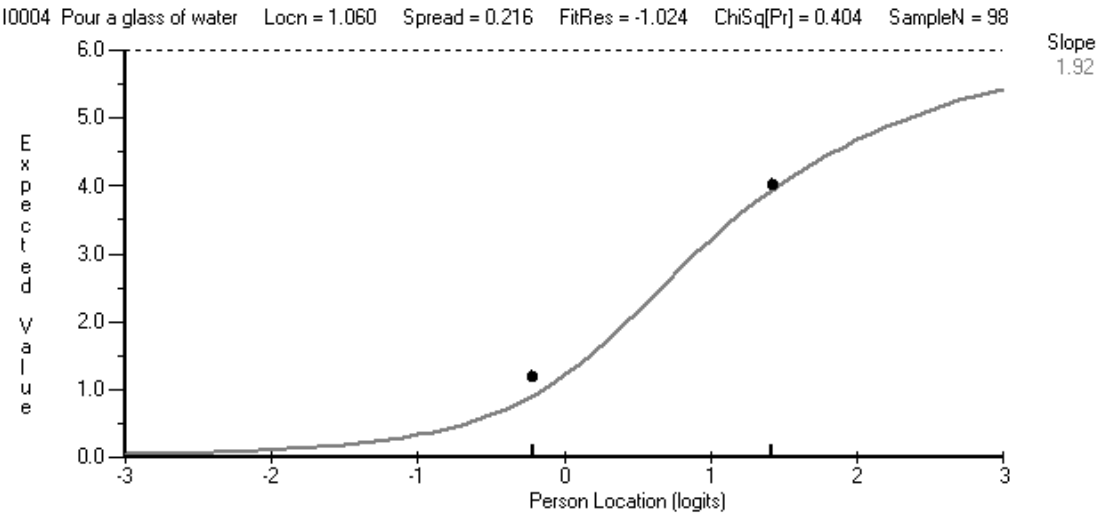


Figure 2. Category probability curves for item 5 (wring out washcloth) in the 11-item CAHAI. (A) In the version using a 7-category scale, the intersections of adjacent curves, or thresholds, were disordered. Threshold parameters were -1.724, -0.993, 0.126, 1.075, 1.294, 0.224 for the 1-2, 2-3, 3-4, 4-5, 5-6, and 6-7 thresholds respectively. (B) The collapsing of score categories to a 4-category scale resulted in ordered thresholds.

(A)



(B)

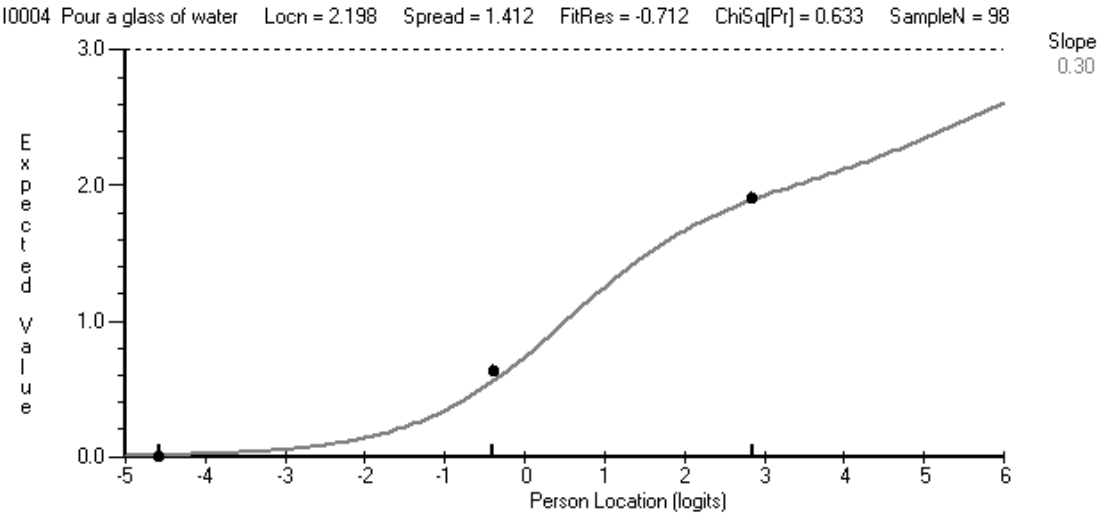


Figure 3. Item characteristic curves of item 4 (pour a glass of water) in the 11-item CAHAI with (A) the original 7-category scoring scale, and (B) the revised 4-category scoring. In both versions, the plots (observed scores) lie close to the curve (predicted scores), indicating fit with the Rasch model. However, the collapsing from 7 to 4 score categories resulted in a decrease in the gradient of the slope.

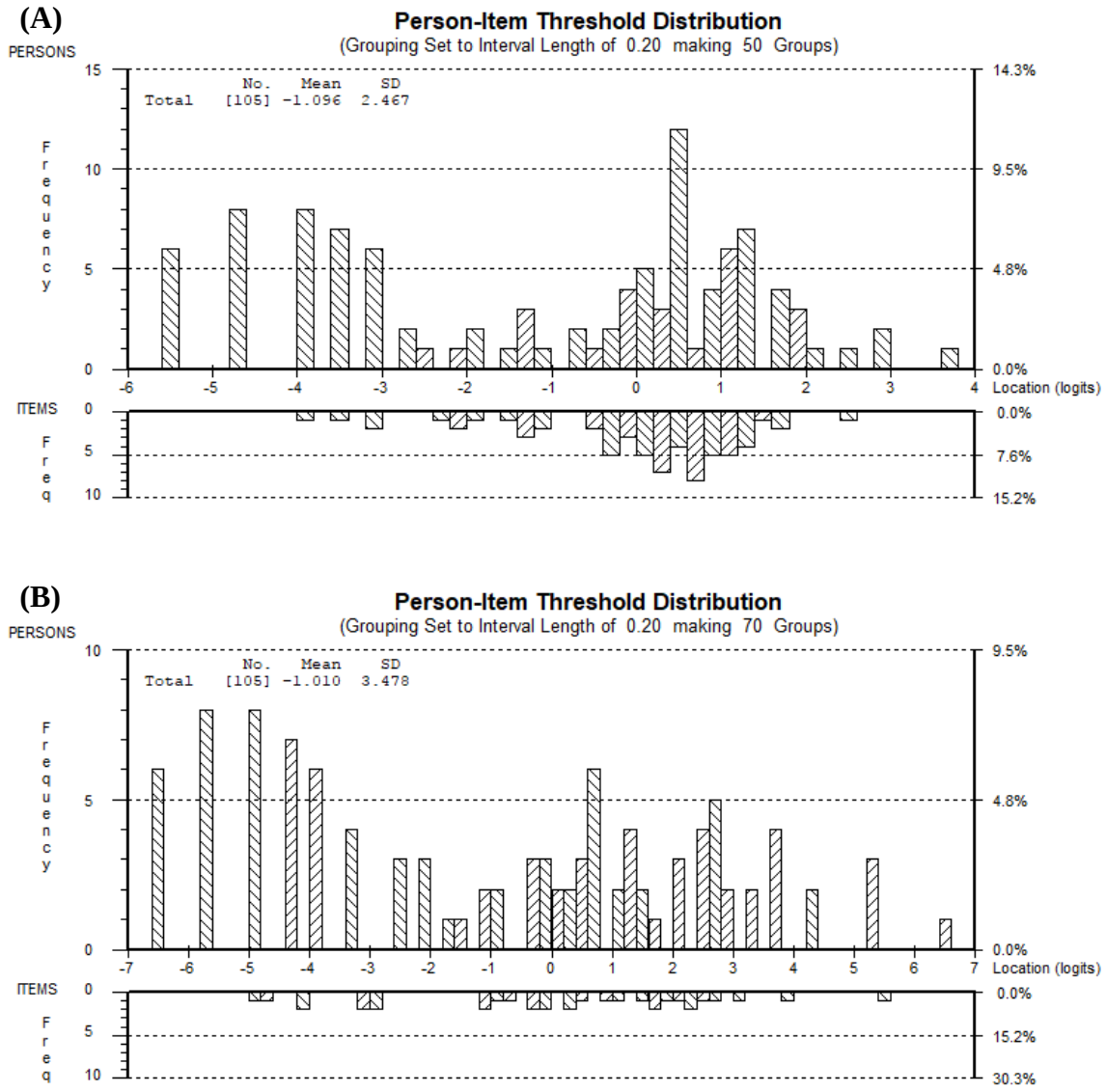


Figure 4. Person-item threshold distributions of the 11-item versions of the CAHAI with (A) the original 7-category scoring scale, and (B) the revised 4-category scale.

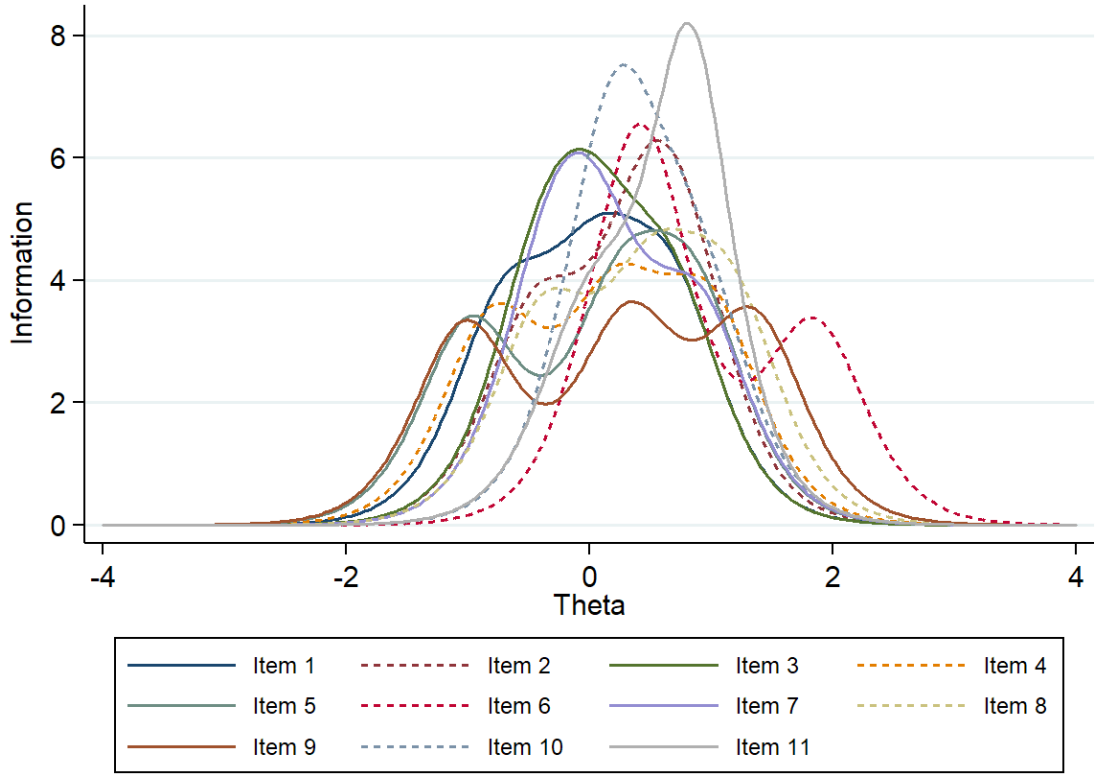


Figure 5. Item information functions of all test items in the 11-item CAHAI with a 4-category scoring scale.

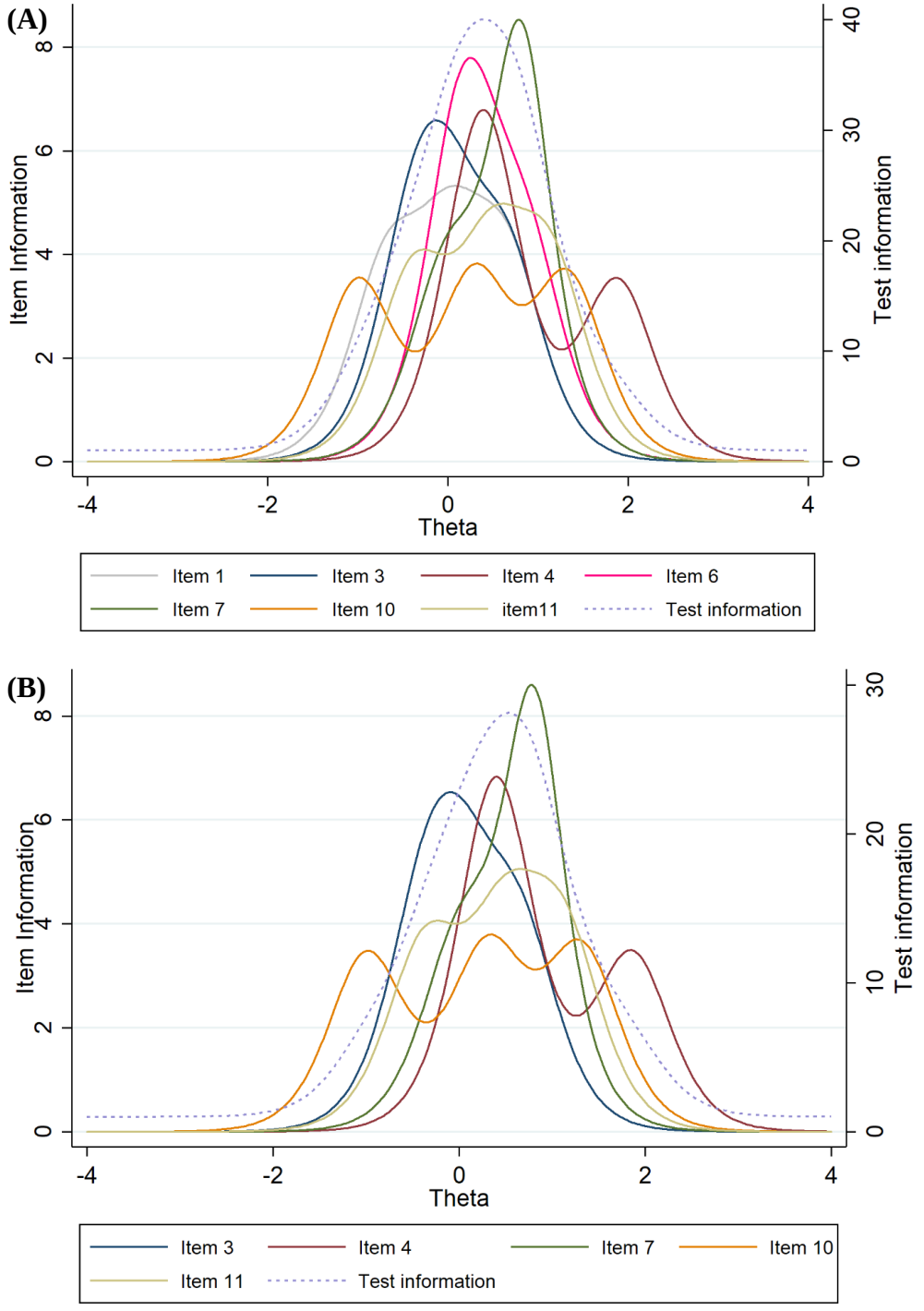


Figure 6. Item and test information functions of the new CAHAI versions. (A) 7-item version, and (B) 5-item version.

Chapter Six: Discussion

Chapter Six: Discussion

This thesis was devoted to the continued evaluation of the Chedoke Arm and Hand Activity Inventory (CAHAI), an outcome measure of post-stroke upper extremity (UE) function (Barreca et al., 2004). There were two overall thesis objectives: (1) to develop a Singapore version of the CAHAI, and (2) to re-evaluate the psychometric properties and clinical utility of the CAHAI using item response theory and Rasch measurement theory. In the first part of this thesis, the evaluation of CAHAI was taken beyond the country where it was developed and was cross-culturally adapted and validated for the stroke population in Singapore. Two manuscripts presented in this thesis (Chapters 2 and 3) reported the translation, cross-cultural adaptation, and psychometric evaluation of the Singapore versions of the CAHAI. The second part of this thesis applied different measurement theories to evaluate the CAHAI and two manuscripts were presented (Chapters 4 and 5). Chapter 4 reports the psychometric evaluation of the original versions of the CAHAI using item response theory and Rasch measurement theory. In Chapter 5, these two measurement theories were then used to refine the original CAHAI, which led to the development of two new versions of the CAHAI. The following sections discuss the study findings of this thesis, the contributions of the body of work, the study limitations, and the recommendations for future directions.

Part I: The Development of the Singapore Version of the CAHAI

The first thesis objective was to develop a Singapore version of the CAHAI and was achieved through three phases: translation, adaptation, and validation. The CAHAI was systematically translated and adapted by following a standardized and validated

eight-step procedure, known as the Translation and Cross-cultural Adaptation of Objectively-Assessed Outcome measures (TCCA-OAO) procedure (Schuster, Hahn, & Ettlin, 2010). The use of a guideline intended for the translation and cross-cultural adaptation of performance-based outcome measures ensured the equivalence between the Singapore and original versions of the CAHAI.

Translation

As Singapore and Canada shared a common language (English), translation of the entire CAHAI manual was not required. However, the Standard Singapore English differs from the Canadian English as the former follows the British English convention (Wierzbicka, 2003). For example, *jug*, rather than *pitcher*, is the term commonly used in Singapore to describe a container with a spout used to hold liquids. The CAHAI administration manual was thus translated from Canadian English to Standard Singapore English and descriptions to seven pieces of equipment were also changed to terms more commonly used in Singapore. Hence, while two cultures/countries may share a similar language, there may be sufficiently meaningful differences that require a thorough review of the outcome measure to ensure relevance to the target culture/country (Guillemin, 1995).

The standard instructions in the CAHAI were also translated into Mandarin and Malay. The quality of translation is highly dependent on the methodology employed, and therefore, a rigorous multi-step process is required (Wild et al., 2005). The translation of the CAHAI followed the steps in the TCCA-OAO procedure (Schuster et al., 2010), which included both forward and back translation processes. Back translation is an

essential step as it serves as a method of validity checking for consistency in the translation (Beaton, Bombardier, Guillemin, & Ferraz, 2000; Guillemin, 1995). By comparing the back-translations with the original CAHAI instructions, the extent to which the translated instructions maintained the content and meaning of the original instructions (i.e., semantic equivalence) were verified. Back translators should be carefully selected; they must be fluent in both the source and target languages and do not have a priori knowledge of the concepts and intent of the outcome measure to minimize biases (Guillemin, 1995; Schuster et al., 2010). For performance-based measures, back translators who are members of the target profession (users of the measure) are preferred, as there is also a need to ensure professional vocabulary is maintained (Schuster et al., 2010).

Adaptation

Two test items in the CAHAI were adapted for the Singapore version; ‘open jar of coffee’ and ‘call 911’ were adapted to ‘open a jar of peanut butter’ and ‘call 995’ respectively. These adaptations were based on the recommendations of the two reviewers to ensure cultural relevance of the test items to the stroke population in Singapore. The relevance of the test items is particularly important for individuals with stroke. As discussed in Chapter 1 of this thesis, familiar daily tasks and objects can be easily understood by individuals with stroke and also facilitate accurate evaluation of UE function (Barreca et al., 2004; Trombly & Wu, 1999; Wu, Trombly, Lin, & Tickle-Degnen, 1998).

While adaptations to outcome measures ensure cultural relevance to the target population, there is also a need to ensure that the constructs measured are not changed in the adaptation process. The adapted measure should measure the same constructs as the original measure, also known as conceptual equivalence (Stewart & Nápoles-Springer, 2000). For post-stroke UE measures, the specific UE functions evaluated in each test item should be similar between the adapted and original measure. Thus, an important step in the adaptation process is to define the constructs measured within the instrument. In the adaptation of the CAHAI, the demands of UE function for successful task performance were examined for test items that required modifications. The reviewers then recommended modifications to the objects that demanded similar UE function and were familiar to individuals with stroke in Singapore (e.g., a jar of peanut butter was selected to replace a jar of coffee).

A more systematic method to define the constructs evaluated within an outcome measure is to use the International Classification of Functioning, Disability, and Health (ICF) (World Health Organization, 2001) as a reference. Constructs within an outcome measure can be identified by linking the measure to the ICF (Cieza et al., 2005). Several post-stroke UE outcome measures, including the CAHAI, have been linked to the ICF (Choo, Ng, Fayed, & Harris, 2017; Velstra, Ballert, & Cieza, 2011). The ICF categories linked to these outcome measures define the specific UE functions evaluated in each test item. For instance, ‘open jar of coffee’ in the CAHAI was linked to five ICF categories: b7300 Power of isolated muscles and muscle groups; b7301 Power of muscles of one limb; d4300 Lifting; d4401Grasping; and d4452 Reaching (Choo et al., 2017). The ICF

categories linked to the measure can be used as a reference in the adaption process; reviewers can provide recommendations for modifications to test items that are consistent with the relevant ICF categories.

Translation and cross-cultural adaptation procedure. It is imperative that the translation and cross-cultural adaptation of any type of outcome measure (performance-based or self-reported) use methodological guidelines to guide the process. In a review that evaluated the quality of translated and cross-culturally adapted post-stroke outcome measures available in Brazil (Lima, Teixeira-Salmela, Simões, Guerra, & Lemos, 2016), only 1 out of the 11 identified measures followed the recommended processes in the methodological guidelines. The quality of the available post-stroke outcome measures in Brazil was affected due to several flaws stemming from the translation and adaptation process (Lima et al., 2016). This was a similar finding in other systematic reviews that evaluated the quality of cross-culturally adapted outcome measures in different clinical populations (Al Zoubi, Eilayyan, Mayo, & Bussières, 2017; Yao et al., 2016). There was poor quality in the translation and adaptation of the majority of outcome measures identified in both reviews. Thus, the use and adherence to methodological guidelines for the translation and cross-cultural adaptation of outcome measures ensures quality in the translation and adaptation process. It also better ensures equivalence between the original and adapted versions of the measure (Beaton et al., 2000).

Validation

The psychometric properties of all Singapore versions of the CAHAI were evaluated in an acute and subacute stroke sample in Singapore. All Singapore versions

demonstrated good inter-rater reliability and construct validity with the Fugl-Meyer Assessment of Upper Extremity and the Action Research Arm Test, which were also comparable with the original versions of the CAHAI. The evaluation of the psychometric properties of a translated and cross-culturally adapted measure is necessary because the reliability and validity of this modified instrument is unknown (Guillemin, 1995). Psychometric evaluation provides evidence for the measure's intended application and target population (Beaton et al., 2000), and allows measurement equivalence with the original measure to be demonstrated (Herdman, Fox-Rushby, & Badia, 1998).

The psychometric properties of the shortened Singapore versions of the CAHAI were also evaluated. This evaluation is necessary as it is erroneous to assume that the shortened versions, which comprise of subset of test items of the full 13-item version, have the same psychometric properties as the full version. Participants' performance on each test item is different, and so, the error associated with the scores on subsets of test items in the shortened Singapore CAHAI versions will be different. Thus, it is necessary to evaluate the psychometric properties of shortened versions of a measure, and also evaluate the extent to which the properties are comparable to that of the full measure.

The design of measurement studies is challenging as there is currently no recognized 'gold standard' study designs, and different study designs and statistical analysis are needed for different psychometric properties. The original COSMIN (Consensus-based standards for the selection of health status measurement instruments) checklist (Mokkink et al., 2010) was used as a reference when designing the study to evaluate the Singapore versions of the CAHAI. The COSMIN checklist describes the

standards for design requirements and preferred statistical methods for studies on the measurement properties of health-related patient-reported outcome measures (Mokkink et al., 2010). By following the standards described in the COSMIN checklist when designing the study to evaluate the Singapore versions of the CAHAI, good quality evidence of the measures' psychometric properties can be better ensured. There are ongoing research projects to develop tools to guide the design and reporting of measurement studies (COSMIN initiative, n.d.). In the interim, quality assessment tools that are currently available, such as the COSMIN risk of bias checklist (Mokkink et al., 2018) which replaced the original COSMIN checklist, can be used to guide the design and reporting of measurement studies.

Part II: Re-evaluation of the CAHAI using Modern Test Theories

The second objective of this thesis was to re-evaluate the psychometric properties and the clinical utility of the CAHAI using modern test theories. The re-evaluation of the CAHAI using different measurement theories than the one employed in the initial development and validation phase contributed extensive knowledge about the measure. There are two key findings from the re-evaluation of the original versions of the CAHAI. First, the scaling issues in the CAHAI were identified, which included disordered score categories in the scoring scale and the lack of interval scaling in the measure. Second, the CAHAI is not a unidimensional measure. Revisions to the CAHAI guided by modern test theories were subsequently made, which resulted in the development of two new versions of the CAHAI: the 11-item and 5-item versions with a 4-category scoring scale.

Score Categories in the Scoring Scale

The CAHAI's original 7-category scale was intended as an adjectival scale, a unipolar scale that measures post-stroke UE function on a continuum from less to more (Barreca et al., 2004; Streiner, Norman, & Cairney, 2015). However, disordered score categories were found, indicating that the original scoring scale did not measure UE function in a progressive manner as it was intended. This result was not surprising as similar problems with the 7-category scoring scale in the Functional Independence Measure, which the CAHAI's scoring scale was based on, were found in three studies (Claesson & Svensson, 2001; Gosman-Hedström & Blomstrand, 2004; Nilsson, Sunnerhagen, & Grimby, 2005). There are several reasons why disordered score categories in the CAHAI's scale may occur. One reason is the labelling of score categories may be confusing or vague to assessors (Cano, Barrett, Zajicek, & Hobart, 2011; Pallant & Tennant, 2007). Another reason is the scale has too many categories and assessors have difficulties discriminating between them (Cano et al., 2011).

The disordering of score categories in the CAHAI's scale was resolved when the seven categories were collapsed into four categories. The collapsing/combining of adjacent score categories is recommended to not only improve the scale of an outcome measure but also to improve the overall quality of the measure (Bond & Fox, 2013; Linacre, 1999). Improvements to the psychometric properties of the CAHAI was demonstrated with the revision from a 7- to a 4-category scoring scale of the 11-item version. Using the original 7-category scale, the 11-item version of the CAHAI violated the assumption of local independence and instability was present with some test items

displaying non-uniform differential item functioning. However, these issues were resolved when a 4-category scoring scale was used.

While the collapsing from seven to four score categories improved both the CAHAI's scoring scale and its psychometric properties, it is important to consider the potential loss of information. For performance-based measures like the CAHAI, the scoring scale is used by assessors who are clinicians or researchers. The information from the scoring scale contributes differently to clinicians' and researchers' understanding of post-stroke UE function. As described in Chapter 1 of this thesis, using post-stroke UE outcome measures facilitates clinical decision-making, such as prescribing appropriate interventions to optimize recovery and discharge planning based on prognosis (Feys et al., 2000; Wolf, Kwakkel, Bayley, McDonnell, & Upper Extremity Stroke Algorithm Working Group, 2016). Clinicians may focus on the total scores on a measure, rather than individual item scores, to make such clinical decisions. Thus, the amount of information potentially lost with the collapsing of score categories in the CAHAI may not substantially alter the clinical decision making process with their clients. Furthermore, clinicians prefer measures that are simple and easy to score due to time constraints in clinical practice (Duncan & Murray, 2012), and thus, may favour lesser score categories. In contrast, post-stroke UE outcome measures are primarily used in research to demonstrate effectiveness of interventions and to predict outcomes of interests (Guyatt, Deyo, Charlson, Levine, & Mitchell, 1989; Veerbeek, Kwakkel, Wegen, Ket, & Heymans, 2011). Although researchers also focus on the total scores on a measure, they may also use scores on subsets of test items to evaluate the effectiveness of an

intervention or examine the extent to which individual test item predict outcomes. The potential loss of information with the collapsing of scores categories in the CAHAI may affect the conduct of such research studies. Hence, it is imperative to balance both clinicians' and researchers' needs when developing and/or refining the scoring scale of an outcome measure.

The optimal number of categories/steps in a scoring scale is complex. There is a loss of information with too few categories, whereas more categories can potentially add 'noise' or error rather than capture more information (Streiner et al., 2015). It is well-recognized that a scale with five to seven categories is recommended, where the level of information captured, the reliability of the measure, and the burden on respondents are optimized (Krosnick, Holbrook, & Visser, 2006; Miller, 1956; Preston & Colman, 2000). However, as Linacre (2000) questioned, "Statisticians can recommend the construction of a scale with many ordered categories. Experts can define them. But can respondents [assessors] discern them?" (p. 617 – 618). To better ensure that both researchers and clinicians' priorities are addressed, both personnel should be involved in the development and refining of the scoring scale of a measure. For example, in the pre-testing stage of an outcome measure, specific feedback from clinicians can be obtained about the number of categories/steps in a scoring scale.

Interval Scaling

The re-valuation of the original versions of the CAHAI using item response theory (IRT) and Rasch measurement theory allowed a formal evaluation of whether the CAHAI has interval scaling properties. Results from both Rasch and IRT analyses

showed that all original versions of the CAHAI do not conform to an interval scale. This finding has several implications. CAHAI scores are manipulated mathematically (e.g., summing individual item scores to obtain a total score) and statistically (e.g. computation of effect size) in clinical practice and research as if they were interval scores. However, results from arithmetic calculations and statistical analyses are not logically valid since the 'distance' between the score categories in the CAHAI's 7-category scoring scale is not equal (Merbitz, Morris, & Grip, 1989). This consequently impacts the validity and quality of clinical decisions and statistical inferences based on these results, and the repeated misuse of the scoring scale in clinical practice and research encourages continued misinformation (Merbitz et al., 1989).

Following the results from both Rasch and IRT analyses, both theories were then used to revise the original CAHAI and two new versions were developed. The new 11-item and 5-item versions of the CAHAI with a 4-category scoring scale demonstrated interval scaling. Thus, both mathematical and statistical manipulations of CAHAI scores are now logically valid and justified (Merbitz et al., 1989; Wright & Linacre, 1989). For clinicians and researchers, this means that simple arithmetic calculations (e.g. summing individual item scores to obtain a total score) can be used for efficient tabulation of CAHAI scores. With empirical evidence of interval scaling, the quality and validity of clinical decisions and statistical inferences based on results from mathematical and statistical manipulation of CAHAI scores can also be indirectly improved.

Interval scaling of the original CAHAI was previously not evaluated, as its initial development and validation were guided by classical test theory. Classical test theory is

the prevalent measurement theory used to guide the development and evaluation of outcome measures in rehabilitation due to its established history and weaker assumptions (i.e., assumptions are relatively easy to meet) (Hobart & Cano, 2009). As discussed in Chapters 1 and 4 of this thesis, one weakness of the classical test theory is the lack of the evaluation of interval scaling: "In CTT [classical test theory], we deal with this problem in a not-overly sophisticated manner – we cover our eyes and hope it will go away (which it never does)" (Streiner, 2010, p. 181). Consequently, this means that many outcome measures used in rehabilitation inherit this limitation of classical test theory and the lack of empirical evidence of interval scaling has several consequences, as described earlier. Therefore, it is necessary to continue to evaluate an outcome measure beyond its initial phase of development and validation. Using item response theory and Rasch measurement theory to re-evaluate outcome measures used in rehabilitation would not only provide evidence of interval scaling but also expand the knowledge on the psychometric properties of the measure.

Unidimensionality

Unidimensionality is an underlying assumption of both IRT and Rasch measurement theory (Bond & Fox, 2013; Hambleton, Swaminathan, & Rogers, 1991). The purpose of the CAHAI is to evaluate function in the affected UE after a stroke (Barreca et al., 2004), thus, it should measure a single trait/ability (i.e. UE function). However, results from both Rasch and IRT analyses found all original versions of the CAHAI not unidimensional. By revising the CAHAI using these two measurement

theories, both new 11-item and 5-item versions of the CAHAI demonstrated unidimensionality.

Both the assumptions of unidimensionality and local independence must be met in IRT and Rasch measurement theory (Bond & Fox, 2013; Hambleton et al., 1991), which also provides the evidence of interval scaling of an outcome measure. Although unidimensionality is one of the main assumptions of IRT and Rasch measurement theory, it is important to consider unidimensionality in relation to the construct of UE function. UE function can be regarded as a single attribute because it needs to work as a synchronized single unit (Carr & Shepherd, 2010). In this thesis, the CAHAI was presumed to be a unidimensional outcome measure. This is because the daily tasks that comprise the CAHAI evaluate the extent to which the UE works as a synchronized unit for successful task completion. However, as presented in Chapter 1, UE function can be categorized into motor and somatosensory subcomponents. Furthermore, UE function can also be considered as different levels of functioning according to the ICF (World Health Organization, 2001). Thus, one may argue that UE function should not be regarded as a unidimensional attribute. The stance on unidimensionality may influence the development and/or refinement of an outcome measure. When UE function is regarded as a unidimensional attribute, the outcome measure would assess UE function as a single unit, like the CAHAI. Otherwise, the outcome measure would have different subscales measuring different components of UE function. For example, the Fugl-Meyer Assessment for Upper Extremity has motor function, sensation, passive joint motion, and joint pain subscales (Fugl-Meyer, Jääskö, Leyman, Olsson, & Steglind, 1975). For post-

stroke UE outcome measures comprised of subscales that separately evaluate components of UE function, judgment is required during the evaluation of the measure to determine if the measure is 'unidimensional enough' (Streiner et al., 2015). To meet the assumption of unidimensionality in IRT and Rasch measurement theory, separate evaluation of each subscale may be needed (i.e., assessing if each subscale measures a single subcomponent of UE function). An alternative consideration is to use multidimensional IRT models to evaluate such post-stroke UE outcome measure.

Contributions of the Thesis

Stroke Rehabilitation in Singapore and Canada

This thesis undertook the first study to translate and cross-culturally adapt the CAHAI for an Asian stroke population. It is also, to the best of my knowledge, the first post-stroke UE outcome measure that was translated, adapted, and validated for the stroke population in Singapore. For clinicians and researchers in Singapore, they now have access to a valid and culturally relevant performance-based post-stroke UE outcome measure. The use of objective measurements from the CAHAI can facilitate several clinical-decision making processes. For instance, the scores provide an indication of the level of UE function, which may be used to determine the frequency of interventions. It may also be used to direct the focus of interventions to train specific UE movements that the individual has difficulties with. The improvements in scores can be used to indicate response to interventions and to also demonstrate the effectiveness of rehabilitation services. From a research standpoint, the development of the Singapore version of the CAHAI can facilitate collaborations for international studies as the CAHAI is also

clinically used in Canada, Australia, and Germany (Gustafsson, Turpin, & Dorman, 2010; Rowland, Turpin, Gustafsson, Henderson, & Read, 2011; Schuster et al., 2010). With translated and cross-culturally adapted versions of the CAHAI, the same outcome measure can be used across different countries. The findings from studies using the CAHAI within one country may help clinicians and researchers in another country. For international multi-centred randomized controlled trials, it also allows data to be pooled and analyzed, and results from each country can be compared.

This thesis also undertook the first study to evaluate the CAHAI using IRT and Rasch measurement theory, which contributed to the body of knowledge about the psychometric properties of the CAHAI. This thesis also demonstrated the novel use of both IRT and Rasch measurement theory to revise the CAHAI. For clinicians and researchers in Canada, there are now two new versions of the CAHAI that demonstrate interval scaling. These two versions support the common practice of arithmetic calculations of raw scores. The raw scores can also be converted to logit scores, which provide a better estimate of the change in UE function in individuals with stroke (Streiner et al., 2015). There are also fewer test items in the revised full (11-item) and shortened (5-item) versions of the CAHAI, with fewer score categories in the revised scoring scale. These new versions are more time efficient than the original CAHAI versions, which improves its clinical utility and may enhance the routine use of the measure in clinical practice.

Knowledge Dissemination

Chapters 2 and 3 of this thesis were published in peer-reviewed journals, and Chapters 4 and 5 are prepared for submission to peer-reviewed journals. The publication and submission of the chapters in this thesis in peer-reviewed journals was not only to disseminate the study findings. It was also intended as knowledge dissemination about topics related to measurement in rehabilitation for clinicians and researchers.

Cross-cultural adaptation of outcome measures. It is well-established that outcome measures must be translated, adapted, and validated for the target setting if the setting differs from the original population in terms of culture, language, and/or country (Beaton et al., 2000; Geisinger, 1994). Unfortunately, clinicians and researchers in Singapore continue to use outcome measures that are not validated for the local clinical population. There is also poor reporting of the translation and adaptation process of outcome measures by researchers in Singapore. The first part of this thesis reinforces the need for outcome measures used in clinical practice and research to be culturally relevant and validated for the intended target setting. It also serves as an exemplar of good quality reporting of research undertaken for the cross-cultural adaptation of performance-based outcome measures.

Modern test theories. IRT and Rasch measurement theory are increasingly applied in rehabilitation. However, many rehabilitation professionals are still unfamiliar with modern test theories and how these theories can be applied. The second part of this thesis disseminates information about modern test theories and their application in a manner that is accessible to clinicians. The manuscripts serve as an entry point for

clinicians to be aware of modern test theories and to gain some understanding of the theories at the conceptual level. The second part of this thesis also highlights to researchers the importance of continued evaluation of existing outcome measures to improve the quality of measurement in rehabilitation. In addition, the publication on the novel use of both IRT and Rasch measurement theory to revise the CAHAI is intended to initiate a scholarly conversation on how these two theories can be used to complement each other.

Limitations and Future Directions

Part I: The Development of the Singapore Version of the CAHAI

The study employed a cross-sectional design and thus, the longitudinal validity of the Singapore version of the CAHAI is currently unknown. Longitudinal validity should be evaluated as one of the measure's purpose is to evaluate the change in UE function over time. Future studies employing longitudinal study designs are recommended to evaluate the test-retest reliability, sensitivity to change, and responsiveness of the Singapore versions.

A second limitation is the use of only classical test theory to guide the evaluation of the Singapore version of the CAHAI. As presented in Chapter 4 of this thesis, the evaluation of the original versions of the CAHAI using modern test theories identified several issues. Accordingly, similar problems with the Singapore version of the CAHAI are expected. Thus, future studies are needed to evaluate the psychometric properties of the Singapore versions using modern test theories. In doing so, three goals can be achieved. First, knowledge about the psychometric properties of the Singapore versions

can be increased. Second, by conducting similar analyses as the re-evaluation of the original CAHAI, the findings in the second part of this thesis can be validated. Lastly, the cross-cultural validity of the Singapore versions can be evaluated using a statistical approach, and the extent to which the CAHAI measures UE function across different cultures can be determined.

Knowledge translation for clinicians. The uptake of evidence in clinical practice, including the routine use of outcome measures, is a complex process requiring considerations of factors at the individual (clinician), managerial, and organizational levels (Duncan & Murray, 2012; Petzold et al., 2012). The dissemination of knowledge about the Singapore versions of the CAHAI is not sufficient to effect a change in the measurement of post-stroke UE function in day-to-day practice. Knowledge translation strategies targeted at embedding and integrating the use of the CAHAI in clinical practice in Singapore need to be developed to support the routine use of the measure.

An active multi-component knowledge translation intervention comprising of interactive educational sessions, printed materials, and educational outreach visits can be developed. These strategies were found to be effective in a systematic review that identified the effective interventions to change the practice behaviours of rehabilitation professionals (Menon, Korner-Bitensky, Kastner, McKibbon, & Straus, 2009). The interactive educational sessions can be one-day training workshops where clinicians learn about the Singapore version of the CAHAI and also receive training to administer and score the measure. The format of the workshop may include lectures, demonstrations, group discussions, and hands-on practice with the administration and scoring of the

measure. Printed materials that include relevant information (e.g., psychometric properties, quick scoring guide, and where to obtain the required equipment) can also be distributed during the interactive educational sessions. A follow-up education outreach visit to workshop participants can be conducted. During these visits, problems regarding the use of the measure can be discussed and any updates about the Singapore version of the CAHAI can also be given.

Part II: Re-evaluation of the CAHAI using Modern Test Theories

Data from 105 participants in a previous validation study of the CAHAI was used in this part of the thesis (Barreca, Stratford, Masters, Lambert, & Griffiths, 2006). This sample size may not be sufficient for IRT analysis of the CAHAI as a two-parameter polytomous model was fitted. The frequencies of score categories used were not distributed uniformly across the CAHAI's scoring scale, with some items having less than 10 observations in specific score categories. The sample size and the distribution of the frequency of score categories may have affected the stability of the estimations using modern test theories. The instability of the estimations may affect the quality of the evidence on the psychometric properties of the CAHAI. Furthermore, the calibration and validation of the new 11-item and 5-item versions of the CAHAI were conducted on the same sample. This limits the generalizability of the study findings. Future studies are recommended to validate the study findings using a larger sample size with a more uniform distribution of score categories and to conduct the calibration and validation using different samples.

The re-evaluation of the CAHAI focused primarily on improving the psychometric properties and clinical utility of the measure. There is also a need to address the interpretation of the CAHAI scores to provide clinically meaningful information that can help to inform or direct patient care. In doing so, it can further promote the routine use of the measure in clinical practice. Thus, future studies that estimate of the minimal clinical important difference of the CAHAI scores and evaluate the extent to which CAHAI scores can predict outcomes (e.g., independence in basic and instrumental activities of daily living) are needed.

Conclusion

The continued evaluation of an outcome measure beyond its initial phase of development and validation can help improve the quality of measurement in rehabilitation. Measurement is a cornerstone in stroke rehabilitation, as well as in the field of rehabilitation science:

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be (Thomson, 1889, p. 73)

This thesis adds to the body of knowledge about an existing outcome measure of post-stroke UE function and contributes to the efforts to improve the quality of measurement in stroke rehabilitation. Continuous improvements to the quality of measurement are imperative to advance scientific knowledge in UE rehabilitation after stroke and to improve patient care and outcomes of individuals with stroke.

References

Al Zoubi, F. M., Eilayyan, O., Mayo, N. E., & Bussières, A. E. (2017). Evaluation of Cross-Cultural Adaptation and Measurement Properties of STarT Back Screening Tool: A Systematic Review. *Journal of Manipulative and Physiological Therapeutics*, *40*(8), 558–572. https://doi.org/10.1016/j.jmpt.2017.07.005

Barreca, S. R., Gowland, C. K., Stratford, P., Huijbregts, M., Griffiths, J., Torresin, W., … Masters, L. (2004). Development of the Chedoke Arm and Hand Activity Inventory: theoretical constructs, item generation, and selection. *Topics in Stroke Rehabilitation*, *11*(4), 31–42.

Barreca, S. R., Stratford, P. W., Masters, L. M., Lambert, C. L., & Griffiths, J. (2006). Comparing 2 versions of the Chedoke Arm and Hand Activity Inventory with the Action Research Arm Test. *Physical Therapy*, *86*(2), 245–253.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186–3191.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: fundamental measurement in the human sciences* (Second). Psychology Press.

Cano, S. J., Barrett, L. E., Zajicek, J. P., & Hobart, J. C. (2011). Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. *Multiple Sclerosis (Houndmills, Basingstoke, England)*, *17*(2), 214–222. https://doi.org/10.1177/1352458510385269

Carr, J. H., & Shepherd, R. B. (2010). *Neurological rehabilitation: optimizing motor performance* (2nd ed). Edinburgh ; New York: Churchill Livingstone.

Choo, S. X., JN Ng, C., Fayed, N., & Harris, J. E. (2017). International Classification of Functioning, Disability and Health Framework: Bridging adapted outcome measures. *International Journal of Therapy and Rehabilitation*, *24*(11), 494–500. https://doi.org/10.12968/ijtr.2017.24.11.494

Cieza, A., Geyh, S., Chatterji, S., Kostanjsek, N., Ustün, B., & Stucki, G. (2005). ICF linking rules: an update based on lessons learned. *Journal of Rehabilitation Medicine*, *37*(4), 212–218. https://doi.org/10.1080/16501970510040263

Claesson, L., & Svensson, E. (2001). Measures of order consistency between paired ordinal data: application to the Functional Independence Measure and Sunnaas index of ADL. *Journal of Rehabilitation Medicine*, *33*(3), 137–144.

COSMIN initiative. (n.d.). COSMIN - Improving the selection of outcome measurement instruments. Retrieved July 9, 2018, from https://www.cosmin.nl/

Duncan, E. A., & Murray, J. (2012). The barriers and facilitators to routine outcome measurement by allied health professionals in practice: a systematic review. *BMC Health Services Research*, *12*, 96. https://doi.org/10.1186/1472-6963-12-96

Feys, H., De Weerdt, W., Nuyens, G., van de Winckel, A., Selz, B., & Kiekens, C. (2000). Predicting motor recovery of the upper limb after stroke rehabilitation: value of a clinical examination. *Physiotherapy Research International: The Journal for Researchers and Clinicians in Physical Therapy*, *5*(1), 1–18.

Fugl-Meyer, A. R., Jääskö, L., Leyman, I., Olsson, S., & Steglind, S. (1975). The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scandinavian Journal of Rehabilitation Medicine*, *7*(1), 13–31.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304–312. https://doi.org/10.1037/1040-3590.6.4.304

Gosman-Hedström, G., & Blomstrand, C. (2004). Evaluation of a 5-level functional independence measure in a longitudinal study of elderly stroke survivors. *Disability and Rehabilitation*, *26*(7), 410–418. https://doi.org/10.1080/09638280410001662978

Guillemin, F. (1995). Cross-cultural adaptation and validation of health status measures. *Scandinavian Journal of Rheumatology*, *24*(2), 61–63.

Gustafsson, L. A., Turpin, M. J., & Dorman, C. M. (2010). Clinical utility of the Chedoke Arm and Hand Activity Inventory for stroke rehabilitation. *Canadian Journal of Occupational Therapy. Revue Canadienne D'ergothérapie*, *77*(3), 167–173.

Guyatt, G. H., Deyo, R. A., Charlson, M., Levine, M. N., & Mitchell, A. (1989). Responsiveness and validity in health status measurement: A clarification. *Journal of Clinical Epidemiology*, *42*(5), 403–408. https://doi.org/10.1016/0895-4356(89)90128-5

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif: Sage Publications.

Herdman, M., Fox-Rushby, J., & Badia, X. (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *7*(4), 323–335.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment (Winchester, England)*, *13*(12), iii, ix–x, 1–177. https://doi.org/10.3310/hta13120

Krosnick, J. A., Holbrook, A. L., & Visser, P. S. (2006). *Optimizing brief assessments in research on the psychology of aging: a pragmatic approach to self-report measurement*. National Academies Press (US). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK83763/

Lima, E., Teixeira-Salmela, L. F., Simões, L., Guerra, A. C. C., & Lemos, A. (2016). Assessment of the measurement properties of the post stroke motor function instruments available in Brazil: a systematic review. *Brazilian Journal of Physical Therapy*, *20*(2), 114–125. https://doi.org/10.1590/bjpt-rbf.2014.0144

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122.

Linacre, J. M. (2000). FIM levels as ordinal categories. *Journal of Outcome Measurement*, *4*(2), 616–633.

Menon, A., Korner-Bitensky, N., Kastner, M., McKibbon, K., & Straus, S. (2009). Strategies for rehabilitation professionals to move evidence-based knowledge into

practice: A systematic review. *Journal of Rehabilitation Medicine*, *41*(13), 1024–1032. https://doi.org/10.2340/16501977-0451

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

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Mokkink, L. B., Vet, H. C. W. de, Prinsen, C. a. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research*, *27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, Lidwine B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., … de Vet, H. C. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Medical Research Methodology*, *10*, 22. https://doi.org/10.1186/1471-2288-10-22

Nilsson, A. L., Sunnerhagen, K. S., & Grimby, G. (2005). Scoring alternatives for FIM in neurological disorders applying Rasch analysis. *Acta Neurologica Scandinavica*, *111*(4), 264–273. https://doi.org/10.1111/j.1600-0404.2005.00404.x

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *The British Journal of Clinical Psychology*, *46*(Pt 1), 1–18.

Petzold, A., Korner-Bitensky, N., Salbach, N. M., Ahmed, S., Menon, A., & Ogourtsova, T. (2012). Increasing knowledge of best practices for occupational therapists treating post-stroke unilateral spatial neglect: results of a knowledge-translation intervention study. *Journal of Rehabilitation Medicine*, *44*(2), 118–124. https://doi.org/10.2340/16501977-0910

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15.

Rowland, T. J., Turpin, M., Gustafsson, L., Henderson, R. D., & Read, S. J. (2011). Chedoke Arm and Hand Activity Inventory-9 (CAHAI-9): perceived clinical utility within 14 days of stroke. *Topics in Stroke Rehabilitation*, *18*(4), 382–393. https://doi.org/10.1310/tsr1804-382

Schuster, C., Hahn, S., & Ettlin, T. (2010). Objectively-assessed outcome measures: a translation and cross-cultural adaptation procedure applied to the Chedoke McMaster Arm and Hand Activity Inventory (CAHAI). *BMC Medical Research Methodology*, *10*(1), 106. https://doi.org/10.1186/1471-2288-10-106

Stewart, A. L., & Nápoles-Springer, A. (2000). Health-related quality-of-life assessments in diverse population groups in the United States. *Medical Care*, *38*(9 Suppl), II102-124.

Streiner, D. L. (2010). Measure for measure: new developments in measurement and item response theory. *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, *55*(3), 180–186. https://doi.org/10.1177/070674371005500310

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use* (Fifth edition). Oxford: Oxford University Press.

Thomson, W. (1889). Electrical units of measurement. In *Popular lectures and addresses* (Vol. 1). London, UK: Macmillan and Company.

Trombly, C. A., & Wu, C. Y. (1999). Effect of rehabilitation tasks on organization of movement after stroke. *The American Journal of Occupational Therapy: Official Publication of the American Occupational Therapy Association*, *53*(4), 333–344.

Veerbeek, J. M., Kwakkel, G., Wegen, E. E. H. van, Ket, J. C. F., & Heymans, M. W. (2011). Early prediction of outcome of activities of daily living after stroke: a systematic review. *Stroke*, *42*(5), 1482–1488. https://doi.org/10.1161/STROKEAHA.110.604090

Velstra, I.-M., Ballert, C. S., & Cieza, A. (2011). A systematic literature review of outcome measures for upper extremity function using the international classification of functioning, disability, and health as reference. *PM & R: The Journal of Injury, Function, and Rehabilitation*, *3*(9), 846–860. https://doi.org/10.1016/j.pmrj.2011.03.014

Wierzbicka, A. (2003). Singapore English: a semantic and cultural perspective. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, *22*(4), 327–366. https://doi.org/10.1515/mult.2003.018

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural

adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural Adaptation. *Value in Health*, *8*(2), 94–104. https://doi.org/10.1111/j.1524-4733.2005.04054.x

Wolf, S. L., Kwakkel, G., Bayley, M., McDonnell, M. N., & Upper Extremity Stroke Algorithm Working Group. (2016). Best practice for arm recovery post stroke: an international application. *Physiotherapy*, *102*(1), 1–4. https://doi.org/10.1016/j.physio.2015.08.007

World Health Organization. (2001). *International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–860.

Wu, C., Trombly, C. A., Lin, K., & Tickle-Degnen, L. (1998). Effects of object affordances on reaching performance in persons with and without cerebrovascular accident. *The American Journal of Occupational Therapy: Official Publication of the American Occupational Therapy Association*, *52*(6), 447–456.

Yao, M., Wang, Q., Li, Z., Yang, L., Huang, P.-X., Sun, Y.-L., … Cui, X.-J. (2016). A Systematic Review of Cross-cultural Adaptation of the Oswestry Disability Index. *Spine*, *41*(24), E1470–E1478. https://doi.org/10.1097/BRS.0000000000001891