A cross-linguistic investigation of reliability of measurement of stuttering severity

Laura Hoffman
BHlthSci(Speech Path)(Hons)

A thesis submitted to Charles Sturt University in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Community Health
Faculty of Science
March 2018
Table of Contents

List of Tables...........................................................................................................viii
List of Figures.............................................................................................................ix
Abbreviations............................................................................................................x
Certificate of Authorship............................................................................................xi
Acknowledgements....................................................................................................xii
Ethical Approvals.......................................................................................................xiv
Publication/s and Presentation Arising from this Research..........................................xv
Co-author Publication Statement................................................................................xvi
Abstract....................................................................................................................xvii

CHAPTER 1
Overview of Stuttering and Linguistic Diversity......................................................1
The Cause and Nature of Stuttering..........................................................................3
   Theoretical Explanations of the Cause and Nature of Stuttering.........................4
   Multifactorial models............................................................................................5
   Cognitive and linguistic processing........................................................................5
   Speech motor control.........................................................................................7
Linguistic Diversity and Stuttering..........................................................................10
   Cross-linguistic Investigation of Stuttering.........................................................11
   Manifestation of stuttering in different languages..............................................12
   Stuttering and Bilingualism..............................................................................17
   Manifestation of stuttering in bilingual speakers.............................................19
Measuring Stuttering in Unfamiliar Languages.......................................................23
The Present Thesis...................................................................................................26
CHAPTER 2

Measurement of Stuttering.................................................................29

Measures used for the Management of Stuttering.................................31

  Measuring Stuttering Frequency.........................................................31
  Percentage syllables stuttered..........................................................31
  Time interval measurement...............................................................32

  Measuring Stuttering Severity.........................................................33
  Severity Rating Scale......................................................................33
  The Stuttering Severity Instrument –Fourth Edition..........................34

  Measuring Speech Rate......................................................................35

  Measuring Speech Naturalness..........................................................35

  Measuring Impact of Stuttering.........................................................36

    The Overall Assessment of the Speaker’s Experience of Stuttering....36

    The Wright and Ayre Stuttering Self-Rating Profile.........................37

  Measuring Anxiety...........................................................................38

    The Unhelpful Thoughts and Beliefs About Stuttering Scales..........39

    The Fear of Negative Evaluation Scale..........................................40

    Spence Children’s Anxiety Scale....................................................41

    The Preschool Anxiety Scale Revised..........................................42

Measurement of Focus for the Program of Research...........................43

CHAPTER 3

Study 1: The Effect of Speech Sample Duration on the Reliability of Measurement of Severity of Stuttering.................................................................48

  Background.....................................................................................49

  Method..............................................................................................53

    Speech Samples...............................................................................53

    Judges............................................................................................53

    Stuttering Measurement Tool.......................................................54

    Procedure.......................................................................................54

    Data Analysis..................................................................................55

  Results...............................................................................................56

  Discussion.........................................................................................60
CHAPTER 4

Study 2: The Reliability of Measuring Severity of Stuttering in a Foreign Language

Research Investigating Identification and Measurement of Stuttering in Unfamiliar Languages

Aim of the Present Study

Method

Speech Samples

Speech sample selection process

Judges

Stimulus Package

Procedure

Data Analysis

Intrajudge agreement

Interjudge agreement

Performance in Vietnamese for the subgroup of judges reliable in English

Logistic regression

Results

Descriptive Statistics

Intrajudge agreement

Interjudge agreement

Logistic Regression

Discussion

Conclusion

CHAPTER 5

Study 3: An Exploration of Reliability of Stuttering Measurement in Vietnam

Speech-language Pathology in Vietnam

Vietnamese stuttering data

Overview of the English and Vietnamese Languages

Aim of the Present Study
CHAPTER 6

Study 4: Comparison of Reliability Across Different Languages..........117

Background.................................................................................118

Aim..........................................................................................123

Method.......................................................................................124

Logistic Regression.....................................................................124

Descriptive Statistics..................................................................125

Intrajudge and interjudge agreement.......................................125

Results........................................................................................126

Logistic Regression.....................................................................126

Does percentage agreement vary over the severity range?.....127

Does language familiarity (familiar versus unfamiliar or other) impact on judges’ performance?.................................129

What is the difference between intrajudge and interjudge agreement?........................................................................ 130
Do any of these three effects vary according to the rating type or language familiarity (familiar versus unfamiliar or other)?

Descriptive Statistics

Intrajudge and interjudge agreement

Discussion

Main Finding One: Judges were more reliable in measuring severity of stuttering using a 9-point stuttering severity rating scale in a familiar language than in an unfamiliar or other language.

Main Finding Two: Judges demonstrated better intrajudge than interjudge reliability.

Main Finding Three: Judges’ performance in the use of the severity rating scale varied depending on stuttering severity.

Conclusion

CHAPTER 7

Conclusion

Summary of Findings of this Research

What this Thesis has Contributed and Implications

Determining Duration of Speech Samples

Improving Clinical Utility of Reliability Research

Expanding Reliability Research into Other Language Families

Identifying Variability of Agreement Across the Severity Rating Scale

Future Directions

Speech Sample Duration

Cross Lingustic Investigation of Other Language Families

Continued Stuttering Research in Vietnam

The Need for Training

The Development of a Stuttering Severity Training Package

Final Remarks

References
# Appendices

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix A</td>
<td>Study 1 Information Sheet</td>
<td>200</td>
</tr>
<tr>
<td>Appendix B</td>
<td>Study 1 Consent Form</td>
<td>202</td>
</tr>
<tr>
<td>Appendix C</td>
<td>Study 1 Instruction Sheet</td>
<td>204</td>
</tr>
<tr>
<td>Appendix D</td>
<td>Study 1 Severity Rating Score Sheets</td>
<td>206</td>
</tr>
<tr>
<td>Appendix E</td>
<td>Study 1 Questionnaire</td>
<td>212</td>
</tr>
<tr>
<td>Appendix F</td>
<td>Study 2 First Rating Session Instruction Sheet</td>
<td>214</td>
</tr>
<tr>
<td>Appendix G</td>
<td>Study 2 Severity Rating Score Sheets</td>
<td>216</td>
</tr>
<tr>
<td>Appendix H</td>
<td>Study 2 Questionnaire</td>
<td>221</td>
</tr>
<tr>
<td>Appendix I</td>
<td>Study 2 Second Rating Session Information Sheet</td>
<td>223</td>
</tr>
<tr>
<td>Appendix J</td>
<td>Study 2 Second Rating Session Consent Form</td>
<td>225</td>
</tr>
<tr>
<td>Appendix K</td>
<td>Study 2 Second Rating Session Instruction Sheet</td>
<td>226</td>
</tr>
<tr>
<td>Appendix L</td>
<td>Study 2 First Rating Session Information Sheet</td>
<td>228</td>
</tr>
<tr>
<td>Appendix M</td>
<td>Study 2 First Rating Session Consent Form</td>
<td>230</td>
</tr>
<tr>
<td>Appendix N</td>
<td>Study 3 First Rating Session Instruction Sheet</td>
<td>232</td>
</tr>
<tr>
<td>Appendix O</td>
<td>Study 3 Severity Rating Score Sheets</td>
<td>234</td>
</tr>
<tr>
<td>Appendix P</td>
<td>Study 3 Questionnaire</td>
<td>239</td>
</tr>
<tr>
<td>Appendix Q</td>
<td>Study 3 Second Rating Session Information Sheet</td>
<td>241</td>
</tr>
<tr>
<td>Appendix R</td>
<td>Study 3 Second Rating Session Consent Form</td>
<td>243</td>
</tr>
<tr>
<td>Appendix S</td>
<td>Study 3 Second Rating Session Instructions</td>
<td>245</td>
</tr>
<tr>
<td>Appendix T</td>
<td>Study 3 First Rating Session Information Sheet</td>
<td>246</td>
</tr>
<tr>
<td>Appendix U</td>
<td>Study 3 First Rating Session Consent Form</td>
<td>248</td>
</tr>
</tbody>
</table>
List of Tables

Table 3.1  Mean, mode, range, and standard deviation of severity rating scale scores assigned to each speech sample by the judges. 56

Table 4.1  Age, gender, and stuttering severity rating scale scores of the adults depicted in the speech samples. 77

Table 4.2  Mean, mode, range, and standard deviation of severity rating scale scores assigned by the judges to each speech sample. Data are presented for the first rating occasion only. 84

Table 4.3  Percentage of judges who achieved criterion intra- and interjudge reliability (≥80% agreement) and ≥70% agreement. 85

Table 4.4  Estimated odds ratio, 95% confidence intervals, and p-values for language, and rating type. 88

Table 5.1  Mean, mode, range, and standard deviation of severity rating scale scores assigned by the judges to each speech sample. Data are presented for the first rating occasion only. 106

Table 5.2  Percentage of judges who achieved criterion intra- and interjudge reliability (≥80% agreement) and ≥70% agreement. 107

Table 5.3  Estimated odds ratio, 95% confidence intervals, and p-values for language, and rating type. 110

Table 6.1  Age, gender, and stuttering severity rating scale scores of the adults depicted in the speech samples in the Hoffman et al. (2014) study and Studies 2 and 3 of the current research. 122

Table 6.2  Estimated odds ratio, 95% confidence intervals, and p-values for language, and rating type for all three studies. 129

Table 6.3  Combined results of percentage of judges who achieved criterion intra- and interjudge reliability (≥80% agreement) and ≥70% agreement. 132
**List of Figures**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Severity rating scale.</td>
<td>45</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Standard deviation of each speech sample by the mean severity for each speech sample for the three durations of speech samples, 1, 3, and 5-minutes.</td>
<td>58</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Curves from the fitted regression model with separate lines for 1-minute (upper line, circles), 3-minutes (middle, crosses) and 5-minutes (lowest, diamonds).</td>
<td>59</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Plot showing the relationship between the three significant variables from the final model. The agreement percentages for each of the sample/time(1,2)/agreement type combinations are also plotted (+ intrajudge, o interjudge). The two panels are for the language categories (English, Vietnamese).</td>
<td>87</td>
</tr>
<tr>
<td>Figure 5.1</td>
<td>Plot showing the relationship between the three significant variables from the final model. The agreement percentages for each of the sample/time(1,2)/agreement type combinations are also plotted (+ intrajudge, o interjudge). The two panels are for the language categories (Vietnamese, English).</td>
<td>109</td>
</tr>
<tr>
<td>Figure 6.1</td>
<td>Plot showing the relationship between the three significant variables from the final model. The agreement percentages for each of the sample/time(1,2)/agreement type combinations are also plotted (+ intrajudge, o interjudge). The panels are for the language categories (for Hoffman et al., 2014), Australian English, Mandarin; for Study 2, English, Vietnamese; and for Study 3, Vietnamese, English.</td>
<td>128</td>
</tr>
</tbody>
</table>
Abbreviations

BFNE Brief Fear of Negative Evaluation Scale
BFNE-II Brief Fear of Negative Evaluation Scale – Version 2
BFNE-R Brief Fear of Negative Evaluation Scale – Revised
BFNE-S Brief Fear of Negative Evaluation Scale – Straightforward Items
EnglishVNr Performance in English for the Subgroup of Judges Reliable in Vietnamese
FNE Fear of Negative Evaluation Scale
GLMM Generalized Linear Mixed Model
LMM Linear Mixed Models
MandarinAEr Performance in Mandarin for the Subgroup of Judges Reliable in Australian English
OASES The Overall Assessment of the Speaker’s Experience of Stuttering
OASES-A The Overall Assessment of the Speaker’s Experience of Stuttering – Adult version
OASES-S The Overall Assessment of the Speaker’s Experience of Stuttering – School aged version (7-12 years)
OASES-T The Overall Assessment of the Speaker’s Experience of Stuttering – Adolescent version (13-17 years)
OR Odds ratio
SD Standard deviation
SR Severity Rating
UTBAS The Unhelpful Thoughts and Beliefs About Stuttering
VietnameseENGr Performance in Vietnamese for the Subgroup of Judges Reliable in English
WASSP The Wright and Ayre Stuttering Self-Rating Profile
Certificate of Authorship

I, Laura Carol-Ann Hoffman, hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma at Charles Sturt University or any other educational institution, except where due acknowledgment is made in the dissertation. Any contribution made to the research by colleagues with whom I have worked at Charles Sturt University or elsewhere during my candidature is fully acknowledged.

I agree that this dissertation be accessible for the purpose of study and research in accordance with the normal conditions established by the Executive Director, Library Services or nominee, for the care, loan and reproduction of theses.

_________________________________  _______________________

Signature                      Date

Laura Hoffman
Printed name
Acknowledgements

There are many people whose support, advice, influence, and contributions have proved invaluable, and I thank God for everyone who was placed in my path and supported me along the way. I would like to express my sincere gratitude to my amazing supervisors, Linda Wilson and Sally Hewat. I cannot thank you enough for your continual support, patience, guidance, encouragement, immense knowledge shared, and time and energy you have both invested in me throughout the process. Linda, I love learning from you and working with you. You are always challenging me and I value your honest feedback, attention to detail, and selfless time and care you provide me with time and time again. Sally, you have a real passion for research and your continual enthusiasm for my research was great motivation to keep going and finish strong.

I gratefully acknowledge Dr Thao Bich Huynh for all of her assistance with recruitment, speech sample collection, translations, and data collection. I have learnt a great amount whilst collaborating together and appreciate all you have done. Many thanks.

I acknowledge Speech Pathology Australia for the financial assistance provided in the form of a Higher Degree Research Grant to undertake this research. An additional thanks to Charles Sturt University for the Australian Postgraduate Award and the assistance it provided.

I am appreciative to Kim Colyvas for his assistance with data analyses and Nguyễn Trần Thảo Ngọc for her assistance with the translation process. I extend my thanks to the staff at Newcastle University speech pathology clinic for all of your help with the collection of our English speech samples.

I sincerely thank the individuals who allowed me to use their speech samples across studies, the bilingual panel members, and speech-language pathologists who graciously
gave their time to participate in this research. This research would not have been possible without each and everyone of you, and for that, I am truly thankful.

To my family, Mum, Grandad, Antoinette, Sandra, Aydan, Gary, Boki, Melissa, Bjay, Ivana, Maddox, Jordan, Katie, Samuel, Alex, Odette, and Emiliya – thank you for all your prayers, unconditional love, support, patience, and encouragement. I am blessed to have all of you in my life and there was no way I would have been able to complete such a task without your motivation, phone/video calls, visits, meals, and photo updates of my beautiful nieces and nephews. You have been supportive of my decision to take on a PhD and have gone above and beyond to help out whenever possible. While we are separated by distance, I have never felt alone. Thank you!
Ethical Approvals

Research presented in this thesis has been approved by Charles Sturt University’s Human Research Ethics Committee (Protocol numbers: 405/2013/08; 405/2013/03; 405/2014/04; 405/2014/03; 405/2011/06) and by The University of Newcastle (Protocol number: H-2013-0256) Human Research Ethics Committee.
Publication/s and Presentation Arising from this Research

Publication arising from this research


Journal articles in progress


Presentation arising from this research

Co-author Publication Statement

As co-authors of the publication arising from this thesis (see previous section) we confirm that Laura Hoffman has made the following contributions to the publication: conception and design, review and interpretation of the literature, design of the stimulus package contents including the questionnaire and editing of speech samples, analysis and interpretation of the findings, preparation of the manuscripts and critical appraisal of contents for the publication, and submission of manuscript for publication.

Signed: Linda Wilson
Date: 08/03/2018

Signed: Sally Hewat
Date: 04/03/2018

Signed: Thao Bich Huynh
Date: 03/03/2018
Abstract

Stuttering measurement is a vital component of assessment and treatment. It is essential for speech-language pathologists to reliably measure stuttering, to enable efficient tracking of clients’ progress and outcomes. Stuttering severity rating scales are widely used in research and clinical practice, and are recommended for measuring severity of stuttering in different languages. With the popularity of severity rating scales worldwide, investigation into the reliability of use of the scale is critical, regardless of whether it is being used for clinical or research purposes. Thus, the present research investigates speech-language pathologists’ reliability of measuring severity of stuttering of adult-speakers who stutter, across languages, using a 9-point severity rating scale.

Previous research has examined speech-language pathologists’ abilities to reliably measure severity of stuttering using a severity rating scale and has shown conflicting results. However, these studies used different analyses and varying durations of speech samples, which could have contributed to the observed differences. Study 1 presents the first study, which examined whether the duration of speech samples influenced the reliability of measurement of severity of stuttering. Results from the study determined the duration of speech samples used for the remainder of studies in the program of research. Specialist speech-language pathologists rated audio, English samples of three different durations (1, 3, and 5-minutes) of adults who stutter, using a 9-point severity rating scale. Judges showed greater variability (poorer reliability) when rating 1-minute speech samples. The 3- and 5-minute samples produced similar results when using the scale to measure stuttering severity.

Studies 2 and 3 used 3-minute speech samples to investigate the reliability of use of the scale across languages. With a rise in globalisation, speech-language pathologists are increasingly tasked with assessing and treating stuttering in languages
that may differ from their own. Study 2 expanded existing knowledge of measuring severity of stuttering in a foreign language that is commonly spoken worldwide: Vietnamese. The aim of Study 2 was to determine whether English-speaking speech-language pathologists could reliably measure severity of stuttering in English (familiar language) and Vietnamese (unfamiliar language); while Study 3 provides the first stuttering measurement data for Vietnamese-speaking speech-language pathologists in Vietnam and investigated whether Vietnamese-speaking speech-language pathologists can reliably measure severity of stuttering in Vietnamese (familiar language) and English (other language, a common second language for Vietnamese-speakers). Results of Studies 2 and 3 exhibited consistent findings: (i) judges performed better measuring severity of stuttering in a familiar language than an unfamiliar/other language; (ii) judges produced greater intrajudge than interjudge agreement, for both languages; and (iii) judges showed better use of the scale for mild and severe stuttering across both languages, with the moderate range of the scale showing greater variability.

The main purpose of Study 4 was to present a re-analysis of data from Hoffman, Wilson, Copley, Hewat, and Lim (2014) to determine whether the consistent findings across the present studies were robust. The re-analysis investigated Australian-English speech-language pathologists’ reliability of use of the scale in Australian English (familiar language) and Mandarin (unfamiliar language). The re-analysis replicated the analyses of Studies 2 and 3. Results of the re-analyses corroborated those of Studies 2 and 3, resulting in three robust findings. The findings of the studies combined are synthesised, with a discussion of the contributions of the research and suggestions for future directions.
CHAPTER 1

Overview of Stuttering and Linguistic Diversity
Effective communication is integral for people of all ages irrespective of linguistic background. Communication is vital to occupational participation, educational attainment, social engagement, emotional health, and contributes immensely to an individual’s quality of life. Yet at least 1% of the world’s population live with a speech disorder known as stuttering (Bloodstein & Bernstein Ratner, 2008). Stuttering is associated with negative consequences across the lifespan, such as preschool age frustration and withdrawal, childhood bullying, stereotypes, occupational and educational disadvantage, and reduced quality of life (Blood & Blood, 2007; Blood, Boyle, Blood, & Nalesnik, 2010; Iverach et al., 2009; Klein & Hood, 2004; Langevin, Packman, & Onslow, 2010; MacKinnon, Hall, & MacIntyre, 2007; Yaruss & Quesal, 2004). If left untreated, the consequences of stuttering can impede the effectiveness of verbal communication, and thus has the potential to induce anxiety, embarrassment, avoidance behaviour, and/or poor self-esteem to an already disabling speech disorder (Bricker-Katz, Lincoln, & McCabe, 2009; Cream, Onslow, Packman, & Llewellyn, 2003; Riley, Riley, & Maguire, 2004). Stuttering is found worldwide and indiscriminate of culture, race, and linguistic diversity. With increasing multiculturalism, speech-language pathologists are highly likely to work with stuttering clients from linguistic backgrounds which differ from their own.

This chapter presents information about the cause and nature of stuttering. Next, an overview of the association between linguistic diversity and stuttering is presented, with commentary on the presentation of stuttering across different languages. The chapter concludes with an explanation of the focus of this research.
The Cause and Nature of Stuttering

Stuttering is a debilitating disorder in which speech disruptions interfere with or prevent normal communication. Speech disruptions, otherwise identified as unambiguous stuttering moments, can be described using the Lidcombe Behavioural Data Language (Teeson, Packman & Onslow, 2003) for stuttering across all ages. There are three categories of stuttering within the taxonomy: repeated movements, fixed postures, and superfluous behaviours. Repeated movements comprise of syllable repetition (e.g., “I-I-I want that”), incomplete syllable repetition (e.g., “I like your d-d-dog”), and multisyllable unit repetition (e.g., “kanga- kanga- kangaroo”). Fixed postures may occur with audible airflow (e.g., “nnnnot now”) which may sound prolonged, or without audible airflow (e.g., “I......[no sounds] won”) whereby the speaker appears to be ‘stuck’ or ‘blocked’ on speech production. Superfluous behaviours include additional verbal or nonverbal behaviour/s that are not part of the intended communication. Verbal superfluous behaviours may include redundant words/phrases or excessive amounts of filler words such as “um”, “like”, or “ah” (e.g., “ah so um-um-um I think, um, that um I will go home”). Nonverbal superfluous behaviours, which include tics, postural adjustments, blinking, grimacing, or other facial expressions, may occur alongside verbal disruptions (Guitar, 2014).

Onset of stuttering can be sudden or gradual, with the average age of onset at 33 months (Yairi & Ambrose, 2012). The age of stuttering onset is typically between 24 and 42 months. However, stuttering may begin prior to 24 months or up to age 5. The distribution of age of onset is similar for girls and boys (Yairi & Ambrose, 2012).

In a review of 44 epidemiological studies of stuttering from several countries, Bloodstein and Bernstein Ratner (2008) identified that stuttering affects approximately 1% of the population at any given time. The majority of the studies were carried out
with preschool, school aged and/or adolescent aged participants, with the exception of Craig, Hancock, Tran, Craig, and Peter’s (2002) research. The latter is the only epidemiology study to investigate prevalence data across the lifespan (i.e., 2 – 99 years of age) and yielded an overall prevalence of 0.72%. Thus, the overall prevalence appears to be lower than the traditional figure of 1%.

Lifetime incidence of stuttering had been universally recognised and accepted in the 20\textsuperscript{th} century at approximately 5% (Bloodstein & Bernstein Ratner, 2008). However, more recent studies (e.g., Dworzynski, Remington, Rijsdijk, Howell, & Plomin, 2007; Felsenfeld et al., 2000; Reilly et al., 2009) suggest a higher incidence, with a central figure of 8% or higher. A community cohort study conducted in Australia revealed a 3 year cumulative incidence of 8.5% (Reilly et al., 2009), and a 4 year cumulative incidence of 11.2% (Reilly et al., 2013). Both of these figures are higher than the conventional figure of 5%. What is encouraging to see, however, is that the differences between prevalence and incidence figures indicate that most people recover from stuttering, either with or without treatment (Yairi & Seery, 2011).

Theoretical Explanations of the Cause and Nature of Stuttering

Despite many theories about the cause and nature of stuttering, the specific cause of stuttering currently remains unknown. The term \textit{theory} refers to the important causal relationships in a phenomenon (Guitar, 2014). Theories are developed as conceptual frameworks to assist and guide research and thinking. Many theoretical perspectives have been proposed over decades to describe the cause and nature of stuttering (for reviews, see Bloodstein and Bernstein Ratner, 2008; Packman & Attanasio, 2017). Packman and Attanasio (2017) reviewed numerous causal theories and categorised them under the following classifications in terms of understanding stuttering as a disorder of: multifactorial models, cognitive and linguistic processing,
and speech motor control. With reference to Packman and Attanasio’s broad classifications, the following section provides a brief overview of the more current theoretical explanations of the cause and nature of stuttering.

Multifactorial Models

Multifactorial models propose that more than one factor needs to be present for stuttering to occur. Hence, a combination of causal factors is believed to vary across individuals rather than stuttering being caused by one single factor. Examples of causal factors include innate (e.g., genetic, emotional, cognitive, and linguistic), and environmental factors. There are a number of multifactorial models, the most clinically influential model being the Demands and Capacities Model (Starkweather & Gottwald, 1990). This treatment model proposes that demands to communicate are greater than the person’s capacity to do so. Demands include time pressures to speak, environmental pressures such as parental demands for increased cognitive functioning, parents’ and peers’ demands for a child to speak, and increased levels of excitement and anxiety. Capacities consist of factors such as motor, linguistic, socioemotional, and cognitive development. If demands on an individual are greater than the individual’s capacity, stuttering will occur.

Cognitive and Linguistic Processing

Several theories propose an association between cognitive and linguistic processing and stuttering (see Packman & Attanasio, 2017; Sasisekaran, DeNil, Smyth, & Johnson, 2006). These theories include the Anticipatory Struggle Hypothesis (Bloodstein, 1995; 1997), the Neuropsycholinguistic theory (Perkins, Kent, & Curlee, 1991), the Covert Repair Hypothesis (Kolk & Postma, 1997; Postma & Kolk, 1993), the Fault Line Hypothesis (Wingate, 1988), and the EXPLAN theory (Howell, 2004). All of
these theories allude that stuttering occurs during the speech production process as a result of a disruption at the level of phonological encoding.

According to the *Anticipatory Struggle Hypothesis* (Bloodstein, 1995; 1997), a child will come to believe that speech is difficult. This makes them anticipate difficulty with speaking, which causes struggle and tension, resulting in stuttering. A few years later, Bloodstein (2001) modified this hypothesis to accommodate emerging evidence of genetic involvement in development of stuttering in early childhood, with anticipatory struggle patterns becoming evident in older childhood and adulthood. Thus, the modified *Anticipatory Struggle Hypothesis* has become more of an explanation of the *nature* of stuttering in later years than a *hypothesis* describing the cause of stuttering at onset.

In recent years, the *Covert Repair Hypothesis* has been the most frequently cited and perhaps most influential theory pertaining to cognitive and linguistic processing. This theory suggests that a disturbance in phonological encoding will lead to a greater number of errors during the formulation of phonetic encoding for speech. The errors are detected by an impaired internal speech monitor for the phonological encoding of speech which then attempts to make corrections that result in stuttered speech. Stuttering, therefore, reflects the need to repeatedly repair speech programs before speech motor execution (Postma & Kolk, 1993).

The *EXPLAN theory* developed by Howell (2004) differs from the *Covert Repair Hypothesis* as it proposes the motor plan to be delayed (issues in timing), instead of being incorrect. *PLAN* refers to the parallel language planning system, while the *EX* relates to the speech execution system (Savage & Howell, 2008). According to *EXPLAN*, speech planning (*PLAN*) and execution (*EX*) are independent processes. However, they do interact with one another. Howell (2004) proposed that stuttering is a
result of an error occurring during the interaction of the linguistic and motoric systems. Errors in timing between planning (e.g., language processing deficits) and/or execution (motor timing deficits) result in stuttering. Thus, it is important that the language and motor systems work together for fluent speech (Howell, 2010). Based on EXPLAN, a plan might not be ready for execution because a word may be too linguistically complex or because the motor speech rate in the preceding sentence is too rapid. When a plan is not ready, the individual may produce one of two responses: (i) repetition or hesitation when producing whole function words as a delaying strategy (stalling) in an attempt to buy time to prepare for the subsequent content word that is not ready; and (ii) problems on part of content words, whereby an individual attempts to execute the content word despite not being ready for it to be produced, resulting in part-word repetitions (Howell, 2004). EXPLAN explains that the planning of the linguistic segments of content words is slow because they are more difficult and take longer to plan and execute than function words (Howell, 2004).

Speech Motor Control

The Interhemispheric Interference Model (Foster & Webster, 2001; Webster, 1985; 1986; 1988; 1998; 2004) is an example of a speech motor control model used to explain the origin of stuttering. This model suggests that the cause of stuttering is associated with irregularities of interhemispheric activity and of the neural mechanism of speech motor control. According to this model, an insufficient supplementary motor area together with a labile system of hemispheric activation is needed for stuttering to occur. The supplementary motor area is particularly fragile and susceptible to other ongoing neural activity. Thus, disturbance of the supplementary motor area is caused by (i) concurrent neural activity; (ii) a lack of left hemispheric activation; and (iii) interference of the over active right hemisphere, arising from increased negative
emotions associated with stuttering (Webster, 2004). These disturbances together affect the systems that regulate speech motor output, resulting in stuttered speech.

There are further causal models and theories that incorporate speech motor control to explain the origin of stuttering. Examples include the Sensory-motor Modelling Theory (Andrews et al., 1983; Neilson & Neilson, 1985, 1987, 2000; Neilson, Neilson, & O’Dwyer, 1992), the Neuroscience Model (Nudelman, Herbrich, Hess, Hoyt, & Rosenfield, 1992; Nudelman, Herbrich, Hoyt, & Rosenfield, 1989), and the Variability Model (Vmodel; Packman, Code, & Onslow, 2007; Packman, Onslow, & Menzies, 2000; Packman, Onslow, Richard, & Van Doorn, 1996).

Although the models and theories cited differ in their explanations of stuttering and in the extent to which they are modelled by speech motor control, they all make reference to the instability in the speech motor system as the basis of stuttering. In brief, the Sensory-motor Modelling Theory and the Neuroscience Model suggest that higher linguistic processes lead to disruptions in the complex monitoring or multiloop control system which subserves speech production. The Vmodel, however, proposes that the motor speech system is susceptible to destabilization because of variability that is inherent in the production of syllabic stress (Packman & Lincoln, 1996). This model suggests that stuttering occurs due to difficulty in the initiation of motor planning for syllable production. Although stuttering has been widely thought to be caused by a deficit in neural processing for speech that is influenced by linguistic and environmental factors (Van Lieshout, Hulstijn, & Peters, 2004), Packman et al., (1996, p. 253) stated that “it is the particular interaction of linguistic and motoric factors inherent in prosody that induces stuttering, and psychological and environmental factors then influence the course of the disorder”. Packman et al., (2007) furthermore linked difficulty with
initiating and sequencing syllables to an underlying dysfunction with the supplementary motor area.

With stuttering currently thought to be due to a deficit or inefficiency in the neural processing underpinning the production of spoken language (Buchel & Sommer, 2004; Chang, Erikson, Ambrose, Hasegawa-Johnson, & Ludlow, 2008; Packman, Code, & Onslow, 2007; Watkins, Smith, Davis, & Howell, 2008), one could consider the most recent theory by Packman (2012), the Packman and Attanasio’s three factor causal model (known as the P&A model). The P&A model was proposed in an attempt to describe the underlying causal aspects of individual stutters. This model is based on the assumption that three causal factors are operating during each moment of stuttering: (i) a deficit in neural processing for spoken language; (ii) triggers for stuttering moments; and (iii) modulating factors. Variable syllabic stress and linguistic complexity are features of speech that can be triggers of stuttering. Whilst the notion that stuttering is associated with linguistic complexity is not new, Packman attributed it to the motoric task demands that the complexity places on the unstable speech system. Modulating factors include physiological arousal, which refers to how an individual will respond to stressful internal and external stimuli. It was concluded that differing stuttering severity across individuals was attributed to the extent of impairment in neural processing across individuals, while the range of modulating factors explain the variability of stuttering within individuals, across communication contexts (Packman, 2012).

It is important to take into account that the range of competing theories presented are predominately based on findings from European studies, which therefore gives a limited perspective, as some theories may not apply to other languages (e.g., Vietnamese, Mandarin, or other Asian languages). Whilst stuttering occurs across all
languages (Guitar, 2014; Shenker, 2011a), there has been some suggestion that stuttering develops differently across languages (Bloodstein & Bernstein Ratner, 2008). Various theories and/or models of stuttering have postulated stuttering to have an association with the level of either phonetic encoding (see Packman et al., 2007), phonological encoding (see Postma & Kolk, 1993; Sasisekaran et al, 2006), or lexical retrieval (see Newman & Bernstein Ratner, 2007). It has been suggested that phonological encoding systems differ across languages (Kureta, Fushimi, & Tatsumi, 2006). Given that stuttering may be associated with a weakness at varying levels of linguistic processing, it is important to consider whether stuttering manifests differently across languages. Thus, the next section will discuss stuttering across various languages.

**Linguistic Diversity and Stuttering**

Stuttering occurs across all languages and cultures (Shenker, 2011a), and exists in both monolingual and bilingual\(^1\) speakers (Finn & Cordes, 1997; Van Borsel, Maes, & Foulon, 2001). This, combined with a rise in globalisation, means that speech-language pathologists are required to work with a linguistically diverse caseload, and therefore are faced with two possible scenarios: (i) needing to work with monolingual-speakers whose language differs from that of the speech-language pathologist; and/or (ii) working with bilingual and/or multilingual-speakers who might share one or possibly no language/s with the speech-language pathologist. It is vital that speech-language pathologists are able to provide appropriate stuttering treatment to all clients, irrespective of their cultural and linguistic background. Thus, speech-language pathologists are increasingly challenged with the task of working with people who

---

\(^1\) There are insufficient data as to whether there are any clinically relevant differences between bilingual or multilingual speakers (Roberts & Shenker, 2007). Thus, in this thesis, the term “bilingual” is also used to refer to “multilingual” speakers.
speak language/s other than their own. This poses a multitude of challenges, including (i) understanding how stuttering manifests in different languages; (ii) providing culturally appropriate services to all clients; and (iii) measuring stuttering in all languages spoken by the client, to monitor progress and outcomes. The latter could include measuring stuttering in a language that may be unknown to them, and this challenge will form a significant part of the present research.

The following section provides a discussion of the loci of stuttering in different languages, along with a summary of what is known to date regarding stuttering in Asian languages. Next, the relationship between stuttering and bilingualism is discussed. The section then concludes with a summary of measuring stuttering across linguistically diverse populations, with a focus on measuring stuttering in unfamiliar languages.

**Cross-linguistic Investigation of Stuttering**

The clinical and theoretical value of investigating stuttering across languages has been identified by many researchers (e.g., Roberts & Shenker, 2007; Shenker, 2011a; Shenker, 2011b; Van Borsel et al., 2001). Cross-linguistic stuttering research is important for research and clinical purposes as it has the potential to divulge common characteristics and underlying attributes of stuttering, enabling researchers to identify what aspects of stuttering are universal, and what aspects may be language specific (Howell & Rusbridge, 2011). Specifically, examining the loci of stuttering across languages has the ability to provide insight into the connections between stuttering and linguistic aspects of the languages (Watson, Byrd, & Carlo, 2011a) as well as possibly enlighten causal factors (Juste, Sassi, & De Andrade, 2012). Furthermore, cross linguistic examinations would also be useful for guiding researchers and clinicians in improving the efficiency of assessing and diagnosing stuttering across languages.
Linguistic influences on stuttering have been examined for many years, dating back to Spencer Brown who, in a range of publications from 1935 to 1945 (e.g., Brown, 1937, 1938a, 1938b, 1938c, 1945; Brown & Moren, 1942), identified several linguistic variables that increased the likelihood of stuttering. Brown showed that English-speaking adults who stutter will stutter more frequently on consonants, on longer words, on sounds in the initial-word position, on nouns, adjectives, verbs and adverbs, on words at the beginning of a sentence, on stressed syllables, and in contextual speech. While Brown’s contribution is critical to our understanding of the loci of stuttering, the research was exclusive to English studies of adult-speakers.

Due to linguistic differences across languages, stuttering loci may vary from one language to another. Bernstein Ratner and Benitez (1985) proposed that more complex languages may require a longer planning interval which may result in stuttering. This, furthermore, supports the notion that stuttering could manifest differently in languages or in phonological systems that are similar or dissimilar.

A commonality reported across the languages for which English language literature was available is that moments of stuttering tend to occur at the beginning of a word, sentence, or clause in English (e.g., Howell & Rusbridge, 2011), German (e.g., Dworzynski, Howell, & Natke, 2003), Spanish (Watson et al., 2011a; Watson, Byrd, & Carlo, 2011b) Japanese (Ujihira, 2011), and Persian (Karimi & Nilipour, 2011). For Persian, stuttering was also found to occur on words starting with glottal stop consonants (Karimi & Nilipour, 2011).

Research pertaining to stuttering loci has often been linked to word classes (i.e., content and function). For example, literature has stated that function words are stuttered more frequently than content words for young children in English (e.g., Au-
Yeung, Howell, & Pilgrim, 1998; Bloodstein & Gantwerk, 1967; Bloodstein & Grossman, 1981; Buhr & Zebrowski, 2009; Howell, Au-Yeung, & Sackin, 1999; Howell & Rusbridge, 2011; Richels, Buhr, Conture, & Ntouro, 2010, Spanish (Au-Yeung, Gomez, & Howell, 2003; Watson et al., 2011b), German (e.g., Dworzynski, Howell, Au-Yeung, & Rommel, 2004), Persian (e.g., Karimi & Nilipour, 2011; Mokhlessin, Shahbodaghi, Mahmoudi-Bakhtiari, Howell, & Faghihzadeh, 2006), and Brazilian Portuguese (e.g., De Britto Pereira, 2011; Juste et al., 2012). This pattern is reversed for older children and adults in English, Spanish, German, Persian, and Brazilian-Portuguese, with content words being stuttered more frequently than function words.

A recent study conducted by Smith and Howell (2013) examined stuttering patterns in Japanese and English-speaking children who stutter. Findings for the Japanese-speaking children did not corroborate those for English, Spanish, and German children who stuttered. Results suggested that Japanese children who stutter, do so more on content words than function words, while the English-speaking children who stuttered demonstrated the opposite tendency. Results of the English-speaking children who stutter confirm previous findings. The authors reasoned the difference between Japanese-speaking children who stutter and English-speaking children was that Japanese words are used differently. For example, the Japanese language does not allow for the stalling type disfluency that is seen in English (e.g., “I…I…I think”), Spanish, and German (Smith & Howell, 2013). The EXPLAN theory (Howell, 2004) proposes that when individuals stutter on function words they are committing the act of stalling, meaning that individuals use delaying strategies on function words (stalling) while planning to execute the production of content words (Smith & Howell, 2013). Trends identified in English, Spanish, and German discussed above are accounted for by the EXPLAN theory. Trends showed that stuttering occurred on function words with
individuals using delaying strategies (stalling) to plan and execute the production of content words. Japanese has a simple phonological structure and function words are rare and have little influence in stuttering (Ujihira, 2011). Hence, the concept of these stalling strategies does not apply to Japanese on function words. As there are few function words used, the distinction between content and function words in stuttering for Japanese is not applicable and cannot be compared to the function and content word class for English. The EXPLAN theory (Howell, 2004) based on the word class dichotomy (content and function words), therefore does not apply to Japanese. However, the majority of stuttering in Japanese occurs at the beginning of sentences or clauses, which is similar to stuttering in the English language (Ujihira, 2011).

In English (e.g., Brown, 1938a), Spanish (e.g., Howell & Au-Yeung, 2007), German (e.g., Dworzynski & Howell, 2004; Dworzynski et al., 2003), Brazilian-Portuguese (e.g., De Britto Pereira, 2011), and Persian (e.g., Azimi, Puladi, Bakhtiari, & Haghani, 2012; Haresabadi, Puladi, Bakhtiyari, & Kamali, 2010; Karimi & Nilipour, 2011), stuttering tends to increase with increased complexity of language. Again, this observation does not appear to apply to Japanese, for which language syntactic complexity is an invalid concept (Ujihira, 2011).

The loci of stuttering have been described in only a few languages (see above), with a strong focus on languages within the Indo-European language family (e.g., English, German, Brazilian-Portuguese, and Spanish). The focus of research on Indo-European languages is a limitation, as aspects of language not seen in Indo-European languages cannot be examined. For example, the English language is not a tone language. Thus, investigation of tonal features is not possible (Howell & Rusbrudge, 2011). Tone languages utilise lexical tones to distinguish minimal individual words that are not differentiated by segmental information (Baudion-Chail, 1986, Yiu & Fok,
Lexical tones are defined as contrastive variations in fundamental frequency or pitch at the syllable level (Keung & Hoosain, 1979; Yiu & Fok, 1995). Every syllable will carry a tone. Thus syllables that are comprised of identical segments but carry different tones will differ in meaning (Hao, 2012). It is the fundamental frequency pattern of the vocalic section of a monosyllable that conveys some of the lexical meaning (Feng, Xu, Zhou, Yang, & Yin, 2012). Furthermore, the fundamental frequency contours of speech concurrently manifest sentence intonation and syllabic tones (Ni, Kawai, & Hirose, 2006).

To date, only two studies have examined the influence of lexical tone on stuttering: (i) in Mandarin (Chou, Zebrowski, & Yang, 2015), and (ii) in Cantonese (Law et al., 2017). Chou and colleagues investigated the relationship between stuttering loci and lexical tone in 20 Mandarin-speaking preschool children. Results indicated significantly more stuttering on syllables carrying Tone 3 and Tone 4 in comparison to Tone 1 and Tone 2. It was suggested that syllables carrying Tone 3 and Tone 4 were associated with more stuttering due to the increasing level of speech motor demand. Finally, stuttering was found to be more frequent under a conflicting tonal context as opposed to a compatible tonal context. A compatible context is defined as being a phonetic environment in which the adjacent fundamental frequency offset and onset values are identical or similar, whereas a conflicting context is a phonetic environment in which they are substantially different (Chou et al., 2015). The authors concluded that lexical tones with a complex tonal contour, particularly when placed under conflicting tonal contexts, contributed to the occurrence of stuttering in Mandarin.

Law et al. (2017) investigated the effects of lexical tone on stuttering in Cantonese adults who stutter. Results indicated no significant differences to be found

\footnote{for which English language literature was available}
for stuttering frequency across all six lexical tones. These preliminary results were inconsistent with those of Chou et al.’s preliminary results, who suggested tones to be a contributing factor to the occurrence of stuttering in Mandarin. It should be noted that the two studies differed methodologically, such as examining children and adults. Thus, findings conducted with children who stutter may not be generalisable to adults who stutter. Furthermore, with differing results arising from these two preliminary studies investigating tone languages (Mandarin investigated by Chou et al., 2015; Cantonese investigated by Law et al., 2017), it is speculated that results of studies about one tone language may not be true for other tone languages. This may be a result of different tone languages having different numbers of tones and different tonal contours across languages. For example, Mandarin has four lexical tones, including one level tone and three contour tones, whereas Cantonese has six lexical tones (Law et al. 2017). Research on the loci of stuttering in tone languages have only been conducted within the Sino-Tibetan language family (i.e., Mandarin and Cantonese). Thus, it is unknown whether any of the current inconsistent results would present in other tone languages across differing language families such as Austro-Asiatic (which includes Vietnamese), and Tai-Kadai (which includes Thai). Further research is needed into the loci of stuttering across different tone languages and various language families.

Research into the manifestation of stuttering in different languages has shown stuttering to vary depending on language spoken (e.g., differences between Indo-European languages and Japanese, and differences within different tone languages). This places pressure on speech-language pathologists in being able to efficiently assess the speech of clients who stutter when clients stutter in languages different from the speech-language pathologist’s native language. Speech-language pathologists are required to utilise appropriate assessments and measures when working with clients
from linguistically diverse backgrounds. Yet, to date, there is no gold standard for assessment and measurement of stuttering in different languages. Assessing stuttering of people who speak different languages is no doubt a complex task. However, the task could be further complicated by the fact that speech-language pathologists may be required to measure stuttering in several languages for clients who are bilingual or multilingual-speakers. Therefore, it is important to understand the interplay between stuttering and bilingualism as well as to determine whether stuttering in monolingual-speakers presents similarly or differently from that of bilingual-speakers. The following section explores the relationship between stuttering and bilingualism and provides an overview of what is know about the manifestation of stuttering in bilingual-speakers.

**Stuttering and Bilingualism**

The relationship between stuttering and bilingualism has been referred to as enigmatic (Karniol, 1992) and not well understood (Shenker, 2004). Given the complexities that present with bilingualism, there may be an assumption that bilingualism is a risk factor for stuttering (e.g., Howell, Davis, & Williams, 2009; Karnoiil, 1992). The prevalence of stuttering among bilingual-speakers remains unknown. An early survey by Travis, Johnson, and Shover (1937) involving 4827 children aged 4 to 17 years, suggested that stuttering is more prevalent among bilingual-speakers (2.8%) than monolingual-speakers (1.8%). Stern (1948) also found that stuttering was more prevalent among bilingual-speakers (2.16%) than monolingual-speakers (1.66%). However, the figures reported by Travis et al. and Stern should be treated with caution, as both studies contained inherent methodological flaws. For example, for Travis et al. and Stern, the presence or absence of stuttering was based on a single assessment only. As stuttering is known to vary in symptoms across time and situation, Van Borsel (2011) argued that a single assessment can hardly be considered
as a valid procedure. Furthermore, both studies did not differentiate between normal speech disfluencies and stuttering. Thus it is difficult to interpret the results of these studies.

Recent studies have not corroborated the early findings of Travis et al. and Stern. For example, a more recent internet survey by Au-Yeung, Howell, Davis, Charles, and Sackin (2000) collated responses from 794 participants from 40 countries, speaking 52 different native languages, and more than 70 different second languages. The study found that the percentage of stuttering speakers were almost identical for bilingual-speakers (21.74%) and monolingual-speakers (21.65%). However, the validity of this study is also highly questionable due to significant methodological flaws. The participants were predominately females, an atypical non-representative sample of people who stutter (Yairi & Seery, 2011); the participants were primarily from the Western world, thus limiting the extent to which their findings can be generalised to the larger bilingual community worldwide; and the study relied on self-report. These methodological flaws significantly undermine the credibility of the study’s findings. Therefore, there is neither sufficient evidence nor a definitive answer as to whether stuttering is more prevalent in bilingual-speakers than in monolingual-speakers.

It is estimated that 50% of the world’s population is bilingual (Ardila, Ramos, & Barracoas, 2011). This suggests that there are millions of bilingual people who stutter worldwide. Speech-language pathologists are required to provide appropriate treatment to all clients irrespective of their linguistic background, which may include people who speak and stutter in more than one language. Thus, it is important for researchers and clinicians to understand the presentation of stuttering in the languages spoken by bilingual people who stutter, as specific patterns of stuttering in each language may possibly be found (Ardila et al., 2011).
The manifestation of stuttering across languages of bilingual people who stutter is of increasing interest. The most common pattern emerging of stuttering in bilingual-speakers is that stuttering occurs in both languages, but varies between languages (Van Borsel, 2011). Variations between languages include differing severity, and patterns and distributions of stuttering. Some studies have shown participants demonstrating varying severity of stuttering but similar patterns and distributions (e.g., Jayaram, 1983; Lim, Lincoln, Chan, & Onslow, 2008), but other studies have shown different patterns and distributions of stuttering, with or without different levels of severity of stuttering (e.g., Bernstein Ratner & Benitez, 1985, Jankelowitz & Borts, 1996; Nwokah, 1988; Schäfer & Robb, 2012). Furthermore, a vast majority of research studies examining stuttering in bilingual people who stutter have indicated a higher rate of stuttering in the less proficient language (e.g., Ardila et al., 2011; Jankelowitz & Borts, 1996; Lim et al., 2008; Schäfer & Robb, 2012).

One aspect of the distribution of stuttering has considered the loci of stuttering. The loci of stuttering have been researched for monolingual-speakers who stutter, for both children and adults who stutter. However, limited research is available that examines the manifestation of stuttering in bilingual people who stutter. Gkalitsiou, Byrd, Bedore, and Taliancich-Klinger (2017) recently conducted a preliminary study, and investigated whether bilingual children who stutter present with the same pattern of stuttering as monolingual children. The study was comprised of four bilingual Spanish- and English-speaking children who stutter aged 46-80 months. Results of the study suggested that the bilingual Spanish- and English-speaking children presented with more stuttering on function compared to content words. These results are consistent with research presented earlier, of monolingual-speaking children who stutter, which
concluded that children appear to stutter more frequently on function than content words in English (e.g., Au-Yeung et al., 1998; Bloodstein & Gantwerk, 1967; Bloodstein & Grossman, 1981; Buhr & Zebrowski, 2009; Howell et al., 1999; Richels et al., 2010), Spanish (e.g., Au-Yeung et al., 2003) and German (e.g., Dworzynski et al., 2004). In contrast, Howell et al. (2004) showed different patterns and distributions of stuttering for a bilingual child. Howell and colleagues conducted a single case study with a bilingual-speaking (Spanish-English) child who stutters, examining the frequency of stuttering on content and function words. Results of the study showed differences in stuttering between the two languages, differences which they speculated were dependent on language proficiency. The speaker stuttered more on content words in the dominant language (Spanish) compared to the non-dominant language (English). The speaker was also found to exhibit a higher percentage of stuttering on function words in the non-dominant language (English) compared to the dominant language (Spanish).

In another attempt to examine linguistic influences of stuttering in bilingual-speakers who stutter, Ardila et al. (2011) investigated an adult bilingual-speaker (English-Spanish) who stutters and was more proficient in English than in Spanish. Whilst the speaker produced more stuttering on function words in both languages, the difference in numbers of stuttering moments between content and function words was significant in the non-dominant language (Spanish) but not in the dominant language (English). Similarly, Schäfer and Robb (2012) examined the frequency of stuttering in adult bilingual-speakers (German-English) who stutter. Results of their study suggested the frequency of stuttering in bilingual-speakers is linked to language dominance, evidenced by stuttering occurring more on content than function words in the dominant language (German), with no difference being observed in the non-dominant language (English). Lastly, a study conducted by Tsai, Lim, Brundage and Bernstein Ratner
(2011) produced conflicting findings to those previously mentioned. Tsai and colleagues examined the content and function word distribution of stuttering in adult bilingual-speakers. One participant in their study was an English-Mandarin bilingual speaker, who identified as an English-dominant speaker. Results for the English-dominant bilingual-speaker indicated consistent frequency of stuttering across languages, with more stuttering on function words than content words. This result contradicts previous findings which suggested language dominance to be a key feature in the presentation of stuttering for bilingual-speakers who stutter, with bilingual-speakers stuttering more in the non-dominant language. In relation to word classes, previous studies showed that adult bilingual-speakers stutter more on content than function words, which was similar for monolingual-speakers (e.g., Schäfer and Robb, 2012). This pattern was reversed for Tsai and colleagues’ results which indicated the English-Mandarin bilingual-speaker to stutter more on function than content words.

The above research focused on bilingual-speakers who stutter, with varying language proficiency (dominant versus non-dominant) in each language. Research of Tsai et al. (2011) also examined the loci of stuttering on content versus function words in adult balanced bilingual-speakers. The latter research has shown different results. Tsai et al. examined four English-Mandarin balanced bilingual-speakers who stutter, as well as an English-Spanish balanced bilingual person who stutters, in two separate analyses. For the English-Mandarin analyses, results for the four balanced bilingual-speakers showed inconsistent cross-linguistic patterns. For example, the first balanced bilingual-speaker stuttered consistently across languages, with stuttering occurring fairly equivalently on content and function words. The second balanced bilingual-speaker stuttered fairly equivalently on content and function words in English, but stuttered more on function than content words in Mandarin. The third balanced bilingual-speaker stuttered more on function than content words in English, but
stuttered equivalently on content and function words in Mandarin. Lastly, the fourth balanced bilingual-speaker stuttered fairly evenly on both content and function words in English, but stuttered more on content than function words in Mandarin. For the English-Spanish analysis, results showed this balanced adult bilingual-speaker to stutter similarly on content and function words in both languages. The observations noted by Tsai et al., for English-Mandarin and English-Spanish bilingual-speakers who stutter, are inconsistent with the notion that adults who stutter do so more on content words than function words. Thus, it appears that there is no consistent pattern of the loci of stuttering in terms of function and content words for balanced bilinguals.

Research to date regarding the loci of stuttering for bilingual-speakers who stutter has yielded inconclusive findings. Given very few studies have been conducted, and using single studies or small participant groups with varying methodologies, more research is needed in this area to have an understanding of which variables are influencing the loci of stuttered words (Tsai et al., 2011). More research is needed across different languages for bilingual-speakers to determine whether any patterns emerge (e.g., within the same or differing language families), to further guide researchers’ and clinicians’ understanding of the presentation of stuttering for people who speak more than one language.

In summary, research has identified that bilingual people who stutter will commonly stutter in both languages (e.g., Bernstein Ratner & Benitez, 1985; Jankelowitz & Bortz, 1996; Shenker, Conte, Gingras, Courcy, & Polomeno, 1998; Van Borsel et al., 2001). Not only is stuttering known to vary between languages (Van Borsel, 2011), studies have shown that severity of stuttering and patterns and distributions of stuttering may also vary between languages of bilingual-speakers. Given inconsistent findings across the presentation of stuttering in bilingual-speakers, speech-
language pathologists need to take caution when assessing and diagnosing stuttering of clients who speak more than one language. It has been recommended that speech-language pathologists obtain and/or measure speech samples in all languages spoken by the client (Finn & Cordes, 1997; Lim et al., 2008; Roberts & Shenker, 2007; Vong, Wilson, & Lincoln, 2011; Watson & Kayser, 1994). If a speech-language pathologist assesses stuttering in only one language spoken by a client, it is possible they might risk misjudging severity of stuttering (Lim et al., 2008). In circumstances where the client’s language is unfamiliar to the speech-language pathologist, obtaining and measuring stuttering in all language/s is unquestionably a difficult task. Thus the next section explores the issue of reliably measuring stuttering in a language not one’s own.

**Measuring Stuttering in Unfamiliar Languages**

With increasing globalisation and mobility, there is a rise of people migrating to countries where the official language of the country differs from their own. Thus, speech-language pathologists worldwide are continuing to be challenged with the task of working with clients who speak an unfamiliar language. For example, Australia is considered one of the most linguistically diverse countries in the world (National Health & Medical Research Council, 2006) with 21% of the population speaking a language other than English at home (Australian Bureau of Statistics, 2016). In Australia, the most common languages spoken at home, other than English, are Mandarin (2.5% of the Australian population using this language), Arabic (1.4%), Cantonese (1.2%), Vietnamese (1.2%), and Italian (1.2%; Australian Bureau of Statistics, 2016). It is also known that languages spoken other than English will vary significantly throughout Australia, depending on location. For example, in Sydney, Mandarin (4.7%) is the most common language other than English spoken at home, followed by Arabic (4%), Cantonese (2.9%), Vietnamese (2.1%) and Greek (1.6%; Australian Bureau of
Statistics, 2016). In Adelaide, Italian (2.1%) is the most common language other than English spoken at home, followed by Mandarin (2.1%), Greek (1.7%), Vietnamese (1.4%) and Persian/Dari (0.9%; Australian Bureau of Statistics, 2016). In Brisbane, Mandarin (2.4%) is the most common language other than English spoken at home, followed by Vietnamese (1%), Cantonese (0.9%), Filipino/Tagalog (0.7%) and Spanish (0.7%; Australian Bureau of Statistics, 2016).

What is problematic is a weak correlation found between the languages spoken by speech-language pathologists and the population. For example, in Australia in 2001, Speech Pathology Australia conducted a membership survey and it was documented that 30.7% of speech-language pathologists spoke a language other than English. However, the most commonly reported language other than English was signed English, which is less commonly used by members of the Australian population. Other common languages reported other than English were French, German, Italian, Afrikaans, Cantonese, Mandarin, Greek, Hebrew, and Spanish (Speech Pathology Australia, 2001). The weak correlation between language spoken by speech-language pathologists and the Australian community almost guarantees that Australian speech-language pathologists will be required to work with clients whose language differs significantly from their native tongue and/or second language. Speech pathologists are required to provide appropriate stuttering treatment to all clients irrespective of linguistic background, which may include people who speak in an unfamiliar language. Thus, speech-language pathologists are required to be able to measure stuttering in languages that may be unknown to them.

This poses a challenge as a speech-language pathologist’s unfamiliarity with a client’s language raises the possibility of misinterpretation of data collected, resulting in possible misdiagnosis of stuttering (Finn & Cordes, 1997). In circumstances where the
speech-language pathologist’s native language is different from the client’s language, caution should be taken when making stuttering judgements and/or measurements. In some situations, it has been suggested that speech-language pathologists could rely on interpreters, family members, or clients themselves (e.g., O’Brian, Packman, & Onslow, 2004a; Roberts & Shenker, 2007) to assist in measuring stuttering severity. However, such an approach raises many issues. For example, Finn and Cordes (1997) have emphasised that untrained, inexperienced interpreters may not provide useful or dependable information about stuttered speech. In therapy (e.g., the Lidcombe Program, Onslow et al., 2017; and the Camperdown Program, O’Brian et al., 2017) when working with clients and/or caregivers, clinicians train clients and/or caregivers to take severity rating measures and need to establish reliability with the client in measurement to track progress and outcomes outside of the clinic setting. Establishing reliability between the clinician and client/caregiver is typically achieved by rating speech samples using the severity rating scale and comparing their score until acceptable agreement is achieved. Acceptable agreement is defined as scores differing no more than one scale value. Achieving reliability is readily possible when the clinician and client share a common language and when the clinician has been able to demonstrate reliability in measurement of severity of stuttering in that common language. If the clinician and client/caregiver establish reliability in the common language, then the client/caregiver can apply their skills in measuring stuttering in any other language spoken. However, if the clinician and client cannot go through the process of establishing reliability because there is no common language, the clinician has few options. Firstly, the clinician could have the client/caregiver provide their own measurement data, which may not be reliable, or the clinician could measure severity of stuttering in the unfamiliar language spoken by the client. For the latter option, little is known whether clinicians can reliably measure stuttering in unfamiliar languages. This
highlights the need for further research in assessing the reliability of stuttering measurement by speech pathologists in an unfamiliar language (Roberts & Shenker, 2007).

The reliable identification of stuttering in one’s own language has been found to be problematic. Regardless of speech-language pathologists’ familiarity with clients’ language, even among trained professionals (e.g., Ingham & Cordes, 1997; Kully & Boberg, 1988) there is evidence of disagreement over stuttering judgements. As such, it is imaginable that measuring stuttering could be even more difficult when a client speaks a language that is different from the speech-language pathologist’s native language. Even with increased mobility and a growing interest from speech-language pathologists in the assessment and management of clients from different linguistic backgrounds, there is still little evidence exploring how well speech-language pathologists are able to make reliable and valid judgements of stuttering in languages other than their own. The investigation of measuring stuttering in an unfamiliar language forms a significant part of this thesis.

The Present Thesis

When working with clients who stutter, speech-language pathologists’ abilities to reliably measure stuttering is an essential component of assessment and treatment, to track client progress and outcomes. With increasing globalisation, speech-language pathologists are being challenged with the task of working with clients who speak unfamiliar languages. This is problematic as there is still a lack of evidence exploring how well speech-language pathologists are able to make reliable measurements of stuttering in languages other than their own. Hence, the focus of this thesis is to investigate the reliability of measurement of severity of stuttering across different languages. The remainder of this thesis is divided into six chapters. Chapter 2 provides
an overview of the literature about measurement of stuttering and concludes with a
discussion of the measure of focus for the program of research: the 9-point severity
rating scale.

Four studies are presented in this thesis. Firstly, it was considered important to
examine the interplay of speech sample duration on stuttering measurement. Thus,
Chapter 3 presents Study 1, which determined whether the duration of a speech sample
influenced experienced speech-language pathologists’ reliability of measuring severity
of stuttering using the 9-point severity rating scale. Results from this study were used to
inform the design of subsequent studies in the research program.

The aim of Study 2 (see Chapter 4) was to examine English-speaking speech-
language pathologists’ reliability of use of the severity rating scale in English (familiar
language) and Vietnamese (unfamiliar language). Investigation of English-speaking
speech-language pathologists’ use of the scale in Vietnamese is particularly apt, given
that Vietnamese is widely spoken in many countries beyond Vietnam (the official
language of Vietnam spoken by over 91 million people), such as the United States,
Australia, Canada, and the United Kingdom. For example, in the United States,
Vietnamese people make up the fourth largest Asian group, with Vietnamese ranking in
the top 10 languages, other than English, spoken at home (Ryan, 2013). In Australia,
Vietnamese also ranks within the top 10 languages, other than English, spoken at home,
with 1.2% of the population speaking Vietnamese (Australian Bureau of Statistics,
2016). In Canada, Vietnamese ranks in the top 25 languages spoken at home (Statistics
Canada, 2012). In the United Kingdom, 15 200 people reported speaking Vietnamese as
their first language (Lewis, Simons & Fennig, 2016). Therefore, it is likely that speech-
language pathologists around the world will be required to work with monolingual
and/or multilingual clients who speak Vietnamese. However, it is unknown whether
English-speaking speech-language pathologists can reliably measure severity of stuttering in Vietnamese. Thus, Study 2 offers new reliability data for a different language and enables comparisons to be made to other unfamiliar languages (e.g., Mandarin, investigated by Hoffman, Wilson, Copley, Hewat, & Lim, 2014).

The third study in this thesis provides the first stuttering measurement data for Vietnamese-speaking speech-language pathologists in Vietnam. The speech-language pathology profession is in its infancy in Vietnam, with only two cohorts (33 speech-language pathologists) of graduates at the time of data collection. No research to date has examined Vietnamese speech-language pathologists’ reliability when measuring severity of stuttering. Thus, Study 3 investigated whether Vietnamese-speaking speech-language pathologists can reliably measure severity of stuttering using a 9-point severity rating scale in Vietnamese (familiar language) and English (other language, a common second language for Vietnamese-speakers). Study 3 is reported in Chapter 5 of this thesis.

Based on the findings of Studies 2 and 3 of the present thesis, Study 4 was developed: Study 4 is presented in Chapter 6 and is a re-analysis of data from Hoffman et al. (2014), to enable comparison to data from Studies 2 and 3, using identical statistical analyses. The purpose of the comparison was to determine whether the consistent findings across the present studies were robust. Study 4 evaluated Australian English-speaking speech-language pathologists’ reliability of use of the 9-point severity rating scale in Australian English (familiar language) and Mandarin (unfamiliar language). The re-analysis enabled the addition of another language (Mandarin) and language family (Sino-Tibetan) to the program of research. Chapter 7 presents a summary of the research findings, followed by a discussion of the contributions of this research, implications, and future directions.
CHAPTER 2

Measurement of Stuttering
Measures of stuttering provide a way to identify and describe the nature, severity, and impact of the disorder. Whilst measurement has been primarily developed to assist in the assessment of the frequency, type, duration, and severity of stuttering, for some clients, it may also be important to assess speech naturalness, speech rate, and concomitant or associated behaviours (Guitar, 2014). Stuttering measurement is a vital component of clinical practice and research, as it assists speech-language pathologists to assess and effectively communicate with clients about their stuttering, enables the development of treatment goals, and aides in establishing the effectiveness of the treatment process and outcome. Furthermore, measurement enables the speech-language pathologist to quantify the impact of the disorder which may be behavioural, related to stuttering moments and how often they occur, and/or non-behavioural (e.g., mental health, relating to anxiety; Onslow, 2017).

In addition to general speech pathology assessment procedures, there are a range of stuttering specific measurements used in clinical practice and research. Within the scope of practice, speech-language pathologists are required to consider clients’ behavioural and mental health domains during clinical assessment of stuttering (Speech Pathology Australia, 2017). Screening for speech related anxiety by speech-language pathologists is recommended, to determine whether a referral to an appropriate health professional (e.g., clinical psychologist) is necessary.

This chapter provides a brief summary of measures of speech, quality of life, and anxiety used in stuttering clinical practice and research. The chapter then outlines the stuttering measure which forms the focus of the program of research, the 9-point severity rating scale, and concludes with a discussion of the reliability of the stuttering severity rating scale.
Measures Used for the Management of Stuttering

Speech-language pathologists working with clients who stutter need to be able to accurately identify behavioural and non-behavioural characteristics of stuttering and reliably measure stuttering severity, frequency, and duration (Guitar, 2014). Whilst there are a number of stuttering instruments and measures used for assessing stuttering, the following section provides a brief summary of some common measures and assessment tools used for the assessment and management of clients who stutter.

Measuring Stuttering Frequency

Measures of stuttering frequency are relatively objective measures (Bloodstein & Bernstein Ratner, 2008) and involve counting of stuttered moments (e.g., counting syllables and/or words which are unambiguously stuttered by the speaker). An advantage of stutter count measures is that they identify all occurrences of stuttering by the individual (Jones, Onslow, Packman, & Gebski, 2006). Percentage syllables stuttered and time interval measures are two examples of measures used to measure stuttering frequency.

Percentage Syllables Stuttered

Percentage syllables stuttered is a widely used measure in clinical practice and research, which is calculated as the number of stuttered syllables divided by the total number of syllables spoken, multiplied by 100. Whilst percentage syllables stuttered measures identify all occurrences of stuttering, it should not be used in isolation, as it is merely a quantitative measure of stuttering frequency, not stuttering severity, and thus does not take account of stuttering typology. For instance, a client might frequently stutter, with each stuttering moment involving long blockages and therefore, largely impacting the client’s communication. In contrast, a different client might stutter at the
same frequency but the stuttering is in short blocks which may not disrupt or impact communication as greatly (Onslow, 1993). This highlights the importance of using other measures, such as speech rate measures and severity ratings, in conjunction with percentage syllables stuttered in order to fully understand the extent of the client’s stuttering.

There have been suggestions that percentage syllables stuttered measures are intrinsically unreliable (Cordes & Ingham, 1994a). Despite these suggestions, there is considerable research substantiating that percentage syllables stuttered measures can be used reliably (e.g., Boberg & Kully, 1994; Craig et al., 1996) and to a high level of interjudge agreement (e.g., Harrison, Onslow, Andrews, Packman, & Webber, 1998; Ingham et al., 2001; Lincoln & Onslow, 1997; O’Brian, Onslow, Cream, & Packman, 2003; Onslow, Costa, Andrews, Harrison, & Packman, 1996; Onslow, Packman, Stocker, van Doorn, & Siegel, 1997).

**Time Interval Measurement**

Time-interval measurement was developed by Cordes, Ingham and colleagues as an alternative to measuring percentage syllables stuttered (Cordes & Ingham, 1994b, 1994c, 1995, 1996; Cordes, Ingham, Frank, & Ingham, 1992; Ingham, Cordes, & Finn, 1993; Ingham, Cordes, & Gow, 1993). In time-interval measurement, speech samples are divided into 4 or 5 second intervals. Listeners then identify whether each interval contains stuttering. Time-interval measures are calculated as the number of stuttered intervals divided by the total number of intervals, multiplied by 100. Cordes and Ingham (1999) identified that speech-language pathologists trained in time-interval measurement could produce reliable interval judgements of stuttering. However, a disadvantage of time interval measurement is that it does not identify all occurrences of stuttering by the individual, nor for practical reasons can it be used clinically to monitor
stuttering severity. Time-interval measurement appears not to have moved beyond the developmental stage, into widespread use for measuring stuttering in research and clinical practice.

**Measuring Stuttering Severity**

Stuttering severity measures are subjective measures based on the perceptions of observers (Bloodstein & Bernstein Ratner, 2008; Huinck & Rietveld, 2007). Stuttering severity refers to the extent of the disruption in the delivery of continuous speech (Yairi & Seery, 2011). When measuring severity of stuttering, speech-language pathologists base their judgement on their overall impression of the speech sample, taking into account stuttering typology and frequency (O’Brian, et al., 2015). The stuttering severity rating scale is an example of a commonly used measure of stuttering severity.

*Severity Rating Scale*

Severity rating scales used for research and clinical practice takes the form of an equal appearing interval scale and varies in the number of scale points (e.g., 5-point, 7-point, 8-point, 9-point, 10-point, 15-point). The scale requires the assessor to listen to speech samples and assign a score on the scale ranging from no stuttering present to extremely severe stuttering (Lincoln & Packman, 2003; O’Brian, Packman, & Onslow, 2010). Severity rating scales require no equipment, are convenient to use when measuring stuttering, and can be used in isolation. Severity rating scale measures are used across various treatment programs (e.g., the Lidcombe Program, Onslow et al., 2017; and the Camperdown Program, O’Brian et al., 2017), and used with both children and adults who stutter. The reliability of the severity rating scale will be discussed later in this chapter.
The above measures examine only a single aspect of stuttering (e.g., percentage syllables stuttered measures stuttering frequency and severity rating scales measure severity). The Stuttering Severity Instrument – Fourth Edition (Riley, 2009), however, incorporates multiple measures. It is a detailed diagnostic assessment tool of stuttering severity designed for clinical and research purposes. This assessment tool is used by speech-language pathologists to determine frequency of stuttering, duration of the three longest stuttering moments, naturalness of the individual’s speech, as well as any physical concomitants. The assessment is a comprehensive and valuable tool to researchers and stuttering specialist clinicians. However, this tool may not be convenient as a routine measure for generalist speech-language pathologists, as the assessment involves an expense to purchase and takes considerable time to complete in comparison to other severity rating measures (e.g., the severity rating scale). Furthermore, Lewis (1995) reported that the Stuttering Severity Instrument – Third Edition (a previous edition) provided no additional information beyond that which could be attained via the severity rating scale.

Most recently, Davidow and Scott (2017) examined the reliability of the Stuttering Severity Instrument – Fourth Edition. Results of the study highlighted limitations in the use of the assessment tool (see Davidow and Scott, 2017). Research from Davidow and Scott did not support the use of the Stuttering Severity Instrument – Fourth Edition, and did not recommend it for the use of measuring stuttering severity. Thus, it would appear the assessment tool should be used only for descriptive purposes. Given the limitations of the tool as a diagnostic assessment, Davidow and Scott have recommended the use of a severity rating scale (9 or 10-point) for clinical purposes, as an alternative measure of severity of stuttering.
Measuring Speech Rate

Obtaining speech rate measures is of value particularly in clinical practice for treatments incorporating speech rate targets. Speech rate can be measured using syllables per minute. Syllables per minute takes into account the number of syllables spoken and the time taken to speak those syllables. Like stuttering frequency measures (e.g., percentage syllables stuttered), speech rate measures should not be used in isolation, as they do not account for or correlate with other elements of stuttering such as severity ratings or speech naturalness (Bloodstein & Bernstein Ratner, 2008).

There has been a recent advance in the measurement of syllables per minute, with research into the development of a smartphone application for monitoring and providing feedback of speech rate for clients (Aharonson et al., 2017). Preliminary results into the reliability of the smartphone application has shown that the application can provide reliable, user-friendly, real-time feedback for speaking rate control practice (Aharonson et al., 2017). It should be noted, however, that this research was conducted using participants who did not have a speech disorder. Thus it is unknown whether reliability results of this study would be shown when used with individuals who stutter.

Measuring Speech Naturalness

Speech naturalness measures do not quantify actual speech events. Measurements of speech naturalness involve the assessor’s subjective perception of how natural sounding is the speech of a client who stutters. Speech restructuring treatment programs (e.g., the Camperdown Program; O’Brian et al., 2017) have been commonly used to help adults reduce stuttering. The drawback is, however, that speech restructuring treatment programs may not produce natural sounding speech. Therefore, a 9-point speech naturalness scale was developed (Martin, Haroldson & Triden, 1984; Ingham, Gow & Costello, 1985; Mackey, Finn, & Ingham, 1997; Onslow, Hayes,
Hutchins, & Newman, 1992) for use by clients and speech-language pathologists to evaluate and measure the client’s speech naturalness throughout the treatment program, and to guide clients in attaining speech that sounds as natural as possible while achieving the desired stuttering reduction (Onslow, 2017). The speech naturalness scale has been shown to be reliable for speech-language pathologists providing feedback to clients about their speech and also for clients’ self-evaluation of speech quality (Ingham, Ingham, Onslow, & Finn, 1989). However, Mackey et al. (1997) identified that naturalness ratings can be affected by presence of foreign accents.

**Measuring Impact of Stuttering**

Most measures of the impact of stuttering incorporate measurement of quality of life. When working with clients who stutter, speech-language pathologists need to consider quality of life as a result of stuttering, as well as recognise and acknowledge the potential interplay between stuttering and any mental health issues that may arise (Speech Pathology Australia, 2017). The Overall Assessment of the Speaker’s Experience of Stuttering (OASES; Yaruss & Quesal, 2010) and the Wright and Ayre Stuttering Self-Rating Profile (WASSP; Wright & Ayre, 2000; Wright, Ayre, & Grogan, 1998) are two examples of instruments that are commercially available to speech-language pathologists for measuring quality of life in relation to stuttering.

*The Overall Assessment of the Speaker’s Experience of Stuttering*

The Overall Assessment of the Speaker’s Experience of Stuttering (OASES) was developed by Yarbus and Quesal (2010) to assess the full impact of stuttering from the perspective of the individual who stutters. This assessment was based on the World Health Organization’s International Classification of Functioning, Disability, and Health (World Health Organization, 2001) to ensure that it would capture the entirety of the disorder. The OASES has useful clinical applications as it can be used to support
speech-language pathologists’ understanding of stuttering, and as a treatment planning and evaluation tool. The OASES contains four sections: (i) general information assessing the individual’s perception of stuttering, (ii) the individual’s affective, behavioural and cognitive reactions to stuttering, (iii) the impact of stuttering on the individual’s communication in daily situations, and (iv) the individual’s overall quality of life, by investigating how stuttering negatively impacts his or her life.

The OASES was originally designed for adults who stutter, 18 years and older, and has been shown to have acceptable reliability and validity (Blumgart, Tran, Yarrus, & Craig, 2012; Yarrus & Quesal, 2006). The adult version is known as OASES-A and is published by Pearson Assessments in English and Spanish. More recently, the OASES-A has been translated into two other languages: Japanese (OASES-A-J; Saki, Chu, Mori, & Yarrus, 2017) and Hebrew (Freud, Kichin-Brin, Ezrati-Vinacour, Roziner, & Amir, 2017), with results indicating reliable and valid data in these two languages. There are now also two other versions available for different ages: The OASES-T was developed for adolescents, 13 – 17 years, and the OASES-S was developed for school aged children, 7 – 12 years. To date, the OASES-S has been translated into Dutch, with preliminary results showing reliable and valid data for this version (Lankman, Yarrus, & Franken, 2015).

The Wright and Ayre Stuttering Self-Rating Profile

The Wright and Ayre Stuttering Self-Rating Profile (WASSP; Wright & Ayre, 2000) was developed initially in the UK as an outcome measure for adults who stutter (Wright et al., 1998). Like the OASES, the WASSP was based on the World Health Organization’s International Classification of Functioning, Disability, and Health framework (World Health Organization, 2001) to reflect the social dimensions of stuttering (Ayre & Wright, 2009). The WASSP has since been used as a clinical
assessment tool to assist with the planning of intervention and to measure change for clients aged 14 years and above; and for research purposes, to evaluate and compare intervention approaches (Ayre & Wright, 2009). The tool is commonly used in the UK. However, international availability appears to be limited (Ayre & Wright, 2009). The WASSP is used to measure stuttering behaviours, thoughts about stuttering, feelings about stuttering, avoidance, and disadvantage due to stuttering. The developers of the WASSP have reported adequate reliability and validity in the use of the tool (Ayre & Wright, 2009). Whilst there have been no formal translations of the WASSP, Ayre and Wright (2009) report informal translations of the tool into Spanish, Finnish and Norwegian.

**Measuring Anxiety**

Anxiety is known to be associated with stuttering. There are many freely available assessments speech-language pathologists can administer without a psychology qualification to detect stuttering related anxiety. Whilst it is not in the scope of practice for speech-language pathologists to attempt to diagnose and manage a mental health disorder, stuttering related anxiety assessments can help a speech-language pathologist to screen for anxiety and determine whether a referral to a clinical psychologist is needed. Unhelpful Thoughts and Beliefs About Stuttering scales (Iverach et al., 2011; Iverach et al., 2016a; St Clare et al., 2009), the Fear of Negative Evaluation Scale (Watson & Friend, 1969), the Spence Children’s Anxiety Scale (Nauta et al., 2004; Spence, 1998), and the parent report Preschool Anxiety Scale Revised (Edwards, Rapee, Kennedy, & Spence, 2010; Macquarie University, n.d.) are examples of measures that are recommended (e.g., see Speech Pathology Australia’s clinical guidelines for stuttering management; Speech Pathology Australia, 2017) for speech-
language pathologists to use when measuring speech related anxiety. Each is outlined briefly below.

*The Unhelpful Thoughts and Beliefs About Stuttering Scales*

Negative feelings can exacerbate behavioural components of stuttering (Bennett, 2006; St Clare et al., 2009). The Unhelpful Thoughts and Beliefs About Stuttering (UTBAS) Scales have been shown to be valid and reliable measures (St Clare et al., 2009), which are useful to speech-language pathologists for identifying how negatively a client perceives their stuttering. Furthermore, the UTBAS can discriminate between stuttering and non-stuttering clients on questions that make no reference to stuttering (St Clare et al., 2009). The UTBAS was developed for adults and is a stuttering specific self-report measure that comprises 66 commonly occurring unhelpful thoughts about stuttering. The scale is freely available and can be downloaded from the website of the Australian Stuttering Research Centre (http://sydney.edu.au/health-sciences/asrc/). A client is presented with a scale ranging from 1 to 5 (1 = never or not at all, 2 = rarely or a little, 3 = sometimes or somewhat, 4 = often or a lot, 5 = always or totally) and indicates how frequently these thoughts and/or beliefs occur, how much the thoughts are believed, and to what degree these thoughts incite anxiety. A Japanese version of the UTBAS is now available, with preliminary results demonstrating good reliability (Chu, Sakai, Mori, & Iverach, 2017).

More recently, a six-item screening version of the UTBAS (known as the UTBAS-6) was developed (Iverach et al., 2016a). An advantage of reducing the number of items makes the UTBAS-6 an ideal brief screening tool for clinicians. The scale was shown to be a reliable scale for screening the unhelpful thoughts and beliefs associated with speech-related anxiety among adults who stutter (Iverach et al., 2016a).
**The Fear of Negative Evaluation Scale**

The Fear of Negative Evaluation (FNE) Scale (Watson & Friend, 1969) is another useful scale for establishing adults’ degree(s) of anxiety in relation to social interaction. The FNE is a self-report questionnaire and comprises 30 statements to which the client answers either true or false. The statements are aimed at identifying and addressing clients’ feelings of anxiety regarding social interactions and how they feel other people view their communication abilities (Watson & Friend, 1969). The FNE has demonstrated good psychometric properties (Watson & Friend, 1969). Since its development, the FNE has been widely used in research and clinical practice (e.g., by psychologists), as well as with stuttering participants in research (e.g., Blumgart, Tran & Craig, 2010; Bricker-Katz et al., 2009; Brundage, Winters & Beilby, 2017; Iverach et al., 2009; Messenger, Onslow, Packman & Menzies, 2004).

Due to the popularity of the 30-item scale over the years, a 12-item version was developed for adults and is referred to as the Brief Fear of Negative Evaluation Scale (BFNE; Leary, 1983). The BFNE consists of eight straightforwardly worded items (e.g., item 9 “I am usually worried about what kind of impression I make”) and four reverse-worded items (e.g., item 4 “I rarely worry about what kind of impression I am making on someone”; see Leary, 1983, for all 12 questions). The BFNE scores demonstrated good internal consistency. Since its development, the BFNE has been extensively researched and modified by researchers, with a further three versions of the BFNE now available:

(i) Brief Fear of Negative Evaluation Scale – Straightforward Items (BFNE-S; Rodebaugh et al., 2004; Weeks et al., 2005). The BFNE-S comprises 8-items, using only the eight straightforwardly worded items from the BFNE (Leary, 1983). The BFNE-S has shown to be a valid tool (Carleton, Collimore, &
Asmundson, 2007; Rodebaugh et al., 2004; Weeks et al., 2005) with excellent internal consistency (Liu & Lowe, 2016).

(ii) Brief Fear of Negative Evaluation Scale – Revised (BFNE-R; Carleton, McCreary, Norton & Asmundson, 2006). The revised tool, encompasses 12-items, which includes the original straightforward items, with an inclusion of modified reverse items of the BFNE (Leary, 1983) to now be straightforwardly worded and presented. The BFNE-R has demonstrated internal consistency (Carleton et al., 2006).

(iii) Brief Fear of Negative Evaluation Scale – Version 2 (BFNE-II; Carleton et al., 2007). The BFNE-II is a revised version of the BFNE-R (Carleton et al., 2006), comprised of 8-items. Seven of the originally straightforwardly worded items (items 1, 3, 5, 6, 8, 9, 12) of the BFNE-R are included, as well as one item that was originally reversed-worded, now revised to be straightforwardly worded (item 10). The BFNE-II has established internal consistency (Carleton et al., 2007).

More recently, research has focused on making comparisons between the modified versions of the BFNE (e.g., Carleton, Collimore, McCabe & Antony, 2011; Lui & Lowe, 2016). Carleton et al.’s results demonstrated superior utility of the BFNE-S and BFNE-II in comparison to the BFNE-R, with the BFNE-S shown to be better at predicting social anxiety in clinical samples than other BFNE alternative measures. Results of Lui and Lowe are consistent with those of Carelton et al., showing that the BFNE-S is superior to the BFNE-II for assessing fear of negative evaluation.

_Spence Children’s Anxiety Scale_

The Spence Children’s Anxiety Scale (Nauta et al., 2004; Spence, 1998) is publically and freely available. It was developed to assess the severity of anxiety
symptoms that can be specifically linked to symptoms of DSM-IV anxiety disorders in children identified in the Diagnostic and Statistical Manual of Mental Disorders. The scale examines the following domains of anxiety: generalised anxiety, panic/agoraphobia, social phobia, separation anxiety, obsessive compulsive disorder, and physical injury fears. There are both parent (for children aged 3 – 5 years) and child versions (for children aged 8 – 15 years) of the Spence Children’s Anxiety Scale. The Spence Children’s Anxiety Scale has been shown to be reliable and valid, with comprehensive normative data available across a diverse range of settings and cultures, with the scale translated into 31 languages. The scale is routinely used by psychologists in clinical practice. However, there is current interest in the use of the scale for research into anxiety and stuttering with children (e.g., Iverach et al., 2016b; Smith et al., 2017). Whilst Spence Children’s Anxiety Scale is not intended for use as a standalone diagnostic tool, it has been recommended that speech-language pathologists could use the scale for screening purposes to determine whether a referral to a clinical psychologist is warranted (Onslow, 2017).

The Preschool Anxiety Scale Revised

The Preschool Anxiety Scale Revised (Edwards et al., 2010; Macquarie University, n.d.) is a publically available scale developed for use as a parent report measure of anxiety for children aged 6 years and under. The scale is available in English and seven other languages (i.e., Arabic, Chinese, Dutch, Icelandic, Indonesian, Norwegian, and Turkish). The scale comprises 28 items over four sections: social anxiety, generalised anxiety, separation anxiety, and specific fears. Parents respond to the items presented using a 5-point scale (0 = not at all true, 1 = seldom true, 2 = sometimes true, 3 = quite often true, and 4 = very often true). Normative data (Edwards et al., 2010) have been obtained for the Australian population, and an elevated score
compared to the mean can assist speech-language pathologists to determine whether a referral to a clinical psychologist is needed. The Preschool Anxiety Scale Revised has been shown to be a reliable and valid measure (Edwards et al., 2010).

When working with clients who stutter, a speech-language pathologist may be required to perform one or more of the above measures discussed in this chapter, to obtain a comprehensive understanding of clients’ stuttering and the impact it has on their lives. Irrespective of the measurement/s chosen, it is crucial that the speech-language pathologist is able to make reliable and valid measures. Thus, investigation into the reliability of stuttering measures used by speech-language pathologists is warranted.

**Measurement of Focus for the Program of Research**

Stuttering measurement is a vital component of assessment and treatment. Thus, it is essential for speech-language pathologists to reliably measure stuttering, to enable efficient tracking of clients’ progress and outcomes. There has been considerable debate among researchers regarding appropriate techniques for measuring stuttering (Yaruss, 1997). With globalisation and the complexity of multilingualism, without a gold standard stuttering measure recommended for use across different languages, speech-language pathologists may not feel comfortable, or have adequate skills to be able to reliably measure stuttering in languages with which they have less proficiency, or languages that they are not familiar with. Roberts and Shenker (2007) identified that there is no consistent methodology to use when assessing the speech of bilingual people who stutter. However, when choosing an appropriate and reliable stuttering measure, Shenker (2004) recommended using severity rating scales to assess and treat bilingual-speakers, as the scale reflects global ratings of stuttering in all languages. Thus, severity
ratings appear to be an appropriate universal tool to measure severity of stuttering, irrespective of language. With the popularity of severity rating scales worldwide, investigation into the reliability of use of the scale is critical, regardless of whether it is being used for clinical or research purposes. Thus, the focus of the present research is to investigate the reliability of speech-language pathologists’ measurement of stuttering severity, across languages.

Severity rating scales are commonly used in research and clinical practice, and are considered to be the most convenient, valid and reliable measure of stuttering severity (Speech Pathology Australia, 2017). Severity rating scales are found to be more valid measures to use than %SS, as they take into account all behavioural features of stuttering rather than merely counting the number of stuttering moments (Onslow, 2017). Additionally, recent research has shown severity ratings to be more reliable than %SS in a clinical setting when needing to detect changes over time with individuals to evaluate progress and outcomes (Karimi, O’Brian, Onslow, & Jones, 2014).

In clinical practice, severity ratings are used to measure client’s stuttering within and outside of the clinic throughout treatment, to measure progress and outcomes. For example, in the Camperdown Program (O’Brien et al., 2017), a treatment program used with adults who stutter, the speech-language pathologist trains the client in the use of a 9-point severity rating scale during the first clinic visit. This enables the client to collect measures in everyday situations, and for the clinician and client to quantify progress during and after treatment (O’Brien et al., 2017). It is essential for the speech-language pathologists and client to confirm agreement between severity rating scores within the first few clinic visits and regularly throughout treatment (O’Brien et al., 2017). Reasonable agreement occurs when the client and speech-language pathologists scores agree or differ by no more than one scale value (O’Brien et al., 2017). Throughout
clinical visits, the speech-language pathologist’s judgement is used as the standard for training the client, and is also used when reaching consensus at the beginning of each clinic visit. Completing measures at the beginning of each visit enables the client to check reliability and refine their scores with the speech-language pathologist until agreement occurs between the client and clinician’s scores. Reaching consensus on a regular basis with the client can be difficult if speech-language pathologists are not reliable with themselves. Thus, it is vital that speech-language pathologists ensure they are able to make reliable measurements of severity of stuttering. This would enable them to provide efficient measurement training and treatment to clients who stutter, and to monitor progress and outcomes.

The present research investigates speech-language pathologists’ reliability in measuring severity of stuttering in adult-speakers who stutter. In clinical practice, a 9-point severity rating scale is commonly used for adults who stutter (e.g., the Camperdown Program, O’Brian et al., 2017). Thus, a 9-point stuttering severity rating scale (see Figure 2.1) was the measure investigated in the program of research. The 9-point severity rating scale is an equal appearing interval scale and requires the speech-language pathologist to listen to a speech sample and assign a score on the scale ranging from 1 (no stuttering present) to 9 (extremely severe stuttering).

![Severity Rating Scale](image)

*Figure 2.1. Severity Rating Scale*
With severity rating scales popularly being recommended for use with clients who stutter, investigation into the reliability of the severity rating scale is critical, regardless of whether it is being used for research or clinical purposes. Early research into the use of severity rating scales was conducted by Sherman and colleagues (Lewis & Sherman, 1951; Sherman, 1952, 1955; Sherman & Trotter, 1956), and focussed on a 9-point scale. Results showed the 9-point severity rating scale to be a reliable tool to use for measuring severity of stuttering. Later research investigated stuttering severity rating scales which differed in the number of scale points (i.e., 5, 7, 9, or 15-points). Results showed that there was little difference of mean scale values or reliability, irrespective of the number of scale points (Cullinan, Prather, & Williams, 1963; Curran & Hood, 1977). Other research into the use of the stuttering severity rating scale has produced conflicting findings. Some research has identified that severity of stuttering can be measured reliably using a severity rating scale by both experienced and inexperienced listeners (e.g., Cullinan & Prather, 1968; Cullinan et al., 1963; Curran & Hood, 1977; Lewis & Sherman, 1951; Martin & Haroldson, 1992; Onslow, Andrews, & Costa, 1990; Sherman, 1952, 1955; Sherman & Trotter, 1956; Young, 1969). Eve, Onslow, Andrews, and Adams (1995) examined the reliability of using a 10-point scale to measure the severity of stuttering in 5-minute audio speech samples. They found a 10-point scale to be a reliable tool when measuring stuttering severity. A more recent study conducted by O’Brian, Packman, Onslow and O’Brian (2004b), examined speech-language pathologists’ reliability of using a 9-point scale with 3-minute video speech samples of adults. The study deemed the 9-point severity rating scale a viable and reliable clinical tool for use with adults to measure stuttering severity.

However, the reliability of stuttering severity measurement has been shown in some studies to be problematic. For example, substantial interclinic discrepancies have been identified when rating severity using a 7-point severity scale (Kully & Boberg,
1988), suggesting that operational definitions of stuttering may be shared within but not across clinics (Cordes & Ingham, 1995). A more recent study by Hoffman et al. (2014) examined, among other things, whether Australian speech-language pathologists could reliably measure Australian-English speech samples using a 9-point severity rating scale. Results identified that only half of the judges were able to reliably measure severity of stuttering using the scale in Australian English. Research showing conflicting findings therefore calls the reliability of use of the scale into question.

As the 9-point stuttering severity rating scale is commonly used in clinical practice and for research purposes, it is concerning that not all research into the reliability of severity rating scales has found that they can be used reliably (e.g., Kully & Boberg, 1988; Hoffman et al., 2014). Thus, further investigation is warranted. The next chapter discusses possible contributing factors that may have led to varying reported reliability levels between studies. It then presents the first study, which sought to explore whether the duration of speech samples influences the reliability of measurement of severity of stuttering by speech-language pathologists.

---

3 A discussion of the results in its entirety will be provided in later sections of this thesis (e.g., see Chapter 4 ‘Research investigating identification and measurement of stuttering in unfamiliar languages’).
CHAPTER 3

Study 1: The Effect of Speech Sample Duration on the Reliability of Measurement of Severity of Stuttering
Speech-language pathologists’ abilities to measure stuttering reliably has been of interest over time. However, speech samples used in stuttering measurement research have varied in duration. This chapter presents the first study of this thesis, which examined whether the duration of speech samples influenced the reliability of measurement of severity of stuttering by speech-language pathologists. Results from the study will determine the duration of speech samples that will be used for the remainder of the studies in the program of research and offer suggestions for future research into the reliability of measurement of stuttering using a 9-point severity rating scale.

**Background**

Over the last two decades, several research papers examining the reliability of use of the severity rating scale to measure severity of stuttering have been published. These papers have adopted varying durations of speech samples. For example, O’Brian et al. (2004b) examined, among other things, the reliability of measuring severity of stuttering using a 9-point stuttering severity rating scale with 3-minute video speech samples of adults. The authors concluded that speech-language pathologists could reliably use the 9-point scale when measuring stuttering severity of adults who stutter. Eve et al. (1995) examined the reliability of a 10-point scale to measure severity of stuttering in 5-minute audio speech samples. They deemed that the 10-point scale could be used reliably when measuring severity of stuttering. A more recent study conducted by Hoffman et al. (2014) examined the reliability of a 9-point stuttering severity rating scale using 3-minute audio speech samples of adults who stutter. In contrast to previous studies, results identified poor agreement of speech-language pathologists when using the scale to measure severity of stuttering. It is evident that consensus is yet to be made regarding the reliability of use of the severity rating scale. Therefore, further research is warranted.
Varying reliability results from the three papers could have been attributed to at least three factors: (i) the use of audio versus audiovisual speech samples. However, research (e.g., O’Brian et al., 2015; Rousseau, Onslow, Packman, & Jones, 2008) has identified that audio-only speech samples are adequate for measuring severity of stuttering; (ii) the use of different analyses; and (iii) the duration of speech samples. Eve et al. (1995) used the pairwise comparison procedure of Martin et al. (1984) to determine grouped data describing participants’ performance. For this procedure, the stuttering severity rating score assigned to each speech sample by each participant was compared to the score assigned to that sample by each of the other participants. Hoffman et al. (2014) also used the comparisons procedure to examine grouped data, and results (intrajudge agreement 81.82%; interjudge rate and re-rate agreement, 68.19% and 72.99% respectively) were comparable to those of Eve and colleagues’ generalist clinicians (intrajudge agreement 75%; interjudge rate and re-rate agreement, 67.6% and 76.9% respectively). However, Hoffman et al. also examined the percentage of individual participants who reached criterion reliability performance. Criterion reliability performance was defined as having assigned a score within plus or minus one scale value of the modal score for the speech sample, for 80% of samples. It was the individual data analyses from Hoffman et al. which identified poorer agreement (intrajudge agreement 50%; interjudge rate and re-rate agreement, 76.92% and 63.64% respectively). Therefore, grouping of data for the pairwise comparison procedure appears to have masked individual variation. For consensus to be made, future research should examine both group data and individual performance.

Another likely factor contributing to the varying reliability levels in previous research is the different durations of speech samples. No research has determined whether duration of a sample affects the reliability of measurement. For this reason, it is
important to investigate whether the duration of speech samples can influence speech-language pathologists’ reliability when using the scale.

Investigating the duration of speech samples is important for researchers due to the necessity of recruiting suitable participant sample sizes. Studies in stuttering measurement use many speech samples and often require a lengthy time to complete. It is commonly known that researchers have difficulty recruiting participants to studies, and time commitment needed to complete studies by participants is often a contributing factor for poor recruitment (Patel, Doku, & Tennakoon, 2003). Small participant sample sizes can be problematic in research, as it can prevent the use of statistical tests to determine the significance of the observed differences in criterion reliability analyses. Additionally, small sample sizes may lead to a non-representative sample of speech-language pathologists. If shorter durations of speech samples could be used without affecting the reliability of their measurement, this would reduce time commitment and likely improve participation in future research.

Likewise, reducing time commitment by using speech samples of a shorter duration might also improve the likelihood of clinicians participating in measurement training procedures to improve their reliability. Both Eve et al. (1995) and Hoffman et al. (2014) recommended training procedures be developed for using the stuttering severity rating scale. If results continue to suggest that reliability needs to be improved, this could be achieved by skill development and training. If shorter samples are found to not affect reliability, and hence could be used in training, this would potentially reduce time commitment for speech-language pathologists to complete training and allow them more frequent feedback on their performance. A reduction in time commitment could be achieved by reducing the duration of each speech sample presented to judges, from 3 or 5-minutes, as used by Eve et al. and O’Brian et al. (2004b) respectively, to 1-minute.
However, no research to date has been conducted to determine specifically whether the duration of speech samples (e.g., 1-minute speech sample versus 3-minutes) influences the reliability of measurement of severity of stuttering. Therefore, the primary aim of this study was to explore the following research question:

Does the duration of speech samples influence the reliability of measurement of severity of stuttering by speech-language pathologists?

The reliability of observational data has been estimated in various ways in stuttering research. One way of estimating reliability is through evaluation of agreement. Point-by-point agreement has been the most frequently used agreement statistic in observational research for speech-language pathology practice when estimating both intrajudge and interjudge agreement (Cordes, 1994). Another way of assessing direct agreement is through examining the variability that occurs between scores assigned by judges to speech samples. Variability is strongly linked with reliability: The greater the variability in scores, the lower will be the reliability. The inverse relationship between the standard deviation (SD) of scores and point-by-point agreement can be demonstrated using a continuous distribution like the normal distribution model, by categorising it into discrete ordinal scores and determining point-by-point agreement. For the present study, reliability was estimated by considering the 9-point scale scores as coming from an underlying continuous distribution and comparing the variabilities between 3 sample durations of 1, 3, and 5-minutes, using variances or SDs. This necessitated a reform of the research question:

Does the duration of speech samples influence the variability of stuttering severity rating scores assigned by speech-language pathologists?
Method

Speech Samples

The study used 27 English audio speech samples of adults who stutter. The speech samples were of three different durations: 9 x 1-minute, 9 x 3-minutes, 9 x 5-minutes. Research has identified audio-only speech samples to be suitable for measuring stuttering severity (O’Brian et al., 2015; Rousseau et al., 2008). The duration of 1-minute was semi-arbitrarily selected. However, 3 and 5-minute durations were chosen to replicate durations in previous stuttering severity rating reliability measurement research (e.g., 3-minutes by Hoffman et al., 2014 and O’Brian et al., 2004b; 5 minutes by Eve et al, 1995). Each audio speech sample was an excerpt taken from a longer sample collected from different speakers. Excerpts were chosen to minimize any conversational partner’s contribution in the sample, and to represent a range of stuttering severity across the different durations.

The adult-speakers ranged in age from 18 to 51 years (mean = 29 years). Twenty-three of the 27 speech samples were from male-speakers and four from female-speakers (eight males and one female for the 1-minute samples, seven males and 2 females for the 3-minute samples, eight males and one female for the 5-minute samples). Scale scores represented a range of severity, as determined by consensus between the author and the primary supervisor: The 1-minute duration samples ranged from 1-7, the 3-minute samples ranged from 2-9, while the 5-minute samples ranged from 1-8.

Judges

The study involved 10 specialist speech-language pathologists for whom a large portion of their caseload is or has been clients who stutter. Specialist speech-language
pathologists were sought, in an attempt to reduce the impact on variability of difference
between judges by recruiting clinicians with extensive experience in using the measure.
Thus, any differences observed can more readily be attributed to the duration of speech
samples. Inclusion criteria were as follows: qualified speech-language pathologists who
practice/have practiced in English; who have more than five years working with clients
who stutter; and who have used the stuttering severity rating scale as part of their work
with clients who stutter. The judges were 10 stuttering specialist speech-language
pathologists aged between 28 and 71 years (mean = 47 years). One of the judges was
male and nine were female. The judges’ clinical experience with stuttering ranged from
7 to 41 years. No additional stuttering severity measurement training was provided to
the judges as part of this research.

**Stuttering Measurement Tool**

A 9-point severity rating scale (see Figure 2.1 in Chapter 2) was used in this
study. The judges were required to make a judgement of the severity of stuttering in
each speech sample by assigning a score ranging from 1 (“no stuttering”) to 9
(“extremely severe stuttering”). No other points on the severity rating scale were
labelled, as research has suggested that this appears to make little difference to
reliability (Cullinan et al., 1963).

**Procedure**

Information sheets (Appendix A) explaining the research and consent forms
(Appendix B) were distributed to potential participants. Once judges signed and
returned the consent form, Stimulus Packages were mailed for completion and return.
The Stimulus Packages included: an instruction sheet (Appendix C); a CD of speech
samples; severity rating score sheets (Appendix D); and a questionnaire (Appendix E).
The judges were asked to complete the rating session alone in a quiet room,
uninterrupted. Judges listened to 27 audio speech samples of different durations (1-minute, 3-minutes, 5-minutes) and were instructed to listen to and judge the severity of stuttering in each sample by assigning a score on the severity rating scale. They were asked to make their severity ratings on their overall impression of the speech sample and not to count stutters. The judges were advised that half scale values (e.g., 6.5) were not allowed. There were no practice items in the rating session.

**Data Analysis**

For the present study, reliability was estimated through evaluation of variability. Linear Mixed Models (LMM) were fit to severity scores with sample code (1 to 27) as a fixed explanatory variable to account for differences between samples and a single residual term. A series of variability structures were then added to this base model with rater code and time group (levels 1, 3 and 5-minutes) to assess the relative magnitudes of between rater variance, within rater variance and whether either of these two depended on duration group. The small sample version of the Akaike Information Criterion $AIC_c$ (Burnham & Anderson, 2004) was used to determine the best fitting variability structure.

A plot of the data suggested the variability might not be constant and there might be a curved relationship between SD and sample mean severity, so a regression analysis of sample SD against sample mean severity, with both a linear and quadratic components, was used to check this. This model was extended by adding the categorical duration variable and interactions with both the linear and quadratic mean severity terms to assess if the durations had different curvature relationships. The approximate normality of the distribution of SD for sample sizes of 10 was confirmed by simulation of normally distributed data, to provide assurance regarding the normality assumption for the linear regression. Proc Mixed from the *Statistical Analysis System (SAS)*
package, version 9.4, was used for the LMM modelling, and JMP, version 11.0, for the regression modelling. Statistical significance was set at \( \alpha = .05 \).

Results

The mean, mode, range, and SD of severity rating scale scores assigned by the judges to each speech sample are presented in Table 3.1. Table 3.1 identifies that the speech samples for each duration of time received a wide range of severity rating scores from the judges. The modal stuttering severity rating score assigned to the 1-minute samples ranged from 1-7, the 3-minute samples ranged from 2-9, while the 5-minute samples ranged from 1-8. These ranges replicated the ranges obtained for the samples by the author and primary supervisor (see Speech samples).

Table 3.1

| Mean, Mode, Range, and Standard Deviation of Severity Rating Scale Scores Assigned to Each Speech Sample by the Judges. |
|---|---|---|---|---|---|---|---|---|
| Sample | 1-Minute | 3-Minute | 5-Minute |
| | Mean | Mode | Range | SD | Mean | Mode | Range | SD | Mean | Mode | Range | SD |
| 1 | 4.50 | 5 | 2-7 | 1.43 | 3.60 | 3 | 3-5 | 0.70 | 5.10 | 5 | 4-7 | 0.99 |
| 2 | 5.00 | 4 | 3-7 | 1.56 | 8.80 | 9 | 8-9 | 0.42 | 5.30 | 6 | 4-6 | 0.82 |
| 3 | 6.70 | 7 | 6-8 | 0.67 | 1.80 | 2 | 1-4 | 0.92 | 3.40 | 3 | 3-4 | 0.52 |
| 4 | 6.70 | 7 | 5-8 | 0.95 | 5.50 | 7 | 4-7 | 1.27 | 1.40 | 1 | 1-2 | 0.52 |
| 5 | 1.30 | 1 | 1-2 | 0.48 | 3.30 | 3 | 2-5 | 0.95 | 7.20 | 8 | 6-8 | 0.79 |
| 6 | 3.60 | 3 | 2-6 | 1.17 | 2.20 | 2 | 1-3 | 0.63 | 3.80 | 5 | 2-7 | 1.69 |
| 7 | 2.70 | 3 | 2-4 | 0.67 | 1.70 | 2 | 1-2 | 0.48 | 6.50 | 7 | 5-7 | 0.71 |
| 8 | 4.30 | 4 | 3-8 | 1.57 | 6.50 | 7 | 5-8 | 0.97 | 1.80 | 2 | 1-3 | 0.63 |
| 9 | 6.40 | 6 | 4-8 | 1.17 | 4.80 | 5 | 3-7 | 1.14 | 7.70 | 8 | 6-9 | 0.95 |

The magnitude of three sources of variation in severity scores was examined using LMMs. Software used was G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007). The best model (based on the lowest AICc) from a series with increasing complexity in
variability structure, identified that there was significant between rater variability, SD, 0.34, 95% CI [0.22, 0.78]. An alternative model with 3 between rater variability components, that is a SD for each duration, was a poorer fit so the single SD estimate applies to all time durations. However, the best fitting model indicated within rater (residual) variability differed between the 3 durations, with 1-minute having the largest (SD=1.13, [0.98, 1.34]), and the other two durations being similar (3 and 5-minutes SDs 0.78, [0.67, 0.93] and 0.84 [0.73, 1.00] respectively). This approximately 40% larger variability for 1-minute samples compared to 3 and 5-minutes suggests the shorter duration (1-minute) is less reliable and would lead to lower agreement levels.

A post-hoc power analysis was carried out to determine what size difference in SD between durations could be detected using the study sample size. A reference variability was chosen that was typical of this study, SD =1 (for the 3-minute condition) and the F test for the ratio of two variances was used to determine that an increase in SD to 1.32 for another duration, or a reduction in SD to 0.75, could be detected with power 0.80 and type 1 error .05 (one sided). The sample size for each variance was set at 82 based on pooling the variances from the 9 samples in a duration (df=(10-1)*9=81).

The variability estimates from the above LMM modelling were based on the assumption that each sample within each of the durations had the same underlying variability and the estimates above were pooled across all 9 samples for each duration. The plot of SD against mean severity rating for each sample in Figure 3.1 suggested this assumption might not be correct. The spline curve in each panel is provided to assist in visualising the patterns in the data.
The final regression model ($R^2_{adj} = 0.41$) was chosen from a series described in the data analysis section. The relationship between the mean severity and SD had two terms, a non-significant main effect ($p = .11$) for the linear component of the model and a significant quadratic component ($p < .001$), indicating that the difference in curvature apparent in Figure 3.1 was statistically significant. Another model tried, tested whether the curvature differed between time durations. It did not ($p = .24$), so the curvature of the relationship between the mean severity and SD was deemed to be the same for all 3 durations. The non-significance of the linear term indicates there was no general rise or fall of SD with mean severity. The other term in the model was the main effect of time duration which was not significant ($p = .45$), indicating no difference in the SDs between the durations. Figure 3.2 shows the fitted regression lines for the 3 durations. The line for 1-minute is the line with the highest SD, with the other 2 durations being similar. This is consistent with the relative sizes from the LMM results, but the regression model did not agree with the significance of the difference between durations found in the LMM. The reason might lie in that the LMM did not account for the variability changing over the severity range. The most important result from the regression model is that there is good evidence to suggest that the variability of severity
rating is highest in the middle of the range, and has a maximum close to a severity rating of 5. This is the region with most variability and hence lowest reliability. Further from the middle of the scale, the variability decreases. For example, for a severity rating of 4 or 6, there is 5% less variability; for a severity rating of 3 or 7, 18% less; and for severity rating of 2 or 8, 41% less.

![Figure 3.2](image)

Figure 3.2. *Curves from the fitted regression model with separate lines for 1-minute (upper line, circles), 3-minutes (middle, crosses) and 5-minutes (lowest, diamonds). The difference between the durations was not significant.*

Previous investigators have assumed agreement levels were constant over the severity range. If the variability changes found here can be verified with additional data, then through the inverse relationship between agreement and variability, so must agreement vary. It would be lowest in the midrange of the scale and highest at the ends of the scale. Previous researchers, by assuming constant agreement levels over the whole range, would be overestimating agreement in the midrange and underestimating it at the ends of the scale if the samples they used to determine agreement were spread approximately evenly over the severity range.
Discussion

Previous research (Eve et al., 1995; Hoffman et al., 2014; O’Brian et al., 2004b) examining the reliability of use of the severity rating scale has shown conflicting results. However, these studies have used different analyses and varying durations of speech samples in their study, which could have been a contributing factor. No research has determined whether the duration of a speech sample affects the reliability of measurement. Therefore, the aim of the current study was to investigate whether the duration (1-minute, 3-minutes, 5-minutes) of speech samples influenced the reliability of measurement of severity of stuttering by specialist speech-language pathologists.

From the results, it appears that 1-minute samples show greater variability than 3 and 5-minute samples. Greater variability translates to poorer reliability. Due to more variability with 1-minute samples, it seems reasonable to assume that the reliability data for 1-minute samples would differ from that for 3 and 5-minute samples. From the results, specialist speech-language pathologists appear to be less reliable in their use of the severity rating scale to measure severity of stuttering in 1-minute speech samples. Therefore, these data indicate that researchers should not use 1-minute samples to evaluate the reliability of measurement of severity of stuttering, using the 9-point severity rating scale, nor for training purposes. What is encouraging to find was that the 3 and 5-minute speech samples generated similar data. This suggests that either 3 or 5-minute speech samples can be used for stuttering severity measurement research and/or training purposes, without impacting on the reliability of the measures.

An interesting finding to highlight was the increased variability evident in the three sample durations shown in the middle ranges of the scale. This has both research and clinical implications. Irrespective of the duration of a speech sample, it appears that speech-language pathologists will likely produce greater variability when rating the
severity of stuttering, using a 9-point scale, when severity is in the moderate range. This is not a fault of speech-language pathologists, but a consequence of the nature of the 9-point scale, with floor and ceiling effects at the ends making variability less. A similar effect is known for the binomial distribution where the SD for the true proportion has greater sampling variability at the middle of the true proportion range than at the edges. Therefore, speech-language pathologists should take steps to check their reliability when clients present with stuttering in the moderate range. Previous researchers have recommended practice and/or training in the use of the stuttering severity rating scale. Results of the current study suggest that training procedures with a focus in the middle range of the severity rating scale might be able to reduce variability.

The results of this study should be interpreted with caution: This study was used for research purposes to inform suitable durations of speech samples for future research in stuttering measurement. The speech sample duration used in research may not be suitable for clinical practice guidelines, as a reduction in the duration of speech samples used for clinical practice has the potential to jeopardize the validity of sampling. This study suggests that 3 and 5-minute speech samples would be appropriate to examine reliability levels in measurement research. If the purpose is to document progress or outcome measures for clinical research or clinical practice, researchers and clinicians should continue to take speech samples of a duration that will provide them with a valid sample of the client’s speech. It is unknown what minimum duration is necessary for making valid measures of severity of stuttering using the 9-point severity rating scale in clinical practice. However, 10-minutes is a commonly recommended speech sample duration for clinical practice.

In conclusion, future studies and the development and evaluation of practice and/or training procedures should avoid samples of 1-minute in duration, due to
increased variability in scores assigned by specialist speech-language pathologists. Current findings support the use of 3 or 5-minute speech samples for research purposes. In particular, the use of 3-minute speech samples (as opposed to 5-minutes) is encouraging, as this could reduce time commitment for future research and training procedures, therefore potentially increasing participation by speech-language pathologists.

Based on the results of Study 1, Study 2 (Chapter 4) and Study 3 (Chapter 5) will use 3-minute speech samples to investigate the reliability of use of the 9-point severity rating scale to measure severity of stuttering across languages.
CHAPTER 4

Study 2: The Reliability of Measuring Severity of Stuttering in a Foreign Language
The challenges arising with measuring stuttering in an unfamiliar language were presented in Chapter 1. Speech-language pathologists are required to provide appropriate treatment to all clients irrespective of their linguistic background, and this may include working with clients from linguistic backgrounds that differ from their own. However, there is still little evidence investigating how well speech-language pathologists are able to make reliable and valid judgements about identifying and measuring stuttering in languages other than their own. Therefore, this chapter presents a discussion of the available evidence for the reliability of the identification and/or measurement of stuttering in unfamiliar languages. Next, the second study is presented, replicating a preliminary study by Hoffman et al. (2014) with a larger sample size and a different unfamiliar language. This study investigated English-speaking speech-language pathologists’ reliability of use of the stuttering severity rating scale in (i) a familiar language (English), and (ii) an unfamiliar language (Vietnamese). The chapter concludes with a discussion of the research findings.

Research Investigating Identification and Measurement of Stuttering in Unfamiliar Languages

To date, only a limited number of existing studies have addressed skills important for judging or measuring stuttering or dysfluencies in unfamiliar languages. The first study was conducted by Humphrey (2004). Thus study examined whether bilingual English-Spanish-speaking undergraduate speech-language pathology student judges were more accurate at identifying the presence or absence of dysfluencies in Spanish speech samples than monolingual English-speaking judges. Results showed that there was no statistically significant difference between the two groups of judges. The study concluded that familiarity with Spanish made no significant difference when judging dysfluencies in Spanish. It should be noted that this study was conducted in
South Florida. Thus it is highly likely that the monolingual English judges had developed some degree of familiarity with features of the Spanish language, even though they did not understand or speak the language. Indeed, Humphrey acknowledged that these results might have been different in an area with different demographics. It is also important to highlight that the study did not establish whether the judges were able to identify stuttering reliably in any language, and it was unclear whether the judges were recording stuttering events or dysfluencies. Furthermore, the judges were students, and thus the results could not be generalised to speech-language pathologists.

A study by Van Borsel and Britto Pereira (2005) reported contradictory findings, identifying that language familiarity does influence stuttering judgements. This study evaluated how well an observer is able to accurately identify the presence or absence of stuttering in the speech of people who speak a language that is not their own. The study involved 28 undergraduate monolingual speech-language pathology student observers (14 Brazilian Portuguese and 14 Dutch-speakers). The observers were presented with twenty 2-minute monologue video speech samples (10 of Brazilian Portuguese-speakers, of which 5 stuttered and 5 did not stutter, and 10 of Dutch-speakers, of which 5 stuttered and 5 did not stutter). Results revealed that the Dutch observers were significantly more accurate in identifying people who stutter in their native language (98.6%) compared to the foreign language (80.0%). The Brazilian Portuguese observers showed no significant differences in identifying people who stutter in the speech samples, regardless of the language spoken (i.e., 81.4% native, and 82.9% foreign). There was no significant difference between the Brazilian Portuguese observers (82.9%) and Dutch observers (80.0%) correctly identifying stuttering in a foreign language. The study concluded that observers were able to identify whether a person stutters or not despite participants stuttering in a different language from the observer’s native language. However, it is important to note that the Dutch observers performed
significantly better when identifying stuttering in their native language (98.6%) than did the Brazilian Portuguese observers (81.4%). This poses a problem, as a potentially confounding variable, skill, has not been controlled. Thus, it is not possible to determine the extent to which the results were dependent on skill versus language.

The above studies should be interpreted with caution, as neither established whether judges were able to identify stuttering reliably. In addition, the judges were students, and thus the results could not be generalised to speech-language pathologists. Finally, no information was provided about intrajudge reliability. It is important to establish intrajudge reliability as it identifies the extent to which the same person can rate the same performance consistently.

In a more stringently controlled study, Einarsdóttir and Ingham (2009) examined whether accuracy of stuttering judgement is impacted by familiarity with the stuttering speaker’s language. The methodology used was time-interval judgement. In this study, 20 experienced speech-language pathologists (10 Icelandic and 10 United States) were presented with multiple, 5 second intervals of speech from nine 3 to 5 year-old Icelandic children speaking Icelandic. They were subsequently asked to identify whether stuttering was present in each interval. The study attempted to eliminate possible variables which could have confounded research findings. Therefore, experienced speech-language pathologists were selected as participants. These participants undertook pre-testing to ensure that they were able to identify stuttering accurately in their own language and that there were no significant differences between groups in terms of skill levels in making stuttering judgements. The pre-testing identified no significant differences between speech-language pathologists within each group, with Icelandic observers (88.5%) and United States observers (90.4%) meeting the 80% accuracy criterion to be included in the research study. Interjudge reliability of the
observers’ judgements was also established prior to commencing the task and intrajudge reliability was established with no significant differences between the Icelandic observers (89.6%) and United States observers (89.3%). This therefore, eliminated concerns of skill level being a variable which could potentially have confounded findings. Results of the study revealed no significant difference between the Icelandic observers (85.1%) and United States observers (86.0%) in identifying whether stuttering was present in the Icelandic speech intervals. The authors concluded that experienced speech-language pathologists were reliable in identifying and distinguishing between stuttered and non-stuttered speech of children who stutter in an unfamiliar language.

These three studies explored the role of language familiarity as a dichotomy (i.e., judge is familiar or unfamiliar) when making judgements of stuttering in unfamiliar languages. However, when making these judgments, it might be important to acknowledge the degree of unfamiliarity with the client’s language. Humphrey (2004) reported that, although an individual may not understand or speak a language, one may have developed some degree of familiarity with the language. Furthermore, Humphrey and Einarsdóttir and Ingham (2009) suggested that their results might have been different if the judgements involved languages that are more remote from the judge’s native language. Thus, it is important to explore the influence of the closeness of the language to the speech-language pathologists’ native language when making stuttering judgements.

A study conducted by Van Borsel, Leahy, and Britto Pereira (2008) extended and replicated Van Borsel and Britto Pereira’s (2005) study with the additional inclusion of an English-speaking panel comprised of 14 English-speaking undergraduate speech-language pathology students. This study examined the closeness of the native language to, and its influence on, making judgements of stuttering in an
unfamiliar language. The methodology was identical to Van Borsel and Britto Pereira’s study, except that the observers judged only the same Dutch samples. The observers did not make any judgements of English samples. Once the English-speaking observers had completed the assessments, the results were then compared with the Brazilian Portuguese and Dutch speech pathology students’ results for identifying the presence or absence of stuttering in the Dutch samples from Van Borsel and Britto Pereira’s study.

The Van Borsel et al. (2008) study revealed a significant difference between the English-speaking and Brazilian Portuguese observers. The English-speaking observers performed significantly better (88.6%) in making judgements of Dutch-speakers where stuttering was absent, than the Brazilian Portuguese observers (74.3%). There was no significant difference between the English-speaking observers (84.3%) and Brazilian Portuguese observers (82.9%) in identifying the presence of stuttering in the Dutch samples. Interestingly, results showed that both Brazilian Portuguese and English-speaking observers were significantly less accurate in identifying the presence of stuttering than the Dutch observers (98.6%). It is important to note, English and Dutch languages share similar characteristics as they are West Germanic branch languages, whereas the Brazilian Portuguese is a Romance branch language. Therefore, the English and Dutch languages are more similar to each other than they are to the Brazilian Portuguese language. The study concluded that the degree of similarity between the native languages to be judged was an influencing factor in accuracy of judgements of stuttering in an unfamiliar language, thus highlighting that, when an unfamiliar language is further remote from the observers’ native language, there is higher risk for false-positive or false-negative judgements. It should be noted, however, that the English observers did not make any judgements of English samples. Thus, the observers’ accuracy of judging stuttering in their native language is unknown. Without knowing how accurate the English observers are in identifying stuttering in their native
language, it is impossible to determine to what extent their results for judgements of stuttering in Dutch were dependent on this skill and to what extent they were determined by degree of similarity between Dutch and English.

The conclusions from the above studies highlight that language familiarity may be a contributing factor and/or cannot be completely excluded as an influence in the identification of stuttering in a foreign language. Furthermore, the closeness of the foreign language to the native language may also contribute when making stuttering judgements. However, it is important to highlight that the identification of stuttering and measurement of severity are related but not identical tasks. The studies reviewed above have not attempted to comment on how identifying the presence or absence of stuttering, whether in 5-second speech intervals or longer speech samples, translates to the reliability of measurement of severity of stuttering in a foreign language. Whilst is it essential to be able to identify the presence or absence of stuttering to determine those who are appropriate for stuttering treatment, once clients are undergoing treatment, it is crucial for speech-language pathologists to be able to reliably measure client progress and outcomes. Only a few recent studies have examined speech-language pathologists’ reliability of measuring severity of stuttering in an unfamiliar language.

A study conducted by Lee, Robb, Ormond, and Blomgren (2014) examined 19 English-speaking speech-language pathologists’ abilities to judge stuttering of two bilingual Spanish-English adults who stutter. These judges were presented with four 1-minute video samples (two from each speaker, one English and one Spanish) of a reading passage (the rainbow passage). Judges were instructed to assess the frequency of stuttering, severity of stuttering (mild, moderate, severe), location (the syllable or word within the reading sample on which the stuttering occurred) and type (repetition, audible fixed posture, inaudible fixed posture, interjection, revision) of individual
stuttering moments, duration of stuttering moments, and physical concomitants (distracting sounds, facial grimaces, head movements, movements of the extremities) of stuttering in both languages spoken by the two speakers. Accuracy of stuttering judgements was determined by comparing their observations of stuttering to those of three expert Spanish-English bilingual speech-language pathologists. Results showed that judges are more reliable in making fine judgements and identifying individual instances of stuttering in English compared to Spanish in terms of ‘location only’ and ‘location and type’ judgements. This confirmed the presence of a language familiarity affect. The results also showed that judges’ overall ratings of severity (mild, moderate, severe), frequency, and type were not different from the bilingual speech-language pathologists’ ratings. This therefore suggests that these factors may not be significantly influenced by language familiarity. Results of the study are consistent with previous findings that suggest speech-language pathologists can accurately identify stuttering from culturally and linguistically diverse backgrounds (e.g., Van Borsel and Britto Pereira’s (2005). It was acknowledged that the study used samples from reading as opposed to spontaneous speech, which is commonly used throughout assessment, therefore limiting the generalizability of the findings.

A second study, conducted by Bosshardt, Packman, Blomgren, and Kretschmann (2015), investigated to what extent groups of speech-language pathologists agree in their stuttering severity ratings when measuring stuttering for seven children speaking different languages. Judges (n = 170) rated speech samples of seven preschool aged children, each of whom spoke one of the following languages: Danish, English, French, German, Greek, Italian, and Persian (Farsi). All languages spoken by the children in the study derive from the Indo-European language family. The judges were grouped by country (i.e., United Kingdom, Denmark, France, Germany, Greece, Italy, Iran, Malta, and United States). Judges were required to listen
to a two-minute audiovisual speech sample and rate the severity of stuttering for each child using a 10-point severity rating scale. To determine whether the stuttering severity ratings were related to the judges’ proficiency in the child’s language, judges self-rated their language proficiency in each child’s language using an 8-point scale. Analyses evaluated agreement of severity rating measures within and between the groups of judges. Findings indicated that the groups of judges did not significantly differ in the extent to which they agreed on the ratings for each of the seven children. Therefore, it appears that ratings were not influenced by familiarity of each group with each child’s language. Results furthermore identified that judges’ proficiency with the child’s language accounted for a maximum of 4.6% of observed within group variances. The authors concluded that language proficiency was not systematically related to variability and agreement of severity ratings. Thus, the influence of language proficiency on measuring stuttering might not be as great as suggested by previous researchers (e.g., Van Borsel et al., 2008).

A study conducted by Hoffman et al. (2014) estimated speech-language pathologists’ reliability when measuring severity of stuttering in a considerably different unfamiliar language (Mandarin) from their familiar language (Australian English). Judges were 26 Australian English-speaking speech-language pathologists who listened to 20 audio-only 3-minute speech samples (10 Australian English and 10 Mandarin) and were instructed to rate the severity of stuttering using a 9-point stuttering severity rating scale. Results identified that judges showed poorer agreement when using the scale to measure stuttering in Mandarin compared to Australian English speech samples for both intra- and interjudge agreement. Judges who achieved criterion performance for the use of the severity rating scale in Australian English were more likely than other judges to demonstrate reliability in Mandarin for both intra- and interjudge agreement. However, demonstration of criterion performance in English did
not guarantee criterion performance in Mandarin. Furthermore, results showed that 50% of individual judges were unable to reliably measure the severity of stuttering in a familiar language. The study highlighted the need for speech-language pathologists to develop both intra- and interjudge agreement in a familiar language as well as unfamiliar languages. The study recommended research into the development of practice and training procedures with the inclusion of familiar and unfamiliar language exemplars to assist speech-language pathologists to develop reliability.

Until recently, studies have focused on examining language familiarity and closeness of the language to the native language when making judgements of stuttering. However, there are other variables that may be factors influencing measurement of stuttering in unfamiliar languages. Firstly, it is possible that a client’s severity of stuttering is likely to influence speech-language pathologists’ judgements. Watson and Kayser (1994) suggested that if the clients stutter severely, identifying stuttering may not be a difficult task. This is because speech-language pathologists will often observe excessive tension and secondary behaviours without difficulty, despite the language being unfamiliar (Van Borsel & Britto Pereira, 2005). Thus, it is likely that assessing mild and moderate stuttering that is not accompanied by these severe behaviours will be a more difficult task for speech-language pathologists. Secondly, it has been documented that the experience and training of a speech-language pathologist may influence the accuracy and reliability of making stuttering judgements in their own language (Brundage, Bothe, Lengeling, & Evans, 2006). Thus, it seems plausible that the task of identifying and measuring stuttering accurately and reliably could be even more difficult for novice and less experienced speech-language pathologists in a familiar and in any unfamiliar languages spoken by the client. Currently, only one study has taken into consideration the interplay of a range of factors involved in the measurement of stuttering severity.
Cosyns, Einarsdóttir, and Van Borsel (2015) expanded on the previous studies and explored different factors involved in the identification of stuttering severity in a foreign language. Factors examined were language familiarity, closeness of the speech-language pathologist’s language to that of the client, stuttering severity of the client, and experience of the speech-language pathologist with assessing stuttering. The study was comprised of 19 speech-language pathologists from five countries (Iceland, Sweden, Norway, Finland, and Belgium). Judges were divided up into two groups: experienced judges, meaning they had 5 or more years of professional experience; and non-experienced, meaning judges had less than 5 years of professional clinical and/or academic experience in the assessment and treatment of stuttering. It was noted that few worked exclusively with clients who stutter. Judges were presented with eight video speech samples (2 x Icelandic, 2 x Swedish, 2 x Norwegian, and 2 x Dutch) and instructed to rate the severity of stuttering on a 10.0-cm visual analogue scale. In addition, judges were asked to score how easy they found it to rate the severity of stuttering of each speech sample, also using a 10.0-cm visual analogue scale. For this study, accuracy was calculated as the absolute difference between stuttering severity scored by the experienced native raters and stuttering severity scored by the judges. Results showed that experienced and inexperienced judges can accurately rate severity of stuttering for severe stuttering. This finding was consistent with Watson and Kayser (1994). When stuttering was in the mild to moderate range, experienced judges were more likely to make accurate measures of stuttering severity in comparison to inexperienced judges. Furthermore, the study showed that, when stuttering was mild to moderate, the following pattern was present: The closer the language was to the judge’s native language, the more accurate an experienced judge will be. This pattern was not present for the inexperienced judges. The study highlighted the need for developing and
evaluating stuttering measurement training in different languages focusing on the identification of stuttering in the mild and moderate range.

**Aim of the Present Study**

It appears that all of the studies discussed above, with the exception of Hoffman et al. (2014), have examined group performance to investigate the reliable identification and/or measurement of stuttering in unfamiliar languages. Presenting grouped data alone enables researchers to determine acceptable levels of agreement for research purposes only, as it can only discriminate between groups of individuals. However, acceptable group reliability ratings do not necessarily indicate that any given individual will achieve acceptable reliability levels (Martin & Haroldson, 1992). To determine the clinical usefulness of measurement, it is important for research to also include data for individual participants, so that researchers and clinicians are able determine whether changes documented over time are due to actual differences in the client’s stuttering or due to error in the clinician’s measurements. Hoffman et al. was the first study to present data describing individual for speech-language pathologists’ reliability when measuring severity of stuttering in a familiar (Australian-English) and unfamiliar language (Mandarin). With only one unfamiliar language investigated using individual analyses, it is unknown whether similar results would be obtained if different unfamiliar languages were investigated. Thus, the aim of this study is to investigate a different unfamiliar language.

With the exception of Hoffman et al. (2014), all studies have examined languages from different branches (e.g., North Germanic, West Germanic, Romance) within the Indo-European language family. Languages that are from the same language family share some linguistic characteristics. Therefore, despite judges identifying and/or measuring stuttering in a language that they do not speak or understand, languages from
the same branch (e.g., Germanic or Romance branch of Indo-European) may not be completely foreign. Thus, it is possible that these judges may in fact have developed at least some familiarity with the language. Van Borsel et al. (2008) suggested that when an individual’s native language is further remote from the foreign language of the client, there is a possibility of a higher risk for false-negative or false-positive identification. This is problematic, as speech-language pathologists nowadays might be required to work with clients whose native language derives from a different language family, which consequently may have considerably different linguistic properties from their native language. Yet it is essential that they are able to make reliable measures of clients’ stuttering, irrespective of language, to monitor progress and outcomes.

It is important to determine whether findings from predominantly Indo-European studies can be generalised to different language families. Thus, the present study aimed to expand the existing knowledge of measuring severity of stuttering in a foreign language and to examine a language from an entirely different family that is commonly spoken worldwide: Vietnamese, a language from the Austro-Asiatic language family.

Vietnamese is spoken by a significant portion of the population, and ranks within the top 20 languages spoken worldwide, with approximately 68 million Vietnamese-speakers (Lewis et al., 2016). Vietnamese is spoken by a significant portion of immigrants in 23 countries (e.g., Norway, with 1.9% of the population speaking Vietnamese; Laos, 1.1%; and New Caledonia, 0.9%; Lewis et al., 2016). Furthermore, Vietnamese is also wildly spoken in many predominantly English-speaking countries. Therefore, it is highly likely that English-speaking speech-language pathologists around the world will be required to work with monolingual and/or multilingual clients who speak Vietnamese. As the patterns of severity of stuttering may vary between speakers’
languages (Shenker, 2011a), it could be argued that it is important to determine whether English-speaking speech-language pathologists can reliably measure severity of stuttering in Vietnamese.

As English and Vietnamese do not share similar language characteristics, it is likely based on previous research (Van Borsel et al., 2008) that English-speaking speech-language pathologists may find it difficult to reliably measure severity of stuttering in Vietnamese. However, it is unknown whether Van Borsel and colleagues’ conclusions can be generalised to other language families such as Austro-Asiatic. Thus, the present study explored the following research question:

Can a 9-point severity rating scale be used by English-speaking speech-language pathologists to reliably measure severity of stuttering in a completely foreign language (Vietnamese)?

The study also investigated English-speaking speech-language pathologists’ ability to reliably measure severity of stuttering in English, as there have been conflicting findings that not all research into the reliability of use of the severity rating scale for English speech samples has found that it can be used reliably (e.g., Hoffman et al., 2014; Kully & Boberg, 1988). Therefore, this study also investigated a subsidiary question:

Can a 9-point severity rating scale be used by English-speaking speech-language pathologists to reliably measure severity of stuttering in a familiar language (English)?
Method

Speech Samples

The study used 20 audio speech samples of adults who stutter (10 English and 10 Vietnamese). The use of audio rather than audio-visual samples was chosen as research (e.g., O’Brian et al., 2015; Rousseau, Onslow, Packman, & Jones, 2008) has identified audio-only speech samples to be adequate for measuring severity of stuttering. Audio speech samples were 3-minute excerpts taken from longer duration speech samples collected from 20 different speakers. Excerpts were chosen to ensure a representation of a range of stuttering severity across the samples, and to minimize any conversational partner’s involvement in the sample. Information about the adult-speakers and the speech samples is presented in Table 4.1.

Table 4.1

*Age, Gender, and Stuttering Severity Rating Scale Scores of the Adults Depicted in the Speech Samples*

<table>
<thead>
<tr>
<th>Sample</th>
<th>English</th>
<th></th>
<th>Vietnamese</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age</td>
<td>Gender</td>
<td>SR</td>
<td>Age</td>
</tr>
<tr>
<td>1</td>
<td>47</td>
<td>Male</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>Male</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>Male</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>Male</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>Female</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>63</td>
<td>Male</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>Female</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>Male</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>42</td>
<td>Male</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>Male</td>
<td>5</td>
<td>26</td>
</tr>
</tbody>
</table>

*Note. SR = severity rating*

The adult-speakers ranged in age from 21 to 63 years (i.e., English-speakers ranged from 21 to 63 years, mean = 35 years; and Vietnamese-speakers ranged from 24
to 31 years, mean = 27 years). Eighteen of the 20 speech samples were from males and two from females (i.e., eight males and two females for English samples, and ten males for Vietnamese speech samples). Severity scale scores represented a range of severity: The English samples ranged from 2 to 9, and the Vietnamese samples ranged from 1 to 8.

_Feedback Sample Selection Process_

The severity rating scale scores used to describe the severity range of the samples were made by a bilingual panel, comprised of two speech-language pathologists. The first panel member was a research collaborator of the study, and the second panel member was independent of the study. The bilingual panel scores were obtained to ensure a range of severity of stuttering in each language, but were not used any further in the research.

The bilingual panel members were native speakers of Vietnamese and fluent in English, and were trained and had experience using the severity rating scale in the assessment and management of stuttering. The bilingual panel members listened to each speech sample in a room together and then each member recorded her rating score. The panel members then compared their rating scores. If the scores were the same, the panel members moved on to the next sample. If the scores differed, the bilingual panel discussed and reached consensus on an appropriate score. The bilingual panel were told to spend as much time as they needed to replay the speech samples and discuss rating scores in order to reach consensus on a score for each speech sample. The panel members also identified any samples they believed to be of poor quality and any samples that contained information that could potentially identify the speakers depicted in the samples.
Samples that were identified by the bilingual panel to be of poor quality and/or to contain any potentially identifying information were discarded. From the remaining 23 samples, 20 samples (10 in English and 10 in Vietnamese) that represented a range of severities were selected. Samples were selected via purposive sampling to ensure representation of a range of severities (Black, 1999). Stratified random sampling was used to select samples for inclusion when multiple available samples fell within the same scale value (Bryman, 2012). Out of the 20 samples, 17 were purposively selected while three were randomly selected. These 20 samples were used to form the Stimulus Packages that were presented to participants.

Judges

The judges were 59 English-speaking speech-language pathologists, aged between 22 and 70 years (mean = 39 years) from 10 different countries (i.e., Australia, United States, Canada, England, Scotland, Northern Ireland, Argentina, South Africa, Portugal, and Germany). Two of the judges were male and 57 were female. The judges’ clinical experience ranged from a minimum of 6 months up to approximately 42 years. The inclusion criteria for the judges were as follows: (i) qualified speech-language pathologist; (ii) who practices or has practiced speech-language pathology in English; (iii) who has worked with a stuttering client within the past 12 months; and (iv) was unfamiliar with the Vietnamese language. ‘Unfamiliar’ was defined as not speaking and not understanding the Vietnamese language. All participants could speak English and 26 judges reported communicating in a language and/or languages other than English (i.e., Afrikaans, Arabic, British Sign Language, Cantonese, Chinese, French, German, Greek, Irish, Italian, Kiswahili, Mandarin, Portuguese, Russian, Spanish, Swedish, Turkish). Out of 59 judges, three judges reported speaking a tone language (i.e., Cantonese, Chinese, and Mandarin). Despite Vietnamese also being a tone language,
Vietnamese derives from a completely different language family (Austro-Asiatic) from Cantonese, Chinese and Mandarin (Sino-Tibetan). To our knowledge, no research has identified that Sino-Tibetan languages (e.g., Cantonese, Chinese and Mandarin) are mutually intelligible with Vietnamese. As the reported tone languages spoken by the three judges are not closely related to Vietnamese, these judges were not excluded from the study. The judges’ recent experience working in the area of stuttering was diverse. Judges reported to have assessed and/or treated between approximately 1 and 500 clients who stutter in the previous 12 months.

**Stimulus Package**

The judges were mailed a Stimulus Package to complete on two separate occasions. The first rating session Stimulus Package included: an instruction sheet (Appendix F); a compact disk (CD) of speech samples, in which the number of each speech sample was announced before the sample was presented, and a 10 second silence was inserted between each sample; severity rating score sheets (Appendix G); and a questionnaire (Appendix H).

For the second rating session, the Stimulus Package, procedure, instructions, and scoring sheets were identical to the first rating session. However, the same 20 speech samples from the first rating session were presented in a different random order for the second rating session. The second rating session Stimulus Package included: an information sheet (Appendix I) explaining the purpose of completing a second rating session; a consent form (Appendix J); an instruction sheet (Appendix K); a CD of speech samples; and severity rating score sheets (Appendix G).
Procedure

Information sheets (Appendix L) explaining the research and consent forms (Appendix M) were distributed to potential judges. Once judges signed and returned the consent form, the Stimulus Package was mailed to them for completion. The study used a 9-point severity rating scale, commonly used in clinical practice when working with adults who stutter, and replicated procedures described by Hoffman et al. (2014). Judges were instructed to listen to and make judgements of the severity of stuttering in each sample by assigning a score ranging from 1 (‘no stuttering’) to 9 (‘extremely severe stuttering’). Judges were advised that ‘Half scale values (e.g., 6.5) are not allowed’. To examine intrajudge agreement, the judges completed the rating sessions on two occasions separated by at least one month. The same 20 audio speech samples from the first rating occasion were presented in a different random order for the second rating occasion. Judges were not initially informed that they were to be asked to complete a second rating occasion. There were no practice items in either session. Of the 59 judges, 40 completed both rating sessions.

Data Analysis

Reliability was estimated through evaluation of agreement. Descriptive and inferential statistics were used to estimate both intrajudge and interjudge agreement.

Intrajudge Agreement

Intrajudge agreement was estimated by examining the percentage of individual judges who met criterion performance. An individual judge was considered to have achieved criterion performance if their severity rating scores fell within ±1 scale value on Time 1 to Time 2 comparison, on 80% of samples (as per Hoffman et al., 2014). The percentage of judges’ severity rating scores that fell within ±1 scale value for 70% or
more of samples was also calculated, to enable some commentary about the degree of improvement required to achieve criterion performance for each language.

*Interjudge Agreement*

Interjudge agreement was estimated by determining the performance of individual judges, to ascertain the percentage of judges who achieved predetermined criterion performance. Criterion performance was defined as having assigned a score within ±1 scale value of the modal score for the sample, for 80% of samples (as per Hoffman et al., 2014). Eighty percent has been identified as minimum acceptable agreement for stuttering measurement research (Cordes, 1994). The percentage of judges who assigned a score within ±1 scale value of the modal score for 70% or more of samples was calculated to provide further commentary of judges’ performance.

*Performance in Vietnamese for the Subgroup of Judges Reliable in English*

A subset of judges (n = 14) achieved criterion intra- and interjudge agreement using the severity rating scale for the English speech samples. These judges were therefore deemed to have made reliable use of the scale for English samples. Data describing their performance on Vietnamese samples were also reported separately (‘Vietnamese\textsubscript{ENG}’) for each of the analyses described above.

*Logistic Regression*

The Generalised Linear Mixed Model (GLMM) modelling process was used (i) to determine whether percentage of agreement within ±1 scale value varied over the severity range, (ii) to determine whether language (English versus Vietnamese) impacted on judges’ performance, (iii) to quantify the difference between intrajudge compared to interjudge agreement, and (iv) to determine whether any of these three
effects depended on the rating type (intrajudge, interjudge) or language (English, Vietnamese).

The GLMM approach was used to fit to a logistic regression fixed effects model, between the outcome agreement within ±1 scale value and the fixed effect explanatory variables: mean severity rating for each sample, rating type, and language. The agreement data were entered into the model in contingency table format as (number of agreements within ±1 scale value, number of pairs used to assess agreement) for each sample. Simulation work showed, however, that for interjudge agreement, where the number of pairs was much larger than the number of judges, there would have been a substantial overestimate of the precision of the agreement estimates. Therefore, sample weights were applied to correct for this, the weights being 1 for intrajudge agreement, and number of judges divided by number of judge pairs for interjudge agreement.

Models examined the curvature of the severity rating effect with polynomials up to degree 4. Two way interactions examined were between severity rating and rating type (intrajudge, interjudge), and between severity rating and languages, and between rating type and languages. A random intercept was used in the model to account for unexplained variability between samples. Possible differences in variability between rating types, or rating sessions (Time 1, Time 2) were assessed by adding grouping variables to the random effect for these variables. The random effects structure was assessed using either Akaike Information Criterion (Burnham & Anderson, 2004) comparisons between models or a homogeneity test for differences in variances between groups, as appropriate. The final GLMM model had one random effect with intercepts based on sample code. Model fit was assessed by plotting Pearson residuals against severity rating by categorical groups, and from an overall goodness of fit index by the Generalised chi-square divided by df ratio, which was 0.87. The model fit was
considered satisfactory. Effect sizes were assessed with odds ratios (OR) and 95% confidence intervals, and statistical significance was set at $\alpha=0.05$. PROC GLIMMIX from the *Statistical Analysis System (SAS) package, version 9.4* was used for the GLMM modelling.

## Results

### Descriptive Statistics

The mean, mode, range, and standard deviation of severity rating scale scores assigned by the judges to each speech sample at the first rating occasion are presented in Table 4.2. Table 4.2 shows that the speech samples for each language received a range of severity ratings from the judges. The modal stuttering severity rating scores assigned to the English samples ranged from 2 to 9, while the Vietnamese samples ranged from 1 to 8. These ranges replicated the ranges obtained for the samples by the bilingual panel.

### Table 4.2

*Mean, mode, range, and standard deviation of severity rating scale scores assigned by the judges to each speech sample. Data are presented for the first rating occasion only.*

<table>
<thead>
<tr>
<th>Sample</th>
<th>English</th>
<th></th>
<th>Vietnamese</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mode</td>
<td>Range</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>2.59</td>
<td>2</td>
<td>1-6</td>
<td>1.08</td>
</tr>
<tr>
<td>2</td>
<td>3.46</td>
<td>4</td>
<td>1-7</td>
<td>1.26</td>
</tr>
<tr>
<td>3</td>
<td>3.69</td>
<td>4</td>
<td>1-7</td>
<td>1.44</td>
</tr>
<tr>
<td>4</td>
<td>6.25</td>
<td>6</td>
<td>3-8</td>
<td>1.17</td>
</tr>
<tr>
<td>5</td>
<td>4.76</td>
<td>5</td>
<td>2-7</td>
<td>1.26</td>
</tr>
<tr>
<td>6</td>
<td>8.39</td>
<td>9</td>
<td>7-9</td>
<td>0.64</td>
</tr>
<tr>
<td>7</td>
<td>2.93</td>
<td>2</td>
<td>1-6</td>
<td>1.20</td>
</tr>
<tr>
<td>8</td>
<td>8.90</td>
<td>9</td>
<td>8-9</td>
<td>0.30</td>
</tr>
<tr>
<td>9</td>
<td>1.98</td>
<td>2</td>
<td>1-4</td>
<td>0.71</td>
</tr>
<tr>
<td>10</td>
<td>5.17</td>
<td>4</td>
<td>2-9</td>
<td>1.52</td>
</tr>
</tbody>
</table>
**Intrajudge Agreement**

Table 4.3 depicts the percentage of judges who achieved criterion level performance for intrajudge agreement, for each language. The table shows that 75% of individual judges achieved the criterion level of intrajudge agreement for English. Only 55% of individual judges were able to achieve the criterion level of agreement for Vietnamese. The large majority of judges (90%) were able to achieve scores within ±1 scale value for 70% or more of samples for English, showing that many judges were approaching criterion agreement. The corresponding figure for Vietnamese was 75%.

A subset of judges who achieved criterion intra- and interjudge agreement using the severity rating scale for the English samples was examined. For those judges who were deemed to have made reliable use of the scale for English samples (n = 14), intrajudge agreement in Vietnamese was examined. The performance for this subgroup was similar to that of the entire group: 57% achieved criterion performance compared to 55% for the entire group, and 71% were approaching criterion performance compared to 75% for the entire group.

Table 4.3

<table>
<thead>
<tr>
<th>Judge Condition</th>
<th>Intrajudge</th>
<th>Interjudge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≥ 80%</td>
<td>≥ 70%</td>
</tr>
<tr>
<td>English</td>
<td>75%</td>
<td>90%</td>
</tr>
<tr>
<td>Vietnamese&lt;sub&gt;all&lt;/sub&gt;</td>
<td>55%</td>
<td>75%</td>
</tr>
<tr>
<td>Vietnamese&lt;sub&gt;ENGr&lt;/sub&gt;</td>
<td>57%</td>
<td>71%</td>
</tr>
</tbody>
</table>

Note. Intrajudge N = 40 for English and Vietnamese<sub>all</sub>; N = 14 for Vietnamese<sub>ENGr</sub>; and Interjudge N = 59 English and Vietnamese<sub>all</sub> and N = 14 Vietnamese<sub>ENGr</sub>.

*Vietnamese<sub>ENGr</sub> presents scores for Vietnamese for the subgroup of judges who achieved both intra- and interjudge agreement in English.
Interjudge Agreement

Table 4.3 also shows the percentage of judges who achieved criterion level performance for interjudge agreement, for English and Vietnamese. Criterion level interjudge agreement was achieved by 66.10% of individual judges for English samples at Time 1 and by 44.07% for Vietnamese. Thirteen of 59 (22%) judges did not achieve criterion performance in either English or Vietnamese. All of the other judges achieved criterion performance in at least one of the languages. The percentages of judges who achieved scores within ±1 of the modal score on at least 70% of samples were 84.75% for English, and 72.88% for Vietnamese.

The results for Time 2 were higher for English, with 70% of judges achieving acceptable criterion performance. Judges’ performance in Vietnamese at Time 2 (42.50%) remained virtually unchanged. Nine of 40 (22.5%) judges did not achieve criterion performance in either English or Vietnamese. All of the other judges achieved criterion performance in at least one of the languages. A greater percentage of judges achieved scores within ±1 of the modal score on at least 70% of samples for Time 2 for both languages.

Logistic Regression

The final GLMM model showed four significant effects: two severity rating terms and main effects for rating type (intrajudge, interjudge) and familiarity with the speech sample language (English versus Vietnamese), and a random intercept for differences between speech samples. None of the effects varied with severity rating, as interactions with and between the two severity rating terms were not significant. Thus, for the present study, all model effects were independent of the levels of other variables. Figure 4.1 shows the agreement percentage levels from the fitted model at all levels of severity rating, for rating type (intrajudge, interjudge) and languages.
On investigation of the percentage of agreement that varied over the severity range, the final GLMM model had two significant severity rating terms, one linear and one quadratic (B(SE) = -1.67 (0.14) and 0.171 (0.014), both p < 0.01). The two higher order polynomial terms tested were not significant. The fitted curved relationship had minimum agreement close to a severity rating scale score of 5, and increased towards the edges of the scale (see Figure 1). This is also seen in and consistent with the findings of Study 1 (Chapter 3). Agreement between the observed data and the fitted curved line identifies that the 9-point stuttering severity rating scale does not have constant agreement over the severity range. When looking at Figure 4.1, for intrajudge agreement, judges reached 80% agreement for mild stuttering (severity rating of 1-3) and severe stuttering (severity rating of 7-9) for English and Vietnamese. Judges showed poorer agreement in the moderate range (severity rating of 4-6) for both languages. Judges’ performance with interjudge agreement for English and Vietnamese
was also similar, with judges achieving agreement for mild and severe stuttering (severity rating of 1-2 and 8-9). Again, judges showed poorer agreement for each language in the moderate range of the scale (severity rating of 3-7).

For the present study, the degree of curvature of the agreement versus severity rating effect was found to be constant across rating types (intrajudge, interjudge), and familiarity of the judge with the language, as none of the interactions between these variables and the linear and quadratic terms in the models were significant.

Logistic regression was also used to determine if there was any relationship between languages (English and Vietnamese), and whether there was any difference between intrajudge compared to interjudge performance. Results are presented in Table 4.4. For this analysis, one group (e.g., Vietnamese) is used as a reference and given a value of 1.00. If the odds ratio for the other group (e.g., English) is also 1.00, then there is no difference between the two groups.

<table>
<thead>
<tr>
<th>Variable value</th>
<th>Odds ratio (OR)</th>
<th>95% confidence interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>1.00*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>1.47</td>
<td>1.19, 1.81</td>
<td>0.001</td>
</tr>
<tr>
<td>Rating Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interjudge</td>
<td>1.00*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrajudge</td>
<td>2.20</td>
<td>1.79, 2.70</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*reference level

Judges had higher agreement levels in English than in Vietnamese, with an odds ratio of 1.47. This shows that English-speaking judges were 1.47 times more likely to reliably rate severity of stuttering using the severity rating scale in English than in a foreign language (Vietnamese). The confidence interval suggests that the difference
could be as little as 1.19 times more likely, or as much as 1.81 times more likely. There is a statistically significant ($p = 0.001$) advantage to judges being able to rate more reliably in a familiar (English) language than in an unfamiliar (Vietnamese) language. To further understand odds ratios, which are similar to comparing ratios of percentages, Figure 4.1 enables one to visually compare the percentage agreement. At a severity rating of 5 for the two interjudge rating curves (lower curves), judges reach approximately 52% agreement for English, and only 43% agreement for Vietnamese. This result suggests that familiarity of language impacts on judges’ ability to reliably measure severity of stuttering using the severity rating scale.

The rating type variable was significant ($p = <0.001$) in the model, with agreement for intrajudge agreement being higher than for interjudge agreement, with an odds ratio of 2.20. This suggests that judges are just over two times more likely to rate severity of stuttering reliably with themselves than with other judges. This result confirms the descriptive statistics showing that judges performed better with intra- than interjudge agreement. The confidence interval estimate in Table 4.4 suggests that odds for measuring stuttering reliably against yourself (intrajudge), could be as little as 1.79 or as much as 2.70 times as great as the odds for measuring stuttering reliably with others (interjudge). Figure 4.1 also enables the visual comparison of percentage agreement between intra- and interjudge agreement. For example, at a severity rating of 5, judges reached around 71% intrajudge agreement (upper curve) and around 52% interjudge agreement (lower curve) for English (familiar language). For Vietnamese (unfamiliar language), judges reached approximately 63% intrajudge agreement (upper curve) and only around 43% interjudge agreement (lower curve) for Vietnamese. This difference is approximately 20% for the intrajudge versus interjudge comparisons. These results suggest that rating type impacts on judges’ ability to reliably measure severity of stuttering using the severity rating scale.
Discussion

Previous studies, with the exception of Hoffman et al. (2014), have examined identification and/or measurement of stuttering in language/s from primarily the Indo-European (Bosshardt et al., 2015; Cosyns et al., 2015; Einarsdóttir & Ingham, 2009; Lee et al., 2014; Van Borsel & Britto Pereira, 2005; Van Borsel et al., 2008) language family. Hoffman et al. was the first study to investigate the Sino-Tibetan (Mandarin) language family. The present study introduced a different language family (Austro-Asiatic) to this field of research. Results of the present study concur with those of previous studies (e.g., Hoffman et al., 2014; Lee et al., 2014; Van Borsel & Britto Pereira, 2005), which suggested that language familiarity plays a pivotal role in speech-language pathologists’ abilities to reliably measure severity of stuttering.

The findings of the present study corroborate those of Hoffman et al. (2014), showing speech-language pathologists to be more reliable in measuring stuttering in a familiar language versus a foreign language. However, Hoffman et al. also suggested that those who were deemed reliable in the use of the scale in a familiar language were more likely to perform better in an unfamiliar language; but for the present study, this pattern was not repeated, with Vietnamese_all and Vietnamese_ENGR judges’ performance being virtually identical. Therefore, Hoffman et al.’s suggestion of speech-language pathologists’ skill development playing a role in reliably measuring severity of stuttering in an unfamiliar language may not be true or may not be true across all languages. The results of both studies suggest that speech-language pathologists should not assume that, if they can produce reliable measures of stuttering in a familiar language, that they will be reliable in measuring stuttering in an unfamiliar language.

Results identified that judges were able to produce more reliable data for intrajudge agreement than for interjudge agreement when measuring severity of
stuttering in a familiar (English) and foreign language (Vietnamese). However, results also highlighted that not all judges were able to achieve criterion intrajudge agreement. Interestingly, results showed that the agreement of many judges was approaching the accepted 80% criterion level, with 90% of judges achieving Time 1 to Time 2 differences (intrajudge agreement) within plus or minus one scale value on 70% of English samples. The corresponding figure for English samples for interjudge agreement, with judges achieving scores within plus or minus 1 scale value for 70% of samples, was around 85% at Time 1 and 90% at Time 2. Judges’ performance in Vietnamese also highlighted that many judges were approaching the reliability criterion of within plus or minus one scale value on 80% of samples of intra- and interjudge agreement. These results have replicated the findings of Hoffman et al. (2014) who also found that many of the judges who did not achieve criterion performance were approaching it. This result is promising, as it seems reasonable to assume that both intra- and interjudge reliability could be readily improved to a criterion level, perhaps via practice or training.

Results identified language familiarity and rating type (intrajudge versus interjudge) played a role in judges making reliable measurements of stuttering severity using the 9-point severity rating scale. However, results showed that the speakers’ severity of stuttering could play an even more important role in reliably measuring severity of stuttering. The present study used logistic regression analyses to examine judges’ performance for each severity rating scale point and it was this analysis that showed judges were able to make reliable use of the scale to measure severity of stuttering, in a familiar and unfamiliar language, for mild and severe stuttering. This is somewhat consistent with Cosyns et al. (2015) who, using a visual analogue scale, reported that both experienced and inexperienced speech-language pathologists could accurately measure severity of stuttering when the speaker is severely stuttering in a
native and foreign language. The present study showed additional findings, highlighting that judges were also more reliable in their use of the scale in the mild range of the scale.

Examining judges’ performance at each severity rating scale point is an important addition to research into the reliable use of the severity rating scale. To the author’s knowledge, Study 1 (see Chapter 3 examining speech sample duration) and the present study are the first to include this analysis. Previous research, including the results of the present study, have estimated judges overall group performance of the severity rating scale points combined. Hoffman et al. (2014) and the present study (see Table 2) estimated how many judges achieved criterion intrajudge and interjudge agreement for a familiar and unfamiliar language. Both studies showed that a group of judges did not achieve criterion reliability (80%) for any rating type or language, although a subset of individual judges did. The inclusion of the logistic regression analysis to examine judges’ performance for each severity rating point showed that previous analyses may have masked some important variation across the scale. In other words, judges’ reliable use in the scale for mild and severe stuttering has served to offset judges’ poorer performance in the moderate range in the scale, resulting in an averaging effect. Thus, when estimating reliability of severity of stuttering in future research, the inclusion of the analysis of performance across the severity rating scale is recommended to prevent the masking and/or averaging effects.

Despite speech samples representing a similar range in stuttering severity, the gender and age of the Vietnamese-speakers was limited in comparison to the English-speakers. However, it is unknown what effect gender and age have on the measurement of severity of stuttering. In addition, the judges self-selected to participate in the study. Therefore, the participant group may reflect those who have a specific interest in the
field of stuttering. This may have led to a biased, non-representative sample of speech-language pathologists. Furthermore, not all judges were monolingual English-speakers, as 26 judges reported speaking a language and/or languages other than English. Although judges’ reported being unfamiliar with Vietnamese, they may not be unfamiliar with rating stuttering samples in other languages. The impact of the inclusion of bilingual participants in the study is unknown.

**Conclusion**

The present study expanded on the existing knowledge of measuring severity of stuttering in an unfamiliar language, with the first inclusion an Austro-Asiatic language, Vietnamese. Despite languages used in the present study being remote from each other in terms of linguistic characteristics and language families (Indo-European and Austro-Asiatic), results concurred with those of previous studies, which examined languages within the Indo-European family: Specifically, language familiarity plays a pivotal role with speech-language pathologists’ abilities to reliably measure severity of stuttering. Results indicated that judges performed better measuring severity of stuttering using the 9-point severity rating scale in a familiar language (English) than an unfamiliar language (Vietnamese). Furthermore, judges produced greater intrajudge agreement in comparison to rating against other judges (interjudge), for both English and Vietnamese.

Whilst judges showed better use of the scale for mild and severe stuttering across both languages, the moderate range of the scale was problematic for judges, irrespective of language, thus highlighting the need for the development of practice and/or training packages across languages, with a focus on training exemplars in the middle range of the severity rating scale. This notion is consistent with recommendations of Hoffman et al. (2014) and Cosyns et al. (2015) into practicing reliable use of the scale.
CHAPTER 5

Study 3: An Exploration of Reliability of Stuttering Measurement in Vietnam
The results of Study 2 have highlighted that it is possible for English-speaking speech-language pathologists to measure severity of stuttering using a 9-point severity rating scale for Vietnamese-speakers who stutter. Whilst this measurement tool is being used worldwide, it would be beneficial to determine the clinical utility of the tool for Vietnamese-speaking speech-language pathologists who speak and practice in Vietnamese. Such research could identify whether the severity rating scale would be an appropriate clinical tool to use for Vietnamese-speaking speech-language pathologists when working with clients who stutter in Vietnam.

This chapter contains an overview of the newly emerging speech-language pathology profession in Vietnam. The chapter discusses what is known about stuttering in Vietnamese and provides an overview of the English and Vietnamese languages. Here, the third study in the program of research is presented, providing the first stuttering measurement data for Vietnamese-speaking speech-language pathologists in Vietnam. The chapter concludes with a discussion of the research findings.

**Speech-language Pathology in Vietnam**

Irrespective of cultural or linguistic background, the ability to communicate effectively is a basic human right for people of all ages (Hoffman, Wilson, Hewat, & Huynh, 2017). Yet in Vietnam, with a population of over 91 million people (General Statistics Office of Vietnam, 2015), it is estimated that over 13 million people have a communication or swallowing disorder (Trinh Foundation Australia, 2015). Prior to September 2010, there were no formal training programs in Vietnam for speech-language pathologists. This suggests that many individuals may not have been able to access appropriate speech therapy services, and therefore may have tried to access

---

4 Data and content from Study 3 have been published as a journal article in *Speech, Language and Hearing* in 2017, by the candidate as first author, and co-authored with Dr Linda Wilson, Dr Sally Hewat, and Dr Thao Bich Huynh. The candidate was the chief investigator in the research described. The author has obtained approval from the Editor of the journal to present the data in the current thesis.
services from abroad. Since the early 1990s, Vietnam relied heavily on international volunteer speech-language pathologists to assist the local health professionals to provide services, and to provide training workshops to the local hospitals, universities and individuals in Vietnam (Atherton, Davidson, & McAllister, 2016; Cheng, 2010; Ducote, 2001; McAllister et al., 2010). Cheng (2010) reported that, since 2002, health professionals in Vietnam had been offered workshops in speech pathology services at hospitals or public health centres, and these health professionals were working with the title of speech therapist. The expertise of the health professionals practicing as speech-language pathologists, as reported by Cheng, is unknown and it is quite possible that they may continue to rely on further training from specialist speech-language pathologist volunteers to provide appropriate services. A country that is dependant on international volunteers is posed with a challenge: that of sustainability (McAllister et al., 2010). Yet within this context, awareness of the benefits of speech therapy and the demand for services have increased over recent years. This placed pressure on the profession in Vietnam for developing and sustaining a service (McAllister et al., 2010).

Vietnam’s first postgraduate speech therapy training program commenced in September, 2010 at Pham Ngoc Thach University of Medicine, Ho Chi Minh City with guidance and support from Trinh Foundation Australia and Australian speech-language pathologist volunteers (Atherton, Dung, & Nhan, 2013). By December 2015, two cohorts had graduated, with 33 qualified Vietnamese speech therapists working in the country, serving a total population of over 91 million people.

**Vietnamese Stuttering Data**

At least 1% of the world’s population live with a communication disorder known as stuttering (Bloodstein & Bernstein Ratner, 2008). Thus, it is likely that 1% (i.e., 910,000 people) of the Vietnamese population will stutter. Over the past few years
qualified speech-language pathologists in Vietnam have been working with clients who stutter, of all ages, and one report demonstrated reduction in stuttering severity for three Vietnamese adults who stutter following treatment based on the Camperdown Program (Hewat, Huynh, Le, & Atherton, 2013). However, when working with clients who stutter, reliably measuring stuttering is an essential component of assessment and treatment, to enable tracking of clients’ progress and outcomes. Yet to date there has been no investigation of the reliability of measurement of stuttering in Vietnamese.

**Overview of the English and Vietnamese Languages**

Currently, there appears to be no research on how stuttering presents in Austro-Asiatic languages (e.g., Vietnamese), and little as to whether stuttering in Vietnamese responds to treatment in the same way as in English. It has been suggested that phonological encoding systems differ between languages (Kureta, Fushimi, & Tatsumi, 2006). Therefore, the question arises as to whether the occurrence of stuttering in English is the same as that of Vietnamese. It is possible that the manifestation of stuttering in Vietnamese may not mimic that observed in English. This is because the Vietnamese language differs from English in almost all facets of linguistic structure. English and Vietnamese do not share similar characteristics as they derive from two completely different language families: Indo-European and Austro-Asiatic (Lewis et al., 2016).

The fundamental difference between English and Vietnamese is that, unlike the phonotactic structure for English, the Vietnamese syllable structure must be associated with the Vietnamese tonal system in order for any syllable to become lexically meaningful. In other words, the pitch of a spoken word is essential to its meaning. Tone languages utilise lexical tones to distinguish minimal individual words that are not differentiated by segmental information (Baudion-Chail, 1986). Lexical tones are
defined as contrastive variations in fundamental frequency or pitch at the syllable level (Yiu & Fok, 1995). It is the fundamental frequency pattern of the vocalic section of a monosyllable that conveys lexical meaning (Feng et al., 2012). Furthermore, the fundamental frequency contours of speech concurrently manifest sentence intonation and syllabic tones (Ni et al., 2006). In Vietnamese, every syllable will carry a tone which is found to be independent of the initial consonant phoneme and produced over half of the syllable length, predominately over the vowel (Tang & Barlow, 2006). There are six lexical tones of the Vietnamese language: (a) high rising, (b) emphasised high-rising, (c) mid-level, (d) mid-rising, (e) low-falling, and (f) low-rising (Hwa-Froelich, 2007).

In English, differences in pitch are not tied to the lexicon in the same way as in Vietnamese. English is a stress-timed language, whereby the stress patterns are commonly used to influence the rhythm and timing of speech (Ladefoged, 1993). Syllable stress is a suprasegmental feature of an utterance and refers to the emphasis given to particular syllables in a word. The production of syllable stress may vary according to the length and context of the utterance produced. In contrast to Vietnamese, which has a tone associated to each word, English has intonation patterns that are associated with sentences and phrases to convey syntactic information (Hwa-Froelich, Hodson, & Edwards, 2002). Syntactic information is the only linguistic information conveyed by pitch in English (Ladefoged, 1993). English intonation patterns may act as a range of functions which may be linguistic, such as having the ability to change the meaning from a directive, to a question or a statement of fact; or non-linguistic, and therefore may be used to convey emotions.

As there are distinct differences in language structure and linguistic processing between English and Vietnamese, it is conceivable that the presentation of stuttering
across English and Vietnamese may differ. As the phonotactic system of Vietnamese is less complex, with fewer syllables than English (Hwa-Froelich et al., 2002), it may be posited that stuttering would present less often among Vietnamese-speakers. Alternatively, both English and Vietnamese-speakers need to process lexical, syntactic, phonological, and prosodic structures in their speech planning process. However, Vietnamese-speakers need to additionally process and incorporate lexical tones (Hwa-Froelich et al., 2002). This addition could add to the complexity of linguistic and motor planning and execution, thus potentially leading to an increase in incidence of stuttering among Vietnamese-speakers in comparison to English-speakers. It is important to highlight that these notions remain unclear and are purely speculation at this point. What is certain, however, is that English and Vietnamese do not share similar language characteristics. Therefore, it is important to investigate the reliability of measurement of severity of stuttering in Vietnamese, to see whether results concur with studies conducted in English.

**Aim of the Present Study**

To date, there is no research related to measuring stuttering in Vietnamese. The reliability of the stuttering severity rating scale has been researched in different languages, yet it is unknown whether this tool can be used by a new profession of novice speech-language pathologists, and whether the scale can be used reliably as a clinical tool in Vietnamese. Therefore, the primary aim of this study was to explore the following research question:

Can a 9-point severity rating scale be used to reliably measure severity of stuttering in Vietnamese by Vietnamese speech-language pathologists?

As noted previously, approximately 50% of the world’s population is bilingual (Ardila et al., 2011), and in Vietnam, Vietnamese is the official language, with English
increasingly favoured as a second language (Central Intelligence Agency, 2013). Accordingly, it is likely that Vietnamese speech-language pathologists will work with clients who also speak English. It is essential that Vietnamese speech-language pathologists are able to reliably measure stuttering, and monitor progress and outcomes, in all the languages spoken by the client. For that reason, research into the reliability of stuttering measurement in both Vietnamese and English by Vietnamese-speaking speech-language pathologists would greatly assist clinicians and researchers to devise appropriate guidelines for reliably measuring the severity of stuttering in Vietnamese and English with Vietnamese clients. Thus, the study investigated a subsidiary question:

Can a 9-point severity rating scale be used by Vietnamese speech-language pathologists to reliably measure severity of stuttering in English?

Method

Speech Samples

The study used the same 20 audio speech samples of adults who stutter (10 Vietnamese and 10 English) as Study 2 (Chapter 4). Scale scores for the adult-speakers represented a range of severity: The Vietnamese samples ranged from 1 to 8, and the English samples ranged from 2 to 9. Chapter 4 (page 75 and Table 4.1) provides a detailed description of the adults depicted in the speech samples, the speech selection process using a bilingual panel, and the severity rating scores assigned for each sample used in the present study.

Judges

The judges were 25 speech-language pathologists from Vietnam, aged between 28 and 55 years (mean = 38 years). The inclusion criteria for the judges were as
follows: (i) qualified practicing speech-language pathologist; (ii) who had worked with a stuttering client within the past 12 months; (iii) was fluent in the Vietnamese language; and (iv) practiced speech therapy in Vietnamese. At the time of this study, only two cohorts of speech-language pathologists (n = 33) had completed a 2-year postgraduate training program at Pham Ngoc Thach University of Medicine, Ho Chi Minh City (18 in 2012, and 15 in 2014; details of the course have been reported by Atherton et al., 2013). Therefore, just over 75% of the trained speech-language pathologists in Vietnam participated. Four of the judges were male and 21 were female.

The judges’ clinical experience ranged from a minimum of 7 months up to approximately 2.5 years. All judges were native Vietnamese-speakers. Sixteen judges reported varied proficiency in a language other than Vietnamese (i.e., 16 reported varied proficiency in English, and 3 judges reported varied proficiency in French in addition to English). Therefore, for the present study, Vietnamese is referred to as the judges’ familiar language, and English being an ‘other’ language. Out of the 25 judges, 17 judges reported using the severity rating scale as part of their work with clients who stutter: 14 reported using the scale routinely or with most stuttering clients, two reported using the scale as an assessment and outcome measure, and one judge reported using the scale intermittently. The judges also rated their own experience using a severity rating scale from no experience, minimal experience, somewhat experienced, very experienced, to highly experienced. All judges rated themselves towards the inexperienced end of the scale. These two pieces of data seem incompatible but can be readily explained by noting that the judges were novice speech-language pathologists. Hence, although most judges reported using the scale, they may not yet have had sufficient experience working with stuttering clients to perceive themselves to be experienced in its use.
**Stimulus Package**

The Stimulus Packages presented to the judges were identical to those of Study 2 (English-speakers). However, judges were Vietnamese-speaking participants. Thus all written documentation was appropriately translated into Vietnamese. The translation process was: (i) the English versions written by the author were translated into Vietnamese by a Vietnamese translator; (ii) the translated Vietnamese versions were then translated back into English by a second translator; and (iii) the author checked that there were no discrepancies between English versions. No discrepancies were identified.

As for Study 2, the judges were mailed a Stimulus Package to complete on two separate occasions. The first rating session Stimulus Package included: an instruction sheet (Appendix N); a compact disk (CD) of speech samples; severity rating score sheets (Appendix O); and a questionnaire (Appendix P). The second rating session Stimulus Package included: an information sheet (Appendix Q) explaining the purpose of completing a second rating session; a consent form (Appendix R); an instruction sheet (Appendix S); a CD of speech samples; and severity rating score sheets (Appendix O).

**Procedure**

The procedure used in the present study (Study 3, Vietnamese-speakers) was identical to that of Study 2 (English-speakers). Information sheets (Appendix T) explaining the research and consent form (Appendix U) were distributed to potential judges. Once judges signed and returned the consent form, the Stimulus Package was mailed to them for completion. Chapter 4 (page 79) provides a detailed description of
the procedure. For the present study, of the 25 judges, 17 completed both rating sessions, and eight completed only the first rating session.

**Data Analysis**

As for Study 2 (see Chapter 4), reliability was estimated through evaluation of agreement. Both descriptive and inferential statistics were used to determine both intrajudge and interjudge agreement.

*Intrajudge Agreement*

Identically to Study 2 (see Chapter 4), intrajudge agreement was estimated by determining the percentage of individual judges who met criterion performance. For the purpose of this evaluation, an individual judge was deemed to have achieved criterion performance if their severity rating scores fell within ±1 scale value on *Time 1* to *Time 2* comparison, on 80% of samples. Like Study 2, the percentage of judges’ severity rating scores that fell within ±1 scale value for 70% or more of samples was also calculated, to enable some commentary about the degree of improvement required by the group to achieve criterion performance for the current study.

*Interjudge Agreement*

Interjudge agreement was estimated by examining the performance of individual judges, to determine the percentage of judges who achieved predetermined criterion performance. As for Study 2 (see Chapter 4), criterion performance was defined as having assigned a score within ±1 scale value of the modal score for the sample, for 80% of samples. Equivalent to Study 2, the percentage of judges who assigned a score within ±1 scale value of the modal score for 70% or more of samples was calculated to provide further commentary of judges’ performance.
Performance in English for the Subgroup of Judges Reliable in Vietnamese

A subset of judges (n = 5) achieved criterion intra- and interjudge agreement using the severity rating scale for the Vietnamese speech samples. Therefore, these judges were deemed to have made reliable use of the scale for Vietnamese samples. Data describing their performance in English were reported separately (“English\textsubscript{VN}”) for each of the analyses described above.

Logistic Regression

The Generalised Linear Mixed Model (GLMM) modelling process was used (i) to see whether percentage of agreement within ±1 scale value varies over the severity range; (ii) to see whether language (Vietnamese versus English) impacted on judges’ performance; (iii) to quantify the difference between intrajudge compared to interjudge agreement; and (iv) to see whether any of these 3 effects depended on the rating type (intrajudge, interjudge) or language (Vietnamese, English). Proc GLIMMIX from the Statistical Analysis System (SAS) package, version 9.4 was used for the GLMM modelling.

As for Study 2 (Chapter 4), the GLMM approach was used to fit to a logistic regression fixed effects model, between the outcome agreement within ±1 scale value and the fixed effect explanatory variables: mean severity rating for each sample, rating type, and language. The agreement data were entered into the model in contingency table format as (number of agreements within ±1 scale value, number of pairs used to assess agreement) for each sample. Simulation work showed, however, that for interjudge agreement, where the number of pairs was much larger than the number of judges, there would have been a substantial overestimate of the precision of the agreement estimates. Therefore, sample weights were applied to correct for this, the
weights being 1 for intrajudge agreement and number of judges divided by number of judge pairs for interjudge agreement.

For the present study, models examined the curvature of the severity rating effect with polynomials up to degree 4. Two way interactions examined were between severity rating (up to degree 2) and both rating type and languages, and also between rating type and languages. A random intercept was used in the model to account for unexplained variability between samples. Possible differences in variability between rating types (intrajudge, interjudge), or rating sessions (Time 1, Time 2) were assessed by adding grouping variables to the random effect for these variables. The random effects structure was assessed using either Akaike Information Criterion comparisons between models or a homogeneity test for differences in variances between groups, as appropriate. The final GLMM model had one random effect with intercepts based on sample code. Model fit was assessed by plotting Pearson residuals against severity rating by categorical groups, and from an overall goodness of fit index by the Generalised chi-square/df ratio, which was 1.23. The model fit was considered satisfactory. Effect sizes were assessed with odds ratios (OR) and 95% confidence intervals, and statistical significance was set at \( \alpha = .05 \).

**Results**

**Descriptive Statistics**

The mean, mode, range, and standard deviation of severity rating scale scores assigned by the judges to each speech sample at the first rating occasion are presented in Table 5.1. Table 5.1 shows that the speech samples for both languages received a range of severity ratings from the judges. The modal stuttering severity rating scores assigned to the Vietnamese samples ranged from 2 to 7, while the English samples ranged from 2
to 9. These results are somewhat similar to the range obtained for the speech samples by the bilingual panel (see Table 4.1).

Table 5.1

Mean, mode, range, and standard deviation of severity rating scale scores assigned by the judges to each speech sample. Data are presented for the first rating occasion only.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Vietnamese</th>
<th></th>
<th></th>
<th></th>
<th>English</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mode</td>
<td>Range</td>
<td>SD</td>
<td>Mean</td>
<td>Mode</td>
<td>Range</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>4.08</td>
<td>4</td>
<td>2-6</td>
<td>1.29</td>
<td>3.08</td>
<td>2</td>
<td>1-6</td>
<td>1.26</td>
</tr>
<tr>
<td>2</td>
<td>7.36</td>
<td>7</td>
<td>4-9</td>
<td>1.25</td>
<td>3.16</td>
<td>2</td>
<td>1-7</td>
<td>1.49</td>
</tr>
<tr>
<td>3</td>
<td>1.88</td>
<td>2</td>
<td>1-3</td>
<td>0.73</td>
<td>5.36</td>
<td>5</td>
<td>3-9</td>
<td>1.78</td>
</tr>
<tr>
<td>4</td>
<td>7.84</td>
<td>7</td>
<td>6-9</td>
<td>0.90</td>
<td>5.80</td>
<td>7</td>
<td>2-9</td>
<td>2.12</td>
</tr>
<tr>
<td>5</td>
<td>3.40</td>
<td>2</td>
<td>1-7</td>
<td>1.68</td>
<td>4.28</td>
<td>3</td>
<td>2-8</td>
<td>1.34</td>
</tr>
<tr>
<td>6</td>
<td>2.32</td>
<td>2</td>
<td>1-7</td>
<td>1.22</td>
<td>8.32</td>
<td>9</td>
<td>2-9</td>
<td>1.57</td>
</tr>
<tr>
<td>7</td>
<td>7.00</td>
<td>7</td>
<td>2-9</td>
<td>1.58</td>
<td>3.60</td>
<td>2</td>
<td>2-6</td>
<td>1.35</td>
</tr>
<tr>
<td>8</td>
<td>4.24</td>
<td>3</td>
<td>2-8</td>
<td>1.71</td>
<td>8.48</td>
<td>9</td>
<td>5-9</td>
<td>1.12</td>
</tr>
<tr>
<td>9</td>
<td>3.84</td>
<td>3</td>
<td>2-7</td>
<td>1.25</td>
<td>2.24</td>
<td>2</td>
<td>1-6</td>
<td>1.23</td>
</tr>
<tr>
<td>10</td>
<td>3.44</td>
<td>2</td>
<td>2-6</td>
<td>1.19</td>
<td>4.79</td>
<td>5</td>
<td>2-8</td>
<td>1.56</td>
</tr>
</tbody>
</table>

Intrajudge Agreement

Table 5.2 depicts the percentage of judges who achieved criterion level performance for each of intra- and interjudge agreement, for each language. The table shows that only 52.94% of individual judges achieved the criterion level of intrajudge agreement for Vietnamese. What was disappointing to see was the poorer performance for English samples, with 29.41% of individual judges achieving the criterion level of agreement. The figure for English for the VNrn subgroup was higher at 40%. However, this result was still less than half of the judges. The large majority of judges were able to achieve scores within ±1 scale value for 70% or more of samples for Vietnamese (76.47%), showing that many judges were approaching criterion agreement.
Performance for English\textsubscript{all} (47.06%), and English\textsubscript{VNr} (60.00%) were increasing, but were still far from approaching acceptable levels of agreement.

Table 5.2

\textit{Percentage of judges who achieved criterion intra- and interjudge reliability (≥80% agreement) and ≥70% agreement}

<table>
<thead>
<tr>
<th>Judge Condition</th>
<th>Intrajudge</th>
<th>Interjudge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≥ 80%</td>
<td>≥ 70%</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>52.94%</td>
<td>76.47%</td>
</tr>
<tr>
<td>English\textsubscript{all}</td>
<td>29.41%</td>
<td>47.06%</td>
</tr>
<tr>
<td>English\textsubscript{VNr}\textsuperscript{a}</td>
<td>40.00%</td>
<td>60.00%</td>
</tr>
</tbody>
</table>

Note. Intrajudge \(N = 17\) for Vietnamese and English\textsubscript{all}; \(N = 5\) for English\textsubscript{VNr}; and Interjudge \(N = 25\) Vietnamese and English\textsubscript{all} and \(N = 5\) English\textsubscript{VNr}.

\textsuperscript{a} English\textsubscript{VNr} presents scores for English for the subgroup of judges who achieved both intrajudge and interjudge agreement in Vietnamese.

\textit{Interjudge Agreement}

Table 5.2 also shows the percentage of judges who achieved criterion level performance for interjudge agreement, for each language. Criterion level interjudge agreement was achieved by 44% of individual judges for Vietnamese samples at \textit{Time 1} and by 36% for English\textsubscript{all}. A considerably higher percentage (80%) of the VNr subgroup achieved criterion level performance at \textit{Time 1}. Seven of 25 (28%) judges did not achieve criterion performance in either Vietnamese or English. All of the other judges achieved criterion performance in at least one of the languages. The percentages of judges who achieved scores within ±1 of the modal score on at least 70% of samples were 60% for Vietnamese, 56% for English\textsubscript{all}, and 100% for English\textsubscript{VNr}. 

107
The results for Time 2 were considerably higher for Vietnamese, with 82.35% of judges achieving acceptable criterion performance. As results from Time 1 to Time 2 had substantially improved, it was considered whether the 8 judges who withdrew before Time 2 had lower agreement, so that their withdrawal improved Time 2 results. When looking at the data for the judges who completed Time 1 and Time 2 using the descriptive statistics described above, it was noted that the difference for Vietnamese still existed and was similar. Therefore, no further analyses of data for this subset of judges were warranted. Judges’ performance in English, all at Time 2 (35.29%) was similar to Time 1 (36%). Three of 17 (17.6%) judges did not achieve criterion performance in either Vietnamese or English. All of the other judges achieved criterion performance in at least one of the languages. A greater percentage of judges achieved scores within ±1 of the modal score on at least 70% of samples for Time 2.

**Logistic Regression**

The final GLMM model showed four significant effects: two severity rating terms and main effects for rating type (intrajudge, interjudge) and language (Vietnamese, English). Neither of the effects varied with severity rating, as their interactions with the two severity rating terms were not significant and neither was the interaction between them. Therefore, all model effects were independent and did not depend on the levels of other variables. The agreement percentage levels from the fitted model at all levels of severity rating for the rating type (intrajudge, interjudge) and language variables are shown in Figure 5.1.
Figure 5.1. Plot showing the relationship between the three significant variables from the final model. The agreement percentages for each of the sample/time(1,2)/agreement type combinations are also plotted (+ intrajudge, o interjudge). The two panels are for the language categories (Vietnamese, English).

On examination of the percentage of agreement that varied over the severity range, the final GLMM model had two significant severity rating terms, one linear and one quadratic ($B(\text{SE}) = -1.01 (0.22)$ and $0.099 (0.021)$, both $p < 0.0001$). The two higher order polynomial terms tested were not significant. The fitted curved relationship had minimum agreement close to a severity rating of 5, and increased towards the edges of the scale (see Figure 5.1). This supports the observation of Study 1 (see Chapter 3 which examined speech sample duration) and Study 2 (English-speakers, see Chapter 4) where the standard deviation of severity ratings was found to be higher in the middle of the severity range and lower at the edges. It is notable that, in all studies, the fitted quadratic model was centred quite close to a severity rating of 5, the middle of the rating scale. The degree of curvature of the agreement versus severity rating effect was found to be constant across rating types (intrajudge, interjudge), and familiarity of the
judge with the language, as none of the interactions between these variables and the linear and quadratic terms in the models were significant.

Logistic regression was used to determine if there was any relationship between languages (Vietnamese and English), and determine whether there was any difference between intrajudge compared to interjudge performance. Results are presented in Table 5.3. For this analysis, one group (e.g., intrajudge) is used as a reference and given a value of 1.00. If the odds ratio for the other group (e.g., interjudge) is also 1.00, then there is no difference between the two groups.

<table>
<thead>
<tr>
<th>Variable value</th>
<th>Odds ratio (OR)</th>
<th>95% confidence interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>1.00*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>1.87</td>
<td>1.29, 2.69</td>
<td>0.001</td>
</tr>
<tr>
<td>Rating Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interjudge</td>
<td>1.00*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrajudge</td>
<td>1.45</td>
<td>1.08, 1.94</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*reference level

Judges had higher agreement levels in Vietnamese than in English, with an odds ratio of 1.87. This shows that judges are almost two times more likely to reliably rate severity of stuttering using the severity rating scale in Vietnamese than in English. The confidence interval suggests that the difference could be as little as 1.29 times more likely, or as much as 2.69 times more likely. There is a statistically significant (p = 0.001) advantage to judges being able to rate more reliably in Vietnamese than in English. To further understand odds ratios, which are similar to comparing ratios of percentages, Figure 5.1 enables one to visually compare the percentage agreement. At a severity rating of 5 for the two interjudge rating curves (lower curves), judges reach
about 65% agreement for Vietnamese, and only 50% agreement for English. This suggests that language familiarity impacts on judges’ ability to reliably measure severity of stuttering using the severity rating scale.

The rating type variable was significant \( (p = 0.01) \) in the model, with agreement for intrajudge agreement being higher than for interjudge agreement, with an odds ratio of 1.45. This suggests that judges are 1.45 times more likely to rate severity of stuttering reliably with themselves than with other judges. This result confirms the descriptive statistics showing that judges performed better with intrajudge agreement than interjudge. The confidence interval estimate in Table 5.3 suggests that odds for measuring stuttering reliably against yourself (intrajudge), could be as little as 1.08 or as much as 1.94 times as great as the odds for measuring stuttering reliably with others (interjudge).

**Discussion**

The present study has provided the first Vietnamese stuttering measurement reliability data for the newly emerging speech-language pathology profession in Vietnam. The aim of the study was to estimate the reliability of Vietnamese speech-language pathologists’ use of a 9-point severity rating scale to measure severity of stuttering in a familiar language (Vietnamese) and in English, which is commonly favoured as a second language in Vietnam. For this study, evaluation of agreement was used to estimate reliability.

Just over half of individual judges (approximately 53%) were able to produce reliable data and achieve criterion intrajudge agreement for the familiar language (Vietnamese). This result is almost identical to the findings of Hoffman et al. (2014), who reported 50% of judges achieving criterion intrajudge agreement in a familiar
language (Australian English). These results highlight that judges should not assume they will be reliable in a familiar language. It is recommended that judges take steps to check reliability with themselves on a regular basis to ensure the data they are collecting for their clients over time are reliable.

Results also showed that judges performed better for intrajudge agreement than for interjudge agreement when measuring severity of stuttering in Vietnamese (familiar language). For interjudge agreement, judges performed poorly at *Time 1*, with only 44% of judges reaching criterion. However, judges achieved criterion interjudge agreement at *Time 2* (approximately 82%), indicating reliable use of the scale at *Time 2*. It is possible that this improvement from *Time 1* to *Time 2* could be attributable to increased exposure and a practice effect. This highlights the important role that practice and training potentially has to skill development for a newly emerging profession of speech-language pathologists in Vietnam.

Results of the present study indicate that judges also performed better at measuring severity of stuttering in a familiar language (Vietnamese) than in English (other language). For intrajudge agreement by individuals: Of the subgroup of judges who were reliable in Vietnamese (VNr), only 40% achieved criterion intrajudge agreement for the English speech samples, compared to approximately 53% for Vietnamese. Although this percentage for English is poor, it is nevertheless greater than the percentage for the entire group (English_all), which was approximately 30% criterion intrajudge agreement. These results suggest that intrajudge agreement might be affected by both language and skill in the use of the severity rating scale. This statement is furthermore supported by data for interjudge agreement for the English speech samples: Of the VNr subgroup, 80% achieved criterion interjudge agreement for the English speech samples at *Time 1*. Thus, most of those judges who were reliable using the scale
in a familiar language (Vietnamese) have also shown reliable use of the scale to measure severity of stuttering in English. This percentage is considerably greater than the percentage of the entire group (36%) who achieved criterion interjudge agreement for English speech samples. Again, it appears that those who have demonstrated criterion performance for use of the severity rating scale in Vietnamese are more likely than others to demonstrate criterion performance in English, further highlighting skill development to be a potentially important factor when using the scale to measure severity of stuttering in a second or other language.

It should be noted, however, that 20% of the VNr subgroup did not achieve criterion interjudge reliability in English. One possible interpretation of this is that language proficiency in English is impacting on interjudge agreement for the English speech samples. However, Bosshardt et al. (2015) examined the contribution of judges’ language proficiency to severity rating measurement, using a 10-point stuttering severity rating scale and speech samples from children in various Indo-European languages, and found that language proficiency was not systematically related to the variability and agreement of severity ratings. Furthermore, they found that language proficiency contributed only a maximum of 4.6% to the observed variance. The current study did not attempt to quantify judges’ degree of language proficiency in English, and thus further commentary regarding the contribution of language proficiency to the results reported here cannot be made.

Judges’ performance in the use of the scale varied depending on the severity rating range. Judges showed better performance in the use of the scale for the upper and lower end scale values (mild and severe stuttering). These results are consistent with those of Cosyns et al. (2015) who, using a visual analogue scale, reported that both experienced and inexperienced speech therapists could accurately measure stuttering
severity when the speaker is severely stuttering. In the present study, for the middle range of the scale (moderate stuttering), judges produced greater variability in Vietnamese and English. This finding has clinical implications, as greater variability translates to poorer reliability. Therefore, it is important for speech-language pathologists to be aware that they might produce poorer results when making measurements of stuttering in the moderate range of the severity rating scale. Thus, it is recommended that speech-language pathologists continue to take steps to check their reliability, particularly in the moderate range of the scale. If speech-language pathologists are not able to achieve acceptable levels of reliability through self-practice, further training may be advised.

Research could assist with the development of procedures and materials to help speech-language pathologists in establishing reliability. Previous studies (e.g., Hoffman et al. 2014; and Van Borsel & Britto Pereira, 2005; and Study 2 from the present thesis) have highlighted the need and/or recommended the development of stuttering exemplars across different languages. Findings of this present study further support the need for such packages to be developed and evaluated. Furthermore, it would be beneficial if future practice and/or training speech sample exemplars focused on the moderate range, for all languages, in an attempt to reduce observed variability across the middle range of the scale.

Speech-language pathology is a developing profession in Vietnam. Therefore, the sample size for Study 3 was small and included only novice judges. Although the speech samples represented a range in severity, the age and gender of the Vietnamese-speakers was limited in comparison to the English speech samples (see Table 4.1). Furthermore, Study 2 (English-speakers, see Chapter 4) investigated judges who were familiar (English) and unfamiliar (Vietnamese) in a language. Any judges who had
some familiarity with (could speak or understand) Vietnamese were excluded from that study. Unfortunately, there was no such clear distinction with judges in the present study (Study 3, Vietnamese-speakers), with Vietnamese as the familiar language, and English as an other language (a common second language). As the profession is in its infancy, the study recruited 25 out of a possible 33 speech language pathologists from Vietnam. Thus, the study incorporated a vast majority of the population of speech-language pathologists in Vietnam. As many identified speaking and/or understanding English, it is unknown what impact their proficiency in English had on the results. However, Bosshardt et al.’s (2015) results indicated that language proficiency (across Indo-European languages) had little impact on measuring severity of stuttering for preschool aged children, contributing a maximum of 4.6% of the variance. Thus, it is likely that judges’ language proficiency in English may not have been a significant variable. Nonetheless, further research into the impact of language proficiency on measuring severity of stuttering for adult-speakers is needed, across different languages (e.g., Vietnamese, Mandarin) and language families (e.g., Austro-Asiatic, Sino-Tibetan).

**Conclusion**

The main purpose of this study was to estimate Vietnamese speech-language pathologists’ reliability when using a 9-point severity rating scale to measure severity of stuttering in a familiar language (Vietnamese) and in English. Results of the present study were consistent with those of Study 2 (English-speakers; see Chapter 4) which revealed that judges performed better in the use of the severity rating scale in a familiar versus an other (present study) or unfamiliar (Study 2) language. The main encouraging finding of the present study was the considerable improvement judges made in a familiar language for interjudge agreement at Time 2, showing reliable use in the scale.
This result suggests that increased exposure and practice may lead to better performance in judges’ familiar language. The second finding that judges produced better intrajudge agreement than interjudge agreement, for both languages in the present study (Vietnamese-speakers) is also consistent with results of Study 3 (English-speakers).

Finally, irrespective of language, judges showed greater variability in their use of the scale in the moderate range of stuttering. This pattern was also evident in Study 2 (English-speakers). The consistent pattern across studies suggests the need for practice and/or training exemplars that target this range. With such practice, it seems reasonable to assume that variability in the moderate range of the scale could be reduced, and overall reliability levels improved for research and clinical practice.

Thus far, results from Study 2 (English-speakers) and the present study (Study 3, Vietnamese-speakers) have produced some consistent findings. These consistencies will be explored in greater detail in the next chapter.
CHAPTER 6

Study 4: Comparison of Reliability Across Different Languages
An integral component of research is to determine whether findings are robust. Results from the current program of research show some consistent findings. Thus, the purpose of the next phase of this research was to complete a re-analysis of the data of Hoffman et al. (2014), using the same statistical analyses as those of Study 2 (English-speakers) and Study 3 (Vietnamese-speakers), to determine whether those data would replicate the findings. This chapter presents the re-analysis of data for the purpose of making comparisons across studies. Afterwards, the main research findings from the program of research are discussed.

**Background**

The current program of research has shown consistent findings across studies. Study 2 (see Chapter 4) investigated English-speaking speech-language pathologists’ reliability of use of the severity rating scale in English (familiar language) and Vietnamese (unfamiliar language). Study 3 (see Chapter 5) examined Vietnamese-speaking speech-language pathologists’ reliability of use of the severity rating scale in Vietnamese (familiar language) and English (an other language, a common second language). The procedure and statistical analyses of Study 3 were identical to those of Study 2. The findings found across both studies were that: (i) judges showed increased reliability of use of the severity rating scale to measure severity of stuttering in a familiar versus an unfamiliar or other language; (ii) judges were more reliable in the use of the scale for intrajudge agreement in comparison to interjudge agreement, in both languages; and (iii) irrespective of language, judges produced greater variability (decreased reliability) in their use of the scale for the moderate range.

As an important component to research is reproducibility, research must be repeated before findings can be well established and accepted as worthwhile and robust. Replications in research involve applying the theory to new situations in an attempt to
determine the generalizability to different races, cultures, or languages (Roberts, 2011). Replications are therefore important for a number of reasons, including: (i) assurance that results are valid and reliable; (ii) determination of generalizability or the role of extraneous variables; (iii) application of results to real world situations; and (iv) inspiration for new research combining previous findings from related studies (https://allpsych.com/researchmethods/replication/). With replications being so valuable, it is unfortunate that, in stuttering research, few research studies are replicated. As the self-correcting nature of science can only be effective with replications, it is important that researchers and clinicians do not accept preliminary single studies at face value. Thus, researchers should not delay too long between the original study and its replication, to avoid any harming of credibility in the field if results of the replication and the original research differ (Roberts, 2011).

To determine whether the consistent findings across the current program of research are robust, the next step is to replicate the analyses of Studies 2 (English-speakers) and 3 (Vietnamese-speakers) with different participants and a different language. Studies 2 and 3 replicated the procedures of that of the Hoffman et al. (2014) study, which investigated Australian speech-language pathologists’ reliability of use of the severity rating scale in English (familiar language) and Mandarin (unfamiliar language).

Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) used identical descriptive statistics to those of Hoffman et al. (2014). However, Studies 2 and 3 included additional statistical analyses (logistic regression) which produced a new finding, not identified in Hoffman et al. Findings showed that judges produced increased reliability in the use of the scale for mild and severe stuttering across both languages, with the moderate range of the scale being more problematic (displaying
increased variability) for judges. The consistent finding of increased variability in the moderate range of the scale for Studies 2 and 3 was also shown in Study 1 (see Chapter 3) across all three speech sample durations (1-minute, 3-minutes, and 5-minutes).

It would be advantageous to run a re-analysis of the Hoffman et al. data, to determine whether the new findings of Studies 2 and 3 can be replicated using different participants and a different language (Mandarin) and language family (Sino-Tibetan). The research aims for Hoffman et al. (2014) were identical (examining whether speech-language pathologists can measure severity of stuttering reliably in a familiar and unfamiliar language) to that of Study 2, and similar to Study 3 (examining a familiar and other language). Thus, to determine the suitability of the re-analysis of Hoffman et al. and comparison to Studies 2 and 3, it is necessary to compare procedures used across studies, to ensure appropriate comparisons can be made. Thus, the following section is a summary of procedures, provided in full in Hoffman et al., to show how data were gathered using similar/identical procedures to those of Study 2 (English-speakers) and Study 3 (Vietnamese-speakers).

The Hoffman et al. (2014) study used 20 audio speech samples of adults who stutter (10 Australian English and 10 Mandarin). Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) used the same number of audio speech samples (10 English and 10 Vietnamese) of adults who stutter. Each audio speech sample was a 3-minute excerpt taken from a longer speech sample. For ease of access, the severity scale scores assigned to each speech sample are presented in Table 6.1. These scores were made by bilingual panels, comprised of four speech-language pathologists (two speech-language pathologists fluent in English and Mandarin for Hoffman et al.; and two speech-language pathologists fluent in English and Vietnamese for Studies 2 and 3). The bilingual panel procedure can be found in Chapters 4 and 5 of this thesis and in the
Hoffman et al. study. The panel scores were obtained to ensure a range of severity of stuttering in each language, but were not used any further in the research. The gender and age distribution of speakers was similar for Study 2 (English-speakers) and Hoffman et al., but narrower for Study 3 (Vietnamese-speakers). Hoffman et al. used speech samples that depicted a range of severity: The Australian English samples ranged from 2 to 9, while the Mandarin samples ranged from 1 to 7. These ranges were almost identical to the ranges in Studies 2 and 3 (English ranged from 2 to 9 and Vietnamese ranged from 1 to 8).

Judges in Hoffman et al. (2014) were 26 speech-language pathologists, aged between 23 and 55 years (mean = 33 years). This mean is somewhat comparable to those from Study 2 (mean = 39) and Study 3 (mean = 38). Two of the judges were male and 24 were female. This ratio between male and female was also similar to that in Study 2 (English-speakers) and Study 3 (Vietnamese-speakers), with a predominance of female judges compared to male judges. The judges’ clinical experience ranged from a minimum of 1 year up to approximately 31 years. This differed from judges’ clinical experience for Study 2 (English-speakers) which ranged from 6 months up to approximately 42 years. Study 2 was comprised of a considerably larger group of judges, 59 judges, compared to 26 for the Hoffman et al. study. Judges’ experience for Study 3 (Vietnamese-speakers) was considerably less diverse, ranging from 7 months to 2.5 years. This is expected due to the Vietnamese speech-language pathologists being novice clinicians making up the first two cohorts of graduates in Vietnam. The judges’ recent experience working in the area of stuttering was both diverse and identical for the Hoffman et al. study and Study 2. For both studies, the judges reported to have assessed and/or treated between approximately 1 and 500 clients who stutter in the previous 12 months.
Table 6.1
Age, Gender, and Stuttering Severity Rating Scale Scores of the Adults Depicted in the Speech Samples in the Hoffman et al. (2014) Study5 and Studies 2 and 3 of the current research

<table>
<thead>
<tr>
<th>Sample</th>
<th>Age</th>
<th>Gender</th>
<th>SR</th>
<th>Age</th>
<th>Gender</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47</td>
<td>Male</td>
<td>3</td>
<td>33</td>
<td>Male</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>Male</td>
<td>7</td>
<td>18</td>
<td>Male</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>Female</td>
<td>5</td>
<td>34</td>
<td>Female</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>Male</td>
<td>5</td>
<td>18</td>
<td>Male</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>Male</td>
<td>3</td>
<td>44</td>
<td>Male</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>Male</td>
<td>9</td>
<td>27</td>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>Female</td>
<td>4</td>
<td>28</td>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>Male</td>
<td>6</td>
<td>18</td>
<td>Male</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
<td>Male</td>
<td>4</td>
<td>47</td>
<td>Female</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>Male</td>
<td>2</td>
<td>18</td>
<td>Male</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Age</th>
<th>Gender</th>
<th>SR</th>
<th>Age</th>
<th>Gender</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47</td>
<td>Male</td>
<td>3</td>
<td>27</td>
<td>Male</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>Male</td>
<td>3</td>
<td>24</td>
<td>Male</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>Male</td>
<td>4</td>
<td>28</td>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>Male</td>
<td>6</td>
<td>28</td>
<td>Male</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>Female</td>
<td>5</td>
<td>31</td>
<td>Male</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>63</td>
<td>Male</td>
<td>7</td>
<td>28</td>
<td>Male</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>Female</td>
<td>4</td>
<td>26</td>
<td>Male</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>Male</td>
<td>9</td>
<td>25</td>
<td>Male</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>42</td>
<td>Male</td>
<td>2</td>
<td>28</td>
<td>Male</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>Male</td>
<td>5</td>
<td>26</td>
<td>Male</td>
<td>4</td>
</tr>
</tbody>
</table>

Note. SR = severity rating, determined by consensus of a bilingual panel.

5 The data for Hoffman et al. (2014) are a reprint of a table from an article published in the International Journal of Speech Language Pathology in 2014, by the candidate as first author, and co-authored with Dr Linda Wilson, Dr Anna Copley, Dr Sally Hewat, and Dr Valerie Lim. The candidate was the chief investigator in the research described.

6 Data from Study 3 are a reprint of data presented from an article published in Speech, Language and Hearing in 2017, by the candidate as first author, and co-authored with Dr Linda Wilson, Dr Sally Hewat, and Dr Thao Bich Huynh. The candidate was the chief investigator in the research described.
The procedure of Hoffman et al. (2014) was identical to that used in both Studies 2 and 3 (refer to Chapter 4 *Procedure* for a detailed description). Judges rated 20 audio speech samples (10 Australian English and 10 Mandarin) of adults who stutter using a 9-point severity rating scale on two separate occasions. The participants were not initially informed that they were to be asked to complete this second rating session.

From the comparison presented above, it is clear that the procedures for the three studies (Hoffman et al., 2014; Study 2 and Study 3) were identical and speech samples and judges somewhat comparable to each other. Hence, completing a re-analysis of the Hoffman et al. data for the purposes of making comparisons across studies is appropriate.

**Aim**

Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) used an analysis (logistic regression), which was not included in the Hoffman et al. study. Thus, the next stage in this program of research is to re-analyse the raw data of the Hoffman et al. study, using an identical analysis (logistic regression) to that used in Studies 2 and 3, to answer the following research question:

Does the data of Hoffman et al. (2014) replicate the findings of those of Study 2 and Study 3?

The following four supplementary questions were used to guide the re-analysis of data from Hoffman et al. (2014):

1. Does percentage agreement within ±1 scale value vary over the severity range?
2. Does language familiarity (familiar versus unfamiliar or other) impact on judges’ performance?
3. What is the difference between intrajudge and interjudge agreement?
4. Do any of these three effects vary according to the rating type or language familiarity (familiar versus unfamiliar or other)?

**Method**

The data analysis completed was a re-analysis of the Hoffman et al. (2014) raw data using logistic regression analyses identical to those completed for Study 2 (English-speakers; see Chapter 4) and Study 3 (Vietnamese-speakers; see Chapter 5). A brief description of the analyses completed for the Hoffman et al. data are provided below.

**Logistic Regression**

The Generalised Linear Mixed Model (GLMM) modelling process was used (i) to see whether percentage of agreement within ±1 scale value varied over the severity range; (ii) to see whether language impacted on judges’ performance; (iii) to determine the difference between intrajudge compared to interjudge agreement; and (iv) to see whether any of these 3 effects depended on the rating type (intrajudge, interjudge) or language. Proc GLIMMIX from the *Statistical Analysis System (SAS) package, version 9.4* was used for the GLMM modelling.

As for Studies 2 and 3, the GLMM approach was used to fit to a logistic regression model between the outcome agreement within ±1 scale value and the fixed explanatory variables mean severity rating for each sample, rating type (intra or inter judge) and speech sample language. The agreement data were entered into the model in contingency table format as (number of agreements within ±1 scale value, number of pairs used to assess agreement) for each sample and time combination. Simulation work showed, however, that, for interjudge agreement, where the number of pairs was much larger than the number of judges, there would have led to a substantial overestimate of
the precision of the agreement estimates. Therefore, sample weights were applied to correct for this, the weights being 1 for intrajudge agreement and number of judges divided by number of judge pairs for interjudge agreement.

For the re-analysis of the Hoffman et al. (2014) data set, models examined the curvature of the severity rating effect with polynomials up to degree 4. Two way interactions examined were between severity rating (up to degree 2) and both rating type and languages, and also between rating type and languages. The possibility of overdispersion due to differences between samples was tested by fitting a mixed model with a random intercept for sample identifier but was found to be unnecessary as the estimate was zero. Possible differences in variability between rating types, languages or sample time period were assessed by adding grouping variables to the residual error terms for these variables. None of these were found to be different. The final model was not a mixed model but a Generalised Linear Model, a standard logistic regression model with no random effects and no special covariance structures for the residuals. Model fit was assessed by plotting Pearson residuals against severity rating by categorical groups and, from an overall goodness of fit index, the Generalised chi-square/df ratio which was 0.64 was underdispersed. The reason for this underdispersion was not clear. However, the model fit was considered satisfactory. Effect sizes were assessed with odds ratios (OR) and 95% confidence intervals and statistical significance was set at \( \alpha = 0.05 \).

**Descriptive Statistics**

**Intrajudge and Interjudge Agreement**

It should be noted that intrajudge and interjudge agreement analyses have already been conducted and published in Hoffman et al. (2014) and also presented in *Chapter 4* for Study 2 and *Chapter 5* for Study 3. Whilst these analyses are not new, the
author has tabulated the data for all three studies combined and will present them together in the results section of this chapter for ease of reference. These data will then be used to help guide the reader through the discussion section of the chapter.

Identically to Studies 2 and 3, reliability for Hoffman et al. (2014) was estimated through evaluation of intrajudge and interjudge agreement. For intrajudge agreement, an individual judge was deemed to have achieved criterion performance if their severity rating score fell within ±1 scale value on Time 1 to Time 2 comparison, on 80% of samples. For interjudge agreement, achieving criterion performance was defined as judges having assigned a score within ±1 scale value of the modal score for the speech sample, for 80% of speech samples.

Results

This section provides results of the re-analysis conducted of the Hoffman et al. (2014) study, and comparison to results of Study 2 (English-speakers) and Study 3 (Vietnamese-speakers).

Logistic Regression

For the re-analysis of Hoffman et al. (2014), the final GLMM model showed four significant effects: two severity rating terms (linear and quadratic) and main effects for rating type (intrajudge, interjudge) and language (Australian English and Mandarin). These effects were identical to those found for Study 2 (English-speakers) and Study 3 (Vietnamese-speakers), but with the substitution of a different language (i.e., the main effects for language for Study 2 were English and Vietnamese, and the main effects for language for Study 3 were Vietnamese and English).
Does Percentage Agreement Vary over the Severity Range?

On examination of the percentage of agreement that varied over the severity range, the final GLMM model for the Hoffman et al (2014) data had two significant severity rating terms, one linear and one quadratic (B(SE) = -1.52 (0.21) and 0.154 (0.022), on the scale of the linear predictor, both \( p < 0.001 \)). The two higher order polynomial terms tested were not significant. The fitted curved relationship had minimum agreement close to a severity rating of 5, and increased towards the edges of the scale (see Figure 6.1a). This result is consistent with the results of Studies 2 and 3 (see Figure 6.1b and Figure 6.1c), showing a very similar effect of severity on agreement level, where the standard deviation of severity ratings was found to be higher in the middle of the severity range and lower at the edges. For the Hoffman et al. study, the degree of curvature of the agreement versus severity rating effect was found to be constant across rating types (intrajudge, interjudge), and familiarity of the judge with the language, as none of the interactions between these variables and the linear and quadratic terms in the models were significant. These results are therefore consistent with those of Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) (see Chapters 4 and 5), which identified that percentage agreement did vary over the severity range with strongest agreement at either ends of the scale and weakest agreement in the moderate range of the scale.
a. Hoffman et al. (2014) Study: *Australian-English* speaking judges rating *Australian-English* and *Mandarin*

b. Study 2: *English* speaking judges rating *English* and *Vietnamese*

c. Study 3: *Vietnamese-speaking* judges rating *Vietnamese* and *English*

Figure 6.1: Plot showing the relationship between the three significant variables from the final model. The agreement percentages for each of the sample/time(1,2)/agreement type combinations are also plotted (+ intrajudge, o interjudge). The panels are for the language categories (for Hoffman et al., 2014), *Australian English, Mandarin*; for Study 2, *English, Vietnamese*; and for Study 3, *Vietnamese, English*.
Does Language Familiarity (Familiar Versus Unfamiliar or Other) Impact on Judges’ Performance?

For the Hoffman et al. (2014) re-analysis, logistic regression was used to determine if there was any relationship between languages (Australian English and Mandarin), and judges’ performance. Results for the Hoffman et al. re-analysis and for Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) are presented in Table 6.2. For this analysis, one group is used as a reference and given a value of 1.00. If the odds ratio for the other group is also 1.00, then there is no difference between the two groups.

Table 6.2
Estimated odds ratio, 95% confidence intervals, and p-values for language, and rating type for all three studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Variable value</th>
<th>Odds ratio (OR)</th>
<th>95% confidence interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoffman et al.  (2014)</td>
<td>Language</td>
<td>Mandarin Australian English</td>
<td>1.00*</td>
<td>1.35</td>
</tr>
<tr>
<td></td>
<td>Rating Type</td>
<td>Interjudge</td>
<td>1.00*</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intrajudge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 2</td>
<td>Language</td>
<td>Vietnamese English</td>
<td>1.00*</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>Rating Type</td>
<td>Interjudge</td>
<td>1.00*</td>
<td>2.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intrajudge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 3</td>
<td>Language</td>
<td>English Vietnamese</td>
<td>1.00*</td>
<td>1.87</td>
</tr>
<tr>
<td></td>
<td>Rating Type</td>
<td>Interjudge</td>
<td>1.00*</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intrajudge</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*reference level

Note
Hoffman et al. (2014) Study: Australian English-speaking judges rating Australian English and Mandarin
Study 2: English speaking judges rating English and Vietnamese
Study 3: Vietnamese-speaking judges rating Vietnamese and English
For the Hoffman et al. (2014) re-analysis, judges had higher agreement levels in Australian English (familiar language) than in Mandarin (unfamiliar language), with an odds ratio of 1.35. This shows that judges are 1.35 times more likely to reliably rate severity of stuttering using the severity rating scale in Australian English than in Mandarin. The confidence interval suggests that the difference could be as little as 1.06 times more likely, or as much as 1.72 times more likely. There is a statistically significant ($p = 0.02$) advantage to English-speaking judges being able to rate more reliably in Australian English than in Mandarin. Table 6.2 shows that results for Study 2 (English-speakers) (OR 1.47, 95% confidence interval [1.19, 1.81], $p = 0.001$) and Study 3 (Vietnamese-speakers) (OR 1.87, 95% confidence interval [1.29, 2.69], $p = 0.001$) were also statistically significant (with Study 3 being notably greater) for judges being able to rate more reliably in a familiar language in comparison to an unfamiliar and/or other language. These consistent findings across the three studies identify that language familiarity impacts on judges’ ability to reliably measure severity of stuttering using the severity rating scale.

*What is the Difference Between Intrajudge and Interjudge Agreement?*

Logistic regression was used to examine whether there was any difference between intrajudge compared to interjudge performance for the Hoffman et al. (2014) study. The data were then used to determine whether these results were comparable to the results of Study 2 (English-speakers) and Study 3 (Vietnamese-speakers). For the Hoffman et al. re-analysis, the rating type variable was significant ($p = <0.001$) in the model, with agreement for intrajudge agreement being higher than for interjudge agreement, with an odds ratio of 1.72. This suggests that judges are 1.72 times more likely to rate severity of stuttering reliably with themselves than with other judges. The confidence interval estimate in Table 6.2 suggests that odds for measuring stuttering
reliably against themselves (intrajudge), could be as little as 1.31 or as much as 2.26 times as great as the odds for measuring stuttering reliably with others (interjudge). Table 6.2 identifies the rating type variable for Study 2 (English-speakers) (OR 2.20, 95% confidence interval [1.79, 2.70], $p = <0.001$) and Study 3 (Vietnamese-speakers) (OR 1.45, 95% confidence interval [1.08, 1.94], $p = 0.01$) to also be statistically significant. These results undoubtedly confirm that rating type (intrajudge versus interjudge) impacts on judges’ ability to reliably measure severity of stuttering using the severity rating scale.

Do any of these Three Effects Vary According to the Rating Type or Language Familiarity (Familiar Versus Unfamiliar or Other)?

In the Hoffman et al. (2014) study, the logistic regression model showed four significant effects: two severity rating terms and main effects for rating type (intrajudge, interjudge) and language (English, Mandarin). None of the effects varied with severity rating, as their interactions with the two severity rating terms were not significant, nor was the interaction between them significant. Therefore, all model effects were independent of other variables. The agreement percentage levels from the fitted model at all levels of severity rating for the rating type (intrajudge, interjudge) and language variables are shown in Figure 6.1. The result of none of the effects varying with severity rating for the Hoffman et al. study is identical to that of Study 2 (English-speakers) and Study 3 (Vietnamese-speakers).

**Descriptive Statistics**

*Intrajudge and Interjudge Agreement*

Table 6.3 depicts combined data of the percentage of judges who achieved criterion performance for each of intrajudge and interjudge agreement, for each language, for Hoffman et al. (2014), Study 2 (English-speakers; see Chapter 4) and
Study 3 (Vietnamese-speakers; see Chapter 5). Discussions of these results have been presented in detail and can be found in Hoffman et al., Chapter 4 for Study 2 (English-speakers) and Chapter 5 for Study 3 (Vietnamese-speakers). These results are existing data and included here for ease of reference.

Table 6.3
Combined Results of Percentage of judges who achieved criterion intra- and interjudge reliability (≥80% agreement) and ≥70% agreement

<table>
<thead>
<tr>
<th>Judge Condition</th>
<th>Intrajudge</th>
<th>Interjudge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≥ 80%</td>
<td>≥ 70%</td>
</tr>
<tr>
<td>Hoffman et al. (2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australian English</td>
<td>50.00%</td>
<td>90.91%</td>
</tr>
<tr>
<td>Mandarin&lt;sub&gt;all&lt;/sub&gt;</td>
<td>50.00%</td>
<td>81.82%</td>
</tr>
<tr>
<td>Mandarin&lt;sub&gt;AE&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt;</td>
<td>70.00%</td>
<td>80.00%</td>
</tr>
<tr>
<td>Study 2&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>75.00%</td>
<td>90.00%</td>
</tr>
<tr>
<td>Vietnamese&lt;sub&gt;all&lt;/sub&gt;</td>
<td>55.00%</td>
<td>75.00%</td>
</tr>
<tr>
<td>Vietnamese&lt;sub&gt;ENGr&lt;/sub&gt;&lt;sup&gt;d&lt;/sup&gt;</td>
<td>57.00%</td>
<td>71.00%</td>
</tr>
<tr>
<td>Study 3&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>52.94%</td>
<td>76.47%</td>
</tr>
<tr>
<td>English&lt;sub&gt;all&lt;/sub&gt;</td>
<td>29.41%</td>
<td>47.06%</td>
</tr>
<tr>
<td>English&lt;sub&gt;VNr&lt;/sub&gt;&lt;sup&gt;f&lt;/sup&gt;</td>
<td>40.00%</td>
<td>60.00%</td>
</tr>
</tbody>
</table>

Note.

<sup>a</sup> Data from the Hoffman et al. (2014) study are a reprint of data presented from an article published in the *International Journal of Speech Language Pathology* in 2014, by the candidate as first author, and co-authored with Dr Linda Wilson, Dr Anna Copley, Dr Sally Hewat, and Dr Valerie Lim. The candidate was the chief investigator in the research described. Hoffman et al. (2014): Intrajudge N = 22 for Australian English and Mandarin<sub>all</sub>; N = 10 for Mandarin<sub>AE</sub>; and Interjudge N = 26 Australian English and Mandarin<sub>all</sub> and N = 10 Mandarin<sub>AE</sub>.

<sup>b</sup> Mandarin<sub>AE</sub> presents scores for Mandarin for the subgroup of judges who achieved both intra- and interjudge agreement in Australian English.

<sup>c</sup> Study 2: Intrajudge N = 40 for English and Vietnamese<sub>all</sub>; N = 14 for Vietnamese<sub>ENGr</sub>; and Interjudge N = 59 English and Vietnamese<sub>all</sub> and N = 14 Vietnamese<sub>ENGr</sub>.

<sup>d</sup> Vietnamese<sub>ENGr</sub> presents scores for Vietnamese for the subgroup of judges who achieved both intra- and interjudge agreement in English.

<sup>e</sup> Data from Study 3 are a reprint of data presented from an article published in *Speech, Language and Hearing* in 2017, by the candidate as first author, and co-authored with Dr Linda Wilson, Dr Sally Hewat, and Dr Thao Bich Huynh. The candidate was the chief investigator in the research described. Study 3: Intrajudge N = 17 for Vietnamese and English<sub>all</sub>; N = 5 for English<sub>VN</sub>; and Interjudge N = 25 Vietnamese and English<sub>all</sub> and N = 5 English<sub>VN</sub>.

<sup>f</sup> English<sub>VN</sub> presents scores for English for the subgroup of judges who achieved both intra- and interjudge agreement in Vietnamese.
Discussion

The purpose of conducting a re-analysis of the Hoffman et al. (2014) study was to include another data set gathered using the same methodology and statistical analyses, to determine whether there are any robust patterns emerging across studies and languages with speech-language pathologists’ use of the 9-point severity rating scale. Three main findings have resulted from the completion of studies in the current program of research and the re-analysis of data from Hoffman et al. The first main finding is that judges were more reliable in measuring severity of stuttering using a 9-point stuttering severity rating scale in a familiar language than in an unfamiliar or other language. The second is that judges demonstrated better intrajudge than interjudge reliability. The final finding is that judges’ performance in the use of the severity rating scale varied depending on the stuttering severity depicted in the samples. These three main findings will be discussed in greater detail, along with their implications.

Main Finding One: Judges were more reliable in measuring severity of stuttering using a 9-point stuttering severity rating scale in a familiar language than in an unfamiliar or other language

Results from Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) and the re-analysis of data from the Hoffman et al. (2014; Australian English-speakers) study have shown that familiarity of language does indeed impact on speech-language pathologists’ reliability of use of the 9-point stuttering severity rating scale. A consistent pattern evident across all studies for both inferential (see Table 6.2) and descriptive (see Table 6.3) statistics reveal that speech-language pathologists performed better in the use of a severity rating scale in a familiar versus an unfamiliar or other language. English-speaking speech-language pathologists in Study 2 were approximately one and a half times (OR = 1.47) more likely to reliably measure severity
of stuttering using the severity rating scale in a familiar versus an unfamiliar language. The comparable figure from Hoffman et al. was 1.35. For Study 3, Vietnamese speech-language pathologists demonstrated greater reliability results and were 1.87 times more likely to reliably measure severity of stuttering using the severity rating scale in Vietnamese (familiar language) than in English (other language). In all, results from Hoffman et al., Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) show that the figures were statistically significant, and consistent with previous research (e.g., Lee et al., 2014; Van Borsel & Britto Pereira, 2005), further indicating that language familiarity impacts on speech-language pathologists’ abilities to reliably measure severity of stuttering. It should be noted however, that the present research did not attempt to quantify judges’ degree of language proficiency, nor to quantify the percentage contribution of language familiarity to the observed variability.

Whilst it was encouraging to see judges’ abilities to perform better in a familiar than an unfamiliar or other language, it should be noted that this pattern was not true for all judges. Results across Studies 2, 3, and 4 identified that not all judges were able to reach criterion performance when measuring severity of stuttering in a familiar language. In Table 6.3, for Study 3 (Vietnamese-speakers), it is shown that just over half of individual judges (approximately 53%) were able to produce reliable data and achieve criterion intrajudge agreement for their native language (Vietnamese). This result is almost identical to the findings of Hoffman et al. (2014; Australian English-speakers), with only 50% of individual judges achieving criterion intrajudge agreement for their native language (Australian English). In Study 2 (English-speakers), a greater percentage (75%) of judges were able to achieve criterion intrajudge agreement for a familiar language (English). A possible reason for this greater percentage could be that this participant sample was comprised of a larger number of judges (n = 59) who had a more extensive range of clinical experience compared to participants in Hoffman et al.
(n = 26) and Study 3 (n = 25). Whilst 75% of judges achieved criterion intrajudge agreement for Study 2 (English-speakers), there were still a percentage (25%) of judges not able to achieve criterion performance in their familiar language. These results are problematic, as clinical decisions based on unreliable data are detrimental and compromise outcomes. Speech-language pathologists should not assume that they will use the scale reliably in a familiar language and, therefore, continue to check their agreement. Hoffman et al. suggested that research into the development of procedures and materials could assist speech-language pathologists to develop acceptable agreement levels in their native and also in other languages. This notion is further supported by the present results of Studies 2 and 3.

A promising pattern to emerge from the program of research was that, for those judges who did not reach criterion performance for intrajudge and/or interjudge agreement, many were approaching the accepted 80% criterion performance level. For example, Study 2 (English-speakers) demonstrated that the group of judges was approaching reliability for English, with 90% of judges achieving $\geq 70\%$ for intrajudge agreement, and approximately 85% for interjudge agreement at Time 1, and 90% intrajudge agreement for Time 2. This pattern of judges approaching criterion performance for Study 2 is consistent with the results of Study 3 (Vietnamese-speakers) and Hoffman et al. (2014; see Table 6.4). This suggests that, with practice and/or training, it is likely that speech-language pathologists can reach acceptable levels of agreement in a familiar language.

Likewise, another consistent result across studies was obtained for the subgroup of judges who demonstrated reliable use of the scale in their familiar language. This relates to their performance in an unfamiliar or other language. Results of Hoffman et al. (2014) indicated that those who had demonstrated criterion performance in the use of
the severity rating scale in their native language were more likely to demonstrate criterion performance in an unfamiliar language (70% intrajudge, 90% interjudge Time 1), than the entire group (50% intrajudge, 61.54% interjudge Time 1). Results of Study 3 (English-speakers) were similar to those of Hoffman et al. For example, of the VNr subgroup, 80% of judges achieved criterion interjudge agreement for the English speech samples at Time 1 (see Table 6.4). This percentage was considerably greater than the percentage of the entire group (36%) who achieved criterion performance for interjudge agreement for English at Time 1.

Results of Study 3 (Vietnamese-speakers) and Hoffman et al. (2014) suggested that agreement appears to be affected by both language and skill in the use of the severity rating scale. However, this pattern was not repeated for Study 2 (English-speakers), with Vietnamese_all and Vietnamese_ENGr judges’ performance remaining virtually identical. This suggests that speculation arising from results of Hoffman et al. and Study 3, that skill development might play a role in the use of the severity rating scale, may not be true or may not be true across all languages. In light of these findings, speech-language pathologists should not assume that they will be reliable in measuring severity of stuttering using the severity rating scale in an unfamiliar or other language, even if they can produce reliable results in a familiar language. Thus, speech-language pathologists should continue to take steps to check their reliability, irrespective of the language spoken by the client.

**Main Finding Two: Judges demonstrated better intrajudge than interjudge reliability**

It is vital for speech-language pathologists to obtain and reliably measure stuttering in speech samples from clients. Whilst interjudge agreement examines the consistency judges have with each other, intrajudge agreement identifies the extent to
which the same person can rate the same performance consistently. There was a consistent pattern evident across all studies, which revealed speech-language pathologists performed statistically significantly better when using the severity rating scale for intrajudge agreement than interjudge agreement, across languages. For Study 2 (English-speakers), judges were more than two times (OR = 2.20, \( p = <0.001 \)) more likely to demonstrate intrajudge than interjudge agreement. Whilst the degree was higher for Study 2 (English-speakers), a consistent pattern was evident for Study 3 (Vietnamese-speakers), as judges were almost one and a half times (OR = 1.45, \( p = 0.01 \)) more likely to achieve intrajudge agreement than interjudge agreement. Results were similar for the re-analysis of Hoffman et al. (2014; Australian English-speakers) (OR = 1.72, \( p = <0.001 \)). Results of Studies 2, 3, and 4 are also consistent with previous research suggesting judges demonstrate better intrajudge agreement than interjudge agreement for measuring and/or identifying stuttering (e.g., Einarsdóttir & Ingham, 2009; Eve et al., 1995).

**Intrajudge Agreement**

Despite judges across studies showing greater intrajudge agreement levels, it is important to note that not all judges were able to produce reliable intrajudge results in a familiar language. Intrajudge agreement is important as it determines whether the clinician’s judgement can be trusted as reliable. For Study 3 (Vietnamese-speakers) and Hoffman et al. (2014; Australian English-speakers), approximately half of the individual judges did not achieve criterion performance. Study 2 (English-speakers) judges performed significantly better, with 75% achieving individual criterion intrajudge performance. However, as previously highlighted in *Chapter 4*, there was still a sizable percentage (25%) who were unable to achieve criterion performance. Therefore, speech-language pathologists should not assume that they will use the scale
reliably, even in their native language. It is recommended that speech pathologists take steps to check their intrajudge reliability, and if necessary, establish their reliability for rating severity of stuttering using the 9-point severity rating scale in a familiar language, as well as that of any unfamiliar and/or other language spoken by clients. Findings of Study 3 (Vietnamese-speakers) show that the VNr subgroup performed better at measuring severity of stuttering using the severity rating scale than the entire group, for English (other language) speech samples. Results of Hoffman et al. showed consistent findings, with the AEr subgroup performing better than the entire group, for Mandarin (unfamiliar language) speech samples. The same pattern was identified in Study 2. Thus, results suggest that speech-language pathologists should first establish reliability of measuring stuttering in their own language before working towards developing reliability in other languages.

Speech-language pathologists could check their intrajudge agreement levels by recording speech samples of their clients in their sessions, rating the sample and then reviewing the sample and re-rating at a later date. This would enable speech-language pathologists to determine whether they reached the criterion used in the program of research; if so, they could repeat the task for other languages.

In clinical practice, it is common for clients to measure their own stuttering severity using the severity rating scale outside of the clinic setting, to help the clinician and client to monitor progress (e.g., in the Camperdown Program, O’Brian et al., 2017). The clinician’s role is to train clients in the first weeks of clinic visits. Thus, checking and establishing their intrajudge reliability could equip speech-language pathologists to more efficiently train clients and caregivers in the use of the severity rating scale. If speech-language pathologists are not able to achieve acceptable levels of reliability through self-practice, further training may be advised.
Interjudge Agreement

Results for interjudge agreement for Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) were comparable to those of Hoffman et al. (2014; Australian English-speakers), highlighting that judges had better interjudge agreement for a familiar language (English for Study 2 and Hoffman et al.; and Vietnamese for Study 3) than for an unfamiliar (Vietnamese for Study 2; Mandarin for Hoffman et al.) or other (English for Study 3) language. However, results also indicated that a minority of judges did not demonstrate reliability with others for familiar language speech samples.

These findings have implications for clinical practice. For example, stuttering therapy is often required on an ongoing basis. Consequently, it is likely that some clients may not consult the same speech-language pathologist for the duration of their therapy course. For example, clients may re-locate to another region, or speech-language pathologists may transfer and leave the workplace. Thus, it is crucial for speech-language pathologists to establish interjudge reliability so that they can communicate effectively about clients and client care. It is therefore recommended that speech-language pathologists make opportunities to practice interjudge reliability in their own workplace with their colleagues. Vong et al. (2011) suggested that this could be achieved by recording speech samples of clients, rating the severity of stuttering, having a colleague listen to and rate the same speech sample, and then compare scores. If speech pathologists are within the clinically accepted level of reliability of plus or minus one scale value on 80% of samples for a familiar language, they could then practice establishing interjudge reliability for other languages relevant to their clinical practice.

In previous research, substantial interclinic discrepancies have been identified when measuring severity or stuttering using a 7-point severity rating scale (Kully &
Boberg, 1988) and when making judgements of stuttered and nonstuttered intervals (e.g., Cordes & Ingham, 1995; Ingham & Cordes, 1992). Cordes and Ingham suggested that operational definitions of stuttering may be shared within but not across clinics, and may be unique to each university and research centre. In Study 3, speech-language pathologists in Vietnam received the same training in the use of the severity rating scale to measure severity of stuttering, at the same university (S. Hewat, personal communication, March 7, 2017). From previous research, it might be expected to see greater interjudge agreement between judges in Study 3 compared to those in Study 2 (English-speaking) and the Hoffman et al. (2014; Australian English-speakers) study, which were comprised of speech-language pathologists from around the world (e.g., Australia, Canada, United States, and Europe) who received formal training from different universities internationally. However, results did not show this, with judges in Study 3 (Vietnamese-speakers) demonstrating poorer interjudge reliability (44.00%) at Time 1 in comparison to Study 2 (66.10%) and Hoffman et al. (76.92%).

Results from the program of research for interjudge agreement from Time 1 to Time 2 showed a lack of consistency across studies. For Study 2 (English-speakers), results at Time 1 for the familiar language (English, 66.10%) and unfamiliar language (Vietnamese, 44.07%) remained similar at Time 2 (70% English, 42.50% Vietnamese). Results for Hoffman et al. (2014) showed a decline in reliability from Time 1 to Time 2 for the familiar language (Australian-English, Time 1, 76.92%; Time 2, 63.64%) and unfamiliar language (Mandarin, Time 1, 61.54%; Time 2, 45.45%). Results of Study 2 and Hoffman et al. (2014) were not comparable to those of Study 3 (Vietnamese-speakers), which showed an apparent greater interjudge improvement for the familiar language. For Study 3 (Vietnamese-speakers), judges performed poorly at Time 1, with only 44% of judges achieving criterion performance in their familiar language (Vietnamese). Interestingly, judges achieved criterion interjudge agreement (82.35%) at
Time 2, thus showing reliable use of the severity rating scale at Time 2. It is speculated that this apparent improvement might be attributable to increased exposure and a practice effect, highlighting the potential importance of practice and training to skill development of novice judges in a newly emerging profession in Vietnam. Why a practice effect did not occur for Study 2 (English-speakers) and Hoffman et al. (2014) is unknown.

**Main Finding Three: Judges’ performance in the use of the severity rating scale varied depending on stuttering severity**

An interesting outcome across the current studies and the re-analysis of the Hoffman et al. (2014) data was the increased variability shown in the moderate range of the severity rating scale. In other words, agreement varied over the range of the severity rating scale. This is to be expected as “variances of rating scales are necessarily related to the medians being smaller towards both ends of the scale and larger in the middle” (Bosshardt et al., 2015, p. 225). Figure 6.1 shows the agreement percentage levels for the severity rating range, for intrajudge and interjudge agreement, for all three studies. When examining the figure, it is evident that percentage of agreement varied considerably over the stuttering severity rating range for all studies. In addition, for Study 2 (English-speakers), judges were able to achieve 80% agreement for severity ratings of 1 to 3 for both English and Vietnamese and also for severity ratings of 7 to 9. Judges did not achieve 80% agreement for severity ratings of 4 to 6. This indicates that judges were able to make reliable judgements (achieve 80% or greater) for mild (severity rating of 1 to a severity rating of 3) and severe (severity rating of 7 to a severity rating of 9) stuttering, for both familiar (English) and unfamiliar (Vietnamese) languages. Judges did not meet the 80% criterion for stuttering in the moderate range. These results are consistent with those of Study 3 (Vietnamese-speakers) and the re-
analysis of Hoffman et al., indicating judges were able to make reliable use of the scale for mild and severe stuttering for both familiar (Australian English for Hoffman et al., 2014; and Vietnamese for Study 2) and unfamiliar (Mandarin for Hoffman et al., 2014) or other (English for Study 2) languages. Likewise, for Study 1, Figure 3.3 (see Chapter 3) showed increased variability across the moderate range of the severity rating scale for all three speech sample durations (1-minute, 3-minutes and 5-minutes). These results are somewhat consistent with those of Cosyns et al. (2015), who reported that speech-language pathologists could accurately measure severity of stuttering when the speaker is severely stuttering. It is also somewhat consistent with the findings of Watson and Kayser (1994) who suggested that, if the client is stuttering severely, identifying stuttering may not be a difficult task. The present studies contributed new findings, identifying also reliable use of the severity rating scale by speech-language pathologists for mild stuttering, irrespective of familiarity with language/s.

Unfortunately, results of all studies in the program of research and the re-analysis of the Hoffman et al. (2014) study showed a consistent pattern of difficulty in measuring severity of stuttering in the moderate range of the severity rating scale, regardless of language. Variability across the severity rating scale range has both research and clinical implications. The severity rating scale is used throughout treatment to monitor client progress. Irrespective of a speech sample duration (1-minute versus 3- and 5-minutes) and language familiarity, it appears that speech-language pathologists will produce greater variability in the moderate range when using the 9-point severity rating scale to measure severity of stuttering. With increased variability (decreased reliability) evident in the moderate range of the scale, it is difficult in clinical practice to determine whether the differences in stuttering measurement (e.g., severity rating score of 6 and a severity rating score of 4) for two speech samples are due to actual changes in the client’s stuttering, or due to measurement error (Yaruss, 1998). Therefore, it is
recommended that speech-language pathologists be aware of this problematic range and take steps to check their reliability when clients’ stuttering is presenting in the moderate range.

It is important that speech-language pathologists provide appropriate stuttering treatment to all clients irrespective of their linguistic background. Therefore, it is encouraging that results from Study 2 (English-speakers) and Study 3 (Vietnamese-speakers) and the re-analysis of Hoffman et al. (2014; Australian English-speakers) showed that a majority of judges were able to make reliable use of the scale for mild and severe stuttering in a familiar, and unfamiliar or other languages. This suggests that language familiarity may not play a large role with measuring severity of stuttering in the mild and severe range. The stuttering severity rating scale is commonly used in treatment programs for both children (e.g., the Lidcombe Program, Onslow et al., 2017) and adults (e.g., the Camperdown Program, O’Brian et al., 2017) who stutter. Being able to make reliable judgements of mild stuttering is critical to determining when clients are ready to enter into maintenance stages of treatment and/or be discharged from treatment programs. Therefore, it is promising to see that speech-language pathologists demonstrated more reliable use of the severity rating scale in the mild range, irrespective of language.

Conclusion

The main purpose of the re-analysis of the Hoffman et al. (2014) data was to replicate the analyses of Studies 2 (English-speakers; see Chapter 4) and 3 (Vietnamese-speakers; See Chapter 5), to determine whether the consistent findings of the program of research could be replicated with data from a different group of participants and another unfamiliar language (Mandarin). Results corroborated those of Studies 2 and 3, highlighting three main robust findings across the studies: (i) Judges
were more reliable in their use of the scale in a familiar versus an unfamiliar or other language; (ii) Judges showed better intrajudge agreement than interjudge agreement; and (iii) Judges produced greater variability (decreased reliability) in their use of the scale in the moderate range. Despite not knowing the underlying cause and nature of stuttering, speech-language pathologists still appear to be able to perceive stuttering across familiar, unfamiliar and other languages. The overall findings highlight the need for speech-language pathologists to work on improving reliability of measuring severity of stuttering in their familiar language, as well as in an unfamiliar or other language.
CHAPTER 7

Conclusion
This program of research contributes to the field of stuttering, specifically stuttering measurement. The overarching purpose of this program of research was to investigate the reliability of speech-language pathologists’ use of a 9-point severity rating scale to measure severity of stuttering in speech samples of different durations, and varying in language familiarity (familiar, unfamiliar, and other) across three languages: English, Vietnamese, and Mandarin. In order to determine the interplay of speech sample duration on stuttering measurement, the first study was undertaken to explore whether the duration of a speech sample (1-minute, 3-minutes, 5-minutes) influenced reliability of measuring the severity of stuttering. Results of Study 1 recommended the use of 3 or 5-minute speech samples for research purposes, with 1-minute samples showing increased variability. Three-minute speech samples were favoured (over 5-minutes), as 3-minute samples reduce time commitment from judges in future studies. Two studies followed, using the recommended duration of 3-minutes, to explore the reliability of measuring severity of stuttering in a familiar language (English for English-speakers and Vietnamese for Vietnamese-speakers), an other language (English, a common second language for Vietnamese-speakers) and an unfamiliar language (Vietnamese for English-speakers). Finally, a re-analysis of the Hoffman et al. (2014) data (Australian-English speakers) was undertaken to examine the robustness of the current findings from the program of research. Findings from these studies were synthesised and discussed in Chapter 6. In this chapter, a summary of the findings are provided, followed by a statement of the contributions of this research and future directions.

**Summary of Findings of this Research**

Research examining speech-language pathologists’ ability to reliably measure stuttering has been of interest over time. Yet, speech samples used in stuttering
measurement research have varied in duration, and results across studies have shown that consensus has yet to be made regarding the reliability of use of the severity rating scale. Investigating the duration of speech samples is important for researchers due to the necessity of recruiting suitable participant sample sizes and reducing time commitment. In an attempt to gather preliminary information, Study 1 (see Chapter 3) was undertaken to examine whether the duration of speech samples influences the reliability of measurement of severity of stuttering by speech-language pathologists. Ten specialist speech-language pathologists rated 27 audio, English speech samples of three different durations (9 x 1-minute, 9 x 3-minutes, 9 x 5-minutes) of adults who stutter, using a 9-point severity rating scale. Variability was found to be larger in the moderate severity range than the mild and severe ranges. In addition, judges showed greater variability (poorer reliability) when rating 1-minute speech samples. The 3-minute and 5-minute samples produced similar results when using the scale to measure severity of stuttering. Samples of 3 and 5-minute durations were found to be equally appropriate for reliability research. Data trends suggested that speech-language pathologists and researchers should focus more attention on practice and training in the middle ranges of the scale, due to increased variability in this range.

Study 2 (see Chapter 4), which investigated English-speaking speech-language pathologists use of the severity rating scale, replicated Hoffman et al.’s (2014) procedure with a larger participant sample and a different unfamiliar language and language family. The aim was to further investigate speech-language pathologists’ reliability of use of the severity rating scale in English (familiar language) and another unfamiliar language (Vietnamese). Fifty-nine English-speaking speech-language pathologists from around the world (i.e., Australia, United States, Canada, England, Scotland, Northern Ireland, Argentina, South Africa, Portugal, and Germany) rated 20 speech samples (10 English, 10 Vietnamese) of adults who stutter, using a 9-point
severity rating scale on two occasions. Findings corroborated some of those of Hoffman et al., identifying judges to be more reliable in measuring stuttering in a familiar language (English) than in an unfamiliar language (Vietnamese). Judges produced greater intrajudge agreement in comparison to rating against other judges (interjudge agreement), for both English and Vietnamese. Finally, judges showed better agreement for use of the scale for mild and severe stuttering across both languages, with the moderate range of the scale being problematic (i.e., showing increased variability).

Study 3 (see Chapter 5) provided the first Vietnamese stuttering measurement reliability data for Vietnamese speech-language pathologists, a newly emerging profession in Vietnam. The study investigated Vietnamese speech-language pathologists’ reliability of use of the severity rating scale for a familiar language (Vietnamese) and an other language (i.e., English, a common second language). Twenty-five Vietnamese-speaking speech-language pathologists from Vietnam rated 20 audio speech samples (10 Vietnamese, 10 English), identical to those used in Study 2, of adults who stutter, using a 9-point severity rating scale on two occasions. Approximately half of the judges were able to measure severity of stuttering reliably in Vietnamese. Judges’ performance in English was poorer than in their native language. Irrespective of language, judges showed greater variability of their use of the scale in the moderate range. Results highlighted the need for judges to develop intra- and interjudge agreement when using the scale to measure stuttering in their native and other languages.

Chapter 6 presented a re-analysis of data from the Hoffman at el. (2014) study. The main purpose of this re-analysis was to enable comparison to data for Studies 2 and 3, to determine whether the consistent findings across the present studies were robust. The re-analysis replicated the analyses of Studies 2 (English-speakers; see Chapter 4)
and 3 (Vietnamese-speakers; See Chapter 5). In Hoffman et al., 26 Australian speech-language pathologists rated 20 audio speech samples (10 Australian English, familiar language; 10 Mandarin, unfamiliar language) of adults who stutter, using a 9-point severity rating scale, on two occasions. Results of the re-analyses corroborated those of Study 2 (English-speakers) and Study 3 (Vietnamese-speakers), with judges producing greater variability in their use of the scale in the moderate range, whilst showing greater agreement for mild and severe stuttering for both languages. Results of the logistic regression analyses also strengthened the descriptive statistics from Hoffman et al., highlighting judges to be more reliable in a familiar language than an unfamiliar/other language, and judges having greater intrajudge agreement compared to interjudge agreement across languages.

Therefore, the program of research has resulted in three robust findings:

1. Judges were more reliable in measuring severity of stuttering using a 9-point stuttering severity rating scale in a familiar language than in an unfamiliar or other language;
2. Judges demonstrated better intrajudge than interjudge reliability;
3. Judges’ performance in the use of the severity rating scale varied depending on the stuttering severity depicted in the speech samples. More specifically, irrespective of familiarity of language (i.e., familiar, unfamiliar or other), judges showed greater variability (poorer reliability) in their use of the scale in the moderate range of the severity rating scale. Judges’ performance in the use of the scale was more reliable in the mild and severe ranges of the scale.
What this Thesis has Contributed and Implications

The program of research has made a number of contributions to the field of stuttering measurement using the 9-point stuttering severity rating scale. It has also provided some guidelines for the measurement of stuttering severity using the scale for research and clinical practice.

Determining Duration of Speech Samples

Research findings of Study 1 (see Chapter 3) offered new research application prospects. Results provide some guidelines for the speech sample duration for future research studies and in the development and evaluation of practice and/or training procedures. These results recommended the use of 3 or 5-minute speech samples for reliability research purposes when measuring severity of stuttering using the 9-point severity rating scale. Based on the findings of Study 1, future research should avoid speech samples of 1-minute duration, due to increased variability in scores assigned by specialist speech-language pathologists, for English samples.

Improving Clinical Utility of Reliability Research

Reliability in the stuttering severity rating scale is crucial for both research and clinical practice. Martin and Haroldson (1992) stated that acceptable group reliability ratings does not necessarily indicate that any given rater will achieve acceptable reliability levels. Research studies that present grouped data alone allow the reader to determine acceptable levels of agreement only for research purposes, as it merely discriminates between groups of individuals. Research that presents individual data is useful for clinical purposes as it is able to discriminate differences within an individual. Research that presents results for the group of judges, has shown to be problematic with reliability studies for stuttering measurement. For example, previous research by Eve et
al. (1995) examined the reliability of speech-language pathologists’ use of a stuttering severity rating scale presenting grouped data describing participant performance. Hoffman et al. (2014) also used grouped data analyses identical to those of Eve et al. Results of Hoffman et al. identified acceptable group intrajudge agreement (approximately 82%) for measuring severity of stuttering in a familiar language (Australian English). The corresponding figures for group interjudge agreement were approximately 68% at Time 1 and 73% at Time 2. The results appeared to have replicated the grouped data findings of Eve et al, who, based on a 10-point scale, reported intrajudge and interjudge agreement figures for the generalist clinician group of approximately 75%, 68% and 77% respectively. Hoffman et al. went one step further, however, and also examined data about the percentage of participants who reached criterion reliability performance, and it was this procedure that revealed that grouping of data may have previously masked individual variation, through an averaging effect. For example, it was found that only approximately half (52.94%) of individual participants reached a criterion of 80% of Time 1 to Time 2 comparisons falling within ±1 scale value in their familiar language (Australian English). Thus, the excellent performance of some judges served to offset the poorer performance of others, resulting in an averaging effect. Hoffman et al. recommended that future research incorporate individual analyses to prevent masking of individual differences and the averaging effects, when estimating both intrajudge and interjudge agreement.

Given the clinical usefulness of reliability research to clinicians, it is imperative for data to be presented in a meaningful way. Thus, Study 2 and Study 3 presented data for a new language family describing the performance of individual judges who achieved criterion intrajudge and interjudge agreement. The importance of this, is that clinicians now have clinical data of individuals’ performance across three languages (English, Vietnamese and Mandarin) and three language families (Indo-European, Austro-Asiatic,
and Sino-Tibetan). Collecting data across different languages, using identical statistical analyses and clear procedures, will eliminate possible confounding variables (e.g., differing analyses, procedures, speech sample duration), and ensure the findings are robust and can be trusted.

**Expanding Reliability Research into Other Language Families**

With the exception of Hoffman et al. (2014), which investigated English (Indo-European) and Mandarin (Sino-Tibetan), previous studies (Cosyns et al., 2015; Einarsdóttir & Ingham, 2009; Lee et al., 2014; Van Borsel & Britto Pereira, 2005; Van Borsel et al., 2008) have investigated only languages from within the Indo-European language family. The current program of research aimed to expand the existing knowledge of measuring severity of stuttering in a language from an entirely different language family that is commonly spoken worldwide: Vietnamese, from the Austro-Asiatic language family.

Study 3 (see Chapter 5) provided the first Vietnamese reliability data for measuring severity of stuttering in Vietnamese by Vietnamese-speaking speech-language pathologists, for the newly emerging speech-language pathology profession in Vietnam. This research has provided Vietnamese speech-language pathologists with an appropriate measurement tool to use when working with clients who stutter in Vietnam. Measuring stuttering in Vietnamese was examined across two studies in the program of research: (i) as a familiar language by Vietnamese-speaking speech-language pathologists; and (ii) an unfamiliar language by English-speaking speech-language pathologists. Investigating a new language family in stuttering measurement research strengthened the importance and uniqueness of the program of research.
Identifying Variability of Agreement Across the Severity Rating Scale

An outcome from the program of research showed that, irrespective of language, speech-language pathologists showed greater reliability for measuring severity of stuttering using the 9-point scale for mild and severe stuttering, for either familiar, other or an unfamiliar language, across three language families: (i) Indo-European (English); (ii) Austro-Asiatic (Vietnamese); and (iii) Sino-Tibetan (Mandarin). What was discouraging, however, was the consistent pattern, of greater variability (poorer reliability) for measuring severity of stuttering in the moderate range of the scale, regardless of familiarity with the language spoken. This problematic range highlights the importance for future research to focus on finding ways to assist speech-language pathologists to improve their reliability in the middle range of the scale. One way to achieve this could be through research and training into the development of a practice and/or training package, with a focus on the moderate range.

Future Directions

There is strong research and clinical impetus to continue to examine the reliability of measurement of stuttering across languages. The current program of research raises a number of topics requiring further investigation in the area of stuttering severity measurement.

Speech Sample Duration

Study 1 (see Chapter 3) was conducted for research purposes to inform suitable durations (1-minute, 3-minutes, and 5-minutes) of speech samples for future research in stuttering measurement. This study was a preliminary study. Thus replication of this study is encouraged, to determine whether the current findings would be repeated.
Furthermore, Study 1 was undertaken in English only. Thus, it is unknown whether the same outcomes would be present for other language/s and/or language families. As stuttering occurs in all languages and cultures, exploring the relationship between speech sample duration and reliability across different languages is important to ensuring that findings are robust and true across languages. This information would be valuable to the future development of practice and/or training packages in the use of the severity rating scale for speech-language pathologists (e.g., determining whether the same or different durations are required across languages).

**Cross Linguistic Investigation of Other Language Families**

The program of research explored the reliability of measurement of stuttering using the 9-point severity rating scale across languages (English, Vietnamese, and Mandarin) that differ significantly in their structure and are from disparate language families (Indo-European, Austro-Asiatic, and Sino-Tibetan). Previous research (Van Borsel et al., 2008) has indicated that identifying stuttering in a language that is further remote from the language of the judge is a more difficult task. Therefore, the inclusion of replication studies for different language families is strongly encouraged. Investigating languages from other language families would determine whether the same outcomes apply to other language families and are universal.

**Continued Stuttering Research in Vietnam**

Speech-language pathology is a developing profession in Vietnam. This thesis presented the first data of Vietnamese speech-language pathologists’ reliability of use of the severity rating scale to measure stuttering. Limited research available to date leaves a lot unknown about stuttering in Vietnamese. For example, it is unknown how stuttering manifests in Vietnamese. Chapter 1 of this thesis discussed what is currently known about the manifestation of stuttering for tone languages (i.e., Mandarin, Chou et
al., 2015; and Cantonese, Law et al., 2017) which produced conflicting findings. Chou et al., investigated the relationship between stuttering loci and lexical tone in Mandarin-speaking preschool aged children and concluded that lexical tones with a complex tonal contour, especially when placed under conflicting tonal contexts, were found to contribute to the occurrence of stuttering in Mandarin children. In contrast, Law et al. investigated adult Cantonese-speakers and results indicated no significant differences to be found for stuttering frequency across lexical tones. Thus, one could speculate that results of studies about one tone language may not be true for other tone languages. Therefore, research into how stuttering manifests and develops in Vietnamese for both children and adult populations would be beneficial, to establish whether the patterns seen in Mandarin and/or Cantonese are similar to patterns in Vietnamese, and therefore further aid in identifying universal features of stuttering; or differ and might therefore be language specific (Howell & Rusbridge, 2011).

It would also be beneficial for further research to investigate the reliability of stuttering in other languages (such as Tay, Khmer, Chinese, and French; Cheng, 2010) which are commonly spoken in Vietnam. To the author’s knowledge, no such research has been published in English. Thus, it is unknown whether the patterns shown in the program of research would extend to these languages. This exploration would help guide future teaching and practice in stuttering measurement for Vietnamese speech-language pathologists.

Since the early 1990s, Vietnam has heavily relied on international volunteer speech-language pathologists. With speech-language pathology being an emerging profession, it is likely that Vietnam may continue to require assistance from international volunteers, to provide professional development to local novice speech-language pathologists and services to clients, for quite some time. However, little is
known about the skill level of volunteers in stuttering measurement, in their own native language, and their ability to measure stuttering in Vietnamese. As the identification and/or measurement of stuttering can be less accurate when assessing stuttering in an unfamiliar language (see Einarsdóttir & Ingham, 2009; Hoffman et al., 2014; and Van Borsel et al., 2005; Chapters 4 and 5 of the present thesis), volunteer speech-language pathologists might benefit from practice in the use of the severity rating scale in Vietnamese, prior to commencing volunteer work in Vietnam. This could encourage consistency between Vietnamese and volunteer speech-language pathologists when working with clients who stutter. Given that there is no research into the role volunteers play with training local clinicians in the use of stuttering measures, it is important to highlight that these notions remain unclear and are purely speculative. However, given the continued need for volunteers, who might be unfamiliar with the Vietnamese language, these notions provide the impetus to further explore the stuttering of Vietnamese-speakers. Further research is desired for a better understanding of how stuttering presents in Vietnamese versus English. Such information would greatly assist speech-language pathologists and researchers to devise appropriate guidelines to reliably identify and measure stuttering in Vietnamese, and to provide suitable stuttering treatment to Vietnamese clients.

The Need for Training

Despite much interest and recent research into the reliability of identifying and/or measuring stuttering, there is still conflicting evidence concerning whether or how well speech-language pathologists are able to make reliable judgements about and/or measurements of stuttering. The identification and/or measurement of stuttering has been shown to be less accurate when assessing stuttering in familiar and foreign languages (see Hoffman et al., 2014; Van Borsel et al., 2005; and the current Studies 2
and 3). Results from the program of research, therefore, reinforce the need for the development of a training package to help speech-language pathologists to develop skills for measuring severity of stuttering using the 9-point scale, to be made available worldwide.

As essential as it is for clinicians to develop reliability in the use of the scale, it is equally vital that speech-language pathology students are provided with appropriate training at the university level in stuttering measurement. Trends in globalisation reveal an increase in cultural and linguistic diversity. Thus, it is critical that universities keep abreast of the ever-changing climate when preparing students to work with a linguistically diverse caseload. Training should embed the development of cultural and linguistic competence in the management of stuttering to facilitate suitable service provision for clients. With the increase of different languages being represented throughout countries, this continues to be a complex task for researchers and universities. Further research is warranted in how best to develop such training. This research should have a multilingual focus, with international collaboration, to ensure students are equipped with suitable skills and resources, for providing appropriate stuttering management to clients, irrespective of the language spoken by the client.

Whilst it is imperative that training centres (e.g., universities) ensure they are teaching appropriate techniques to students, and researchers investigate what training techniques might have the best impact in improving speech-language pathologists’ reliability in the use of the scale, it is vital that clinicians also accept responsibility for improving their ability to make reliable judgements and measures of their clients’ stuttering. It is recommended that clinicians commit to self-practice in their use of the scale and continue to train until they are able to achieve intrajudge reliability. This could be achieved via audio or video recording their clients’ speech, which allows for
playback and re-scoring after online measures have been made. It is also recommended that clinicians take the time to consult colleagues and practice in their use of the scale to improve interjudge reliability.

As training appears to be commonly recommended for measuring stuttering severity in English and/or other languages (e.g., Cosyns et al., 2015; Van Borsel & Britto Pereira, 2005; Hoffman et al., 2014), it is important to consider what is currently available for researchers and clinicians for improving their reliability of measuring severity of stuttering. Thus, a description of available training in stuttering measurement is provided.

A number of training packages have been developed to assist speech-language pathologists to develop skills for measuring stuttering. Among these is the Stuttering Counts: Measuring Stuttering and Speech Rate CD-Rom (Block & Dacakis, 2002), and the Stuttering Measurement System (SMS) program (Ingham, Bakker, Moglia, & Kilgo, n.d.). The Stuttering Counts: Measuring Stuttering and Speech Rate CD-Rom provides speech examples and samples for practice of (i) identifying and counting stutters, (ii) counting syllables, (iii) measuring stutter counts (i.e., percentage syllables stuttered), and (iv) measuring speech rate (i.e., syllables per minute). The SMS program is a freely available online training program to assist speech pathologists to identify stuttering event judgements. This program provides speech examples and samples with which to practice (i) identifying stuttering, (ii) measuring the percentage of syllables stuttered, (iii) measuring speech rate, and (iv) measuring speech naturalness.

The literature also describes a Stuttering Measurement Assessment and Training (SMAAT) program for adults (i.e., Ingham, Cordes, Kilgo, & Moglia, 1998) and children (i.e., Ingham et al., 2006). The SMAAT program was developed to study stuttering event judgements in the speech of adults and children who stutter. The
program was developed for research purposes only, under a grant held by the University of California Santa Barbara, and is not available for public or free use (R. J. Ingham, personal communication, May 27, 2013).

It appears that all publically available stuttering measurement training resources have been developed in English only. Furthermore, the stuttering measurement training resources have failed to incorporate training in the use of the stuttering severity rating scale, a scale commonly used to assess and measure severity of stuttering in clinical practice (O’Brian et al., 2004b). Thus, research into the development and evaluation of training packages for reliably measuring severity of stuttering is warranted.

Currently, without training packages available in the use of the stuttering severity rating scale, it can only be speculated that practice and training would help. However, it is still unknown how best to train speech-language pathologists to improve their reliability in their use of the scale. Given the rise in multiculturalism worldwide, there is an urgent need for this to be researched. Collaboration amongst international researchers would make this aim more viable. Therefore, the next section will explore considerations in the development of a stuttering severity training package for improving speech-language pathologists’ reliability of the use of the scale.

**The Development of a Stuttering Severity Training Package**

Whilst training has been recommended for quite some time, there is a lack of empirical evidence on how best to improve speech-language pathologists’ abilities to reliably measure severity of stuttering in familiar and any languages other than their own. Currently, it is unknown what type of training is needed and whether the training packages would have an impact on increasing reliability for clinicians and students.
Results of the current program of research provide us with some initial guidelines for the development of practice and/or training packages. There are many considerations when developing a training package. Firstly, results from Study 1 (see Chapter 3) suggested avoiding speech samples of 1-minute in duration, due to increased variability in scores assigned by specialist speech-language pathologists. The use of 3 or 5-minute speech samples is recommended in the development of training packages. In particular, the use of 3-minute speech samples (as opposed to 5-minutes) is favoured, as this could reduce time commitment for future training procedures, therefore potentially increasing participation by speech-language pathologists.

Next, the inclusion of agreed stuttering exemplars in the severity rating training package could eliminate confusion caused by conflicting descriptors of stuttering (Cordes & Ingham, 1995). The recommendation of agreed stuttering exemplars has been made across various stuttering research studies about the identification of stuttering. For example, Einarsdóttir and Ingham (2005, 2008, 2009) recommended that stuttering exemplars from various languages might assist speech pathologists to identify universal features of stuttering. Einarsdóttir and Ingham recommended exemplar-based training across languages, to aid the development of the skill of correctly identifying stuttering occurrences. Brundage et al. (2006) also recommended the inclusion of stuttering exemplars when training students and generalist clinicians to accurately identify stuttering. An important consideration in clinical practice is to determine whether a clinicians’ judgement is reliable and valid. Thus is has been recommended that clinicians check that their judgements match those of expert clinicians (Yaruss, 1998). Therefore, to ensure there is uniformity in the use of the severity rating scale, the training package should incorporate the inclusion of agreed exemplars, made by a panel of experienced judges.
It is unknown whether speech pathologists would benefit from the use of stuttering exemplar packages alone or would also require stuttering severity practice packages, to assist in improving their reliability of measuring severity of stuttering. Brundage et al. (2006) suggested the need for the development of a standardised measurement protocol to be taught explicitly to students training to be clinicians. Brundage et al. recognised, however, that knowledge of stuttering through theory and generalist clinical practice, does not appear to be sufficient to ensure reliable and valid measures of stuttering. It was concluded that a skill, such as stuttering measurement, needs to be taught and practiced to be performed accurately.

It is unknown how much practice is needed to reach criterion reliability levels for measuring severity of stuttering using the severity rating scale. However, research which investigated the identification of stuttering could be drawn upon, to guide the development of practice recommendations for measuring severity of stuttering. Practice is an integral component of achieving reliable judgements of stuttering, and results from research conducted by Jani, Huckvale, and Howell (2013) identified that judges achieved more accurate judgements of stuttering when allowed to replay a speech sample compared to those who made a judgement after listening to the speech sample only once. Thus, the development of a training package should allow opportunities for speech-language pathologists to repeat speech samples prior to making severity rating scores.

To ensure intrajudge reliability, speech-language pathologists need to demonstrate that they can achieve acceptable reliability levels for repeated judgements across different speech samples (Yaruss, 1998). To achieve this, it has been recommended that speech-language pathologists continue to practice stuttering frequency measurements until they achieve acceptable agreement, for at least 10
different speech samples (Yaruss, 1998). Results of the current research support the need for a focus on training to be in the moderate range of the scale. This recommendation arises from the robust observation of increased variability (decreased reliability) across all current studies, irrespective of familiarity of language. It remains unknown whether the recommended 10 different speech samples (e.g., for identifying stuttering; Yaruss, 1998) would be sufficient for measuring severity of stuttering in speech samples representing the moderate range of the scale, or whether additional exemplars are needed for this range when developing the severity rating training package.

There is a need for research to also investigate how to most efficiently provide training across different languages. For examples, research could investigate whether speech pathologists would benefit from the use of packages for each language (e.g., a Mandarin, Arabic, or Italian package), or for each language family sharing similar characteristics (e.g., developing a Chinese package encompassing Mandarin, Cantonese, and Hokkien stuttering exemplars). It is unknown whether training the client in one language could then generalise to other languages in the same language family, or indeed beyond. Thus, further research is needed regarding the interaction of familiarity of language with reliability of use of the severity rating scale.

Given the complexity of measuring severity of stuttering using the scale in different languages, there is much to be learnt about how best to develop training. Collaboration between international research partners would undoubtedly assist with increasing the database of languages included in the training package. Thus, the initial training package would need to allow for the inclusion of additional samples from other languages as time progresses and collaboration increases. For the inclusion of samples from a range of languages to occur, a stringent procedure would need to be developed.
and the inclusion of a pool of experienced judges in the use of the scale is recommended. Whilst the development and set up would be time consuming and challenging, it would also have long term benefits for all. An internationally collaborative training package would enable researchers from different countries to work together, and support clinicians of different language backgrounds to work together to improve their ability to reliably measure severity of stuttering.

**Final Remarks**

Irrespective of linguistic background, the ability to communicate effectively is a basic human right. Amidst the rise in globalisation, speech-language pathologists are increasingly required to work with clients who stutter in languages that may differ from their own. With distinct differences in structure and linguistic processing between languages, the ability to reliably measure severity of stuttering in different languages remains a challenging and complex task, particularly when the language spoken by a client differs from that of the speech-language pathologist. Yet speech-language pathologists are required to provide appropriate stuttering management to all clients, irrespective of their cultural and linguistic background. Thus, an auspicious finding across the current research was that the majority of speech-language pathologists were able to reliably measure severity of stuttering for mild and severe stuttering, regardless of familiarity with language (i.e., familiar, unfamiliar and other). Thus, an advantage of the severity rating scale is that it can be used for languages with very dissimilar structure compared to English (i.e., Vietnamese and Mandarin). Another consistent, yet problematic, finding across studies was the increased variability found in the moderate range of the severity rating scale across all languages. Findings of the present research support the recommendation of further research to be undertaken to investigate ways to improve speech-language pathologists’ reliability of use of the severity rating scale to
measure severity of stuttering in a familiar, other, and unfamiliar language, particularly in the moderate stuttering range. Such research into practice and/or training packages could provide resources that would assist speech pathologists to improve their skills when measuring severity of stuttering. Increased globalisation highlights the urgent need for researchers and clinicians to have best practice measurement guidelines for uniformity in the use of the severity rating scale across languages. It is hoped that future research with international collaborators will continue to build upon the existing data and resources, so that researchers and clinicians will have clearer guidelines for measuring severity of stuttering across different languages.
References


Appendix A: Study 1 Information Sheet

Information Sheet

THE EFFECT OF SPEECH SAMPLE DURATION ON THE RELIABILITY OF MEASUREMENT OF SEVERITY OF STUTTERING: A PILOT STUDY

This project will be conducted by the following Research Team:

Laura Hoffman (lhoffman@csu.edu.au)
Principal Investigator, Speech Pathologist & PhD Candidate
School of Community Health, Charles Sturt University, PO BOX 789, Albury, NSW, 2640, Australia

Dr Linda Wilson (liwilson@csu.edu.au; 02 6051 9257)
Research Supervisor, School of Community Health, Charles Sturt University

Dr Sally Hewat (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Research Supervisor, School of Humanities and Social Sciences, The University of Newcastle

Stuttering measurement has been developed to assist in the assessment of the frequency and severity of stuttering. Stuttering severity rating scales are commonly used to assess and measure severity of stuttering. A preliminary study, conducted by the current research team and colleagues, estimated the reliability of English-speaking speech pathologists’ use of the severity rating scale to measure severity of stuttering. Results suggested that not all participants were reliable when measuring the severity of stuttering in English. This is inconsistent with conclusions of previous research, which concluded that the severity rating scale could be used reliably with English samples by English-speaking participants. Therefore, further investigation into the use of this scale is warranted.

However, researchers and clinicians are increasingly faced with time constraints, thus potentially reducing their availability for participation in research studies, professional development sessions, and training and practice opportunities designed to improve clinical and research skills. Reducing time requirements for research and training may improve participation. In the case of measurement research and training, time reductions could be achieved by reducing the duration of each speech sample presented to participants. However, it is unknown whether such a reduction would impact on measurement reliability. Thus, the aim of this study is to investigate whether the duration of speech samples influences the interjudge reliability of measuring the severity of stuttering.

This research study will involve speech pathologists rating 27 speech samples varying in length (9 x 1 minute; 9 x 3 minute; 9 x 5 minute), using a 9-point severity rating scale. Your role in this study would be to listen to the audio speech samples, and assign a severity rating score to each of the speech samples using a 9-point scale.
If you agree to take part in this research study, please sign the consent form and return it to the principal investigator or research supervisors. After signing and returning the consent form, a Stimulus Package will be sent out to you. The package will include a CD containing audio speech samples, an instruction sheet, severity rating scale score sheets, a short questionnaire and a pre-paid envelope to return the package once completed. The Stimulus Package should take approximately 1½ hours to complete.

If you agree to participate in this study, you will be required to uphold the confidentiality of the clients in the speech samples by not showing the CD to other people and will be required to return the CD once you have completed the task.

If you are a qualified speech pathologist, have experience working with clients who stutter, and used the stuttering severity rating scales with clients, you are eligible to participate in the research study. If you have met the participant criteria and are interested in participating in this research study, please read and sign the consent form and return it to us by email or at the postal address provided above, by 2 May, 2014.

Participation in this research study is entirely voluntary. If you choose to take part in the study, you may withdraw at any time and will not be subjected to any penalty or discriminatory treatment for doing so. No identifying information about you will be revealed or published and confidentiality will be maintained throughout the entire research process. The data from this study may be used in future research.

If you have any questions, queries, or require further information, please do not hesitate to contact the Principal Investigator, Laura Hoffman, by phone, email or at the postal address provided on page 1.

Thank you for your time. We look forward to hearing from you.

NOTE: Charles Sturt University’s Human Research Ethics Committee has approved this project (405/2013/03). If you have any complaints or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer:

Executive Officer Human Research and Ethics Committee
School of Community Health, P.O Box 789, Albury, NSW, 2640

Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.

Please keep this information sheet for future reference.

Laura Hoffman
Speech Pathologist & PhD Candidate
Charles Sturt University

Dr Linda Wilson
Research Supervisor,
Charles Sturt University

Dr Sally Hewat
Research Supervisor,
The University of Newcastle
Appendix B: Study 1 Consent Form

Consent Form

THE EFFECT OF SPEECH SAMPLE DURATION ON THE RELIABILITY OF MEASUREMENT OF SEVERITY OF STUTTERING: A PILOT STUDY

This project will be conducted by the following Research Team:

Laura Hoffman (lhoffman@csu.edu.au)
Principal Investigator, Speech Pathologist & PhD Candidate
School of Community Health, Charles Sturt University, PO BOX 789, Albury, NSW, 2640, Australia

Dr Linda Wilson (liwilson@csu.edu.au; 02 6051 9257)
Research Supervisor, School of Community Health, Charles Sturt University

Dr Sally Hewat (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Research Supervisor, School of Humanities and Social Sciences, The University of Newcastle

1. I have read and understood the information sheet given to me and I have had the opportunity to ask questions about the research project and received satisfactory answers.
2. The purpose of this research project has been explained to me, including the (potential) risks/discomforts associated with the project.
3. I understand that any information or personal details gathered about me in the course of this research are confidential and that neither my name nor any other identifying information will be used or published without my written permission.
4. I agree not to discuss my rating scores with anyone else participating in the research project.
5. I understand that I am free to withdraw my participation in the research project at any time, and that if I do I will not be subjected to any penalty or discriminatory treatment.
6. I will not copy or show the contents of the CD to any other person and will return the CD immediately upon completion via the pre-paid envelope provided.

Print Name: ________________________________
Signature: ___________________________ Date: ______________

Please provide an address to which you want the Stimulus Package delivered:

Address: ........................................................................................................................................
NOTE: Charles Sturt University’s Human Research Ethics Committee has approved this project (405/2013/03). I understand that if I have any complaints or concerns about this research I can contact: Executive Officer Human Research and Ethics Committee School of Community Health, P.O Box 789, Albury, NSW, 2640
Appendix C: Study 1 Instruction Sheet

Instruction Sheet

THE EFFECT OF SPEECH SAMPLE DURATION ON THE RELIABILITY OF MEASUREMENT OF SEVERITY OF STUTTERING: A PILOT STUDY

Within 2 weeks of receipt please:

1. Review the Information Sheet
2. Complete the Questionnaire
3. Complete the Stimulus Package

Materials:
- CDs containing 27 speech samples
- Severity Rating Score Sheet

Complete this package alone in a quiet room, uninterrupted. You may wish to break completion of the package into multiple sessions. It is extremely important that you make your rating scores of each speech sample independently. Please do not consult with anyone or ask others for their opinion as to what score to assign to each speech sample.

Use the severity rating scale score sheet supplied in the package to record your responses. There are 27 severity rating scales presented on the score sheet, each numbered “1” through to “27”, to correspond with the total number of speech samples.

You will see and hear a number of samples of speech differing in length (1-minute; 3-minute; 5-minute). The order of presentation will be randomised. Each sample will be followed by 10 seconds of silence. Your task is to rate the severity of each sample using a 9-point scale. If the speech in a sample contains “extremely severe stuttering” circle the “9” on the scale. If the speech in a sample contains “no stuttering” circle the “1” on the scale. If the speech in the sample is somewhere between “extremely severe stuttering” and “no stuttering”, circle the appropriate number on the scale (see example below). Half scale values (e.g., 6.5) are not allowed. Do not hesitate to use either end of the scale if appropriate. “Severity” will not be defined for you. In making your judgements do not try to count stutters. Base your severity rating on your overall impression of the speech sample.
Example:

*Speech Sample 1*

1 2 3 4 5 6 7 8 9

<table>
<thead>
<tr>
<th>No stuttering</th>
<th>Extremely severe stuttering</th>
</tr>
</thead>
</table>

4. **Return Package**

Once you have completed rating all speech samples in the Stimulus Package, **please return:**

- [ ] the CD
- [ ] questionnaire
- [ ] score sheets

Return these items in the pre-paid envelope provided to:

Laura Hoffman  
School of Community Health,  
Charles Sturt University,  
PO BOX 789  
Albury, New South Wales, 2640

Thank you very much for your participation in the research study. Your contribution is greatly appreciated and it will assist us to provide evidence about whether the duration of a speech sample will influence the interjudge reliability of measuring the severity of stuttering.

**NOTE:** Charles Sturt University’s Human Research and Ethics Committee has approved this research project (405/2013/03). If you have any complaints or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer:

The Executive Officer  
Human Research and Ethics Committee  
School of Community Health  
Charles Sturt University  
Albury NSW 2640  
Tel: (02) 6051 9230  
Fax: (02) 6051 9230

Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.

Many thanks,

**Laura Hoffman**
Speech Pathologist & PhD Candidate  
Charles Sturt University

**Dr Linda Wilson**
Research Supervisor,  
Speech Pathology Lecturer  
Charles Sturt University

**Dr Sally Hewat**
Research Supervisor,  
Speech Pathology Senior Lecturer  
The University of Newcastle

www.csu.edu.au
CIRCO5 Provider Numbers for Charles Sturt University are 00005F (NSW), 01947G (VIC) and 02900B (ACT).  
ABN: 83 0178 708 551
Appendix D: Study 1 Severity Rating Score Sheets

Stimulus Package

SEVERITY RATING SCORE SHEET

THE EFFECT OF SPEECH SAMPLE DURATION ON THE RELIABILITY OF MEASUREMENT OF SEVERITY OF STUTTERING: A PILOT STUDY

Circle the appropriate number on the scale (see example below). Half scale values (e.g., 6.5) are not allowed.

Example Speech Sample

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 1

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 2

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering
Speech Sample 3

1 2 3 4 5 6 7 8 9
No stuttering  | Extremely severe stuttering

Speech Sample 4

1 2 3 4 5 6 7 8 9
No stuttering  | Extremely severe stuttering

Speech Sample 5

1 2 3 4 5 6 7 8 9
No stuttering  | Extremely severe stuttering

Speech Sample 6

1 2 3 4 5 6 7 8 9
No stuttering  | Extremely severe stuttering

Speech Sample 7

1 2 3 4 5 6 7 8 9
No stuttering  | Extremely severe stuttering
Speech Sample 8

1 2 3 4 5 6 7 8 9

No stuttering    Extremely severe stuttering

Speech Sample 9

1 2 3 4 5 6 7 8 9

No stuttering    Extremely severe stuttering

Speech Sample 10

1 2 3 4 5 6 7 8 9

No stuttering    Extremely severe stuttering

Speech Sample 11

1 2 3 4 5 6 7 8 9

No stuttering    Extremely severe stuttering

Speech Sample 12

1 2 3 4 5 6 7 8 9

No stuttering    Extremely severe stuttering
Speech Sample 18

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 19

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 20

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 21

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 22

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering
Speech Sample 23

1 2 3 4 5 6 7 8 9

No stuttering
Extremely severe stuttering

Speech Sample 24

1 2 3 4 5 6 7 8 9

No stuttering
Extremely severe stuttering

Speech Sample 25

1 2 3 4 5 6 7 8 9

No stuttering
Extremely severe stuttering

Speech Sample 26

1 2 3 4 5 6 7 8 9

No stuttering
Extremely severe stuttering

Speech Sample 27

1 2 3 4 5 6 7 8 9

No stuttering
Extremely severe stuttering

Thank you for completing this Stimulus Package.
Your participation is greatly appreciated.
Please return the entire Stimulus Package in the pre-paid envelope to:

Laura Hoffman, School of Community Health, Charles Sturt University,
PO BOX 789, Albury, NSW, 2640.
Appendix E: Study 1 Questionnaire

THE EFFECT OF SPEECH SAMPLE DURATION ON THE RELIABILITY OF MEASUREMENT OF SEVERITY OF STUTTERING: A PILOT STUDY

Name (Please print): …………………………………………………………………………………………………………………

Age: …………..

Gender: □ Male □ Female

1. Are you a native English speaker?
   □ Yes  □ No
   If no, please specify native language…………………………………………………………………………………………

2. Do you/or have you practiced speech pathology in English?
   □ Yes  □ No
   If you have NOT practiced speech pathology in English, DO NOT proceed. Please return the CD with the contents of the Stimulus Package in the pre-paid envelope.

3. Do you speak any language other than English?
   □ Yes  □ No
   If yes, please specify language/s ……………………………………………………………………………………………

4. How many years have you been a qualified speech pathologist? ………………………………………

5. Please estimate how many years you have worked with clients who stutter………………

6. Are you currently working with clients who stutter?
   □ Yes  □ No
   If so, please estimate how many clients you have assessed and/or treated with stuttering in the:
   Last 12 months ……………….  Last 3 months ………………. 
7. Are you familiar with the Severity Rating Scales used to measure stuttering?

☐ Yes  ☐ No

8. Have you used Severity Rating Scales as part of your work with clients who stutter?

☐ Yes  ☐ No

If you are NOT familiar and/or have NOT used the Severity Rating Scale as part of your work with clients who stutter, DO NOT proceed. Please return the CD with the contents of the Stimulus Package in the pre-paid envelope.

Thank you for taking the time to complete this questionnaire. Your confidentiality will be assured.

Please return this questionnaire in the pre-paid envelope along with the Stimulus Package to:

Laura Hoffman,
School of Community Health,
Charles Sturt University,
PO BOX 789, Albury, NSW, 2640
Participant Instruction Sheet

THE RELIABILITY OF A SEVERITY RATING SCALE TO MEASURE STUTTERING IN A FAMILIAR AND AN UNFAMILIAR LANGUAGE

Within 2 weeks of receipt please:

5. review the information sheet
6. complete the questionnaire
7. complete the stimulus package

Materials:
- CD containing 20 speech samples
- Severity Rating Score Sheet
- Questionnaire

Complete this package alone in a quiet room, uninterrupted. It is extremely important that you make your rating scores of each speech sample independently. Please do not consult with anyone or ask others for their opinion as to what score to assign to each speech sample.

Use the severity rating scale score sheet supplied in the package to record your responses. There are 20 severity rating scales presented on the score sheet, each numbered “1” through to “20”, to correspond with the total number of speech samples.

You will hear a number of samples. Each sample will be followed by 10 seconds of silence. Your task is to rate the severity of each sample using a 9-point scale. If the speech in a sample contains “extremely severe stuttering” circle the “9” on the scale. If the speech in a sample contains “no stuttering” circle the “1” on the scale. If the speech in the sample is somewhere between “extremely severe stuttering” and “no stuttering”, circle the appropriate number on the scale (see example below). Half scale values (e.g., 6.5) are not allowed. Do not hesitate to use either end of the scale if appropriate. “Severity” will not be defined for you. In making your judgements do not try to count stutters. Base your severity rating on your overall impression of the speech sample.
Example:

Speech Sample 1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>6</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No stuttering</td>
<td>Extremely severe stuttering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8. Return Package
Once you have completed rating all speech samples in the Stimulus Package, please return:

☐ the CD
☐ questionnaire
☐ score sheets

Return these items in the pre-paid envelope provided, to:
Laura Hoffman
School of Community Health,
Charles Sturt University,
PO BOX 789
Albury, New South Wales, 2640

Thank you very much for your participation in the research study. Your contribution is greatly appreciated and it will assist us to provide evidence about whether stuttering in a familiar and unfamiliar language can be reliably measured using severity rating scales.

NOTE: Charles Sturt University’s Faculty of Science Minimal Risk Human Ethics Committee has approved this project (405/2014/04). If you have any complaints or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee,
Locked Bag 49, Dubbo, NSW 2830.
Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.

Many thanks,

Laura Hoffman  
Principal Investigator  
Charles Sturt University

Dr Linda Wilson  
Research Supervisor  
Charles Sturt University

Dr Sally Hewat  
Research Supervisor  
The University of Newcastle
Appendix G: Study 2 Severity Rating Score Sheets

SEVERITY RATING SCORE SHEET

THE RELIABILITY OF A SEVERITY RATING SCALE TO MEASURE STUTTERING IN A FAMILIAR AND AN UNFAMILIAR LANGUAGE

Circle the appropriate number on the scale (see example below). Half scale values (e.g., 6.5) are not allowed.

Example Speech Sample

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>No stuttering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Extremely severe stuttering</td>
</tr>
</tbody>
</table>

Speech Sample 1

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>No stuttering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Extremely severe stuttering</td>
</tr>
</tbody>
</table>

Speech Sample 2

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>No stuttering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Extremely severe stuttering</td>
</tr>
<tr>
<td>Speech Sample 3</td>
<td>Speech Sample 4</td>
<td>Speech Sample 5</td>
<td>Speech Sample 6</td>
<td>Speech Sample 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 2 3 4 5 6 7 8 9</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No stuttering</td>
<td>No stuttering</td>
<td>No stuttering</td>
<td>No stuttering</td>
<td>No stuttering</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremely severe stuttering</td>
<td>Extremely severe stuttering</td>
<td>Extremely severe stuttering</td>
<td>Extremely severe stuttering</td>
<td>Extremely severe stuttering</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Speech Sample 8

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 9

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 10

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 11

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering

Speech Sample 12

1 2 3 4 5 6 7 8 9

No stuttering Extremely severe stuttering
<table>
<thead>
<tr>
<th>Speech Sample 13</th>
<th>Speech Sample 14</th>
<th>Speech Sample 15</th>
<th>Speech Sample 16</th>
<th>Speech Sample 17</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

No stuttering | Extremely severe stuttering

No stuttering | Extremely severe stuttering

No stuttering | Extremely severe stuttering

No stuttering | Extremely severe stuttering

No stuttering | Extremely severe stuttering
Speech Sample 18

1 2 3 4 5 6 7 8 9
No stuttering

Extremely severe stuttering

Speech Sample 19

1 2 3 4 5 6 7 8 9
No stuttering

Extremely severe stuttering

Speech Sample 20

1 2 3 4 5 6 7 8 9
No stuttering

Extremely severe stuttering
Appendix H: Study 2 Questionnaire

THE RELIABILITY OF A SEVERITY RATING SCALE TO MEASURE STUTTERING IN A FAMILIAR AND AN UNFAMILIAR LANGUAGE

Name (Please print): ………………………………………………………………………………………………………………………………

Age: …………………

Gender: □ Male    □ Female

1. Do you/or have you practiced speech pathology in English?
   □ Yes      □ No

   If you have NOT practiced speech pathology in English, DO NOT proceed. Please return the CD with the contents of the Stimulus Package in the pre-paid envelope.

2. Are you a native English speaker?
   □ Yes      □ No
   If no, please specify native language……………………………………………………………………………………………………..

3. Can you speak the Vietnamese language?
   □ Yes      □ A little     □ No

4. Do you understand the Vietnamese language?
   □ Yes      □ A little     □ No

   If you answered ‘Yes’ or ‘A little’ to question 2 or 3, DO NOT proceed. Please return the CD with the contents of the Stimulus Package in the pre-paid envelope.

5. Do you speak any language other than English?
   □ Yes      □ No
   If yes, please specify language/s …………………………………………………………………………………………………………

6. How many years have you been practising as a speech pathologist? ……………………………

7. Please tick the descriptors that apply to your current employment (tick as many that apply):

   Caseload  Setting
   □ Adult    □ Hospital
   □ Paediatric □ Rehabilitation units
   □            □ Private practice
   □            □ Community health centre
   □            □ Preschool/Kindergarten
   □            □ School – primary and/or secondary
   □            □ Specialist stuttering clinic
   □            □ Specialist disability services
   □            □ Other, please specify ……………………………………………
8. Please estimate how many clients you have assessed and/or treated with stuttering in the:

<table>
<thead>
<tr>
<th>Last 12 months</th>
<th>Last 3 months</th>
<th>Last week</th>
</tr>
</thead>
</table>

If you have NOT assessed and/or treated a stuttering client in the last 12 months, DO NOT proceed. Please return the CD with the contents of the Stimulus Package in the pre-paid envelope.

9. Have you used Severity Rating Scales as part of your work with clients who stutter?

- [ ] Yes
- [ ] No

If yes, please describe how you use the scale (tick as many that apply):

- [ ] Case load
- [ ] Frequency
  - [ ] Adults
  - [ ] Paediatrics
  - [ ] Routinely in every session with every client
  - [ ] With most clients
  - [ ] Intermittently with clients
  - [ ] Assessment and/or outcome measure only
  - [ ] Other, please specify………………………………………

10. Please rate yourself and identify what best describes your experience with using a stuttering severity rating scale (e.g., 10-point or 9-point scale) to measure the speech of people who stutter:

- [ ] No experience at all
- [ ] Minimal experience
- [ ] Somewhat experienced
- [ ] Very experienced
- [ ] Highly experienced

Thank you for taking the time to complete this questionnaire. Your confidentiality will be assured.

Please return this questionnaire in the pre-paid envelope along with the completed Stimulus Package to:

Laura Hoffman,
School of Community Health,
Charles Sturt University,
PO BOX 789, Albury, NSW, 2640.
Appendix I: Study 2 Second Rating Session Information Sheet

THE RELIABILITY OF A SEVERITY RATING SCALE TO MEASURE STUTTERING IN A FAMILIAR AND AN UNFAMILIAR LANGUAGE

This research will be conducted by the following Research Team:

**Laura Hoffman** (lhoffman@csu.edu.au)
Principal Investigator, Speech Pathologist & PhD Candidate
School of Community Health, Charles Sturt University, PO BOX 789, Albury, NSW, 2640, Australia

**Dr Linda Wilson** (liwilson@csu.edu.au; 02 6051 9257)
Research Supervisor, School of Community Health, Charles Sturt University

**Dr Sally Hewat** (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Research Supervisor, School of Humanities and Social Sciences, The University of Newcastle

One month ago you completed a Stimulus Package which was seeking to identify whether stuttering in a familiar and unfamiliar language could be measured reliably by speech pathologists. Please accept our sincere thanks for your participation thus far. This research study has examined interjudge reliability. Interjudge reliability involves examining the extent to which speech pathologists assign the same score to each speech sample. Your completion of the first rating session has assisted us greatly to determine interjudge reliability. This research study also aims to examine intrajudge reliability of speech pathologists’ ratings of speech samples. Intrajudge reliability is different from interjudge reliability as it refers to the extent to which the same person can rate the same sample consistently. Intrajudge reliability is important to establish as it is known to be a variable that can be related to judgement accuracy of stuttering. To help us examine intrajudge reliability, we ask that you please take the time to complete a second rating Stimulus Package sent out with this information sheet.

The second rating session will involve you rating the same 20 speech samples (10 English and 10 Vietnamese) as in the first rating session using a 9-point severity rating scale. Each sample is of three minutes duration. Your role in this study would be to listen to audio speech samples, and assign a rating score to each of the speech samples using the 9-point scale. This Stimulus Package is identical to the first rating session, except that the order of the speech samples has been changed. The Stimulus Package should take approximately one hour to complete.

No identifying information about you will be revealed or published and confidentiality will be maintained throughout the entire research process. The data from this study may be used in future research. Participation in this research study is entirely voluntary. If you decide that you do not want to continue and complete the second rating session, you may withdraw from the study and will not be subjected to any penalty or discriminatory treatment for doing so. If you have chosen to withdraw your participation, please return the Stimulus Package immediately via the pre-paid envelope.

If you have any questions, queries, or require further information, please do not hesitate to contact us by phone, email or at the postal address provided.

This research study is funded by a Speech Pathology Australia Higher Degree Research Grant.
We are especially grateful for your participation in this study. Your contributions will assist us to identify whether a 9-point severity rating scale is a reliable clinical tool to use when working with clients from different linguistic backgrounds.

**NOTE:** Charles Sturt University's Faculty of Science Minimal Risk Human Ethics Committee has approved this project (405/2014/04). If you have any complaints or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer: Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830, Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327. Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.

Please keep this information sheet for future reference.

Laura Hoffman  
Principal Investigator  
Charles Sturt University

Dr Linda Wilson  
Research Supervisor  
Charles Sturt University

Dr Sally Hewat  
Research Supervisor  
The University of Newcastle
Appendix J: Study 2 Second Rating Session Consent Form

THE RELIABILITY OF A SEVERITY RATING SCALE TO MEASURE STUTTERING IN A FAMILIAR AND AN UNFAMILIAR LANGUAGE

This research will be conducted by the following Research Team:

**Laura Hoffman** (lhoffman@csu.edu.au)
Principal Investigator, Speech Pathologist & PhD Candidate
School of Community Health, Charles Sturt University, PO BOX 789, Albury, NSW, 2640, Australia

**Dr Linda Wilson** (liwilson@csu.edu.au; 02 6051 9257)
Research Supervisor, School of Community Health, Charles Sturt University

**Dr Sally Hewat** (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Research Supervisor, School of Humanities and Social Sciences, The University of Newcastle

I have read and understood the information sheet given to me and I have had the opportunity to ask questions about the research project and received satisfactory answers.

The purpose of this research project has been explained to me, including the (potential) risks/discomforts associated with the project.

I understand that any information or personal details gathered about me in the course of this research are confidential and that neither my name nor any other identifying information will be used or published without my written permission.

I agree not to discuss my rating scores with anyone else participating in the research project.

I understand that I am free to withdraw my participation in the research project at any time, and that if I do I will not be subjected to any penalty or discriminatory treatment.

I will not copy or show the contents of the CD to any other person and will return the CD immediately upon completion via the pre-paid envelope provided.

Print Name: ____________________________
Signature: ____________________________ Date: __________________

**NOTE:** Charles Sturt University’s Faculty of Science Minimal Risk Human Ethics Committee has approved this project (405/2014/04). If you have any complaints or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830, Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.
Appendix K: Study 2 Second Rating Session Instruction Sheet

Participant Instruction Sheet:

Second Rating Session

THE RELIABILITY OF A SEVERITY RATING SCALE TO MEASURE STUTTERING IN A FAMILIAR AND AN UNFAMILIAR LANGUAGE

Within 2 weeks of receipt please:

1. review the information sheet
2. sign the consent form (if you choose to consent)
3. complete the stimulus package

Materials:
- CD containing 20 speech samples
- Severity Rating Score Sheet

Complete this package alone in a quiet room, uninterrupted. It is extremely important that you make your rating scores of each speech sample independently. Please do not consult with anyone or ask others for their opinion as to what score to assign to each speech sample.

Use the severity rating scale score sheet supplied in the package to record your responses. There are 20 severity rating scales presented on the score sheet, each numbered “1” through to “20”, to correspond with the total number of speech samples.

You will hear a number of samples. Each sample will be followed by 10 seconds of silence. Your task is to rate the severity of each sample using a 9-point scale. If the speech in a sample contains “extremely severe stuttering” circle the “9” on the scale. If the speech in a sample contains “no stuttering” circle the “1” on the scale. If the speech in the sample is somewhere between “extremely severe stuttering” and “no stuttering”, circle the appropriate number on the scale (see example below). Half scale values (e.g., 6.5) are not allowed. Do not hesitate to use either end of the scale if appropriate. “Severity” will not be defined for you. In making your judgements do not try to count stutters. Base your severity rating on your overall impression of the speech sample.
Example:

**Speech Sample 1**

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

No stuttering | Extremely severe stuttering

4. **Return Package**

Once you have completed rating all speech samples in the Stimulus Package, please return:

- ☐ the CD
- ☐ the consent form
- ☐ score sheets

Return these items in the pre-paid envelope provided, to:

Laura Hoffman  
School of Community Health,  
Charles Sturt University,  
PO BOX 789  
Albury, New South Wales, 2640

Thank you very much for your participation in the research study. Your contribution is greatly appreciated and it will assist us to provide evidence about whether stuttering in a familiar and unfamiliar language can be reliably measured using severity rating scales.

**NOTE:** Charles Sturt University’s Faculty of Science Minimal Risk Human Ethics Committee has approved this project (405/2014/04). If you have any complaints or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830.  
Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.

Many thanks,

Laura Hoffman  
Principal Investigator  
Charles Sturt University  

Dr Linda Wilson  
Research Supervisor  
Charles Sturt University  

Dr Sally Hewat  
Research Supervisor  
The University of Newcastle
Participant Information Sheet

THE RELIABILITY OF A SEVERITY RATING SCALE TO MEASURE STUTTERING IN A FAMILIAR AND AN UNFAMILIAR LANGUAGE

This research will be conducted by the following Research Team:

Laura Hoffman (lhoffman@csu.edu.au)
Principal Investigator, Speech Pathologist & PhD Candidate
School of Community Health, Charles Sturt University, PO BOX 789, Albury, NSW, 2640, Australia

Dr Linda Wilson (liwilson@csu.edu.au; 02 6051 9257)
Research Supervisor, School of Community Health, Charles Sturt University

Dr Sally Hewat (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Research Supervisor, School of Humanities and Social Sciences, The University of Newcastle

The aim of this research study is to determine whether stuttering in a familiar (English) and unfamiliar (Vietnamese) language can be measured reliably by speech pathologists. With increasing multiculturalism, speech pathologists are likely to work with stuttering clients from linguistic backgrounds that differ from their own. It is important for speech pathologists to be able to accurately and reliably assess and measure the speech of people who speak another language and stutter.

This research study will involve speech pathologists rating 20 speech samples (10 English and 10 Vietnamese) of three minutes duration, using a 9-point severity rating scale. Your role in this study would be to listen to audio speech samples, and assign a severity rating score to each of the speech samples using a 9-point scale. Your time commitment would be approximately one hour to complete the Stimulus Package tasks.

No identifying information about you will be revealed or published and confidentiality will be maintained throughout the entire research process. The data from this study may be used in future research.

If you are a qualified practicing speech pathologist who has practiced in English, have worked with a stuttering client in the past 12 months, and do not speak or understand Vietnamese, you are eligible to participate in the research study. If you have met the participant criteria and are interested in participating in this research study, please read and sign the consent form and return it to us by email or at the postal address provided above.

Participation in this research study is entirely voluntary. If you choose to take part in the study, you may withdraw at any time and will not be subjected to any penalty or discriminatory treatment for doing so. If you have any questions, queries, or require further information, please do not hesitate to contact us by phone, email or at the postal address provided.

This research study is funded by a Speech Pathology Australia Higher Degree Research Grant.

Thank you for considering this invitation.
NOTE: Charles Sturt University's Faculty of Science Minimal Risk Human Ethics Committee has approved this project (405/2014/04). If you have any complaints or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer: Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830.
Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327
Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.

Please keep this information sheet for future reference.

Laura Hoffman  
Principal Investigator  
Charles Sturt University

Dr Linda Wilson  
Research Supervisor  
Charles Sturt University

Dr Sally Hewat  
Research Supervisor  
The University of Newcastle
Appendix M: Study 2 First Rating Session Consent Form

THE RELIABILITY OF A SEVERITY RATING SCALE TO MEASURE STUTTERING IN A FAMILIAR AND AN UNFAMILIAR LANGUAGE

This research will be conducted by the following Research Team:

**Laura Hoffman** (lhoffman@csu.edu.au)  
Principal Investigator, Speech Pathologist & PhD Candidate  
School of Community Health, Charles Sturt University, PO BOX 789, Albury, NSW, 2640, Australia

**Dr Linda Wilson** (liwilson@csu.edu.au; 02 6051 9257)  
Research Supervisor, School of Community Health, Charles Sturt University

**Dr Sally Hewat** (Sally.Hewat@newcastle.edu.au; 02 4921 5159)  
Research Supervisor, School of Humanities and Social Sciences, The University of Newcastle

I have read and understood the information sheet given to me and I have had the opportunity to ask questions about the research project and received satisfactory answers.

The purpose of this research project has been explained to me, including the (potential) risks/discomforts associated with the project.

I understand that any information or personal details gathered about me in the course of this research are confidential and that neither my name nor any other identifying information will be used or published without my written permission.

I agree not to discuss my rating scores with anyone else participating in the research project.

I understand that I am free to withdraw my participation in the research project at any time, and that if I do I will not be subjected to any penalty or discriminatory treatment.

I will not copy or show the contents of the CD to any other person and will return the CD immediately upon completion via the pre-paid envelope provided.

Print Name:  
Signature:  
Date: 

Please provide the address to which you want the Stimulus Package delivered:
NOTE: Charles Sturt University’s Faculty of Science Minimal Risk Human Ethics Committee has approved this project (405/2014/04). If you have any complaints or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830, Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.
Bản thông tin dành cho người tham gia

ĐỘ TIN CẬY CỦA VIỆC SỬ DỤNG THANG ĐO ĐO NĂNG ĐỂ ĐO LƯỢNG MỤC NÓI LẤP Ở VIỆT NAM

Trong vòng 2 tuần sau khi nhận được, xin vui lòng:

1. Xem lại bản thông tin
2. Hoàn thành bảng câu hỏi
3. Hoàn thành gói mẫu kích thích

Từ liều:
- CD có 20 mẫu lời nói
- Bản cho điểm độ nặng

Hoàn thành gói mẫu này một mình trong một phòng yên tĩnh, không bị xao lạc. Việc bạn cho điểm mỗi mẫu nghiên cứu cách đọc là rất quan trọng. Xin không hỏi ý kiến hoặc tư vấn với ai về ý kiến của họ về việc nên cho mấy điểm cho mỗi mẫu đánh giá.

Sử dụng bản cho điểm độ nặng được cung cấp trong gói mẫu để ghi lại các đánh giá của bạn. Có 20 thang được trình bày trong bản cho điểm, mỗi thang được đánh số từ “1” đến “20”, tương ứng với tổng số mẫu lời nói.

Vì dụ:

Mẫu 1

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Không lập</td>
<td>Lập rất nặng</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Gửi lại gói mẫu kích thích

Khi bạn đã hoàn thành việc cho điểm tất cả các mẫu lỗi nói trong Gói mẫu Kích thích, xin hãy gửi lại:

- CD
- bằng câu hỏi
- bản cho điểm

Gửi lại các vật này qua phong bì được trả tiền trước đến:
Laura Hoffman
School of Community Health,
Charles Sturt University,
PO BOX 789
Albury, New South Wales, 2640, Australia

Cảm ơn bạn rất nhiều vì sự tham gia trong nghiên cứu. Đồng góp của bạn rất giá trị và sẽ hỗ trợ chúng tôi về cung cấp chứng cứ cho việc xem xét việc đo lường nói lập bằng thang đo độ nặng có đáng tin cậy hay không.

LUU Y: Ứy ban Đạo đức về Giám thiều Nguy cơ cho Con người của Bộ môn Khoa học thuộc Đại học Charles Sturt đã chấp thuận để an nghiên cứu này (405/2014/03). Nếu bạn có bất cứ thắc mắc hoặc nghi ngờ về tính đạo đức trong quá trình tiến hành nghiên cứu này, bạn có thể liên hệ Ứy ban qua Trường phong:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830,

Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Bất kỳ vấn đề nào mà bạn nếu ra sẽ được giữ kin và tìm hiểu toàn diện và bạn sẽ được thông báo kết quả.

Xin cảm ơn rất nhiều,

Laura Hoffman
Nghiên cứu viên chính
Charles Sturt University

Dr Linda Wilson
Người hướng dẫn nghiên cứu
Charles Sturt University

Dr Sally Hewat
Người hướng dẫn nghiên cứu
The University of Newcastle

Dr Thao Huynh
Công sự nghiên cứu
Cho Ray Hospital
Gợi mẫu kích thích

BÀN CHO ĐIỆM ĐỘ NẲNG

ĐƠN TIN CẢY CỦA VIỆC SỬ DỤNG THANG ĐO ĐỘ NẲNG ĐỂ ĐO LƯỢNG MỨC NỔI LẤP Ở VIỆT NAM

Khoanh tròn số phù hợp trên thang điểm (xem ví dụ bên dưới). Không được phép dùng các trị số ở khoảng giữa (như 6.5).

Ví dụ mẫu lối nói

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khổng lập</td>
<td>Lập rất nặng</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mẫu 1

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khổng lập</td>
<td>Lập rất nặng</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mẫu 2

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khổng lập</td>
<td>Lập rất nặng</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mẫu 3

1 2 3 4 5 6 7 8 9

Không lập                  Lập rất nằng

Mẫu 4

1 2 3 4 5 6 7 8 9

Không lập                  Lập rất nằng

Mẫu 5

1 2 3 4 5 6 7 8 9

Không lập                  Lập rất nằng

Mẫu 6

1 2 3 4 5 6 7 8 9

Không lập                  Lập rất nằng

Mẫu 7

1 2 3 4 5 6 7 8 9

Không lập                  Lập rất nằng
Mẫu 8

1 2 3 4 5 6 7 8 9

Không lặp                                      Lập rất nằng

Mẫu 9

1 2 3 4 5 6 7 8 9

Không lặp                                      Lập rất nằng

Mẫu 10

1 2 3 4 5 6 7 8 9

Không lặp                                      Lập rất nằng

Mẫu 11

1 2 3 4 5 6 7 8 9

Không lặp                                      Lập rất nằng

Mẫu 12

1 2 3 4 5 6 7 8 9

Không lặp                                      Lập rất nằng
### Mẫu 13

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
</table>

Không lập | Lập rất năng

### Mẫu 14

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
</table>

Không lập | Lập rất năng

### Mẫu 15

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
</table>

Không lập | Lập rất năng

### Mẫu 16

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
</table>

Không lập | Lập rất năng

### Mẫu 17

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
</table>

Không lập | Lập rất năng
**Mẫu 18**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Không lập</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lập rất năng</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Mẫu 19**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Không lập</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lập rất năng</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Mẫu 20**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Không lập</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lập rất năng</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix P: Study 3 Questionnaire

BẢNG CÂU HỎI

DỌ TIN CẤY CỦA VIỆC SỬ DỤNG THANH DO ĐỘ NĂNG ĐẾ DO LƯỠNG MỨC NỘI LẤP Ở VIỆT NAM

Tên (Xin viết hoa): ........................................................................................................

Tuổi: ............

Giới: ☐ nam ☐ nữ

1. Bạn có / hoặc đang thực hành âm ngữ trĩ liệu với ngôn ngữ tiếng Việt không?
   ☐ có ☐ không
   
   Nêu bạn KHÔNG thực hành âm ngữ trĩ liệu với ngôn ngữ tiếng Việt, XIN KHÔNG tiếp tục hoàn thành bảng câu hỏi. Xin gửi lại CD với các thành phần của Gói màu kích thích trong phong bì đã được trả tiện trước.

2. Bạn có phải là người nói tiếng Việt bản xứ không?
   ☐ Có ☐ Không
   Nếu không, xin ghi rõ ngôn ngữ bản xứ (tiếng mẹ đẻ) của bạn................................................

3. Bạn có nói tiếng nào khác ngoài tiếng Việt không?
   ☐ Có ☐ Không
   Nếu có, xin ghi rõ ngôn ngữ nào (một hoặc nhiều) ............................................................

4. Bạn đã thực hành âm ngữ trĩ liệu bao nhiêu tháng / năm rồi?
   .................................................................................................................................

5. Xin đánh vào các mô tả phù hợp với công việc hiện tại của bạn (chọn nhiều lựa chọn nếu thích hợp)
   
   Các bệnh
   ☐ Người lớn
   ☐ Trẻ con
   ☐ Người lớn và trẻ con

   Bối cảnh
   ☐ Bệnh viện
   ☐ Các đơn vị phục hồi chức năng
   ☐ Thực hành tự nhiên
   ☐ Trung tâm sức khỏe cộng đồng
   ☐ Nhà trẻ/ Trường mẫu giáo
   ☐ Trường học – tiểu học và/hoặc cấp hai
   ☐ Phường khám chuyển về nơi cấp
   ☐ Các dịch vụ chuyển bị tổ chức kết thúc
   ☐ Khác, xin ghi cụ thể .................................................................
7. Hãy ước tính bao nhiêu bệnh nhân mà bạn đã khám và/hoặc điều trị về nơi cấp trong vòng:

12 tháng qua .....................  3 tháng qua .....................  Trước đây .....................

8. Bạn đã sử dụng Thang do độ nặng trong công việc của bạn khi gặp bệnh nhân nơi cấp?
☐ Có  ☐ Không

Nếu có, xin mô tả bạn sử dụng thang này như thế nào:

- Tần suất
  ☐ Sử dụng thường xuyên trong mỗi buổi làm việc với mỗi bệnh nhân nơi cấp
  ☐ Thường xuyên với bệnh nhân nơi cấp
  ☐ Chỉ sử dụng tại chỗ hoặc do hướng kết quả
  ☐ Khác, xin ghi cụ thể..........................
  ........................................................................

- Ca bệnh
  ☐ Người lớn
  ☐ Trẻ con
  ☐ Người lớn và trẻ con

9. Xin tự đánh giá và xác định điều nào mở tả tốt nhất về kinh nghiệm của bạn trong việc sử dụng thang do độ nặng nơi cấp ( như là thang 10 điểm hoặc thang 9 điểm ) để đo lường lợi nơi của người nơi cấp:

☐ Hoàn toàn không có kinh nghiệm
☐ Có ít kinh nghiệm
☐ Một vài kinh nghiệm
☐ Rất kinh nghiệm
☐ Chuyên sâu

Cảm ơn bạn đã dành thời gian hoàn thành bảng câu hỏi này. Thông tin về bạn sẽ hoàn toàn được bảo mật.

Xin gửi lại bảng câu hỏi này cùng với các phần khác của Gói mẫu kích thích đến:
Laura Hoffman, School of Community Health, Charles Sturt University, PO BOX 789, Albury, NSW, 2640
Bàn thông tin cho người tham gia lần đánh giá thứ hai

ĐÔ TIN CẬY CỦA VIỆC SỬ DỤNG THANG ĐO ĐỘ NANG ĐẾ ĐO LƯỢNG MỨC NÓI LẤP Ở VIỆT NAM

Nghiên cứu này sẽ được thực hiện bởi Nhóm nghiên cứu sau đây:

Laura Hoffman (lhoffman@csu.edu.au)
Người nghiên cứu chính, Chuyên viên Âm ngữ trị liệu và Nghiên cứu sinh tiến sĩ
Trường Sức khỏe Công động, Đại học Charles Sturt, PO BOX 789, Albury, NSW, 2640, Australia

Tiến sĩ Linda Wilson (liwilson@csu.edu.au; 02 6051 9257)
Người hướng dẫn nghiên cứu, Trưởng Sức khỏe Công động, Đại học Charles Sturt

Tiến sĩ Sally Hewat (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Người hướng dẫn nghiên cứu, Trưởng Khoa học xã hội và nhân văn, Đại học Newcastle

Bác sĩ Huỳnh Bích Thảo (huynhbthao@gmail.com; 84 919 000 369)
Công sự nghiên cứu, Bác sĩ và Chuyên viên Âm ngữ trị liệu, Bệnh viện Chợ Rẫy

Một tháng trước, bạn đã hoàn thành một Gói mẫu kích thích được dùng để xác định xem việc đo lượng mức nói lớp 13 của chuyên viên Âm ngữ trị liệu có đáng tin cậy hay không. Xin hãy nhận lời cảm ơn chân thành của chúng tôi về sự tham gia của bạn cho đến nay. Nghiên cứu này kiểm tra độ tin cậy giữa các nhà đánh giá. Độ tin cậy giữa các người đánh giá liên quan đến việc kiểm tra xem mức độ mà các chuyên viên Âm ngữ trị liệu cho điểm giống nhau cho mỗi mẫu lời nói. Việc bạn hoàn thành lần đánh giá đầu tiên đã hỗ trợ chúng tôi rất nhiều để xác định độ tin cậy giữa các nhà đánh giá. Nghiên cứu này cũng hướng đến độ tin cậy nội tại của mỗi chuyên viên Âm ngữ trị liệu khi cho điểm các mẫu đánh lời nói. Độ tin cậy nội tại của người đánh giá thì khác với độ tin cậy giữa các nhà đánh giá vì nó liên hệ đến mức độ mà cùng một người có thể cho điểm cùng một mẫu một cách hàng định. Việc xác định độ tin cậy nội tại của một người đánh giá thì liên quan trong ví như là một biểu số mà có thể ảnh hưởng đến tính chính xác khi đánh giá nói lớp. Để giúp chúng tôi kiểm tra độ tin cậy nội tại của một người đánh giá, chúng tôi yêu cầu bạn hãy vui lòng đánh thời gian để hoàn thành lần đánh giá thứ hai cho Gói mẫu kích thích.

Lần đánh giá thứ hai sẽ bao gồm việc bạn đánh giá cùng 20 mẫu như lần đầu với thang đo lượng đo nằng từ 1 đến 9 điểm. Mỗi mẫu dài 3 phút. Vai trò của bạn trong nghiên cứu này là nghe các mẫu ghi âm lời nói, và cho điểm cho mỗi mẫu sử dụng thang 9 điểm. Gói mẫu kích thích này cũng là gói mẫu kích thích của lần đánh giá trước nhưng thứ tự của các mẫu đã thay đổi. Gói mẫu kích thích này mất khoảng 1 giờ đồng hồ để hoàn thành.
Không có thông tin xác định danh tính của bạn sẽ được tiết lộ hoặc công bố và sự bảo mật sẽ được duy trì qua suốt quá trình nghiên cứu. Đủ liệu từ nghiên cứu này có thể được sử dụng cho các nghiên cứu trong tương lai. Sự tham gia của bạn trong nghiên cứu này là hoàn toàn tự nguyện. Nếu bạn quyết định không muốn tiếp tục và hoàn thành lần nghiên cứu thứ hai, bạn có thể rút ra khỏi nghiên cứu và sẽ không bị phạt hoặc phân biệt đối xử cho việc rút khỏi nghiên cứu.

Nếu bạn có bất kỳ câu hỏi, thắc mắc hoặc cần thêm thông tin, xin vui lòng liên hệ chúng tôi mà không phải ngăn gì, qua điện thoại, email hoặc địa chỉ email đã được cung cấp.

Nghiên cứu này được tài trợ bởi một Quỹ Nghiên cứu Cao học của Hội Âm ngữ tri liệu Úc.

Chúng tôi rất biết ơn sự tham gia của bạn. Các đóng góp của bạn sẽ hỗ trợ chúng tôi xác định xem tháng 9 điểm để đánh giá độ nặng có phải là công cụ làm sàng đáng dùng hay không.

**LUU Ý:** Ủy ban Đạo đức về Giảm thiểu Nguy cơ cho Con người của Bộ môn Khoa học thuộc Đại học Charles Sturt đã chấp thuận đề án nghiên cứu này (405/2014/03). Nếu bạn có bất cứ phiền hoắc e ngại nào về tính đạo đức trong quá trình tiến hành nghiên cứu này, bạn có thể liên hệ Ủy ban qua Trưởng phòng:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830,
Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Bất kỳ vấn đề nào mà bạn nêu ra sẽ được giữ kín và tìm hiểu toàn diện và bạn sẽ được thông báo kết quả.

Xin giữ bản thông tin này để tham khảo trong tương lai.

---

Laura Hoffman  
Nghiên cứu viên chính  
Charles Sturt University

Dr Linda Wilson  
Người hướng dẫn nghiên cứu  
Charles Sturt University

Dr Sally Hewat  
Người hướng dẫn nghiên cứu  
The University of Newcastle

Dr Thao Huynh  
Công sứ nghiên cứu  
Cho Ray Hospital
Appendix R: Study 3 Second Rating Session Consent Form

Mẫu chấp thuận tham gia nghiên cứu

ĐO TIN CẬY CỦA VIỆC SỬ DỤNG THANG ĐO ĐỘ NĂNG ĐỂ ĐO LUONG MỨC NỘI LẤP Ở VIỆT NAM

Nghiên cứu này sẽ được thực hiện bởi Nhóm nghiên cứu sau đây:

Laura Hoffman (lhoffman@csu.edu.au)
Người nghiên cứu chính, Chuyên viên Âm ngữ trị liệu và Nghiên cứu sinh tiến sĩ
Trường Sức khỏe Cộng đồng, Đại học Charles Sturt, PO BOX 789, Albury, NSW, 2640, Australia

Tiến sĩ Linda Wilson (liwilson@csu.edu.au; 02 6051 9257)
Người hướng dẫn nghiên cứu, Trường Sức khỏe Cộng đồng, Đại học Charles Sturt

Tiến sĩ Sally Hewat (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Người hướng dẫn nghiên cứu , Trường Khoa học xã hội và nhân văn, Đại học Newcastle

Bác sĩ Huỳnh Bích Thảo (huynhbthao@gmail.com; 84 919 000 369)
Công sự nghiên cứu, Bác sĩ và Chuyên viên Âm ngữ trị liệu, Bệnh viện Chợ Rẫy

Tôi đã đọc và hiểu bản thông tin được đưa cho tôi và tôi đã có cơ hội để đặt câu hỏi về dự án nghiên cứu và đã nhận được câu trả lời từ nhóm nghiên cứu.

Mục tiêu của nghiên cứu đã được giải thích cho tôi, bao gồm các nguy cơ/ bất tiện ( có thể có ) đi kèm với dự án nghiên cứu.

Tôi hiểu rằng bất cứ thông tin hoặc chi tiết về cá nhân tôi có được trong lúc làm nghiên cứu là bảo mật và không có phần nào kể cả tên hoặc thông tin định danh nào sẽ được dùng hoặc công bố mà không có sự cho phép tôi trên văn bản giấy tờ.

Tôi đồng ý không thảo luận số điểm mà tôi đánh giá với bất kỳ ai khác tham gia trong nghiên cứu này.

Tôi hiểu rằng tôi có tự do để rút khỏi nghiên cứu bất kỳ lúc nào, và nếu tôi rút ra thì tôi sẽ không chịu bất kỳ hình phạt hay sự phân biệt đối xử nào.

Tôi sẽ không sao chép hoặc đưa cho bất kỳ ai xem nội dung của CD và sẽ gửi lại CD ngay khi tôi hoàn tất qua phong bì đã được trả tiền trước.

Tôi đã có cơ hội để hỏi và có câu trả lời thỏa đáng.
Tên: __________________________________________________________

Chữ ký____________________________________Ngày:__________________

LUU Y: Ủy ban Đạo đức về Giảm thiểu Nguy cơ cho Con người của Bộ môn Khoa học thuộc Đại học Charles Sturt đã chấp thuận đề án nghiên cứu này (405/2014/03). Nếu bạn có bất cứ than phiền hoặc e ngại nào về tính đạo đức trong quá trình tiến hành nghiên cứu này, bạn có thể liên hệ Ủy ban qua Trường phong:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830,
Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Bất kỳ vấn đề nào mà bạn nêu ra sẽ được giữ kín và tìm hiểu toàn diện và bạn sẽ được thông báo kết quả.
Appendix S: Study 3 Second Rating Session Instructions

Lần chấm điểm thứ hai cho nghiên cứu

Dear ...

Tôi là Laura Hoffman, chuyên viên âm ngữ trị liệu và là nghiên cứu sinh tiến sĩ ở Đại học Charles Sturt, Australia.

Một tuần trước, anh/chị đã hoàn thành một bảng đánh giá cho nghiên cứu về độ lượng nói lập cho các mẫu tiếng Anh và tiếng Việt. Cảm ơn anh/chị rất nhiều vì đã đóng góp cho đến nay. Tôi rất quan trọng sự tham gia của anh/chị.

Chúng tôi hiện đang cố gắng hoàn thành giai đoạn cuối cùng của dự án. Nghiên cứu này yêu cầu anh/chị hoàn thành bảng thứ hai và cũng là bảng cuối cùng về đánh giá các mẫu giọng nói được sắp xếp theo thứ tự khác. Việc đánh giá lần thứ hai và cũng là lần cuối cùng này sẽ giúp xem xét độ tin cậy của chuyên viên âm ngữ trị liệu theo thời gian – độ tin cậy nội tại.

Tôi xin chân thành cảm ơn anh/chị có thể dành thời gian để hoàn thành bảng đánh giá lần thứ hai này.

Xin xem trong phần định kèm 4 tài liệu:
1. Bản thông tin
2. Phiếu chấp thuận (xin kí tên và gửi lại)
3. Bản hướng dẫn
4. Bảng chấm điểm (xin gửi lại)

20 mẫu giọng được tạo từ thư mạng đến sau, xin vui lòng bấm vào:


Khi đã hoàn thành bảng chấm điểm, xin scan và gửi email Phiếu chấp thuận và Bảng chấm điểm cho tôi (lhoffman@csu.edu.au) hoặc BS.Thảo.

Nếu có bất kỳ câu hỏi nào, xin liên hệ tôi hoặc BS.Thảo.

Sự tham gia trong nghiên cứu của anh/chị là điều chúng tôi rất trân trọng

Xin cảm ơn,
Laura Hoffman
Speech Therapist and Research Student
Charles Sturt University
Bản thông tin dành cho người tham gia

ĐỘ TIN CÁY CỦA VIỆC SỬ DỤNG THANG ĐO ĐỘ NẰNG ĐỀ ĐO LƯƠNG MỤC Ô VIỆT NAM

Nghiên cứu này sẽ được thực hiện bởi Nhóm nghiên cứu sau đây:

Laura Hoffman (lhoffman@csu.edu.au)
Người nghiên cứu chính, Chuyên viên Âm ngữ trị liệu và Nghiên cứu sinh tiến cử Trường Sức khỏe Công đồng, Đại học Charles Sturt, PO BOX 789, Albury, NSW, 2640, Australia

Tiến sĩ Linda Wilson (liwilson@csu.edu.au; 02 6051 9257)
Người hướng dẫn nghiên cứu, Trường Sức khỏe Công đồng, Đại học Charles Sturt

Tiến sĩ Sally Hewat (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Người hướng dẫn nghiên cứu, Trường Khoa học xã hội và nhân văn, Đại học Newcastle

Bác sĩ Huỳnh Bích Thảo (huynhbthao@gmail.com; 84 919 000 369)
Công sự nghiên cứu, Bác sĩ và Chuyên viên Âm ngữ trị liệu, Bệnh viện Chợ Rẫy

Nghiên cứu trong lĩnh vực nói lập và dự đoán tình tin cậy của việc đo lường đã được thực hiện với người chì nói tiếng Anh. Không có nghiên cứu nào được thực hiện để khám phá việc sử dụng sự đo lường mức độ nói lập ở Việt Nam. Do đó, mục tiêu của nghiên cứu này là xác định xem nói lập trong tiếng Việt và tiếng Anh có thể được đo lường một cách đáng tin cậy bởi các chuyên viên âm ngữ trị liệu hay không.

Nghiên cứu này bao gồm việc các nhà âm ngữ trị liệu đánh giá 10 mẫu nói tiếng Anh và 10 mẫu nói tiếng Việt, là các mẫu lời nói dài 3 phút, sử dụng thang đo độ nhật từ 1 đến 9 điểm. Vai trò của bạn trong nghiên cứu này là nghe các mẫu nói, rồi đánh giá mức độ ngữ âm, và cho điểm độ nặng cho mỗi mẫu bằng thang 9 điểm.

Không có thông tin xác định danh tính của bạn sẽ được tiết lộ hoặc cung cấp và sự bảo mật sẽ được duy trì qua suốt quá trình nghiên cứu. Dự liệu từ nghiên cứu này có thể được sử dụng cho các nghiên cứu trong tương lai.

Nếu bạn là một chuyên viên âm ngữ trị liệu được chứng nhận và/hoặc đang thực tập trong lĩnh vực âm ngữ trị liệu mà sử dụng tiếng Việt, thì bạn thích hợp để tham gia nghiên cứu này.

Nếu bạn đủ điều chuẩn tham gia nghiên cứu và có quan tâm tham gia trong nghiên cứu này, xin liên lạc trực tiếp với thành viên của nhóm nghiên cứu qua thông tin liên lạc bên trên.
Sự tham gia của bạn trong nghiên cứu này là hoàn toàn tự nguyện. Nếu bạn quyết định không muốn tiếp tục và hoàn thành lần nghiên cứu thứ hai, bạn có thể rút ra khỏi nghiên cứu và sẽ không bị phạt hoặc phân biệt đối xử cho việc rút khỏi nghiên cứu. Nếu bạn có bất kỳ câu hỏi, thắc mắc hoặc cần thêm thông tin, xin vui lòng liên hệ chúng tôi mà không phải ngần ngại, qua điện thoại, email hoặc địa chỉ email đã được cung cấp.

Nghiên cứu này được tài trợ bởi một Quỹ Nghiên cứu Cao học của Hội Âm ngữ trị liệu Úc.

Cảm ơn bạn đã cân nhắc về việc tham gia nghiên cứu.

**Lưu ý:** Ủy ban Đạo đức về Giảm thiểu Nguy cơ của Bộ môn Khoa học thuộc Đại học Charles Sturt đã chấp thuận đề án nghiên cứu này (405/2014/03). Nếu bạn có bất cứ phiền hoắc e ngại nào về tính đạo đức trong quá trình tiến hành nghiên cứu này, bạn có thể liên hệ Ủy ban qua Trường phòng:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830,
Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Bất kỳ vấn đề nào mà bạn nêu ra sẽ được giữ kín và tìm hiểu toàn diện và bạn sẽ được thông báo kết quả.

Xin giữ bạn thông tin này để tham khảo trong tương lai.

**Laura Hoffman**
Nghiên cứu viên chính
Charles Sturt University

**Dr Linda Wilson**
Người hướng dẫn nghiên cứu
Charles Sturt University

**Dr Sally Hewat**
Người hướng dẫn nghiên cứu
The University of Newcastle

**Dr Thao Huynh**
Công sử nghiên cứu
Cho Ray Hospital
Mẫu chấp thuận tham gia nghiên cứu

ĐỒ TIN CÂY CỦA VIỆC SỬ DỤNG THANG DO ĐỘ NĂNG ĐỆ ĐO LƯỠNG MỤC NỔI LẤP Ở VIỆT NAM

Nghiên cứu này sẽ được thực hiện bởi Nhóm nghiên cứu sau đây:

Laura Hoffman (lhoffman@csu.edu.au)
Người nghiên cứu chính, Chuyên viên Âm ngữ trị liệu và Nghiên cứu sinh tiến sĩ
Trưởng Sức khỏe Cộng đồng, Đại học Charles Sturt, PO BOX 789, Albury, NSW, 2640, Australia

Tiến sĩ Linda Wilson (liwilson@csu.edu.au; 02 6051 9257)
Người hướng dẫn nghiên cứu, Trưởng Sức khỏe Cộng đồng, Đại học Charles Sturt

Tiến sĩ Sally Hewat (Sally.Hewat@newcastle.edu.au; 02 4921 5159)
Người hướng dẫn nghiên cứu, Trưởng Khoa học xã hội và nhân văn, Đại học Newcastle

Bác sĩ Huỳnh Bích Thảo (huynhbthao@gmail.com; 84 919 000 369)
Cộng sự nghiên cứu, Bác sĩ và Chuyên viên Âm ngữ trị liệu, Bệnh viện Chợ Rẫy

Tôi đã đọc và hiểu bản thông tin được đưa cho tôi và tôi đã có cơ hội để đặt câu hỏi về dự án nghiên cứu và đã nhận được câu trả lời thỏa đáng.

Mục tiêu của nghiên cứu đã được giải thích cho tôi, bao gồm các nguy cơ/ bất tiện ( có thể có ) đề kèm với dự án nghiên cứu.

Tôi hiểu rằng bất cứ thông tin hoặc chi tiết về cá nhân tôi có được trong lúc làm nghiên cứu là bảo mật và không có phần nào kết quả hoặc thông tin dữ liệu nào sẽ được dùng hoặc công bố mà không có sự cho phép tôi trên văn bản giấy tờ.

Tôi đồng ý không thảo luận số điểm mà tôi đánh giá với bất kỳ ai khác tham gia trong nghiên cứu này.

Tôi hiểu rằng tôi có tự do để rút khỏi nghiên cứu bất kỳ lúc nào, và nếu tôi rút ra thì tôi sẽ không chịu bất kỳ hình phạt hay sự phân biệt đối xử nào.

Tôi sẽ không sao chép hoặc đưa cho bất kỳ ai xem nội dung của CD và sẽ gửi lại CD ngay khi tôi hoàn tất qua phong bì đã được trả tiền trước.

Tôi đã có cơ hội để hỏi và có câu trả lời thỏa mãn.
Tên: ____________________________________________________________

Chữ ký____________________________________Ngày:__________________

Xin ghi bằng tiếng Anh địa chỉ mà bạn muốn gói mẫu kích thích được gửi đến:

________________________________________________________________________________________

LƯU Ý: Ủy ban Đạo đức về Giảm thiểu Nguy cơ cho Con người của Bộ môn Khoa học thuộc Đại học Charles Sturt đã chấp thuận đề án nghiên cứu này (405/2014/03). Nếu bạn có bất cứ than phiền hoặc e ngại nào về tính đạo đức trong quá trình tiến hành nghiên cứu này, bạn có thể liên hệ Ủy ban qua Trường phòng:

Ingrid Stuart, Executive Officer, Faculty of Science Minimal Risk Human Ethics Committee, Locked Bag 49, Dubbo, NSW 2830, Email: scienceFHEC@csu.edu.au, Phone: 61 2 68857327

Bất kỳ vấn đề nào mà bạn nêu ra sẽ được giữ kín và tìm hiểu toàn diện và bạn sẽ được thông báo kết quả.