

PART ONE: INTRODUCTION

Chapter 1: The Nature and Treatment of Stuttering

Introduction

Stuttering is a speech disorder that is characterised by a number of verbal and non-verbal behaviours. The verbal behaviours include syllable repetition (e.g., “a-a-a-about this time last week”) and multi-syllable repetition (e.g., “a big-a big-a big dog was standing at the gate”) as well as fixed posture with audible airflow (e.g., “SSSSam said I could do it”) or without audible airflow (e.g., “I ... [no sound] ... don’t want to”) (Teesson, Packman, & Onslow, 2003). More severe stuttering can also be associated with superfluous non-verbal behaviours such as tics, blinking and facial grimacing (Teesson, Packman, & Onslow, 2003).

The age at which this disorder begins varies, but it is mostly between 2 and 5 years. Onset of stuttering can be sudden or gradual, with parents reporting onsets that range from 1 day to several months (Onslow, Harrison, & Jones, 1993). Early stuttering is usually characterised by the repetition of vowels and words (Yairi, 1997) and the child will often recover naturally without any formal treatment. It is thought that about one in 20 children will experience some form of stuttering in their early life but that up to 80% of that population will recover naturally, leaving about one in a 100 with a long-term disorder (Bloodstein, 1995). Boys are more at risk than girls with an approximate ratio of 3:1. Girls also appear to be more likely to recover naturally (Bloodstein, 1995).

There is strong evidence suggesting a hereditary link, with children of people who experienced stuttering being more likely to develop the disorder (Andrews, Craig, Feyer, Hoddinott, Howie, & Nielson, 1983). Children whose relatives did not recover naturally are also more likely to have a long-term disorder (Yairi, Ambrose, Paden, & Throneburg, 1996).

Chronic stuttering is debilitating because it disrupts or prevents normal communication. The earliest signs of its ill effects have been measured in 6 and 7-year-olds (DeNil & Brutten, 1991; Vanryckeghem & Brutten, 1997). Almost all adults who stutter report experiencing the disabling effects of the condition during childhood and early adolescence (Hayhow, Cray, & Enderby, 2002). Stuttering children suffer bullying (Langevin, Bortnick, Hammer, & Wiebe, 1998), are perceived negatively by peers (Langevin & Hagler, 2003) and are rejected by peers more often than children who do not stutter (Davis, Howell & Cooke, 2002).

In adulthood, stuttering can cause social maladjustment and failure to attain occupational potential (Craig & Calver, 1991; Hayhow, Cray, & Enderby, 2002). Stuttering is known to be associated with social anxiety, with half of those seeking clinical help warranting a co-morbid psychiatric diagnosis of social phobia (Stein, Baird, & Walker, 1996).

The cause of stuttering is unknown, however there have been many theories proposed on the origin of the disorder (Packman & Attanasio, 2004). Stuttering is currently thought to be underpinned by some deficit or inefficiency in neural processing, with a neuroanatomical component a possibility (Büchel & Sommer, 2004.) An extensive range of treatment methods have been tried throughout history with differing levels of success (Bloodstein, 1995) and many treatments for stuttering are currently available. The present thesis deals with treatment of early stuttering in children younger than 6 years, therefore the following section will describe the characteristics of early stuttering.

Early stuttering

All people speak with at least some disfluencies and there have been a number of studies that have examined this in non disabled speakers. A summary of

the findings of these studies is given in Onslow (1996). The essence of these findings is that:

1. Disfluencies occur more in preschool children than older speakers.
2. The frequency of disfluencies decreases during the school years.
3. There is considerable variation in the disfluencies of early life.

With regard to the last point the variation comprises both the frequency and types of disfluencies at various times within the same child as well as considerable variation between children.

There have also been a number of studies that have investigated the speech of children exhibiting attributes of early stuttering. An early study by Johnson et al. (1959) compared disfluencies in the speech of 68 stuttering boys with 68 non-stuttering boys matched for age, gender and socioeconomic status of the family. The mean age of the children was 5 years with a range from 2.5 to 8 years. The stuttering children had been stuttering for between 1 month and 3 years. Samples of speech used for analysis were on average 500 words in duration. Table 1.1 summarises the findings of the study.

The results suggest overlapping distributions with non-stuttering children having all the same disfluencies as the stuttering children but with less frequency in most cases. Three types of disfluencies were equally common for both groups: interjections, revisions and incomplete phrases. What appears to distinguish children who stutter from those who do not is the increased frequency of repetitions of words, phrases, and syllables and to a lesser extent prolonged sounds and broken words. It was also observed that the stuttering children had a greater number of repetition units per disfluency. These results suggest that stuttering is not a dichotomy but is rather a continuous scale where children who display an increased frequency of certain types

of disfluencies are considered to be stuttering. However some caution should be exercised when interpreting these data because some of the included children were too old to be classified as displaying early stuttering (Bloodstein, 1995). Further, the validity of these disfluency categories has been questioned empirically (Onslow, Gardner, Bryant, Stuckings, & Knight, 1992).

Table 1.1: Mean number of disfluencies per 100 words of 68 stuttering boys and 68 non-stuttering boys (Johnson et al., 1959)

Type of disfluency	Stutterers	Non-stutterers	<i>p</i>-value
Interjection	3.62	3.13	NS*
Sound and syllable repetitions	5.44	0.61	0.01
Word repetitions	4.28	1.07	0.01
Phrase repetitions	1.14	0.61	0.01
Revisions	1.30	1.43	NS*
Incomplete phrases	0.34	0.23	NS*
Broken words	0.12	0.04	0.05
Prolonged sounds	1.67	0.16	0.01
All categories	17.91	7.28	0.01

*NS = non-significant ($p>0.05$)

Other research has also shown that part-word repetitions are probably the most distinguishing feature of early stuttering (Yairi, 1997). However early stuttering may involve head, face or body movements and stuttering preschool children may show more “clusters” of disfluencies compared to non-stuttering preschool children (Onslow, 1996).

Natural Recovery from Early Stuttering

There have been very few prospective studies examining natural recovery from early stuttering. An early, prospective study of the onset and course of stuttering was included in the “1000 family study” conducted in England between 1946 and 1962 (Andrews & Harris, 1964). From that study it was estimated that 34 of 43 of children who started stuttering in the preschool years recovered. This represents a natural recovery rate of 79%, however some of these children received treatment for their stuttering.

A more recent study of 1,021 preschool children was conducted in Denmark (Mansson, 2000). Speech and language evaluations occurred when the children were 3, 5 and 9 years of age. At the second assessment 71% of the 51 children identified as stuttering at the initial evaluation were no longer stuttering and at the third assessment 85% were no longer stuttering. The treatment history of the children was not provided. These two studies are the only epidemiological studies completed to date where the children were seen prior to onset of stuttering. A further study by Kloth (1995) also saw children prior to onset of stuttering however the cohort consisted only of children of stutterers.

Currently underway is a further large-scale, prospective study of natural recovery from early stuttering. The Illinois Early Childhood Stuttering Project (Yairi & Ambrose, 1999) is recruiting children who began stuttering within the previous 12 months from the general population of Illinois, USA. A preliminary report of 84 children followed for 4 years or more since the onset of stuttering (Yairi & Ambrose, 1999) indicated that 74% recovered naturally, although the authors suggest that this is a conservative estimate and that the true natural recovery rate is likely to be higher.

The results from these three studies suggest that around 75-85% of children who begin stuttering will recover naturally without formal treatment.

For clinical populations it is not appropriate to assume estimates of natural recovery from the studies summarised above as participants were recruited from the general population rather than from speech clinics. Parents would not normally present to the clinic immediately following onset of stuttering in their child. As a number of children only experience transient stuttering it is unlikely these children would present to a speech clinic also. For example, in the Andrews and Harris (1964) study, 16 of the 43 children identified as stuttering stuttered for less than 6 months. Consequently natural recovery rates for children who present to a speech clinic are likely to be lower than population recovery rates. Therefore the proportion of children who present to a speech clinic but would naturally recover if left untreated is likely to be somewhat lower than 75%. How much lower is unknown at present.

Measurement of Stuttering

Measurement of stuttering is an important issue both clinically and from a research point of view however there is a general lack of consensus on how it should be done and this is mainly due to the complexities of the task. There are a number of reasons why this task is so complex and these reasons will be described in the following paragraphs. Firstly, as illustrated above in Table 1.1, non disabled speakers often speak with some disfluency however these disfluencies are regarded as normal speech. People who stutter also exhibit these normal disfluencies however they also exhibit disfluencies that are regarded as stuttering. The distinction between these normal disfluencies and disfluencies that are not normal is not always obvious, especially to the untrained listener.

The second reason why measurement of stuttering is not straightforward is a lack of consensus on how frequency of stuttering should be reported. There are a number of different types of stuttering as described in the first paragraph of this chapter. Which of these should be reported and how they should be reported is an important methodological question. Early studies reported just three types of repetition but later Johnson et al. (1959) introduced a reporting system that included eight types of disfluencies and this has been used in many subsequent studies, often with some modification (Yairi, 1997). The different types of disfluencies are often combined into one overall measure and then reported as a frequency by dividing the total number of disfluencies by the total number of words or syllables. There is also a lack of consensus on how multiple types of stuttering on the same word should be reported. Yairi (1981), for example, suggested that multiple types of disfluencies on a single word should be counted separately but others regard this to be one stuttered word. Yairi (1997) illustrates the different ways of reporting frequency of stuttering with a simple example. In a sample of 300 words containing 400 syllables there were 25 examples of disfluency observed in 15 words. This could be reported as 5% disfluent ($15/300$), or 8.3 disfluencies per 100 words ($25/300$), or 6.3 disfluencies per 100 syllables ($25/400$).

The terms “disfluency,” “dysfluency” and “stuttering” are often used interchangeably in the literature to refer to speech perturbations. Strictly speaking, the first term refers to normal speech events and the second refers to speech events that are abnormal in a general sense. The term “stuttering” refers to abnormal speech perturbations specifically associated with the disorder. Throughout this thesis, the term “stuttering,” when it is used to refer to a speech perturbation associated with the

disorder, refers to such an event that is perceived as such by an observer, and when such a perception is made unambiguously.

The third reason why measurement of stuttering is problematic relates to the infinite possibilities of how one can sample an individual's speech. Frequency of stuttering tends to vary within individuals and this variation can be substantial. Therefore the method used in sampling speech could have a large effect on the results obtained. Variation can occur from hour-to-hour, day-to-day, from situation-to-situation, and can occur over time. It may depend on how the individual is feeling at the time or it may depend on the level of difficulty in the speech attempted. Therefore, to obtain a representative assessment of the frequency that stuttering occurs in a particular individual, samples should be larger rather than smaller, should be taken at various times throughout the day, should be taken on several days, and should be taken in different speaking situations. Ideally, speech should be monitored continuously over a substantial period of time however this would be impractical on the majority of occasions.

In the chapters that follow, unless otherwise stated, frequency of stuttering is reported as percentage of syllables stuttered (%SS). This is simply the calculation of the total number of syllables containing unambiguous stuttering of any type divided by the total number of syllables assessed, and then multiplied by 100 to obtain a percentage. Unless otherwise stated, changes in frequency of stuttering will be reported in relative terms. For example if a group of individuals were stuttering with an average frequency of 10% prior to treatment but then with an average frequency of 5% post treatment, their change in frequency of stuttering would be reported as a reduction in stuttering of 50%. And finally, without exception, for all treated groups compared with untreated groups, the true treatment effect is the difference in

improvement between the groups. This is especially important for comparisons between treated and untreated groups of young stuttering children where untreated children are likely to spontaneously recover, and in placebo controlled trials where there may be a placebo effect.

Treatment Programs for Early Stuttering

Prior to the 1930s treatment for early stuttering consisted of general advice to parents such as telling their young stuttering children to slow down, take a deep breath and think before speaking (Bloodstein, 1995). The situation changed in the 1930s with the development of the speech pathology profession in the United States and two influential theories on early stuttering. Bluemel's (1932) model of primary and secondary stuttering described primary stuttering as sound, syllable and word repetitions unaccompanied by signs of effort or emotion that would often disappear if parents did not interfere. Secondary stuttering was described as the awareness of the impediment, the physical effort and struggle to conceal stuttering, and especially the fear of certain letters, words or situations. Secondary stuttering was thought to develop from primary stuttering if the parents made the child conscious of stuttering as a social deficit.

A second influential theory of stuttering, the diagnosogenic theory (Johnson, 1942), went somewhat further than Bluemel's model and suggested that primary stuttering was in fact normal disfluency and that the chronic disorder is caused by parents' abnormal reactions to normal disfluency in the child's speech. This caused a widespread belief that direct treatment of stuttering in the preschool years would do more harm than good. In fact, it was recommended that attention should not even be drawn to the child's speech. This belief was maintained for decades until the 1980s when prominent clinicians and researchers of the time came to a consensus that early

intervention was not only desirable but also essential. This pronouncement was made at the conclusion to a conference at the University of Washington in 1982 and subsequently published as Prins and Ingham (1983). This proved to be hugely influential and many interventions for early stuttering were subsequently proposed. A number of these interventions are summarised in the following pages.

Stuttering Prevention and Early Intervention: A Multi-Process

Approach

Starkweather, Gottwald, and Halfond (1990) developed a method of early intervention for stuttering in the 1980s. This method was based on the Demands and Capacities model of the onset and development of stuttering (Starkweather, 1987). The intervention involved reducing demands on the speech of the stuttering child while also attempting to increase the child's capacity for fluent speech. This was achieved by counselling the parents of the child to bring about appropriate change in their attitudes and, in some cases, behaviour. Examples of behaviour change thought to be appropriate in some instances included reducing parental speech rate, simplifying parental language, changing their negative reactions to the child's stuttering, avoiding interrupting the child during speech and introducing daily special talking time. Direct intervention was also introduced for some children. This involved modifying the child's speech using a number of techniques such as speech modelling or fluency shaping.

The results of this treatment are presented in general terms without data to illustrate the actual level of stuttering before and after treatment (Starkweather, Gottwald, & Halfond, 1990). Of the 55 families coming to the treatment clinic, 16 were briefly counselled and not admitted into the full treatment. The children of these families were thought to be at minimal risk of stuttering. Of the remaining 39

families, seven withdrew from treatment before it had concluded for a variety of reasons and four of the children were still stuttering. The remaining 32 families either had completed the treatment successfully (29) or were still in therapy (3). Success was reported as fluency approaching normal levels although some children were still demonstrating residual stuttering on discharge from treatment. The average duration of treatment for the 29 completed cases was 12 clinic sessions but a few children required 30 or 40 sessions.

A subsequent publication (Gottwald & Starkweather, 1999) states that from 1993 through to 1996, a further 15 families were treated with 14 families successfully discharged after an average of 15 treatment sessions. One family withdrew from the program and the child continued to stutter. Although the results of this treatment program suggest the majority of children achieve fluent speech, there are no data to back up these claims. In addition, as stated earlier in this chapter, a number of young stuttering children would be expected to naturally recover from the disorder. Therefore it is unclear whether the results achieved are any better than what would have been expected without intervention.

The Parent-Child Group Approach to Stuttering in Preschool Children

Assessment in the parent-child group approach (Conture, 1990) involves a case history form filled out by the caregivers of the child, an interview of the child's parents and a direct examination of the child. From previous experience, out of every 100 children evaluated initially, 10 will not require treatment, 40 will require further evaluation before treatment may or may not be initiated and 50 will proceed directly to treatment. The treatment program focuses on communicative interactions, speech production behaviours and attitudes about speech in general, with emphasis on parents' attitudes to their child's speech and their own speech. Although direct

approaches to modifying a child's speech production are used, emphasis is on indirect methods and direct methods are used only on a minority of children.

The indirect approach for the child involves instruction in listening when someone else is talking, waiting for their turn to speak, and not talking when someone else is talking. The parents are told that interrupting their child, talking for them, use of complex sentence structures, fast speaking and correcting the child's speech do not facilitate fluent speech in their child. The parents are then asked to observe their own speech behaviours and correct as necessary. More than 200 children have been treated using this parent-child group approach and it is estimated that 70% become ready for dismissal within 12 to 36 weeks after beginning (Conture & Melnick, 1999). Criteria for dismissal are an average of 3% or less stuttering per 100 words across an 8-week period. The remaining 30% of children required direct treatment before success was achieved. Relapse is estimated to be less than 10%. Once again, these claims are not substantiated with objective data. For example, it is not known what the baseline level of stuttering was for the children who entered the treatment program or whether this treatment program is more effective than natural recovery.

Developmental Intervention: Differential Strategies

Another treatment program for early stuttering employs differential strategies for attaining fluency (Gregory, 1999). When young children present with a stuttering problem, an evaluation is undertaken to determine whether there is a concern about the child's speech. If there is a concern then a number of strategies can be used to treat the child. A typically disfluent child (less than 2% stuttering) will normally lead to the parents undertaking preventive counselling. This involves discussion with the parents on the child's disfluency. The clinician will attempt to convey concepts that

are thought to enhance fluency, including having a realistic expectation on the child's communication, taking turns when speaking and not interrupting, and giving support and positive feedback to the child when they are speaking.

For a child diagnosed as being borderline, atypically disfluent (2-3% stuttering) a different treatment strategy is implemented, namely prescriptive parent counselling with limited involvement of the child. The parent counselling is similar to that described previously except that the parents are trained to identify stuttering and to chart stuttering frequency over time. In addition, the clinician sees the child for 30-50 minute sessions twice weekly. During these sessions, the child and clinician take part in activities that enhance fluency. The clinician models an easy, relaxed, slow speaking manner with an emphasis on pausing between phrases. This program of treatment is reported to be almost 100% successful in removing parental concern over their child's speech in 2 to 4 weeks.

A comprehensive therapy program is usually undertaken for those children diagnosed as atypically disfluent (3% or more stuttering). The clinician sees these children two to four times a week for 30 to 50 minute sessions and the parents are counselled weekly. The treatment given to the children is based on slower more relaxed speech (Gregory & Hill, 1998). The comprehensive treatment program normally takes 6 to 12 months to complete and 5% of children have a persistent problem that requires longer follow up. The reported outcomes from this treatment program are also not substantiated with objective measures of speech. It is therefore unknown whether the differential approach to treatment of early stuttering is more effective than letting nature take its course.

The Stuttering Intervention Program

The Stuttering Intervention Program (Pindzola, 1987) can be broken up into four distinct parts. Part 1 involves determining whether the child is stuttering or just exhibiting normal disfluency. In this regard, speech behaviours are analysed and baseline speech data are collected. In Part 2 the parents are involved in discussions with the clinicians on the child's speech. The parents are counselled and it is suggested they use short sentences, developmentally appropriate vocabulary and an unhurried rate of talking to their child. Fluency of the child's speech is enhanced in Part 3 by reinforcement of fluent speech and punishment of disfluent speech. Actual speech is manipulated by the use of slow, soft and smooth speech techniques. Short, simple sentences are initially encouraged with longer utterances introduced later on in the treatment process. The fourth part of the program is implementation in the schools where the speech pathologist works with the classroom teacher to ensure the child reaches the goals of an "Individualised Education Program".

Evaluation of the Stuttering Intervention Program has been conducted with just three participants (Pindzola, 1999). All three were of preschool age and were evaluated before and after the program using a numerical rating scale that was first administered on 43 normal children, then on seven stuttering children as well as matched controls. A higher score on the rating scale indicates more behaviour associated with stuttering. The average score on the scale was 17 for normal children and 27 for stuttering children. For the three children the average score was 30 at pre-intervention and 18 at post-intervention. All three children reduced their score by at least 10 points. Thus, although there are data reported for this treatment program, only three children are included, the outcome is based on a rating scale rather than

being objective, and there is no control group. It is unknown whether these three children would have recovered from their stuttering without intervention.

Speech Motor Training

This program was developed over a period of 20 years and its purpose is to improve speech motor production, thus improving fluency of speech (Riley & Riley, 1999). In the program, the child models desirable speech behaviours as displayed by the treating clinician. Three rates of speaking are used. These are very slow, slow, and normal with the child moving onto the higher rates as the treatment program progresses. The complexity of the speaking tasks is also gradually increased throughout the program. The child can only move to the next level of complexity once he or she has mastered the current level. Reinforcement of desirable speech behaviour is achieved through the use of tokens, moves in a game, or points towards a toy. In addition approximately half of the children need to be trained in breath control at some stage during the treatment process.

Retrospective data on nine children and an initial pilot study of three children given speech motor training showed encouraging results hence a systematic controlled experiment was conducted to evaluate the treatment program (Ingham & Riley, 1998). In the experiment, six children of age 49 to 77 months were given the treatment program over 8 weeks. Frequency of stuttering was reduced by about half throughout the study and this was a statistically significant reduction. Speech rate increased slightly and speech naturalness improved but not to an acceptable normal level. However this improvement was statistically significant. After the experiment, four of the six children went on to receive another (different) treatment program and five of the six children were stutter-free 2 years after the initial treatment. While the results from these initial studies are encouraging, a control group was not included,

sample size was small and duration of follow up was limited. Further research is required to establish the efficacy of this treatment.

The Fluency Rules Program

Runyan and Runyan (1993) developed their treatment program for young children based on methods found to be effective for adults. Originally there were 10 “rules” designed to teach children to speak fluently using a natural sounding pattern. These 10 rules were then refined and reworded in a child-friendly way and some of the rules were dropped because they did not apply to young children. The five remaining rules were then categorised into two types of rules: universal rules that were used for all children and primary rules that were only used if necessary. The universal rules are: “speak slowly,” “say a word one time,” and “say it short.” The primary rules are “use speech breathing” and “start Mr Voice Box running smoothly / gently” (Runyan & Runyan, 1999, pp. 161).

Initially the treatment clinician imitates speech that violates each of the rules in an attempt to have the child understand the concept of each rule. Next the child is taught how to self-monitor and identify when a rule is broken in their own speech. Once this is mastered the children attend 60-minute sessions where they practice using the rules with the treating clinician. Generalisation to the home and school is achieved by involving the parents and teachers in the treatment program. The parents and teachers are taught hand signals that they use to keep the child aware of the rules when speaking.

Evaluation of the program was initially conducted on 23 young stuttering children, of which six were preschool age (Runyan & Runyan, 1999). Five of the six were treated successfully and finished the program with fluent speech. The remaining child pulled out of the program very early on and was not able to be

contacted. Since this early study a further five children had been through the program successfully however one child did show signs of relapse and was monitored accordingly. Objective speech data on the 11 preschool children are not presented. The effectiveness of this treatment compared to recovery without intervention is unknown.

The Monterey Fluency Program

The Monterey fluency program (Ryan & Van Kirk, 1978) was developed in the late 1960s and early 1970s and is based on operant conditioning. There are three phases to the program: establishment (within-clinic fluency), transfer (beyond-clinic fluency) and maintenance (fluency over time). A fluency interview is first conducted to establish whether the child does indeed have a stuttering problem. A criterion of greater than 3 stuttered words per minute is the cut-off imposed to indicate stuttering. In addition, the child is monitored over a number of months to determine whether spontaneous recovery occurs. If persistent stuttering is indicated, the child is accepted into the treatment program. Progress through each stage of the program only occurs when the child has passed a criterion test based on monologue and conversation with a cut-off of less than 0.5 stuttered words per minute. Within-clinic fluency is normally established using Gradual Increase in Length and Complexity of Utterance (GILCU) but occasionally Delayed Auditory Feedback is used.

The primary goal of treatment is to reduce the frequency of stuttering but data on speech rate are also collected. The Monterey fluency program has been evaluated on a number of clients including 20 that were of preschool age (Ryan & Ryan, 1999). Six children had multiple speech and language problems including persistent stuttering. All six achieved normal fluency and this was maintained over time. Of the remaining 14 children, five achieved fluency by parent counselling alone. A further

nine children had persistent stuttering and were given the full treatment program. All nine children achieved normal fluency and maintained this fluency over time. An average of 20 hours of clinician input was required for clients of all ages but is not reported for the preschool children specifically. Results for preschool children appear to be promising but once again are not backed up with objective data or compared to a no-treatment control group.

Extended Length of Utterance (ELU)

The ELU treatment program (Costello, 1983 and for a review see Davidow, Crowe, & Bothe, 2004) is based on response contingent stimulation and progressively longer or more difficult speaking tasks. The program begins with the production of 1-syllable words with progression to 6-syllable utterances then monologues of various durations and finishing with monologues of 5-minute duration. Duration of conversation is then systematically increased up to 5 minutes. Progression through this increasingly difficult array of speaking tasks is dependent on 10 consecutive fluent responses. Positive reinforcement for each fluent response is provided with tokens and punishment is used for each moment of stuttering in the form of the clinician requesting the child to “stop”.

Outcome data for the ELU treatment program are limited with two reports based on one and four children respectively however ELU was only used after another treatment had already been undertaken. Two further reports are based on one and six children respectively however the complete treatment program was not employed for the single participant and the second report was based on an experiment with all six children receiving 24 hours of treatment regardless of whether more was required to achieve minimal stuttering. Although results are based on objective data and suggest ELU treatment is associated with significant reductions

in stuttering, the studies reported involved small samples and did not employ a natural recovery control group.

In summary, the treatment programs for early stuttering described above have a number of similarities and there is some overlap of methods applied. However, they can be broadly categorised into two groups: those that emphasise the modification of the child's environment (often by way of parent counselling in an attempt to encourage fluency) and those that employ more direct methods that attempt to encourage fluency by positive reinforcement and/or some kind of speech modification. Some employ both. One thing is common to all of these treatment programs, however, and that is a lack of objective data to determine whether they are effective at eliminating stuttering or whether they are better than natural recovery. None of the studies reported included a comparison with a no-treatment control group.

The Lidcombe Program of Early Stuttering Intervention

Since the Lidcombe Program is the subject of this thesis, it is described here in considerable detail. The origins of the Lidcombe Program can be traced to single-subject laboratory experiments that showed response contingent stimulation reduced stuttering in preschool-age children (Martin, Kuhl, & Haroldson, 1972; Reed & Godden, 1977). Attempts to replicate Martin, Kuhl, and Haroldson's (1972) experiment in Sydney speech pathology clinics in the early 1980s were unsuccessful (Onslow, 2003), however researchers and clinicians at the Stuttering Unit in Sydney experimented clinically for a number of years to develop a combination of components that formed an effective stuttering treatment program for preschool-age children. An early version of the Lidcombe Program had been developed by the late 1980s (Harrison, 2002).

The Lidcombe Program is a parent-administered, behaviour therapy for early stuttering. Treatment occurs in everyday speaking situations. It incorporates weekly clinic visits by children and parents, during which parents are trained to present verbal contingencies for their children's stutter-free speech, and verbal contingencies for unambiguous stuttering. The contingencies for stutter-free speech are acknowledgment (e.g., "That was smooth"), praise (e.g., "That was good talking"), and request for self-evaluation (e.g., "Were there any bumpy words then?"). The contingencies for unambiguous stuttering are acknowledgment (e.g., "That was a bit bumpy") and request for self-correction (e.g., "Can you say that again?").

When parents have learned to administer these verbal contingencies in structured speaking situations, the verbal contingencies are presented during everyday conversations. Stage 1 of the Lidcombe Program is concluded when children achieve for 3 consecutive weeks a stuttering rate of at or below 1.0 percent syllables stuttered in the clinic, and daily parental stuttering severity scores of 1 (no stuttering) or 2 (extremely mild stuttering) on a 10-point scale. Speech measures continue during Stage 2 and are used to determine whether the child is maintaining the treatment effects. During Stage 2, the family visits the clinic less frequently if the child maintains treatment benefits, and eventually they stop visiting altogether. During Stage 2, the parent gradually withdraws the verbal contingencies as visits to the clinic decrease in frequency. Any departure from criterion speech performance, as specified with the clinical measures at the end of Stage 1, is acted upon immediately. The Lidcombe Program is described in Onslow, Packman, and Harrison (2003) and the treatment manual can be downloaded from the website of the Australian Stuttering Research Centre (2001).

A number of studies have suggested that intervention with the Lidcombe Program is a promising treatment for early stuttering. In a preliminary study of four children younger than 5 years, Onslow, Costa, and Rue (1990) used an early version of the program. Stuttering frequency was measured in several within- and beyond-clinic speaking situations for a 2-month period before and a 9-month period after the treatment. Results showed that all four children reduced their stuttering to near zero levels. However, there were limitations; the study involved only four participants and there was no control group.

A second study (Onslow, Andrews & Lincoln, 1994) was designed to replicate and expand the Onslow, Costa, and Rue (1990) findings by using a larger group of children and employing a no-treatment control group. Twelve children in the experimental group achieved median percentage of syllables stuttered (%SS) scores below 1.0 at 12-months post-treatment. The treatments were completed in a median of 11 1-hour clinic sessions and a median of 85 days from the start of treatment. The majority of parents of the control group children withdrew from the study and elected to have treatment begin for their child. Therefore a comparative analysis could not be performed.

Since these two initial studies, there has been further medium and long-term outcome data reported in a number of publications. Lincoln and Onslow (1997) undertook a long-term follow up study of 42 children who received the Lidcombe Program. Results indicated that stuttering frequency reached near-zero levels and that these levels were maintained in everyday speaking situations for up to 7-years post-treatment. The report by Lincoln, Onslow, and Reed (1997) indicated that unsophisticated listeners did not perceptually distinguish between the speech of

preschool children treated with the Lidcombe Program and the speech of the control children.

A study by Woods, Shearsby, Onslow, and Burnham (2002) investigated whether the Lidcombe Program might be associated with negative psychological effects for the children receiving the treatment and also for the relationship between children and their parents. Results suggested that over the pre-treatment to post-treatment period there was no evidence of negative psychological impact. In fact, the data suggested post-treatment improvement however statistical significance for this result was not obtained. This result was consistent with the view that the Lidcombe Program is a safe treatment for children and their families.

The Lidcombe Program has been shown to reduce stuttering rates beyond that of natural recovery over a 3-month period (Harris, Onslow, Packman, Harrison, & Menzies, 2002). In the Harris et al. (2002) study the Lidcombe Program was compared with a wait-list control in a randomised, parallel-group study of 23 preschool-aged children. The primary outcome was %SS, assessed blindly and subjected to intra- and inter-judge reliability checks. Measures were made before treatment and after 12 weeks, in a variety of within- and beyond-clinic speaking situations. The results of this study showed %SS reduced by approximately 60% in the Lidcombe Program group and approximately 30% in the control group. A comparison of the two groups indicated a statistically significant difference in the reduction of %SS.

In contrast to the other treatments reviewed earlier, the Lidcombe Program is associated with an extensive program of research. This program of research has been ongoing for the past 10-15 years and has shown the Lidcombe Program to be a promising treatment for early stuttering. However it is possible that the apparent

success of the program is predominantly due to spontaneous recovery of the children rather than the treatment program itself. The study by Harris et al. (2002) has contributed to demonstrating that the Lidcombe Program is more effective than natural recovery. However that study included only 23 participants and followed them for only 3 months. Therefore a second randomised study is critical to confirm the results of the earlier study. In addition, such a study should follow the children for longer, employ a larger sample, and follow the strict guidelines for the conduct of a pragmatic phase III randomised clinical trial (Moher, Schulz, & Altman, 2001).

It should be clarified here that it was not the intention in this chapter to compare the efficacy of the Lidcombe Program with the other treatments for early stuttering summarised earlier in the chapter. It could be argued why should the Lidcombe Program be studied further and not one of the other treatments for early stuttering? The fact that the Lidcombe Program has been extensively researched in earlier phase trials means that it is appropriate for a phase III randomised trial to be undertaken. If one of the other treatments was to be studied further then it would not be appropriate to conduct a phase III trial. The potential of the other treatment programs would need to be established in earlier phase trials before a phase III trial could be deemed appropriate.

The Present Thesis

The topic of this thesis is an evaluation of the effectiveness of the Lidcombe Program of Early Stuttering Intervention. This was achieved primarily through a randomised controlled trial of the program, in which children receiving the treatment were compared to a no treatment control group over a 9-month period. The remainder of this thesis is divided into seven chapters. The following chapter presents a systematic review of randomised controlled studies of treatment for

stuttering. Chapter 3 is a review of power in stuttering research and the implications of failing to consider power appropriately in a body of research. The next two chapters describe two retrospective studies of the Lidcombe Program of Early Stuttering Intervention: the first undertaken in Sydney, Australia and the second in Norwich, England. A meta-analysis combining the results of these two studies is also presented. Chapter 6 presents guidelines on how to analyse stuttering frequency data appropriately and Chapter 7 reports a randomised controlled trial of the Lidcombe Program. Chapter 8 then concludes the thesis with discussion and implications for future research.

Chapter 2: Review of randomised studies of treatment for stuttering

Introduction

Given that this thesis is an evaluation of the effectiveness of the Lidcombe Program of Early Stuttering Intervention it is appropriate at this point to examine past research that has sought to examine the efficacy of stuttering treatments. The gold standard method for treatment efficacy research is the randomised controlled trial (Piantadosi, 1997). In fact, the highest level of evidence for treatment efficacy is a systematic review of all randomised controlled trials (National Health and Medical Research Council, 1999). The aim of this chapter is to systematically review and present all the randomised evidence of efficacy for treatment of stuttering. A detailed description of randomised controlled trials is included at the end of this Introduction for anyone unfamiliar with this type of study.

Although this systematic review only includes randomised studies, it should be noted that there have been many non-randomised studies of treatments for stuttering that have contributed to knowledge in this area. An excellent example is the many studies devoted to treatment using prolonged or smooth speech. That evidence base was reviewed extensively in the context of a clinical trial (Onslow, Costa, Andrews, Harrison, & Packman, 1996), and was added to with two subsequent clinical trials (Craig, Hancock, Chang, McCreedy, Shepley, McCaul, et al, 1996; O'Brian, Onslow, Cream, & Packman, 2003). However only randomised studies were included in the present review as there is good evidence that non-randomised studies tend to over-estimate treatment effects (Kunz & Oxman, 1998).

Data for treatments that have been rigorously tested on multiple occasions will be combined where appropriate. Then conclusions will be formed concerning which treatments are efficacious, which are not, and where the evidence is inconclusive. Of particular importance will be determining which, if any, treatments

for early stuttering have been evaluated with a randomised controlled trial. The following section is a brief tutorial on randomised controlled trials.

Randomised Controlled Trials (RCTs)

The Trial Protocol

The first management task in a RCT is to develop a trial protocol. This protocol outlines in detail what the trial will set out to accomplish and how it is to be accomplished. It is the procedural manual for the trial and outlines the various rules inherent in its conduct. The trial protocol contains a detailed description of the treatments that will be administered throughout the trial. It is essential that all treatments in the trial are thoroughly operationalised in this document, so that any clinician with requisite training, who presents the treatment according to the document, can be deemed to have delivered the treatment appropriately.

Choice of control treatment

It is usual practice in phase III RCTs that the subjects in the control arm of the trial receive the standard practice treatment for the disease or condition under study. If there is no standard practice treatment for the disease or condition under study then the subjects in the control arm normally receive a placebo or no treatment.

Eligibility Criteria

Potential subjects for a trial are assessed initially against eligibility criteria. There is a trade-off between strict and liberal eligibility criteria. Strict criteria can select subjects that are most likely to benefit from the new treatment, thus giving the new treatment the best possible chance to be shown efficacious. However, if the new treatment is shown to be efficacious with strict eligibility criteria, that result may not be generalisable to a wider population. On the other hand, liberal selection criteria

that admit subjects who are less likely to benefit from the treatment will reduce the size of the treatment effect. Ideally, eligibility criteria will be designed that give the new treatment a good chance to be shown efficacious and at the same time provide a result that is generalisable to a clinically important portion of the population with the disorder or disease.

Random Allocation

Once eligibility has been assessed and the subject meets all criteria, and informed consent has been obtained, the subject is allocated to one of the treatment arms. The process of random allocation is essential to a RCT because it eliminates any bias that might occur with allocation of subjects to treatment. Any difference shown between the two groups will then be due to either chance or treatment effect. To prevent any such biases from the treatment personnel involved in a RCT, the random allocation to treatments is conducted by personnel who are completely independent of the clinical activities of the trial. Ideally, for the duration of the trial, neither the treatment personnel nor the subjects become aware of the treatment arm to which subjects are allocated. The concealment of the allocated treatment is called blinding, and the case above is referred to as double-blinded. Blinding is an excellent protection against many sources of bias that might arise, because clinicians do not know which treatment is allocated to subjects. For example, if the clinicians are investigators in a study, they may well have a vested interest in an outcome that shows the experimental treatment to be superior.

Only treatments that can be completely disguised can be double blinded, the classic example being pharmaceutical trials, where different drugs can be delivered without the treatment personnel or the subjects being able to identify which they receive. However, this is often not possible. Examples would be a comparison of

angioplasty with thrombolysis for the treatment of acute myocardial infarction, or oncology studies where chemotherapy, radiotherapy and surgical procedures are compared. In many cases, however, a compromise is possible with a single blinded trial, where the subjects do not know which treatment has been allocated but the clinician does. Examples would include trials of counselling psychology interventions where it is possible to conceal the identity of the intervention from the subjects. In RCTs of non-pharmacological treatments for stuttering, it would normally be impossible to blind either the subjects or clinicians, and this arrangement is referred to as an open study. In such cases where blinding is not possible, it is particularly important that outcome assessment is blind. In other words, outcome assessment is performed without knowledge of treatment group assigned.

In addition to eliminating biases that can occur with treatment allocation, it is essential in a RCT to balance other biases between treatment groups. This can be achieved through stratification, which ensures that each group has the same balance of factors that could influence treatment outcome. Random allocation alone is not always sufficient to ensure that this occurs, especially when the sample size is small (in RCT methodology, a “small” sample is fewer than 1,000 subjects). If the sample size is small, then by chance alone the two groups can be quite different with respect to important prognostic factors. As the sample size increases, these imbalances tend to even out. Stratification ensures that the groups are matched on important prognostic factors. If, for example, it is thought that boys are less responsive to treatment than girls, then stratification by gender ensures that approximately equal numbers of boys and girls are allocated to different treatments. Stratification reduces the need to perform statistical adjustment at the analysis stage—a procedure that is less desirable than having a reasonable balance between treatments within strata.

Once the strata of an RCT have been agreed upon, the method of treatment allocation needs to be considered. Allocation needs to achieve an adequate compromise between the conflicting demands for randomisation and the need to ensure balance between groups and within each stratum of each group. A method of allocation needs to have a substantial random component otherwise allocation can be predictable, which can make it possible for investigators to introduce bias by influencing the treatment that is assigned. Two of the most common allocation methods are blocked randomization and minimization (Taves, 1974).

Study Outcomes and Hypotheses

In order to minimize statistical Type I errors or Type II errors, hypotheses need to be formulated a priori—in other words, before the trial begins. Such errors will become most prominent if a hypothesis is developed and tested post hoc—after the data are collected. A Type I error occurs when, from the sample, it is declared that a difference exists, when in the population this is not the case. A Type II error occurs when no difference is found whereas in the population a real difference exists. This is illustrated in Table 2.1, with reference to whether or not an effect exists in reality and whether or not an effect is detected. It is usual for RCTs to have only one primary hypothesis but there may be several secondary hypotheses. The primary hypothesis is the fundamental focus of the study, and is used to set the subject numbers when the sample size of the study is calculated (see Sample Size, below). Consequently, the primary hypothesis needs to be quantitatively specific. For example, a hypothesis "treatment A will be superior to treatment B" cannot be used to design a study, but the hypothesis "treatment A will be better than treatment B by at least 20%" allows a sample size to be determined and defines the minimum clinically accepted benefit based on prior evidence.

Table 2.1: Conditions under which Type I and Type II errors occur

The Experimental Result		
Reality	Effect Detected	No effect detected
There is an effect	Correct	Type II error
There is no effect	Type I error	Correct

The primary hypothesis of a RCT is normally tested using outcome data. Ideally, the outcomes should be commonly used, easy to measure, reliable and appropriate for the hypothesis to be tested. Outcomes may be behavioural, physiological or psychological. In medical trials, mortality is a common outcome. Behavioural and psychological outcomes require great care in their consistency of measurement as many of these outcomes suffer from measurement bias even though standard psychological tests are used.

Sample Size

The sample size of a study is usually calculated with reference to the primary hypothesis, in other words, the main purpose of the study. The sample size obtained can also be used to calculate the power associated with secondary hypotheses. Sample size is an important issue in the design of an RCT because it contributes to the power to detect a statistically significant difference if one exists. Too few subjects will risk a Type II error of failing to find such a difference that exists, and too many subjects constitute wasted resources. In practice, however, under-powered studies with too few subjects are more common than over-powered studies. To calculate sample size, the difference between treatments as specified in the primary

hypothesis is considered, and the sample size required to detect this difference (if it exists) needs to be realistically obtainable in order for the trial to proceed.

Analysis by intention to treat

Analysis by intention to treat is a critical factor in the design of a scientifically rigorous RCT. Analysis by intention to treat means that subjects will be statistically analysed by the treatment to which they were originally assigned, regardless of the treatment they actually received. For example a subject allocated to Treatment A that subsequently received Treatment B would be analysed as if that subject remained in the Treatment A group. Also, drop-out subjects are included in the analysis. Intention to treat analysis preserves the randomisation and ensures an unbiased statistical analysis, although it may result in a diluted estimate of the actual treatment effect.

Interim Analysis

It is common for data to be statistically analysed at various times throughout the duration of a trial. This is called interim analysis, and the final analysis then occurs at the completion of the trial. The benefits of interim analyses are that they: (1) provide a formal means of monitoring progress of the trial, (2) allow judgments about stopping the trial early if the safety of the subjects is of concern or if it is clear that one treatment is superior or inferior to the other, and (3) allow adjustment of the sample size in the event that estimated event rates or population standard deviations are incorrect. However, of themselves, interim analyses deplete the level of significance available at the end of the study, and some adjustments to significance level may be required at that time.

Stopping rules

Stopping rules are often included in the trial protocol, for implementation in the event that it becomes unethical to continue the trial. For example, it may become clear early in the trial that the experimental treatment is vastly superior or, at the other extreme, that it is harmful. Normally, interim data analyses provide information on which to base such a decision.

Data Monitoring and Trial Management

Informed consent and confidentiality are ethical considerations for most research involving humans. However, data monitoring is an additional consideration with RCTs. A Data Monitoring Committee (DMC) and a Trial Management Committee (TMC) are normally established to oversee the trial. The DMC is a group of two or more people independent of the trial who have expertise relevant to the study. All interim analyses are reviewed by the DMC. This process provides independent monitoring, and is critical where subjects are receiving a potentially harmful treatment or a treatment that is demonstrably inferior to another experimental treatment. The TMC normally comprises the investigators involved in the trial and they are responsible for the conduct of the trial.

Methods

The inclusion criteria for accepting a study into this review are straightforward and are detailed below:

1. Randomised controlled trials;
2. Participants were children or adults with developmental stuttering;
3. At least five participants evaluated;
4. Results reported in English.

Important quality indicators of clinical trials that were used to evaluate the included studies were concealment of allocation, blinding and all participants accounted for in an intention-to-treat analysis (Schulz, Chalmers, Hayes, & Altman, 1995). If double blinding is not possible, as it typically is not for non-drug treatments of stuttering, then blinded outcome assessment should be performed. As well as assessing these general clinical trial quality indicators for each study, quality indicators specific to stuttering research (Bloodstein, 1995) were also used to evaluate the included studies. These stuttering specific quality indicators included:

1. Outcomes should include reliable and objective measures of speech behaviour made before, during and after treatment;
2. Speech measures should be based on repeated evaluations and adequate samples of speech;
3. Improvement in speech should be demonstrated outside the clinic in everyday speaking situations;
4. Improvements should be maintained at long-term follow-up.

The primary outcome of interest in this review was reduction in level of stuttering while secondary outcomes of interest included speech rate, speech naturalness and side effects. An effective treatment for stuttering would be one that significantly reduces the level of stuttering, leaves speech sounding natural, maintains or increases speech rate, and is associated with minimal or no side effects. The criteria for accepting studies into the review were simple and broad and hence unlikely to require much debate on which studies were eligible, and which were not.

A systematic search for identification of studies was conducted using the following resources:

1. The Cochrane Central Register of Controlled Trials (CENTRAL) in the Cochrane Library (Issue 2, 2004);
2. MEDLINE (January 1966 to April 2004);
3. EMBASE (1988 to 2004);
4. Reviews of citations in publications identified by the above strategies.

Any article containing the word “stuttering” or “stammering” was included in the initial pool of studies. The abstract was then obtained and examined, and articles that obviously did not relate to studies of treatment for stuttering were discarded. The full manuscripts of all the remaining studies were then obtained and only those adhering to the inclusion criteria retained for further analysis. In all, 27 studies are included in this systematic review.

Results

Of the numerous manuscripts that were rejected, three were reviews of studies of stuttering treatment. The first of these reviews by Andrews, Guitar, and Howie (1980) included 42 studies, published in 29 papers covering the treatment of 756 people who stutter. The majority of these studies were non-randomised so are not included in the current review, and all were of non-drug treatments. A meta-analysis combining data from all studies, irrespective of the treatment studied, led Andrews et al. to conclude that the average effect size was an improvement of 1.3 standard deviations on pre-treatment stuttering scores. The treatments that appeared to be most effective were prolonged-speech and gentle onset techniques, and to a lesser extent, attitude and airflow techniques.

The second review was by Brady (1991) who performed a critical review (without meta-analysis) of pharmacological interventions for stuttering. Although a

section was devoted to non-pharmacological treatments, the emphasis of the review was drug treatments. The review identified that many drugs had been used to treat stuttering, and this was probably due to the many causal theories of this poorly understood disorder. However, it was also found that most studies included in the review did not use adequate designs. Brady concluded that promising pharmacological treatments that should be researched further include calcium channel blocking agents and cholinergic drugs. A number of the studies included in Brady's review were also included in the current review.

Unlike the previous two reviews, the third review identified (Woods, Twohig, Fuqua, & Hanley, 2000) concentrated on just one treatment of stuttering, that being treatment with regulated breathing. Regulated breathing, originally proposed by Azrin and Nunn (1974), encompasses a variety of components, including awareness training, relaxation, competing response training, motivation training, and generalisation training. The originally proposed treatment was subsequently simplified but both the original and simplified approaches appeared to reduce stuttering frequency and increase speech rate. However, Woods et al. (2000) identified a number of limitations associated with the research on regulated breathing. A number of the studies included in the review by Wood's et al. were randomised studies and are included in the current review, however the remaining studies were non-randomised hence are ineligible for the present review and are not included.

There were five randomised studies rejected for inclusion in the current review. Two studies reported the results of fewer than five participants (Costa & Kroll, 1995; Hays, 1987); one study was reported in German (Wille, 1999) and two studies were experiments (rather than trials) that were completed within a single day

(James, 1983; Moore & Ritterman, 1973). All the remaining studies that were rejected were non-randomised studies.

Of the 27 studies included in the present review, 16 examined a form of drug therapy, nine examined non-drug therapy, and two included both drug and non-drug therapies. Only three treatments were studied more than once: haloperidol (8 studies), regulated breathing (5 studies) and verapamil (2 studies). All of the drug therapy studies were double blind and most were cross-over designs. Cross-over designs are often appropriate for drug therapy studies as the effect of a drug will normally “wear off” in a short time period so that the effect of a second drug can be assessed on the same patient in the same trial. This fact also suggests that any drug effect on stuttering will quickly wear off if the person ceases to take the drug. The non-drug studies were all parallel group designs without blinding. There was a wide range of sample sizes but all studies were small (8 to 66 participants). Most studies included either adults alone or adults and children, with just three studies of children exclusively. The duration of the studies ranged from just 4 days through to one study that was conducted over 24 months. Table 2.2 below provides a summary of the 27 studies included in the present systematic review.

Haloperidol Studies

Haloperidol is the most commonly studied drug in the treatment of stuttering and there are eight such studies in this review. Most of these studies included a second drug, in combination with haloperidol, to offset the Parkinson-like symptoms caused by haloperidol. A summary of these eight studies is presented in the following paragraphs. Unfortunately, a meta-analysis cannot be performed because six of the eight studies did not include estimates of variation in their reports.

Table 2.2: Summary of 27 included studies in chronological order

Study	Experimental treatment	Sample size	Age of participants	Duration of study
Edgren et al. (1970)	Diazepam	29	Not stated	1 week
Wells et al. (1971)	Haloperidol	36	Adults and adolescents	8 weeks
Peins et al. (1972)	"Tape recorded method"	36	Adults and adolescents	6 months
Swift et al. (1975)	Haloperidol	22	Adults	3 weeks
Rosenberger et al. (1976)	Haloperidol	8	Adolescents	12 weeks
Moleski et al. (1976)	Rational-emotive therapy	20	Adults	1 month
Rantala et al. (1976)	Haloperidol	66	All ages	8 weeks
Murray et al. (1977)	Haloperidol	26	Adults	3 months
Andrews et al. (1977)	Haloperidol	27	Adults	8 weeks
Burns et al. (1978)	Haloperidol	12	Adults	4 days
Azrin et al. (1979)	Regulated breathing	38	All ages	3 months
Prins et al. (1980)	Haloperidol	21	All ages	8 weeks
Ladouceur et al. (1981)	Regulated breathing	16	Adults and adolescents	1 month
Rustin et al. (1981)	Oxprenolol	31	Adults and adolescents	12 weeks

Ladouceur et al. (1982)	Regulated breathing	21	Children	1 month
Eversham et al. (1985)	Speech reconstruction	48	Not stated	24 months
Saint Laurent et al. (1987)	Regulated breathing	40	Adults	10 months
Brumfitt et al. (1988)	Verapamil	14	Adults	6 weeks
Waterloo et al. (1988)	Regulated breathing	32	Adults	8 months
Brady et al. (1989)	Verapamil	10	Adults	3 weeks
Kampman et al. (1993)	Bethanechol	10	Adults	2 weeks
Gordon et al. (1995)	Clomipramine	17	Adults and adolescents	10 weeks
Althaus et al. (1995)	Clonidine	25	Children	28 weeks
Stager et al. (1997)	Pimozide	11	Adults	18 weeks
Maguire et al. (2000)	Risperidone	16	Adults	6 weeks
Franklin et al. (2002)	Olanzapine	24	Adults	3 months
Harris et al. (2002)	Lidcombe Program	23	Young children	12 weeks

Wells and Malcolm's (1971) 8-week trial was a double blind, parallel group study of males and females aged 15 to 60. There were six groups of six participants receiving haloperidol (0.75 to 1.5 mgs per day) or placebo, the muscle relaxant orphenadrine or placebo, speech therapy or no speech therapy. The orphenadrine was used to avoid the side effects of haloperidol which would potentially un-blind the

investigators. Speech measures were taken at 0, 4 and 8 weeks post-randomisation, and were based on three recorded samples of spontaneous speech of 100 words each. Assessments were performed by a speech pathologist but reliability checks were not reported. Also it is not reported whether the speech recordings were made inside or outside the clinic.

Results suggested participants not taking haloperidol showed no significant changes in stuttering or speech rate at 4 or 8 weeks. Of the 12 participants taking haloperidol, 10 improved on all measures of stuttering and speech rate at 4 weeks but two participants were marginally worse ($p < 0.05$ for all measures). At 8 weeks, four of the improvers at 4 weeks had pulled out of the trial however the remaining participants all maintained their gains at 8 weeks without further improvement. Side effects reported (dry mouth, blurred vision, drowsiness, depression) were similar for participants on haloperidol or placebo. At the close of the trial, most participants elected to go onto the active preparation (haloperidol).

As stated in Chapter 1 it is important to remember that the true treatment effect is estimated by the difference in improvement observed between the treated group and the untreated group. Therefore in the Wells and Malcolm trial it is the (not reported) difference between the haloperidol groups and the groups not taking haloperidol that estimates the true treatment effect. The (reported) improvement in stuttering observed in the haloperidol groups is an estimate of the true treatment effect plus any placebo effect or spontaneous recovery that occurred during the trial. Please keep this in mind when reading the remainder of this Chapter and also the remainder of this thesis.

The next randomised study of haloperidol was undertaken by Swift, Swift, and Arellano (1975). Twenty two adult participants began the parallel group, 3-week,

double blind study, however only 17 participants completed. Haloperidol (1.5 to 3.5 mgs per day) was compared to placebo. However benztropine, an antiparkinsonian agent, was given to both groups to minimise the differentiation of side effects in the two groups. Speech measures were taken on multiple occasions before, during and after treatment using adequate samples of spontaneous speech and while reading specific passages. Reliability checks were also conducted however speech was not measured outside the clinic.

The results of this study showed six of the seven participants given haloperidol improved with respect to speech dysfluency and rate and this improvement was statistically significantly greater than for the 10 participants receiving placebo. Side effects were similar between the two groups except that when the dose of haloperidol reached 2.5 mgs per day, three participants reported constant restlessness. This restlessness was largely brought under control by decreasing the dose of haloperidol or increasing the dose of benztropine. Three months after the trial, three participants were still taking haloperidol and maintaining their improvements in disfluency and speech rate. Participants who did not continue with the treatment returned to their baseline levels of stuttering.

Rosenberger, Wheelden, and Kalotkin (1976) compared haloperidol (3 mgs per day) with placebo in eight young adult participants in a double-blind crossover trial over 12 weeks. As for the previous study, objective speech measures were based on adequate samples taken on multiple and varied occasions throughout the study. Reliability checks were also conducted but once again speech was not measured outside the clinic. Results from this study suggested a definite response for four participants, a probable response for two participants and no evidence of response for the final two participants. Outcomes measured included speech dysfluency and rate,

but side effects were not reported. Participants were advised to reduce their dose if side effects became intolerable, however it was not reported whether this actually occurred. Results were marginally statistical significant and it seemed the participants with higher levels of baseline stuttering showed greater improvements.

A study by Rantala and Petri-Larmi (1976) is the largest in this review. Sixty six adult and child participants were studied in a double blind, crossover study of haloperidol (up to 3 mgs per day for adults) versus placebo over an 8-week period. Like the previous two studies, speech measures were not based on samples from outside the clinic but objective and reliable measures were taken on multiple occasions throughout the study and were based on adequate samples of reading, conversation and spontaneous speech. Eight participants discontinued the treatment for various reasons, including four due to side effects of the drug.

Outcomes observed were stuttering symptoms and speech rate, and results showed statistically significant improvements in both compared to baseline after treatment with haloperidol and for stuttering symptoms but not speech rate after placebo. However, improvements in stuttering symptoms were greater after haloperidol compared with placebo but this difference was not statistically significant. The improvement in stuttering symptoms after haloperidol was of the order of 25-30%. Side effects were generally minor and almost as common after placebo, however one participant experienced extrapyramidal symptoms and discontinued medication. Only a few of the participants continued with haloperidol for more than 1 year but it was reportedly of great help for a few adults with severe stuttering who had not been treated previously.

Murray, Kelly, Campbell, and Stefanik (1977) undertook a double blind, cross-over study of 26 adults comparing haloperidol (3 mgs per day) with placebo

over 3 months. Speech measures appeared to be of sufficient quality as for the previous three studies, but once again recordings of speech were not made outside the clinic. Side effects were a significant problem during the course of the study, the worst being lethargy, occurring in 17 of the 26 participants. Eight participants discontinued the medication due to side effects, despite an improvement in speech in five of the eight. Unlike the previous studies, no anti-parkinsonian drug was taken in combination with haloperidol. Of the remaining 18 participants, both speech rate and number of dysfluencies improved for 11 participants on haloperidol compared to placebo, three participants benefited equally from both treatments and four participants were unchanged. Overall, improvements on haloperidol versus placebo were marginally statistically significant at best. The majority of participants in the study chose not to continue medication and two years later only one person was still taking haloperidol.

Andrews and Dozsa (1977) conducted another double blind, crossover comparison of haloperidol (3 mgs per day) with placebo in 27 adults. Only spontaneous speech was recorded for analysis and no measures were taken outside the clinic. However, reliability checks were conducted and recordings of adequate size were taken throughout the duration of the study. Once again, no antiparkinsonian drug was used in combination with haloperidol and 12 participants dropped out, eight due to side effects. One participant was on placebo at the time. Of the remaining 15 participants, there was a modest benefit due to haloperidol over placebo that was not statistically significant. Five participants had reduced the frequency in their stuttering and five had increased their speech rate by 50% or more while on haloperidol. Only one participant whose speech rate had increased by 400% continued on treatment after follow-up on the trial ceased.

Burns, Brady, and Kuruvilla (1978) published a study of 12 adults in a double-blind, cross-over design comparing the effects of haloperidol and apomorphine, an antiparkinsonian agent, with placebo over 4 days. Haloperidol was administered as a one-off injection of 0.5 mgs and the dose of apomorphine was 0.5 to 0.75 mgs, depending on age. The aim of this study was to determine whether significant reduction in dysfluency could be obtained at a dose (of haloperidol) that was unlikely to produce significant side effects. Apart from a lack of data from outside the clinic, speech measures appeared to be of reasonable quality and quantity. Results showed haloperidol was associated with modest but statistically significant reductions in dysfluency, and speech rate was maintained or even increased in some participants. Apomorphine had an inconsistent effect on the speech outcomes that was not statistically significant. The number of side effects appeared to be increased after apomorphine and haloperidol was associated with more instances of dry mouth and drowsiness or yawning.

The final study of the effects of haloperidol was conducted by Prins, Mandelkorn, and Cerf (1980). This study began with 21 adults and children in a double blind, cross-over design of duration 8 weeks, but only 14 participants completed the study. A detailed description of the speech measures taken throughout the study is provided in the report. This description indicates measures of adequate quality and quantity except out-of-clinic recordings were not included. Haloperidol (3 mgs per day) resulted in statistically significant improvements in disfluency and speech rate for those completing the study, but the sizes of the improvements were regarded to be clinically non-significant. Although benztropine was administered to counteract the side effects of haloperidol, five participants discontinued treatment while on haloperidol due to side effects.

There is a consistency among these eight randomised studies of the effect of haloperidol on stuttering. They were all randomised double blind studies and all participants were accounted for, however a number of participants did not complete the trial, usually while taking haloperidol. Speech measures were generally objective, based on repeated and adequate samples of speech taken throughout the trial process. Reliability checks were generally conducted but none of the studies included assessments on speech taken from outside the clinic. Long-term outcomes were also not measured in this group of trials. Due to the double-blinded design of these trials, it is assumed that all assessments of speech outcome were performed in a blinded fashion.

All eight studies showed a modest overall effect of the drug on stuttering. Some participants appeared to benefit quite considerably while others did not benefit at all. This treatment effect was in the form of reduced stuttering and increased speech rate and was evident with reading and spontaneous speech. Apart from the first study by Wells and Malcolm, which showed a more marked treatment effect, improvements in speech outcomes after haloperidol were in the vicinity of 20-30%, meaning the majority of participants were still stuttering significantly. Just as consistent as the improvements in speech outcomes were the reported side effects of haloperidol. These side effects were sufficiently severe to stop all but a few of the participants from continuing with the treatment indefinitely. However this was despite the use of additional drugs to negate the significant side effects of haloperidol. Thus it could be summarised that the modest improvements in speech outcomes obtained while on haloperidol were not sufficient to offset the significant side effects for most participants.

Other Trials of Pharmacological Treatments

The earliest trial identified (Edgren, Leanderson, & Levi, 1970) compared Diazepam, a muscle relaxant, with placebo in a double blind, cross-over design with 29 participants taken from the waiting list for stuttering therapy. In the 1-day trial, diazepam and placebo were administered on 2 days, 1 week apart. A full analysis had not been performed at the time the publication was written, however preliminary analysis suggested a small improvement in self-ratings of stuttering tendency (11 point scale) after taking diazepam. A second, 54-day treatment trial of 16 participants was also reported but analysis on speech performance had not been completed and was not presented. Side effects were not reported for either study.

Only a preliminary report was available for these trials and subsequent publications were not identified. However, the preliminary report stated that the first trial was a double blind, cross-over design. The design of the second trial was not stated. For the first trial, only 20 of the 29 participants randomised were included in the analysis and the speech outcome measured was self-reported stuttering tendency before and during public speaking after short-term treatment. Hence the outcome was not objective, not based on everyday speaking situations and not measured after long-term follow-up of the participants.

Rustin, Kuhr, Cook, and James (1981) undertook a complicated trial of speech therapy in conjunction with drug therapy in 31 adults over 12 weeks. The effect of oxprenolol, a beta-adrenergic-blocking agent, was assessed in double blind, crossover fashion as well as the effect of speech therapy in single blind, parallel group fashion. Speech therapy involved progressive relaxation and a slowed speech technique delivered over a 6-week period. There is lack of detail on the speech measures however it appeared that there were no baseline measures or beyond-clinic

data and only number and duration of blocks and speech rate were assessed.

Oxprenolol lowered pulse rate and blood pressure but appeared to have no effect on speech outcomes. Speech therapy, on the other hand, was associated with significant reductions in frequency and duration of stuttering. Reductions were in the order of around 50% and were enhanced with a 6-week maintenance phase however significant stuttering of around 3% remained at the end of follow-up.

The next two studies examined verapamil, an antihypertensive drug that has pharmacological effects on both smooth and striated muscle. The first of these two studies was by Brumfitt and Peake (1988). Fourteen adult participants were studied in a double blind, crossover study comparing verapamil (80 mg twice daily) with placebo over 6 weeks. The outcomes observed included self-reported and speech pathologist-reported severity of stuttering, heart rate, blood pressure, side effects, as well as objective measures of speech rate and frequency of stuttering using tape recordings of speech. Speech was assessed under three speaking situations, all apparently within the clinic, of 1-minute duration each. Measures were taken at baseline and after the two treatment conditions but reliability was not assessed. The results suggested no useful effect of verapamil on any of the speech outcomes measured. There was also no evidence of significant side effects caused by this treatment.

The second study of verapamil was by Brady, Price, McAllister, and Dietrich (1989) who conducted a double blind, crossover trial of 10 adults receiving either verapamil (80 mg three times daily) or placebo over 3 weeks. Outcomes assessed in this trial included frequency of stuttering, speech rate, and side effects as well as psychological outcomes. Assessments were made before and after each treatment condition using speech samples of 1000 words or 3-minutes duration. Within-clinic

spontaneous speech as well as reading was assessed in blinded fashion and reliability checks were conducted. The results for this trial were more positive than for the previous trial. While taking verapamil, frequency of stuttering was reduced and speech rate increased, compared to placebo. However these differences were only statistically significant during reading of standardised passages. No difference was observed during spontaneous speech tasks. Also, the reductions in frequency of stuttering were very modest (around 20%) as was the increase in speech rate. Reported side effects were similar while on verapamil or placebo. Eight of the 10 participants continued to take the active medication after the trial was completed, however, 6-8 months later only two participants were still taking verapamil.

Kampman and Brady (1993) compared bethanechol (cholinergic agent) with placebo in a double blind, crossover study of 10 adults over 2 weeks. Multiple objective measures of stuttering frequency and speech rate based on 1,000 syllable or 3 minute speech samples were taken throughout the study and later scored blindly. All 10 participants completed the study and findings suggested bethanechol was not superior to placebo in reducing frequency of stuttering. Results on all measures of speech taken were very similar while taking bethanechol or placebo. However, two participants in the trial believed they were more fluent while on active medication and continued with the medication after the completion of the trial. These two participants were still taking bethanechol one year later and believed their speech remained fluent. There was no difference between the active medication and the placebo in frequency and severity of side effects.

Althaus, Vink, Minderaa, Goorhuis-Brouwer, and Oosterhoff (1995) studied the effect of clonidine on stuttering in children. Clonidine is an antihypertensive drug, effective in the treatment of ADHD and used in the treatment of Tourette

syndrome. Twenty five children, aged 6-13 years were studied in a placebo-controlled, double blind, crossover study over 28 weeks. Parents and teachers took various subjective and objective measures of frequency of stuttering based on three different speaking situations throughout the trial. Speech rate and side effects were not mentioned in the report. The results provided no evidence of effect of clonidine on the frequency of stuttering but children did appear to be less hyperactive, more task oriented and more approachable by others.

Gordon, Cotelingam, Stager, Ludlow, Hamburger, and Rapoport (1995) compared two antidepressants, clomipramine and desipramine in a double blind, crossover design with 16 adults and one adolescent over 10 weeks. Speech measures were taken before, during and after each treatment condition. Clomipramine was associated with statistically significant reductions in self-reported stuttering severity as well as speech rate compared to desipramine. However, objective measures of severity were not reported and the self-reported improvements were 20-30%. Side effects were not directly reported but it appeared that they were not severe as half of the participants were still taking clomipramine 12 months after the trial had concluded.

Stager, Calis, Grothe, et al. (1997) conducted a double-blind, placebo-controlled, crossover trial to examine the effects on fluency of paroxetine and pimozide in adults over 18 weeks. Pimozide is a highly selective dopamine antagonist while paroxetine is a highly selective serotonin reuptake inhibitor. Eleven participants were included in the study however the study was stopped early due to significant side effects for two participants while withdrawing from paroxetine. One participant suffered from behavioural dyscontrol and the other from suicide ideation. Prior to stopping the trial, six participants completed the pimozide phase, six

completed the placebo phase, and five completed the paroxetine phase. Results were based on self-ratings of fluency but also within clinic change in percent fluency and suggested four participants had improved fluency on pimozone (13 to 46% improvement in percent fluency) while two participants had worse fluency (9 to 17% decrease in percent fluency). No significant clinical response was found for any participants while taking paroxetine.

Risperidone, an antipsychotic used to treat the symptoms of schizophrenia, was the subject of a trial by Maguire, Riley, Franklin, and Gottschalk (2000). This parallel group, placebo-controlled, double blind study of 16 adults was conducted over a 6-week period. Speech was assessed prior to and throughout the trial process during within-clinic conversation and oral reading of standardised passages of 500 syllables. Reliability checks were conducted on the various speech measures. All measures of stuttering severity reduced on both active drug and placebo but only significantly so for those participants on risperidone. Of the outcomes measured only stuttering frequency was statistically significant when the two treatment groups were compared. Reductions in %SS were around 50% for the group on active drug compared to around 25% for the placebo group. Risperidone was generally well tolerated however three participants reported sedation. This was resolved in all three cases when the dose of active drug was reduced.

The final study in the group of pharmacological treatments is a trial of another antipsychotic drug, olanzapine, by Franklin, Riley, Maguire et al (2002). This trial was a double blind, placebo-controlled, parallel-group study of 24 adults over 3 months. Participants were rated on multiple objective and subjective speech measures. The results showed statistically significant reductions in objective measures of stuttering frequency for those participants on active drug compared to

those participants on placebo. Side effects appeared to be minimal and all participants agreed to go on to active drug in a long-term open label study after completion of the randomised trial. The size of the treatment effect or magnitude of the reduction in stuttering severity while on olanzapine was not reported. The information from this study was obtained from conference proceedings (Franklin, Riley, Maguire et al, 2002) as a full manuscript was not available at the time of writing.

Regulated Breathing

The next group of studies involve the evaluation of regulated breathing for treatment of stuttering. This treatment was introduced by Azrin and Nunn (1974) and has been the subject of a review (Woods Twohig, Fuqua, & Hanley, 2000), as described previously (pp 51). Regulated breathing in its original form, and after simplification, has been the subject of many studies, of which five were randomised controlled trials and so are included in the current review. Because this is the only non-drug treatment that has been the subject of more than one randomised study, it is of great interest. The following paragraphs will describe each of the five studies in chronological order and then a summary paragraph will be presented. Only one of the five studies reported estimates of variation hence a meta-analysis could not be performed.

The first randomised study of regulated breathing was by Azrin, Nunn and Frantz (1979). Thirty-eight adults and children were randomly assigned in parallel to either regulated breathing (N = 21) or abbreviated desensitisation (N = 17) and followed for 3 months. Desensitisation (Wolpe, 1958) involves the manipulation of anxiety through imagery and the use of deep muscular relaxation. Regulated breathing involved teaching the speaker to breathe smoothly and deeply, pause at

natural speech junctures, plan ahead for the content of the speech, and relax chest and neck muscles. Behavioural procedures were also employed such as relaxation training, self-correction for errors, social support, daily home practice, and response awareness. The results of this study showed an impressive reduction in stuttering for the regulated breathing group of around 95% but only a 10% reduction for those participants receiving abbreviated desensitisation. However, the main outcome of this study was self-reported frequency of daily stuttering episodes. No statistical comparison was reported and data at 3 months were only available for 25 of the 38 randomised participants.

A second study by Ladouceur, Boudreau, and Theberge (1981) compared regulated breathing with or without awareness training in a parallel group study of 16 adults and adolescents over one month. Awareness training was achieved in a 2-hour session where participants allocated to this group were taught to become aware of their stuttering by pressing a counter whenever they stuttered. Training in regulated breathing, as described previously, was then undertaken. The control group received training in regulated breathing without the awareness training. Frequency of stuttering and rate of speech was assessed during five within-clinic assessment periods (baseline 1, post awareness training, baseline 2, post regulated breathing training, and follow-up) by an observer who was unaware of the nature of the research. Only conversation was assessed and reliability checks were performed on 30% of the speech samples.

Results showed that the awareness training reduced stuttering frequency compared to no awareness training ($p < 0.05$) and that regulated breathing reduced stuttering frequency even more markedly ($p < 0.05$) however there was no difference between the two groups at both follow-up occasions. The size of the reduction in

stuttering frequency was more than 50% compared to baseline and was maintained at the second follow-up occasion however both groups still had significant stuttering of around 4%. Speech rate was not significantly different for either group at follow-up compared to baseline.

Ladouceur and Martineau (1982) assessed regulated breathing training in 21 children (ages 5 to 16 years) in their parallel group study over one month. Three groups were compared: regulated breathing training with or without parental assistance and a waiting list control group. Parental assistance involved the parents of the participants learning the concepts of regulated breathing in a 2-hour session, then applying them to their children in two 45-minute sessions per week under supervision of the clinician. Tape recordings of within-clinic speech at baseline, post-test and follow-up were assessed for frequency of stuttering and speech rate by an observer who was unaware of the nature of the research. Four-minutes of speech were assessed on each occasion and reliability checks were performed on 30% of the speech samples. No statistically significant differences were found, however results suggested both groups receiving training in regulated breathing reduced their stuttering by around 40% and these reductions were maintained at 1-month follow-up. The control group, in contrast, increased their level of stuttering by about 20%, on average, over the follow-up period.

A fourth study, by Saint-Laurent and Ladouceur (1987), involved another variation on the regulated breathing method of treatment for stuttering. An intensive method was introduced as well as a maintenance phase in a parallel group study of 40 adults over 10 months. The five groups under study consisted of intensive regulated breathing, with and without a maintenance phase, standard regulated breathing, with and without a maintenance phase, and a no treatment group. The

intensive group received three daily, 8-hour training sessions, as opposed to the standard group who received 8 weekly, 3-hour sessions. Maintenance involved an extra eight 2-hour clinic sessions 2 weeks apart and 20 minutes of daily practice. Speech was recorded under four different conditions including conversation in a non-clinic setting and assessed by an independent observer for speech disfluency and rate. Assessments occurred at baseline, during treatment and at 10-month follow-up using adequate samples of speech. Reliability checks were performed on 10% of the recordings.

The results of this study showed that those participants receiving any form of regulated breathing training reduced their frequency of stuttering significantly more than the no treatment group ($p < 0.05$). There was no difference between those receiving intensive treatment and those receiving standard treatment and the maintenance phase provided no additional benefit. Reductions in stuttering were in the vicinity of 50% for the participants who received regulated breathing training and were largely maintained at 10 months follow-up, however stuttering frequency was still around 5%, on average, indicating significant stuttering at the final follow-up point. Speech rate increased for all groups, significantly so for the groups that received regulated breathing training.

The final study in the group of regulated breathing was by Waterloo and Gotestam (1988). This parallel group study of 32 adults compared regulated breathing with a waiting-list control group over 8 months of follow-up. Regulated breathing training was delivered in one session of between 2 and 3 hours and speech was assessed for speaking rate and frequency of stuttering during a variety of speaking situations at baseline and immediately after treatment as well as 2, 3 and 8 months later. Two independent observers assessed the speech samples of 200 words

and reliability checks were conducted. Results indicated an impressive and statistically significant reduction in frequency of stuttering for those participants in the regulated breathing group compared to controls who remained at similar levels throughout the study. Reductions were initially approximately 90% for the regulated breathing group but levels of stuttering gradually increased over the 8 months of follow-up, reaching an average of around 5% at the 8-month point. Speech rate also increased in the treatment group but was unchanged in the controls.

The five trials of regulated breathing employed quite different designs and were of differing quality but appeared to give consistent results. The quality of the speech measures employed improved with each subsequent trial. Results obtained from all five trials indicated a moderate effect of treatment on frequency of stuttering. Reductions were in the range of 40-90% however participants retained significant levels of stuttering after treatment of around 4-5% of syllables spoken. Reductions in frequency of stuttering were maintained for up to 10 months after treatment. This reduction was achieved with quite variable amounts of clinician time ranging from 3 to 40 hours of training in total. There was also evidence that speech rate was increased. Unlike the haloperidol trials, participants did not experience side effects from regulated breathing training. This is one significant advantage of speech therapy treatments over drug treatments. It was also found that regulated breathing appears to be effective for all age groups from 5 years upward.

Other Non-Pharmacological Treatments

The earliest randomised trial of a non-drug treatment identified was by Peins, McGough, and Lee (1972). This parallel-group study of adults and adolescents compared three groups of 12 participants over 6 months. One group received standard therapy based on the procedures and principles of Van Riper (1963, pp.

380-412). This therapy included 1-hour, weekly sessions with an experienced speech clinician and involved avoidance and anxiety reduction in a symptomatic therapy system. The experimental group received one tape-recorded, 30 minute, highly structured lesson each month and saw a clinician for 30 minutes twice monthly. Each taped lesson included a number of speaking tasks and instructions were given to speak more slowly and smoothly in a manner of speaking termed legato style. The third group was a waiting list control group and did not receive any speech therapy.

The main outcome of the trial was measured on a 7-point scale of severity of stuttering that was rated by an experienced speech pathologist blinded to the identity and hence treatment group assigned to the participant. Reliability was checked and multiple 1-minute samples of speech were taken at baseline and at 3-monthly intervals from within the clinic. Results were obtained for all 36 participants and suggested statistically significant reductions in severity for the two groups receiving speech therapy but not for the control group. A comparison of both therapy groups combined showed a statistically significant difference compared to the control group. However, improvements were modest with the average score reducing from about 5 at baseline to just over 4 at 6 months for the two therapy groups. The greatest improvement for any of the participants was a reduction of 2 points on the 7-point scale. All participants had significant stuttering at the 6-month end-point.

Moleski and Tosi (1976) investigated the efficacy of comparative psychotherapy. This parallel group trial of 20 adult participants over one month compared rational-emotive therapy (with and without in vivo tasks) with systematic desensitisation (with and without in vivo tasks) and there was a no-treatment control group. Systematic desensitisation involves the manipulation of anxiety through imagery and the use of deep muscular relaxation (Wolpe, 1958) whereas rational-

emotive therapy employs a cognitive restructuring of specific irrational ideas underlying self-defeating emotional or behavioural states (Ellis, 1961). The in vivo behavioural tasks were telephone calls to friends and spontaneous conversations with strangers.

Reliability checks were performed on speech measures but it was unclear whether ratings were performed in a blinded fashion. Ratings were not performed on speech from outside the clinic but repeated evaluations were made for reading as well as 5-minute samples of spontaneous speech. Results were presented as *p*-values (<0.05 or <0.1) from a number of comparisons using analysis of variance. No other details were presented however conclusions suggested rational-emotive therapy was superior to both systematic desensitisation and no-treatment in terms of reducing stuttering behaviour. Systematic desensitisation was more effective than no-treatment in reducing speech dysfluencies.

Eversham and Fransella (1985) conducted a randomised study of relapse in a group of 48 people who stutter over 24 months. The participants received speech therapy in the form of prolonged-speech training (slowed speech incorporating extended vowels and continuous voicing) with or without a psychological reconstruction programme in a parallel group study. The reconstruction programme involved helping the participants to reconceptualise themselves as fluent speakers so that they were less likely to fall back into bad habits after speech therapy had finished.

Recordings of within-clinic spontaneous speech were taken at baseline as well as a number of follow-up occasions and 5-minute samples were used for the assessments. Speech outcomes were not assessed blindly however reliability checks were performed by independent speech pathologists. Results suggested significant

decreases in frequency of stuttering and increase in speech rate for both groups. They also suggested the group that received psychological reconstruction had significantly less relapse at the 24 month end-point. However, absolute changes in speech outcomes were not presented. Outcomes were dichotomised into significant/non-significant improvement and chi-square significance tests performed. Hence it was not known how much improvement was achieved. Also the two groups had quite different levels of stuttering at baseline.

The only study to evaluate treatment of preschool aged children was conducted by Harris, Onslow, Packman, Harrison, and Menzies (2002). In this study, the Lidcombe Program (see Chapter 1) was compared with a wait-list control in a parallel-group study of 23 children over 12 weeks. This study did not present the Lidcombe Program treatment to completion. The primary outcome was %SS, assessed blindly and subjected to intra-rater and inter-rater reliability checks. Assessments of speech were made before and during treatment in a variety of within and beyond-clinic speaking situations. The results of this study showed %SS reduced by approximately 60% in the Lidcombe Program group and approximately 30% in the control group. A comparison of the two groups indicated a statistically significant difference in the reduction of %SS in favour of the Lidcombe Program.

Discussion

A systematic review of randomised, controlled trials of treatments for stuttering found that 27 trials met the inclusion criteria. These 27 trials can be divided into studies that examined the effectiveness of drug treatments and those that examined non-drug therapies. The drug trials were generally double blind, placebo-controlled and often crossover designs whereas the non-drug studies were generally parallel group designs unable to be double-blinded. However, in the majority of the

non-drug trials it was stated that outcome assessment was performed in a blinded fashion and reliability checks were often conducted. Presentation of means and standard deviations for the speech outcomes assessed was rare hence meta-analysis could not be performed.

Only three treatments were studied in more than one trial. Those treatments were haloperidol (eight studies), regulated breathing (five studies) and verapamil (two studies). Sample sizes ranged from eight to 66 participants but most studies included 10-30 participants. Sample size calculations were not included in the methods section of any of the trial reports. The duration of the studies ranged from the effect of treatment over a few days to follow-up of 24 months in one study of relapse. Most studies assessed the participants over a few weeks, or months in some cases. Beyond-clinic measures of speech were rarely included in the outcome assessments.

In the majority of cases the trials in this systematic review appeared to be of reasonable quality. However there are a number of ways that the quality of future trials of stuttering treatments could be improved. Recommendations for future trials are outlined in the following paragraphs:

1. *Participants are truly randomised into the treatment groups and the method for doing this is clearly stated in the methods section of the trial report so that allocation concealment can be determined.* Although all of the trials in this review reported allocation of treatment to be randomised, the method of allocation was generally not well described.

2. *An appropriate treatment is chosen for the control group such as placebo, wait-list control or standard practice therapy.* This was done for the majority of trials but in some cases no justification was provided for why a particular treatment

was chosen as the comparison group therapy. If there is no standard practice treatment then the control group would normally receive placebo or no treatment.

3. *Trials of drug treatments are double blinded.* If double blinding is not possible as is generally the case for non-drug treatments, then outcome assessment could be performed blinded to treatment allocated. The method of blinding the outcome assessment needs to be clearly stated in the methods. Most of the trials in the review stated they had employed either double blinding or blinded outcome assessment however details of how the blinded outcome assessment was achieved were rarely sufficient to indicate whether it had been done appropriately.

4. *All randomised participants are accounted for in the analysis and the primary analysis is by intention-to-treat.* This appeared to have been done in most studies however it is becoming more and more common to report the flow of participants through the various stages of a randomised trial in a flowchart diagram. With this method it is clear how many participants failed to complete the trial and the reasons why this occurred. The effect of any drop-outs or missing data on the results of the trial should also be assessed via appropriate imputation methods, sensitivity analyses, and comparisons of characteristics of subjects who dropped out with subjects who remained in the trial.

5. *A sample size calculation is performed at the design stage of the trial and details included in the methods section of the trial report.* Sample size calculations were not presented in any of the trials reported in the present review.

6. *Appropriate summary statistics, such as means and standard deviations, are reported for all speech outcomes.* This rarely occurred in the reports of trials in the present review which caused the problem that inclusion in a meta-analysis was not possible.

7. *Participants are followed for a reasonable period so that the long-term effects of treatment can be assessed.* The duration of follow up of the included studies was generally modest and only a few weeks in most cases which is sufficient to determine short-term benefits of treatment only. Because relapse is a significant problem with stuttering, it is desirable that follow up be extended to a number of months or even years if possible.

8. *Recordings of speech are taken during beyond-clinic as well as within-clinic speaking situations.* Beyond-clinic recordings of speech were rarely included in the group of studies reported in this review. Therefore generalisation of treatment effects beyond the treatment clinic could not be determined in most cases.

More details on how to report the results of randomised trials appropriately can be found in the CONSORT statement revised recommendations for improving the quality of reports of parallel group randomised trials (Altman, Schulz, Moher, Egger, Davidoff, Elbourne, et al., 2001).

Summarizing all the treatments assessed in this systematic review, only two have reasonable evidence to suggest they are effective in reducing stuttering: haloperidol and regulated breathing. Haloperidol is associated with a modest reduction in frequency of stuttering of around 20-30% and increased speech rate. Its effect is inconsistent as some participants receive a large benefit but some no benefit at all. Further, side effects prohibit the long-term use of haloperidol in the majority of people who stutter. Therefore this drug would not be recommended for the treatment of stuttering except perhaps for those with very severe stuttering who show a marked benefit without significant side-effects. However the long-term side effects of haloperidol may prohibit its long-term use in this patient group.

Two other antipsychotic medications, risperidone and olanzapine, show some promise as they appear to be as effective as haloperidol but without the significant side-effects, although three participants on risperidone did experience sedation and the full manuscript was not available for the olanzapine trial at the time of writing. They have, however, only been subjected to one randomised study each, and the reduction in frequency of stuttering appears to be only modest at best. Other drug treatments included in this review included anti-depressants and anti-hypertensive drugs however none appeared to be effective in reducing stuttering.

The majority of the non-drug therapies assessed in randomised studies appeared to reduce stuttering by around 50% on average. These treatments were not associated with significant side effects such as those of the drug treatments. It must be pointed out that only regulated breathing was studied more than once and most of the trial designs did not allow a thorough assessment of the treatment to be performed. For example, two studies examined only the immediate effect of treatment while the method of analysis for some of the other studies was somewhat crude. An appropriate control group was not included in some of the studies and the quality of speech measures was sub-standard in some cases.

Although the studies of regulated breathing were of mixed quality and design, the consistency of the evidence suggests regulated breathing is an effective treatment for the reduction in frequency of stuttering and increase in speech rate. This treatment appears to reduce stuttering by approximately 50% on average, with no apparent side effects. Participants from 5 years and upwards were included in the studies reported. On the negative side, however, is evidence that the participants appear to be left with significant stuttering post-treatment of around 4-5% of

syllables spoken. The long-term effects of the treatment were also not adequately presented in this group of trials.

Of the 27 studies included in the review, only three examined treatment of children exclusively. Clonidine, an anti-hypertensive drug, was assessed in 25 children, aged 6-13 years, but showed no evidence of effectiveness. Regulated breathing was assessed in 21 children, aged 5-16 years, and appeared to be effective in reducing stuttering by around 40%. However, statistical significance was not reached. The final study of children evaluated the Lidcombe Program in preschool children (2-5 years). The treatment was not run to completion in this experimental study. Nonetheless, the Lidcombe Program was effective in reducing stuttering by 60% in 12 weeks and was statistically significantly superior to a wait-list control group which improved by 30%. This was the only randomised study identified that evaluated a treatment for early stuttering.

Chapter 3: Statistical Power in Stuttering Research

Introduction

It was apparent from the previous chapter that randomised studies of treatments for stuttering to date are of modest size. The largest trial included 66 participants however the remaining studies were much smaller, comprising just 10-30 participants in most cases. Hence this chapter reviews one of the most critical aspects of research design, namely sample size and statistical power. Prior to the review itself the importance of designing research studies of sufficient size and power is discussed, along with an overview of the mathematics behind the calculation of statistical power. This chapter provides valuable background information to the design of the randomised controlled trial of the Lidcombe Program presented in Chapter 7.

The ultimate goal of statistical inference in health research is to arrive at inductive statements about populations. The stuttering literature has many examples of this goal being achieved over a number of decades. The effects of rhythm, chorus reading and associated speaking conditions, the adaptation effect, and the effects of operant methods on stuttering are now accepted attributes of the condition (for reviews, see Bloodstein, 1995; Ingham, 1984; Van Riper 1982), and replicated findings have led to inductive statements about prevalence and incidence, and about the natural course of stuttering (for reviews, see Andrews, 1984; Yairi & Ambrose, 1999). Replicated findings on response contingent stimulation and prolonged-speech, both of which can ameliorate stuttering, have allowed confident general statements to be made (for reviews see Bloodstein, 1995, Ingham, 1984; Onslow, 1996). These examples of inductive statements all relate to positive findings. A positive finding, replicated by independent researchers, can lead to conclusions about the population under study.

However, the accumulation of knowledge through research is not constrained to positive findings. It is also important to accumulate null findings about populations. Much current stuttering research is directed at areas where null findings would provide important information for understanding the nature and causes of the condition. Examples include the role of lexicalization, phonological encoding and other linguistic variables. It is as critical to make null statements in those domains of inquiry as it is to make positive statements. For if there are no significant effects related to these variables, then we need to establish that information and move on to other areas of theoretical and clinical interest.

In health research, and the social sciences generally, it is rare for every individual in a population to display a characteristic that can generally be attributed to the population. Statistical inference is essential not only to the establishment of a significant result, but also to the establishment of a null finding about a population. As pointed out by Young (1994), null results achieved in studies with adequate statistical power permit more extensive scientific reasoning than significant results. Group difference designs which are commonly used in stuttering research are, in essence, studies of the correlation of variables. Consequently, a significant finding of group difference does not suggest anything about causal relationships between variables, however a null result in a group study with sufficient power can rule out the possibility of a causal link between variables (Young, 1994).

Statistical power is a critical methodological issue in the design of a study. Despite this, consideration of power has been generally neglected in the social sciences (Cohen, 1992). In this chapter the statistical concept of power is described, parameters of research that influence it are outlined, and calculation of power is demonstrated.

Power and errors in statistical inference

In most instances it is not possible to measure a characteristic in an entire population thus a sampling procedure is used to select a subset of the population. Once samples have been selected, inference can be made from the sample and then generalised to the wider population. Thus, whenever statistical inference is made from a sample, it is subject to some error. With an adequate sampling procedure (e.g., one without selection bias), systematic error is reduced so that only random error remains. Once inferences are made about the population at large, two main types of such error can occur. As stated in the previous chapter Type I error occurs when, from the sample, it is declared that a difference in outcome between groups exists, when in the population this is not the case. A Type II error occurs when no difference is found whereas in the population a real difference exists. These two sources of error are not independent of each other, and later it is demonstrated how they may cause statistical problems with power in research.

The power of a study is its capacity to detect a difference if a difference truly exists in the wider population. This is usually represented as a probability. Consider a comparison of anxiety in speaking situations between a control group and a group of people who stutter in a study powered at 0.80. This would mean that, if the population of stuttering participants and the population of control participants indeed do differ in speech-related anxiety, then the study would have an 80% chance of detecting this difference. A number of studies producing non-significant results, each at a power of 0.80, would make a compelling case that indeed no differences in speech anxiety between the populations of stuttering and control participants exist.

The power of a study is equal to the complement of the probability of a Type II error ($1 - \text{Type II error}$). This is because a Type II error occurs when a statistically

significant difference is not detected when it is present, and power is the probability to detect a statistically significant difference if it exists. Cohen (1988) recommends minimum power of 0.80 and this has become accepted as a benchmark, just as 0.05 has been accepted as a benchmark for the probability of a Type I error in a study. However, just as investigators are free to impose more stringent criteria for detecting Type I errors in determining statistical differences—such as .01—power may be set higher than 0.80 in an experiment.

Fundamental Parameters that Influence Power

Nature of the Dependent Variable

The nature of the dependent variable will affect power. For example, when the outcome of a study is a mean power is calculated differently from a study where the outcome is, say, a proportion. This is illustrated in the formulae and examples below.

Effect Size

The effect size is the size of the difference that the study is designed to detect. The “minimum worthwhile difference” is the smallest difference between two or more groups that would be clinically worthwhile detecting. Such a difference might be quite small in a theoretical speculation about a condition, but in a clinical study the minimal worthwhile difference is often considerable. In calculating the power of a study, effect size can be estimated from available data that indicate the likely size of the actual population difference being investigated, or in terms of the least worthwhile difference. An estimated effect size using a pilot or previous study or studies could be clinically trivial or clinically substantial. A clinically trivial effect

size estimated from a previous study or studies would provide little motivation to conduct another similar study.

Cohen (1988) categorises effect sizes into “small,” “medium,” and “large.” For example, if the sample means of two normally distributed populations are to be compared, such as mean anxiety scores for stuttering and non-stuttering participants, using a 2-sample t-test, then the effect size is given by $[\text{mean 1} - \text{mean 2}] / \text{SD}$, where SD is the estimated, pooled standard deviation of the two groups. A small effect size is 0.2, medium is 0.5 and large 0.8. Although these notional effect sizes simplify sample size calculation, if a more accurate estimate of effect size is available, it is better to calculate sample size based on this more accurate estimate. For example, assume that a study involves a comparison of two normally distributed samples as above, that power is set at .8, and 2-tailed significance level is set at 0.05. The three required total sample sizes for the proposed 2-group study will be 52 (large effect), 128 (medium effect) and 788 (small effect). However, a realistic estimate of the effect size could be somewhere between 0.2 and 0.5 (perhaps 0.35) thus the sample size required to produce a power of .8 could be somewhere between 128 and 788. Hence if a sample size of 128 is used, the study power will be below .8. But if a sample size of 788 is used the study power will be greater than .8, resulting in an inefficient study, where more participants are recruited than are necessary to answer the question.

Population Variability

Power decreases with increase in variability of the population to be sampled. For normally distributed data, the variability is usually expressed as the standard deviation (SD). An estimate of the likely standard deviation and difference in means

under study can be determined from a pilot study or previous research. This enables a measure of effect size to be established with the formula $[\text{mean 1} - \text{mean 2}] / \text{SD}$.

Significance Level

The significance level specified determines the probability of committing a Type I error. A lower level of significance results in a smaller probability of making a Type I error. With other variables remaining constant, power reduces if a smaller significance level is specified. A level of significance of 0.05 is most commonly used when designing studies. However, if it is important that a Type I error not be committed—such as with a study that is unlikely to be replicated in future—then a value of less than 0.05 may be specified.

One-tailed or 2-tailed Testing

A decision whether to specify a 1-tailed or 2-tailed test will affect power because a study that assumes a 1-tailed test will have more power than a study that assumes a 2-tailed test (all other things being equal). This is because a 1-tailed test assumes that any effect of interest can only be in one direction. In practice 2-tailed tests are generally used in preference to assuming that an effect in the opposite direction does not exist.

Sample Size

Once the parameters outlined above are determined, the larger the sample size, the greater the power: if the other parameters remain constant then the larger the sample size, the more information about the true difference is obtained. This results in the study having a greater capacity—that is, power—to detect a difference if it actually exists in the population. Smaller sample sizes will have less power to detect the same difference. A general formula in words is:

Sample size per group (n) =

[Function of significance level (α) + Function of power ($1-\beta$)] / [Size of difference (d)]²

The sample size (n) and the size of the difference (d) are related as follows:

$$n \propto 1/d^2$$

This means that sample size is inversely proportional to the effect size squared, as illustrated in Table 3.1. This relationship can be written as:

$$n = \{(Z_{\alpha/2} + Z_{1-\beta}) / \Delta\}^2$$

where $Z_{\alpha/2}$ is the Z-value for a 2-sided α level test (Z-values can be obtained from a table of the standard normal distribution which can be found in most elementary statistics texts)

$Z_{1-\beta}$ is the Z-value corresponding to the power

Δ is the standardised difference (effect size)

Comparing two proportions: p_1 vs p_2

$$\Delta \approx (p_1 - p_2) / \sqrt{p_1(1-p_1) + p_2(1-p_2)}$$

Comparing two means: $\mu_1 - \mu_2$

$$\Delta = (\mu_1 - \mu_2) / \sigma\sqrt{2}$$

where σ is the population standard deviation

Table 3.1: Illustration of sample size (n) being inversely proportional to the effect size (Δ) squared

Sample Size (n)	Effect Size
100	Δ
400	$(1/2) \Delta$
1600	$(1/4) \Delta$
6400	$(1/8) \Delta$

It is not suggested that the formulae given above should be used to calculate sample size in practice because computer software packages can do that. The formulae are provided to illustrate the mathematics behind a basic sample size calculation. There are a number of software programs available for the calculation of power and sample size including GPOWER (Faul & Erdfelcher, 1992), STPLAN (Brown, 1993), and PS (Dupont & Plummer, 1990), all of which are available on the internet for free. A number of statistical software packages include a sample size calculator; for example S-Plus (MathSoft, 2000). There are also a number of texts that include sample size tables such as Machin, Campbell, Fayers, and Pinol (1997) and Lemeshow, Hosmer Jr, Klar, and Lwanga (1990). Textbooks that describe power analysis include Kraemer and Thiemann (1987) and Lipsey (1990).

In designing studies, it is critical to recognise that sample size calculation yields a broad estimate of the number of participants required. As is the case with the statistical analysis procedure itself, the calculation rests on various assumptions. These assumptions are ideally based on the best information available but this information may not be completely accurate. Also worth noting is that different sample size calculators may give slightly different results. This could be due simply

to rounding of fractions, or it could be caused by the calculations being based on different statistical tests or different assumptions. For example, a sample size calculation for the difference between two proportions could be based on Fisher's exact test or uncorrected chi-square. Hence, it is advisable to be conservative when determining what sample size should be used. It is possible to design a study so that any subsequent information becoming available during the running of the study can be used to adjust the sample size accordingly however this concept is beyond the scope of this thesis.

Other Parameters that Influence Power

Allocation Ratio

Allocation ratio is the ratio of participants to be recruited to each group of the study. An allocation ratio of one implies there will be equal numbers of participants recruited to each group. Power decreases as the allocation ratio moves away from one. Sometimes allocation ratios other than one are chosen when, for example, one of the treatments is very expensive or complicated to administer (Kirby, Gebski, & Keech, 2002).

Compliance

This is a particularly important issue with clinical trials comparing different treatments. Compliance is the proportion of participants who remain in the study receiving treatment as specified in the trial protocol. Consider a trial comparing a new treatment with a control group. In such a study, there will typically be drop-ins and drop-outs. Drop-ins are participants who are assigned to the control group but who receive active treatment. Drop-outs are participants who are assigned to the

active treatment but do not complete this treatment. Drop-ins and drop-outs tend to dilute the treatment effect: they tend to make the observed treatment effect less than the actual treatment effect. Power decreases as non-compliance increases. To adjust for non-compliance in a 2-group study, the sample size needs to be increased by a factor F , where

$$F = 1/(c_1 + c_2 - 1)^2$$

and C_1 and C_2 are the proportions of participants who remain in their allocated groups (or take the medication as per protocol) for the duration of the study (Piantadosi, 1997). Table 3.2 shows F values, indicating the proportional increases in sample size required to maintain power for compliance rates of c_1 and c_2 . Note that the proportional increases given in Table 3.2 are only accurate to one decimal place.

Table 3.2: Approximate proportional increases in sample size required to maintain power for compliance rates of c_1 and c_2

c_2	c_1	100%	90%	80%
100%		1.0	1.2	1.6
90%		1.2	1.6	2.0
80%		1.6	2.0	2.8

Issues Associated with Failure to Consider Power a Priori

To illustrate some of the issues associated with failure to consider power a priori, a body of research was considered with reference to sample size and power. Papers dealing with stuttering published in the Journal of Speech, Language and Hearing Research (Vol 39, No. 1 to Vol 40, No, 4) and the Journal of Fluency

Disorders (Vol 21, No. 1 to Vol 22, No, 3) were selected for analysis with the criterion that they used groups to search for at least one possible effect. Forty-five papers dealing with stuttering were published in those journals over that period, and 26 met the above criterion. Papers that were not included were case studies, descriptive studies, or literature reviews. Table 3.3 presents the results of this analysis of the 26 studies. The number of groups studied as well as the number of participants in each group is presented along with the main conclusion of each study and whether researchers reported that a power analysis was performed. Of the 26 included studies, four were single-group studies, 17 were 2-group studies, three were 3-group studies and two were 4-group studies. The numbers of participants in each group ranged from five to 55 and in one case more than 600. Most were in the range 10 – 20 participants.

For each study, an approximate power was determined based on the sample size and the type of statistical analysis used to summarise the data, and small, medium, and large effect sizes as described by Cohen (1988). Assumptions for the power calculations were a significance level of 0.05, a 2-tailed test, and equal variances for the groups. Power was determined using tables from Cohen (1988) and the computer package STPLAN (Brown, 1993). This procedure is intended to illustrate the importance of considering power prior to beginning a study.

Post hoc power calculations are sometimes used to aid in the interpretation of statistically non-significant results however this approach is fundamentally flawed because the observed difference in outcome between the groups is used to estimate the clinically relevant difference which is incorrect (Hoenig & Hersey, 2001).

Table 3.3: Details of papers dealing with stuttering published in the Journal of Speech, Language and Hearing Research (Vol. 39, No. 1 to Vol. 40, No. 4) and the Journal of Fluency Disorders (Vol. 21, No. 1 to Vol. 22, No. 3)

Study	No. of groups	Parti- pants per group	Main Conclusion	% Power (small, medium & large effect)
1. Craig et al. (1996)	4	20-27	EMG & smooth speech treatments similarly effective but better than controls	10, 50, 90
2. Finn (1996)	3	14	Past speech behaviours of recovered stutterers consistent with persistent stutterers but different from controls	8, 27, 61
3. Ingham et al. (1996) ^a	2	10, 19	Resting state brain blood flow similar for stutterers and controls	8, 25, 55
4. Paden & Yairi (1996)	3	12	Early phonological characteristics (EPC) of persistent stutterers differ from controls but EPC of recovered stutterers do not differ from controls	8, 21, 46
5. Bosshardt & Fransen (1996)	2	14	Stutterers retrieve semantic information more slowly than controls but there is no difference with respect to speed of word identification	8, 25, 54
6. Van Lieshout et al. (1996a)	2	12	Stutterers are not less able to assemble abstract motor plans for short verbal responses than controls	7, 22, 47
7. Yaruss & Conture (1996)	2	9	Stutterers who exhibit normal phonology and stutterers who exhibit disordered phonology are comparable in terms of basic speech disfluency	7, 16, 34
8. Prins et al. (1997)	2	12	During lexicalisation, slow processing could serve to disrupt fluency in some persons who stutter	7, 22, 47

9. Ingham & Cordes (1997)	2	10, 15	Comparisons of self-judgements and observer judgements of stuttering gives inconsistent results	7, 20, 46
10. Blood et al. (1997).	2	12	Variations in stuttering could be related to multiple, minor, daily stressors in some stutterers	7, 20, 46
11. Mackay et al. (1997)	3	10	Speech naturalness ratings distinguish among stutterers, controls and controls who spoke a different dialect	7, 20, 45
12. Van Lieshout et al. (1996b)	2	12	Stutterers may not have problems in creating abstract motor plans for speech	7, 22, 47
13. Stuart et al. (1996)	1	12	Slight alterations in the frequency of auditory feedback in stutterers may be fluency-enhancing	10, 40, 80
14. Ambrose et al. (1997)	3	22	Persistent & recovered stuttering possess a common genetic etiology but persistence is due, in part, to additional genetic factors	41, 95 ^β , 95
15. Finn et al. (1997)	2	10	Speech of children recovered from stuttering is perceptually indistinguishable from normal controls	7, 17, 38
16. Stager et al. (1997)	2	10, 12	Aerodynamic variables modified by stutterers & controls when speaking under conditions were also the variables related to changes in fluency for the stutterers	7, 20, 44
17. Finn (1997)	2	15	Speech of recovered stutterers was perceptually different from controls	8, 25, 56
18. Silverman & Berstein -Ratner (1997)	2	7	Normal disfluencies but not stuttering increased as syntactic complexity increased for stutterers & controls	6, 14, 28

19. Logan & Conture (1997)	2	14	Changes in the number of clausal constituents within an utterance may influence the number of speech disfluencies within that utterance	8, 24, 52
20. Robb & Blomgren (1997)	2	5	Stutterers have greater or quicker dimensional changes in vocal tract behaviour compared to controls	6, 9, 17
21. Max et al. (1997)	1	10	The adaptation effect in stuttering is a result of motor learning	9, 33, 69
22. McClean (1996)	1	14	Reduced attenuation of mechano-receptor input at the time of speech-movement initiation contributes to speech disfluency	8, 24, 52
23. Yaruss (1997)	1	45	It is important to evaluate more than one speaking situation when diagnosing a child who stutters	29, 92, 100
24. Vanryckeghem & Brutten (1996)	2	55	The difference in communication attitude between children who stutter & controls is due to more than just the level of speech disruption	11, 60, 98
25. Wenker et al. (1996)	4	40	More favourable personality traits were assigned to stutterers when presented live rather than on audiotape	15, 74, 99
26. Janssen et al. (1996)	2	>600	Relatives of female stutterers are not more likely to stutter than relatives of male stutterers	45, 99, 100

^a This was the only study that reported that a power analysis was performed.

^b Power was the same for a medium and large effect because two proportions were being compared and for a medium or greater effect, the “treatment” proportion was 1 (100%) and therefore could not be increased further.

It is apparent from Table 3.3 that consideration of power a priori is not common in the research sampled. Only one of the 26 studies selected for analysis (Ingham, Fox, Ingham, Zamarripa, Martin, Jerabek, et al, 1996) included a statement about study power however this appears to be a post-hoc power calculation. The majority of studies appeared to be underpowered for the detection of even a large effect. No study appeared to have sufficient power to detect a small effect.

For this, and any other body of research, it can be problematic if power is not considered a priori. Such a practice may introduce an inherent ambiguity of interpretation. Consider, for example, the Stuart, Kalinowski, Armson, Armson, Stenstrom, and Jones (1996) study in Table 3.3, where there is a power of 0.1 for detecting a small effect and a power of 0.8 for detecting a large effect. If such a study has not considered power a priori then there is no indication of what size of effect is clinically worthwhile. It may be that only a large effect would be clinically worthwhile, in which case the study is adequately powered. Alternatively, it may be that a small effect would be clinically worthwhile, in which case the study is underpowered. Consideration of power a priori makes the researcher consider the size of effect that is clinically or theoretically worthwhile detecting. The experiment is then designed to determine whether such an effect exists.

In determining whether a study has found a true null result, the reporting of confidence intervals is essential because it indicates the likely size of a true effect and the range of plausible values for the size of the effect (Lang & Secic, 1997). A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain

percentage (confidence level) of the intervals will include the unknown population parameter. Confidence intervals are usually calculated so that this percentage is 95%, but 90%, 99% confidence intervals can be produced for the unknown parameter. The size of the effect should be reported, along with the confidence interval, and the *p*-value.

For example consider a study comparing two treatments for a life threatening disease. If a survival difference of 5% is reported between the two treatments, with a 95% confidence interval of 1% to 9%, and a clinically worthwhile difference is regarded to be at least 10%, then the study could confidently conclude a null result. In other words, there is no clinically worthwhile difference between the two treatments with regard to survival. However, if a similar study reported a 5% difference in survival between two treatments but a 95% confidence interval of -5% to 15%, then the study is inconclusive because a 10% clinically worthwhile survival difference between the two treatments cannot be ruled out. In other words, it is quite plausible that a clinically worthwhile survival difference truly exists.

It would have been useful if confidence intervals had been included for the main result for all studies reported in Table 3.3. However, this was not possible because, for most studies, it was not clear what the main result was. For example, the Ingham et al. (1996) study reported 37 comparisons of mean cerebral blood flow but it was not stated which was the most important. Consequently, instead of including confidence intervals on all results appearing in Table 3.3, some examples are presented below of how confidence intervals can give additional, useful information about some of those results.

Consider the Finn (1996) report in Table 3.3. When comparing 14 persistent stuttering participants (PS) with 14 participants who had unassisted recovery from

stuttering (URS) using a speech behaviour checklist, it was found that the average total score was similar (21.82 for PS versus 22.75 for URS). The difference of 0.93 (22.75 – 21.82) was statistically non-significant, however if a 95% confidence interval is calculated it shows that the true difference could be as low as –0.81 or as high as 2.67. A difference of 2.67 could be clinically significant. If it is clinically significant, then the PS versus URS result is inconclusive because it has failed to rule out the possibility of a clinically significant difference.

The Craig, Hancock, Chang, McCready, Shepley, McCaul, et al (1996) study in Table 3.3 compares four groups of stuttering children, three treated differently and a control group. One of the groups received EMG feedback (n = 25) and at 3 months post-treatment, these children had, on average, 2.3% stuttered syllables (SS) during clinic conversation. Another group received intensive smooth speech (n = 27) and these children also had, on average, 2.3% SS during clinic conversation at 3 months post-treatment. Hence the difference between these two groups is 0% SS at 3 months post-treatment during clinic conversation. In other words, their rate of stuttering is the same. However if a 95% confidence interval is calculated it shows that the true difference could be as large as 1.35% SS in either direction. In other words, it is possible that children treated with one of these two treatments could achieve 1.35% fewer stuttered syllables than children treated with the other treatment at 3 months post-treatment. While this difference is small it could be considered clinically significant.

The way around such ambiguity in findings is to calculate, a priori, the power of the study either with reference to the effect size that is thought to be present, or with reference to the “least worthwhile difference.” The study then proceeds until the participant numbers required by the power analysis have been recruited. At the

conclusion of the study, the size of the effect is reported along with a confidence interval indicating the range of plausible sizes of the true effect, and, of course, the p -value.

An important point to note, however, is the possible inaccuracy of sample size calculations. Because these calculations are often based on limited amounts of data—sometimes no data—the estimated variability and size of effect may be quite different from reality. The actual size of the effect may be much lower or higher than expected. This can be true of the variability as well. The possibility of a real effect much larger than that expected and/or variability much less than that anticipated warrants the use of interim analyses. There are occasions when interim analyses of data are essential, the clearest example being a clinical trial. In the event that the effect is larger than anticipated and one treatment appears more efficacious than the other, it may be ethically necessary to stop the trial before the projected participant numbers have been recruited. In this case, evidence for such a course of action would be based on interim levels of significance set well below 0.05 to minimise the likelihood of a Type I error with lower than projected participant numbers.

The following two illustrations show the importance of considering power a priori and why recruitment of participants should continue unless there is a very strong reason not to continue. Figure 3.1 illustrates the importance of calculating power a priori with reference to a clinical trial comparing two chemotherapy regimes for the treatment of patients with advanced breast cancer (Bishop, GebSKI, Simes, Dewar, Snyder, Byrne, et al., 1999). The study incorporated the dependent variable “quality of life,” and Figure 3.1 shows p -values for a comparison of quality of life between the two treatments as participants were recruited to the study. The a priori calculation for the study determined that 390 participants were required to achieve

adequate power. Figure 3.1 shows that, if analysis had occurred with 50 participants, a p -value less than 0.05 would have been attained, and a conclusion could have been formed that one chemotherapy regime resulted in better quality of life for patients. However, such a conclusion would be quite likely to be wrong—a Type I error—because of the low chance of detecting a real effect with so few participants. Nonetheless, the finding would be ambiguous because it would also be possible that a large effect, detectable with lower participant numbers, was identified.

As Figure 3.1 illustrates, the only way to clarify this situation is for data to be analysed with the appropriate participant numbers for adequate power, as determined with an a priori power calculation. Had Bishop et al. conducted a preliminary analysis of their data they would have had little justification for concluding that an effect was present after they had recruited 50 participants.

Figure 3.2 illustrates, again with reference to the Bishop et al. data, another problem associated with underpowered research, which is the opposite of the problem discussed above. This is a straightforward Type II error, where a real effect exists but is not detected because of an underpowered analysis. Figure 3.2 represents changing p -values over the course of the study for the dependent variable “time to first treatment failure.” Time to first treatment failure is the time taken from randomisation till the patient's treatment stops due to either progression of their disease, excessive toxicity or death.

The p -value at the end of the study for 390 patients was 0.004, showing a high likelihood that a real effect existed. On the face of it, around 200 participants seems a respectable number for an analysis. However, Figure 3.2 shows that if an analysis had been done with that number of participants, the effect would not have

been detected, and time to first treatment failure would have been found not to be significantly different for the two treatment arms.

Figure 3.1: p -values for a measure of “quality of life” during the course of the Bishop et al. (1999) study at the times that various numbers of participants were recruited

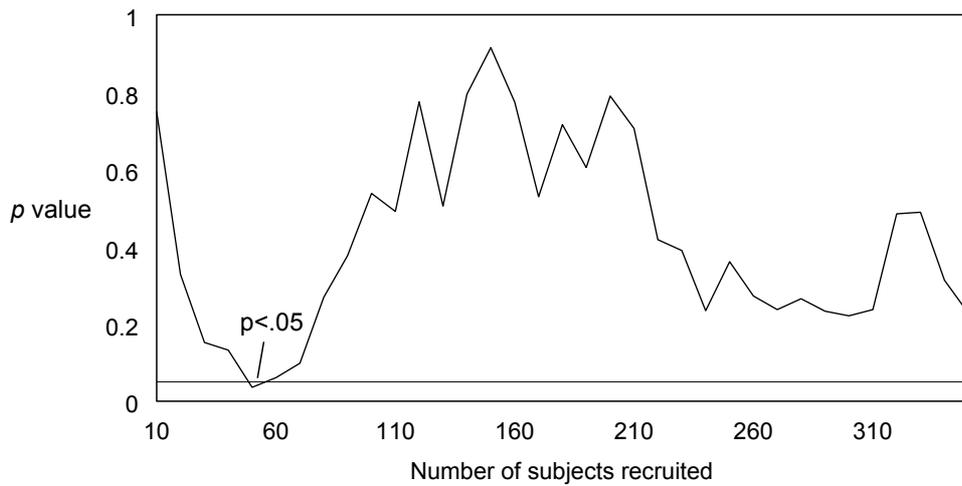
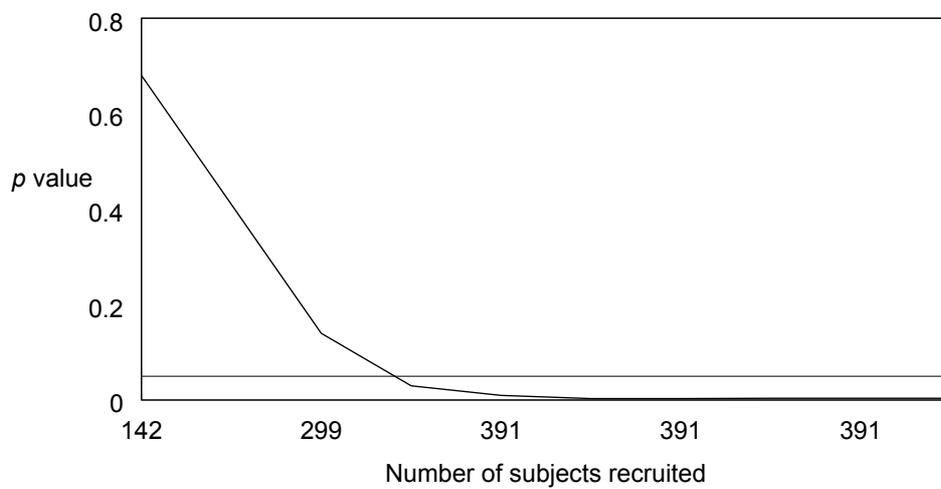


Figure 3.2: p -values over the course of the Bishop et al. (1999) study for the dependent variable “time to first treatment failure”



Example Sample Size Calculations

Three hypothetical sample size calculations are provided below to illustrate how sample size calculations are performed in practice. These examples were chosen for three reasons. First, they illustrate issues encountered in research situations. Second, the examples include the most common dependent variables in research (proportions and continuous outcomes), and finally, they reflect the type of studies that occur in speech pathology research.

Hypothetical Sample Size Calculation (1)

An example sample size calculation is given for a hypothetical trial of a speech pathology treatment for early stuttering, showing how parameters considered thus far are taken into account to determine the sample size. This calculation is based on uncorrected chi-square, although it is arguable that the more conservative Fisher's exact test should be used. Consider a randomised controlled trial to compare a new treatment for stuttering preschool children with no treatment, by means of a waiting list control. A definition of recovery could be less than 1% stuttered syllables (%SS). The first step for the calculation is to determine the primary hypothesis to be tested. Say that 50% of a clinical cohort of preschool-age children is likely to recover naturally from stuttering within 12 months of onset of the condition. Also say that pilot data collected on children treated with the new treatment show that 80% of cases are recovered within 12 months of beginning treatment. Based on these data, the primary hypothesis could be that the proportion of children recovered in the treatment group 12 months after commencing treatment will be 30% greater than the control group.

Now that the primary hypothesis has been determined, other assumptions for the sample size calculation need to be made. Other assumptions could be a 2-sided

test, 80% power, and a level of significance of 0.05. A 2-sided test is selected to allow for the possibility that the control children do better than the children who receive the new treatment. Using the formula given previously, a sample size of 39 in each group is required:

$$n > \{(Z_{\alpha/2}\sqrt{2pq} + Z_{1-\beta}\sqrt{p_1q_1 + p_2q_2})/(p_1 - p_2)\}^2$$

where: $p = (p_1 + p_2)/2$ and: $q = 1 - p$

$$n > \{(1.96\sqrt{2*0.65*0.35} + 0.842\sqrt{0.8*0.2 + 0.5*0.5})/(0.8 - 0.5)\}^2$$

$$n > 38.5$$

Now, refer to Table 3.2 that gives the proportional increase in sample size required to allow for various rates of drop-ins and drop-outs (see Compliance). If we assume a drop-in rate of 10% and a drop-out rate of 10%, then the requisite sample size becomes 62 in each group.

Hypothetical Sample Size Calculation (2)

Consider an investigation of social anxiety in people who stutter. The primary hypothesis could be that people who stutter are, as a group, more anxious in social situations than people who do not stutter. Specifically, consider a primary hypothesis that mean Subjective Units of Distress (SUD) scores in a social situation are 10 points higher for stuttering participants than for non-stuttering participants. Assume that the SUD scale ranges from 0-100, with normative data indicating with confidence that the population mean is 50 and the distribution is approximately normal. Again, pilot data have been gathered in the development of the study, and those data suggest that the stuttering participants have an average SUD of 60 with a

standard deviation of 15. If 80% power is assumed, with a 2-tailed significance level of 0.05, then the sample size required is 36 in each group:

$$n > 2\{(Z_{\alpha/2} + Z_{1-\beta})/(\delta/\sigma)\}^2$$

$$n > 2\{(1.96 + 0.842)/(10/15)\}^2$$

$$n > 35.3$$

However, since there were few participants in the pilot study, the investigators do not consider that the estimate of mean and SD for the stuttering participants is particularly robust. Because of these limits on confidence in the pilot data, the investigators might be interested to know how the sample size changes with slightly different assumptions with respect to the size of effect anticipated. This is called a sensitivity analysis, and the results are given in Table 3.4. This table shows the sample sizes required for different population standard deviations and population mean SUD scores.

Table 3.4 shows an extraordinary range in sample sizes with variation of the population estimates of mean and SD for stutterers SUD scores. This illustrates the benefits of reliable pilot data in order to make the sample size calculation as accurate as possible. Such accuracy avoids the problems associated with having to increase sample size mid way through a study.

In the present example, the value of a sensitivity analysis is demonstrated because the sample size is sensitive to small changes in the assumed size of effect. Because of this result, it would be advisable in this hypothetical study for the investigators to monitor the sample standard deviation throughout the study with interim analyses to ensure that adequate power is achieved.

Table 3.4: The sample sizes required for different population standard deviations and population mean subjective units of distress SUD anxiety scores

Standard Deviation (SD)	Mean SUD = 55	Mean SUD = 60	Mean SUD = 65
SD = 15	142	36	17
SD = 20	252	64	29
SD = 25	394	100	45

Randomisation is impossible in this study, because stuttering cannot be randomly assigned. However, a variation in the design above could control bias. The investigators could match participants on variables such as age and gender. Matching tends to reduce bias so that any difference observed between the two groups is likely to be due to stuttering. If matching is used then it is appropriate to use a paired analysis and this can be taken account of in the sample size calculation. Paired analysis will result in a reduced sample size because it is assumed that pairs of participants are correlated; that is, more similar to each other than to the other participants. Hence, variation is reduced resulting in fewer participants required to maintain equivalent power. If it is thought that the correlation between the matched pairs is 0.5 then the sample size required for the original scenario given above is 18 in each group.

Hypothetical Sample Size Calculation (3)

Consider a study designed to explore language skills in children who stutter. Assume that 10% of the general population has some kind of language impairment ($p_1 = 0.1$) and the hypothesis for the study is that the stuttering children are more

likely to have this impairment. A minimal worthwhile difference could be 10%, thus the study should be powered to find a statistically significant difference if 20% of the stuttering children have language impairment ($p_2 = 0.2$). A possible design for this study would be a cross-sectional survey. If it is again assumed that the investigators require power of 0.8 and a 2-tailed significance level of 0.05, then the sample size required will be 200 stuttering children and 200 non-stuttering children.

One reason why the sample size is large in this study is that only 10% of the non-stuttering children are expected to have the outcome of interest—language impairment. Hence, a large number of participants are required in order to be confident that the assumed proportion is indeed accurate. There is a consequence to this large sample size that is pertinent in the planning of the study. Because this study requires a very large sample size, it may not be feasible to expect a replication of the result in the future. Therefore, in order to make the study definitive, a power greater than 80% may be warranted. If 90% power is required for this study then the sample size becomes 266 in each of the two groups (532 participants in total). Considering the resource commitment for a study of such magnitude, it would be necessary for the investigators to be confident from pilot data that the stuttering children are indeed more likely (by 10%) to have language impairment.

The investigators for this hypothetical study may well believe that the information to be gained from it is worth having. However, it is likely that many investigators would not be able to resource the study or recruit sufficient participants. In this case, it is an option to estimate the effects on the study of reducing the total sample size to 300. If power is to be maintained at 80% then the difference that could be detected with 300 participants is 12%. Alternatively, if the difference is to remain at 10%, then with 300 participants, the study will have a power of 68%, which means

that, if there is a difference in language impairment rates between stuttering and non-stuttering children, the study has a 0.68 chance of detecting it.

Discussion

It is fundamental to determine if effects are not present in a population, because such a null result can rule out any causal link between variables. The establishment of null results requires statistical inference, and a priori calculation of power is essential to that process. Without a priori power consideration in the formulation of a study, an erroneous null conclusion is a strong possibility. In a program of systematic research, the long-term effect of underpowered studies is the possible compounding of misinformation through replicated errors of Type II. A program of underpowered research is quite likely to replicate erroneous null findings. Underpowered research is a wasteful use of research resources because there is little chance of detecting an effect. Further—in the social sciences at least—such research is unethical because it exposes human participants to inconvenience, discomfort, and/or risk without justifiable benefits.

That being said, and this review having identified an underpowered body of research, it needs to be noted that it simply is difficult, if not impossible in many cases, to obtain high participant numbers for disorders of low prevalence. Further, much current research into the nature and treatment of stuttering is resource-intensive, making the acquisition of statistically powerful participant numbers impractical because of time and cost issues. Examples of such research include neural imaging, acoustic, and kinematic studies, and clinical trials. Hence, three ways are suggested to redress the research problems that we have identified that are associated with statistical power:

The first is that prior to embarking on studies involving statistical inference in new domains of inquiry, it may on some occasions be appropriate to undertake pilot studies that do not involve statistical inference in order to collect information necessary to determine whether a larger study is feasible. In some situations, this may be a desirable beginning to a potential program of research rather than presenting an initial, underpowered statistical study with 5-10 participants. Researchers should be encouraged to publish the estimates of standard deviations from these pilot studies to assist other researchers in planning future studies. Alternatively, within-subject experimentation provides an ideal method to generate evidence for the possibility of population effects with certain variables. The purpose of such preliminary studies would be to determine the feasibility of recruiting larger numbers of participants for group design investigation of the area of interest. In the case of non-clinical research where a “least worthwhile effect” analysis is not viable, such studies would provide some information about the possible effect size so that a subsequent study could be designed with an accurate power calculation.

The second approach to redressing power problems in stuttering research is for authors to publish and interpret power analyses and confidence intervals in all research involving statistical inference. This seems a reasonable and viable suggestion, considering that software for calculating power is now readily available, much of it available as shareware on the internet. One benefit of such routine publication and interpretation of power calculations and confidence intervals would be to assist readers in establishing the appropriate caveats in interpreting results. This would be particularly beneficial when authors present statistical analysis of data with limited participant numbers. This would be particularly important in interpreting

reports where both significant and null results occur in the same, low powered analysis or analyses.

Finally, although it imposes additional practical difficulties in conducting research, the problem of underpowered studies could be overcome by collaboration across research centres. For if a domain of inquiry is of sufficient interest to the field it should be possible to recruit various research centres in a single study. Then the burden of recruiting large numbers of participants can be spread amongst several centres. This occurred, for example, in the Craig et al. (1996) study. A bonus from this model would be to establish, in one report, the generalisability of a research result beyond one research centre. With studies comparing stuttering participants with controls in which matching was required, such a model would not present problems because each participating centre could enrol matched pairs of participants into a study. The randomised, controlled trial reported in Chapter 7 is an example of recruiting subjects from more than one location.

A variant of the model above is the application of meta-analysis to combine the data from several studies to make a more powerful statement than is possible from any of the individual studies themselves (Petitti, 1994). Such a procedure has been attempted for research on stuttering treatment (Andrews, Guitar, & Howie, 1980) but attracted justifiable criticism on the grounds that the studies concerned were so methodologically different that the exercise had questionable value (Ingham, 1984). However, such a problem could be offset if various research centres published reports independently, but agreed a priori on some basic commonalities in their research design so that a prospective meta-analysis could be performed. In the interest of a unified, powerful program of research, such collaborative efforts by research groups in similar areas would appear to be worthwhile and achievable.

**PART TWO: PREDICTORS OF TREATMENT TIME WITH THE
LIDCOMBE PROGRAM**

Chapter 4: Predictor Study One

Introduction

This chapter presents research conducted in the late 1990s in the form of a retrospective file audit of children treated with the Lidcombe Program over an 8-year period. The main purpose was to provide much needed clinical data on the Lidcombe Program, such as information pertinent to the widely debated topic of when to begin treatment for early stuttering. The information generated from this retrospective study was used in the design of the randomised controlled trial of the Lidcombe Program presented in Chapter 7.

During the past two and a half decades the merits of treating stuttering soon after onset have become obvious (see Chapter 1). Predictably, then, there has been much debate about exactly when treatment should begin during the first years of stuttering, with recommendations ranging from immediate intervention for all children to delays of 6 months, 1 year, and even several years (Andrews, 1984; Curlee, 1993; Ingham, 1994; Packman & Lincoln, 1996; Riley & Riley, 1983; Starkweather, 1990). Ultimately, this issue is probably best resolved by the judgment of the consulting speech pathologist for each case individually, rather than with set guidelines for use with all stuttering children (Onslow & Packman, 1999b; Packman & Onslow, 1998). However it has been suggested that clinicians take into account factors such as severity of stuttering, whether the stuttering is improving, family history, age and gender of the child, time since onset, the reaction of parents and child to the stuttering, and whether parents wish treatment to proceed (Bernstein Ratner, 1997; Curlee & Yairi, 1997; Packman & Lincoln, 1996; Packman & Onslow, 1998; Zebrowski, 1997). A critical consideration in this issue, however, is whether delaying intervention for some period in the hope that natural recovery will occur will render a child less responsive to treatment (see Ingham & Cordes, 1998). In

other words, is treatment recovery put at risk in those children who do not recover naturally during such a delay?

However, at the time this study was conducted (it has since been published) almost nothing was known about responsiveness to treatment and factors that might predict it. In other words, little was known about whether some children respond more quickly to treatment than others, and why that might occur. It is tempting to assume that factors that influence natural recovery also influence tractability. Much pertinent information about the natural course of stuttering has been generated from the Illinois Early Childhood Stuttering Project. It has been reported that having a family history of recovery and being a girl predict natural recovery (Ambrose, Cox, & Yairi, 1997). Also, it is self evident that the shorter the period since onset, the more likely a child is to recover without intervention. On the other hand, in terms of stuttering severity, children who recover naturally appear not to differ from those whose stuttering persists (Ambrose et al., 1997). These findings suggest, then, that gender and period since stuttering onset may be relevant to clinical tractability, yet this has never been studied empirically.

There are preliminary data to suggest that stuttering severity and time since onset may influence treatment outcome. Starkweather and Gottwald (1993) reported that children with more severe stuttering required more treatment time, and they also found a significant correlation between time elapsed from the onset of parental concern to the start of treatment, and the duration of treatment. The idea that the longer the time since onset the more intractable the condition is also supported by the fact that, relative to adults, preschool children show rapid establishment and generalisation of treatment effects (e.g., Adams, 1984; Bloodstein, 1995; Costello, 1983; Curlee, 1984; Ingham, 1984; Prins, 1983). Existing data for the Lidcombe

Program suggest that while a median of 10.5 clinic visits is sufficient for medium and long term control of stuttering in preschool-age children, slightly more visits are needed to achieve the same results with school-age children (Lincoln, Onslow, Wilson, & Lewis, 1996; Onslow, Andrews, & Lincoln, 1994; Onslow, Costa, & Rue, 1990).

Of the variables cited above that might influence treatment time, age and time since onset are potentially the most clinically useful as it is possible to influence the point in the natural course of the disorder at which treatment begins. If older children who have been stuttering longer are less tractable then clinicians could improve treatment effectiveness and efficiency and reduce service delivery costs by introducing treatment soon after onset.

In short, the timing of early stuttering intervention is a pressing issue in speech pathology. While it has been suggested that the decision to intervene may be delayed for some period after onset to determine whether natural recovery occurs, at the time this study was conducted it was not known if such a delay makes stuttering more difficult to treat in those children who do not recover.

Medium and long term outcome data and social validity data have been reported for the Lidcombe Program in a series of publications (Lincoln & Onslow, 1997; Lincoln et al., 1997; Lincoln et al, 1996; Onslow et al., 1994; Onslow, Costa, & Rue, 1990). These prospective studies deal with a total of 61 children. However, that small number of participants does not provide a statistically powerful means to address any of the foregoing issues about the timing of early stuttering intervention. During the period in which these formal treatment studies were conducted, a much larger number of children were treated with the Lidcombe Program in public clinics. The present report deals with case data for 250 children for whom this treatment was

judged by their clinicians to be successful. The present report is not an outcome study but is an investigation of whether age, time since onset, gender, and stuttering severity relate systematically to the time required for treatment with the Lidcombe Program.

Method

Participants

Participants were 261 children who were diagnosed as stuttering by speech pathologists and who began treatment before 6 years of age. With permission from the South Western Sydney Area Health Service and the University of Sydney, the clinic files for the children were accessed and systematically reviewed. These children began their treatment between September 1989 and September 1997. Of these children, 11 (4.2%) did not complete the treatment. For these children, Table 4.1 presents their age, gender, onset-to-treatment interval, percent syllables stuttered (%SS) at first treatment session, the number of treatment sessions completed, the reasons why the treatment was not completed, and their within-clinic %SS scores at the time of their last clinic session. Table 4.1 shows the reasons for failing to complete treatment are unremarkable for a speech clinic, and, with the exception of Participant 6, that each child had shown clinical progress at the time of their last clinic session. However, Participant 6 completed only three sessions.

A total of 250 children were deemed to have completed Stage 1 of the program after achieving zero or near-zero stuttering. Mean age at the first treatment session was 46 months (SD = 9.4 months, median 46 months, min. 27 months, max. 71 months). There were 192 boys and 58 girls.

Table 4.1: Details of the 11 children who failed to complete Stage 1

Gender	Age in months at initial treatment	Months from onset to initial treatment	%SS at initial treatment session	Number of clinic sessions completed	Reason why LP not completed	%SS at last treatment session
M	59	29	11.4	15	Language delay, transferred to local clinic	3.1
M	44	12	18.4	8	Family relocated interstate	8.8
F	48	6	3.0	4	Family circumstances changed, referred to local clinic	0.6
M	62	14	7.2	14	Referred to local clinic due to behavioural problems	2.1
M	35	15	5.0	10	Complex family problems, withdrew from treatment	3.0
F	43	25	10.9	3	Problems with transport to clinic, referred to local clinic	12.0
M	44	9	16.6	24	Developed language problems, transferred to generalist clinic	4.1
M	59	29	11.4	15	Language delay, transferred to local clinic	3.1
M	49	6	4.5	5	Family relocated overseas	1.4
M	66	42	5.1	4	Unable to attend clinic due to changed family circumstances	0.6
F	29	5	6.4	18	Offered treatment at local clinic	0.4

Diagnoses of stuttering were made using the positive identification component of a consensus diagnostic procedure described by Onslow (1992). First, for a diagnosis to be made, it was necessary for one or both parents to believe the child to be stuttering. This is a commonly used diagnostic sign (e.g., Conture & Caruso, 1987; Costello, 1983; Curlee, 1993; Starkweather, Gottwald, & Halfond, 1990; Riley & Riley, 1983; Yairi & Ambrose, 1992a). Second, two clinicians with extensive experience in stuttering management agreed that the child was stuttering. Subsequent to diagnosis, each child was placed on the treatment waiting list and treated according to the order of that list. The mean waiting list period from assessment to the first treatment session for the participants was 3.9 months.

The Treatment

Participants were diagnosed and treated at either, the Stuttering Unit, Bankstown Health Service, Sydney, or at a specialist stuttering clinic at the University of Sydney. These clinics are devoted exclusively to the management of stuttering. They provide a range of treatment services for adults and children. The children were treated by seven clinicians with extensive experience in stuttering management.

All participants were treated using the Lidcombe Program. The essential features of the program did not change during the period of study. During the first stage of the program, the parent(s) and child attend the speech clinic once a week and during these visits the parents are trained how to administer verbal response contingent stimulation in the child's everyday environment. The second, or maintenance, stage of the program commences when the child achieves the following criteria: less than one percent syllables stuttered (1%SS) as measured by the speech pathologist during within-clinic conversations and a weekly mean severity score of

2.0 or less on the 10-point rating scale (see Chapter 1). The conduct of the maintenance stage is based on a schedule of within-clinic and beyond-clinic assessments developed by Ingham (1980; 1981). Scheduled assessments decrease in frequency contingent on the child meeting the criterion speech performance described above over a substantial period.

Dependent Variable

During Stage 1 of the Lidcombe Program, the child and parents attend the clinic each week for 45-60 minutes. For the purposes of this study, the number of clinic visits to complete Stage 1 was recorded, and is referred to as number of clinic visits. This was used as a measure of treatment time. Based on the median number of treatment hours in a previous report (Onslow et al., 1994), this variable was dichotomised to fewer than 10 clinic visits (short treatment time) and 10 or more clinic visits (long treatment time) so that logistic regression could be used to predict which children would be treated in a shorter time and which children required a longer treatment time. This analysis was chosen because treatment time as a continuous variable did not meet the requisite assumptions for least squares regression. This categorisation resulted in 110 children with a shorter treatment time and 137 children with a longer treatment time.

Predictor Variables

The following predictor variables were ascertained for all children from their clinical records: gender, age at first treatment session, onset-to-treatment interval (time from reported onset of stuttering to the first treatment session), and stuttering frequency (percentage of syllables stuttered [%SS]) at first treatment session. In the Lidcombe Program, %SS scores are obtained routinely for children at the start of

each clinic session. The clinician converses with the child and records non-stuttered and stuttered syllables on-line using a button-press electronic counting device. In most cases, 300 syllables of the child's speech are sufficient for a representative speech sample, but a longer speech sample is needed in cases of mild stuttering. Details of the procedures for collecting the within-clinic %SS measure can be found in Onslow, Packman, and Harrison (2003).

In contrast to the measure of Stuttering-Like Disfluencies (SLD) used by Yairi and colleagues (e.g., Yairi & Ambrose, 1992a; 1992b) the %SS measure is based on the number of speech events thought by an experienced clinician to be unambiguous stuttering. Thus, while Yairi and colleagues require at least 3 SLD per 100 syllables for a child to be considered to be stuttering, a rate of 1.0 %SS or more is considered clinically significant. For example, if a child stutters at an average of 2.0% SS at a speech rate of 100 syllables per minute then two observable stutters per minute will occur on average when the child is talking.

Clinical measures of %SS were considered to be valid and reliable for the purposes of the present study, for several reasons. First, in numerous clinical trials with preschool children, this measure has been shown to document changes in stuttering from clinically significant levels to zero or near-zero levels, and those changes have been substantiated by social validity data (see Chapter 1). Second, %SS scores have been shown to be reliable in those clinical studies, and the clinicians in the present report have all participated in generating those data. Further, the clinicians concerned have used the %SS measure routinely during specialised clinical practice over long periods.

Family history of stuttering was not used in the regression analyses, because this is a questionable measure even in prospective studies that incorporate systematic

interviews (Felsenfeld, 1997; Hedges et al., 1995; Yairi, Ambrose, & Cox, 1996) and therefore unsatisfactory in a retrospective study such as this.

The predictor variables were categorised before the regression analysis was conducted, in order to facilitate the identification of any relationships with the dependent variable and to allow further data interpretation using odds ratios (see results). Additionally, categorised rather than continuous data avoided the undesirable assumption that any relation to treatment time would be linear and, in the case of onset-to-treatment interval, categorisation improved the reliability of this variable by reducing its dependence on parent recall. In the case of stuttering severity, categorisation improved the reliability of this variable by reducing its dependence on one retrospective measurement in the clinic. This a priori categorisation of predictor variables was guided empirically, where possible, or if not possible, by what seemed to be clinically appropriate. Thus, onset-to-treatment interval was categorised as shorter than 12 months (130 children) and greater than or equal to 12 months (113 children), because it seems that around 50% of natural recovery takes place within one year of onset; age at first treatment session was categorised into 2-3 years (143 children) and 4-5 years (107 children), to represent early preschool-age and late preschool-age; and stuttering severity was categorised intuitively into less severe or below 5.0 %SS (144 children) and more severe or 5.0 %SS or greater (71 children).

Results

Descriptive Statistics

Means, standard deviations, medians, and ranges for each predictor variable are presented in Table 4.2. There are some missing measures of %SS at first

treatment session because, on some occasions, the clinician declined to record this speech measure for reasons of poor validity. For example, some children will not converse freely with either a clinician or parent at the first treatment session, resulting in truncated utterances that have little validity as speech samples. The following are specific quotes from the clinic files which illustrate the various reasons why the %SS measure was not collected at the first treatment session:

“Very shy, said fewer than 20 spontaneous syllables.”

“(within clinic) sample not valid. Not well/lethargic for past 2 days.”

“Mother reports (child) is very shy with strangers, so thinks she is not likely to talk here until more at ease.”

“No (within clinic) speech measures taken - he didn't talk spontaneously at all.”

“Fell asleep in the waiting room and hardly awake during session, so no speech measures.”

Table 4.2: Descriptive statistics for the 250 children who completed Stage 1

Variable	N	Mean	SD	Median	Min	Max
Baseline %SS	215	4.2	4.1	3.5	0	28
Age in months	250	46	9.4	46	27	71
Months from onset to treatment	243	12.4	7.9	11	0	36
Number of clinic sessions	247	12.5	9.1	11	1	85

Absence of a within-clinic %SS measure at the first visit was deemed missing data. The investigator did not collect a within-clinic %SS score from a subsequent clinic session because such measures would be potentially confounded by treatment effects. Some measures of number of clinic visits and onset-to-treatment interval are missing because of ambiguous clinical records. In the case of onset-to-treatment-interval, one common cause of such ambiguity was when parental reports of the time of onset were not sufficiently specific for the present analyses. Another common cause was when the mother and father, or some other person present at the interview, presented differing reports of the time of onset. Other instances included reports that the parent could not recall or the absence of mention in the file about the time of onset. The three missing records or number of clinic visits occurred because of incomplete records kept in files.

At the first treatment session, the median %SS score was 3.5. In other words, at the first treatment session half the children were below 3.5 %SS and half were above 3.5 %SS. The median of 11 clinic visits for these 250 children to complete Stage 1 is consistent with data reported in clinical trials with smaller numbers of participants (Onslow et al., 1994; Onslow, Costa, & Rue, 1990).

Thirty nine children had %SS scores of less than 1.0 at the first treatment session. However, while they did not appear to stutter significantly during the speech sample collected at the first clinic visit, these children all met the criteria for stuttering described above. Consequently, as this is a retrospective study, they were not removed. To investigate whether these low-scoring cases truly required treatment for their stutter the number of clinic visits for the 39 children was analysed. They took a median of 6 clinic visits to complete treatment, with a range of 2-28 visits. Thus, these children were clearly stuttering and needed treatment.

Logistic Regression

Logistic regression was used to determine if there was any relationship between outcome (shorter or longer treatment time) and the predictor variables; that is, between number of clinic visits and gender, age at first treatment session, onset-treatment interval and %SS at first treatment session. In this analysis, one group is used as a reference and given a value of 1.0. If the odds ratio for the other group is also 1.0, then there is no difference between the groups.

The results of univariate logistic regression are presented in Table 4.3. The predictor %SS at first treatment session was highly statistically significant ($p < 0.001$). The odds ratio of 4.1 indicates that the odds of the more severe children needing 10 or more clinic visits to attain maintenance were 4.1 times greater than the odds of the less severe children needing 10 or more clinic visits to attain maintenance. Using intervals shorter than 12 months as the reference category for onset-to-treatment interval, the children with intervals of 12 months or longer required fewer clinic visits but this was not statistically significant at the conventional level (odds ratio 0.62, $p = 0.064$). The other two predictor variables, gender and age, did not have a significant relationship to treatment time.

A multivariate logistical regression was also performed and the results are given in Table 4.4. The p -values and odds ratio's for both %SS at first treatment session and onset-to-treatment interval remain similar to those of the univariate logistic regression however the p -value for onset-to-treatment interval was statistically significant at the conventional level in this multivariate analysis. The odds ratio of 0.56 indicates that the odds of the children with onset-to-treatment intervals of 12 months or longer needing 10 or more clinic visits to attain maintenance were approximately half that of the odds of the children with onset-to-

treatment intervals shorter than 12 months needing 10 or more clinic visits to attain maintenance. Hence it can be concluded that %SS at first treatment session and onset-to-treatment interval are independently associated with treatment time.

Table 4.3: Results of univariate logistic regression of treatment time

Variable	Odds ratio (OR)	95% confidence interval	<i>p</i>-value
Gender: Boys	1.0*		
Girls	0.69	0.38 – 1.24	0.21
Baseline %SS: < 5%SS	1.0*		
5%SS +	4.1	2.1 – 7.8	<0.001
Time since onset: < 12 months	1.0*		
12 months +	0.62	0.37 – 1.03	0.064
Age at baseline: 2 or 3 years	1.0*		
4 or 5 years	0.81	0.49 – 1.34	0.40

* reference category

Table 4.4: Results of multivariate logistic regression of treatment time

Variable	Odds ratio (OR)	95% confidence interval	<i>p</i>-value
Time since onset: < 12 months	1.0*		
12 months +	0.56	0.31 – 1.0	0.05
Baseline %SS: < 5%SS	1.0*		
5%SS +	3.7	1.9 – 7.2	<0.001

* reference category

Kaplan-Meier Plots and Log-rank Tests

Kaplan-Meier plots (Kaplan & Meier, 1958) were used to show patterns of recovery within groups. Figure 4.1 is a Kaplan-Meier plot, for all cases, of the cumulative proportion of participants who attained maintenance against number of clinic visits. Plots for the boys against the girls are presented in Figure 4.2. The median number of clinic visits is 11 for boys and 9 for girls. The log-rank statistic (Harris & Albert, 1991) tests whether there is any evidence of a difference between two or more Kaplan-Meier curves and the p -value of .99 indicates that there is no significant difference in recovery for boys and girls.

Figure 4.3 presents plots for the more severe children ($\%SS \geq 5.0$) against the less severe children ($\%SS < 5.0$). The less severe children required a median of 9 clinic visits, and the more severe children required a median of 12 visits. The highly significant log-rank test ($p < 0.001$) and the distinctly different plots for the two groups confirm the results of the regression analysis that more severe children require a longer treatment period for control of their stuttering.

Figure 4.4 presents Kaplan-Meier plots for the two age groups: 2-3 years and 4-5 years. The median number of clinic visits is 11 for the younger children and 10 for the older children. The two age groups are not significantly different according to the log-rank test ($p = 0.77$) and this is supported by the crossing-over of the plots. Figure 4.5 presents a Kaplan-Meier plot for shorter (< 12 months) and longer (≥ 12 months) onset-to-treatment interval. Medians for the two groups are 11 for the children with intervals less than 12 months and 9 for children with intervals 12 months or greater. The two groups are similar according to the log-rank test ($p = 0.094$) although the p -value approaches statistical significance.

Figure 4.1: Number of clinic sessions required to complete Stage 1

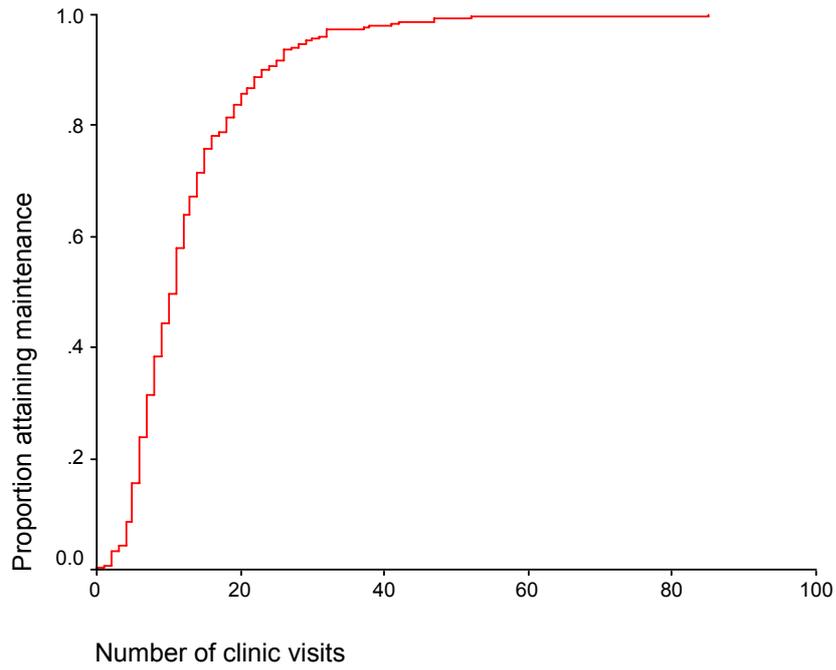


Figure 4.2: Number of clinic sessions required to complete Stage 1 by gender

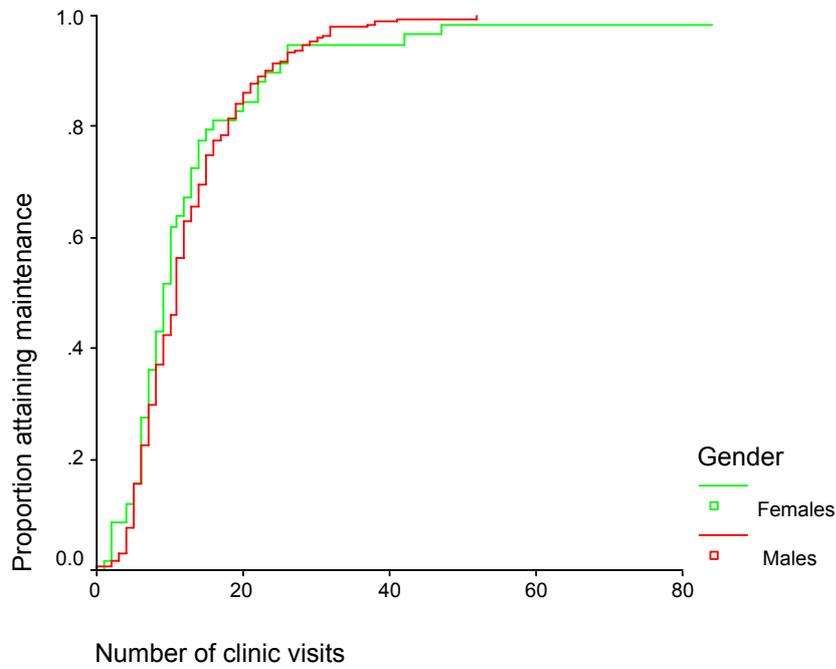


Figure 4.3: Number of clinic sessions required to complete Stage 1 by baseline severity of stuttering

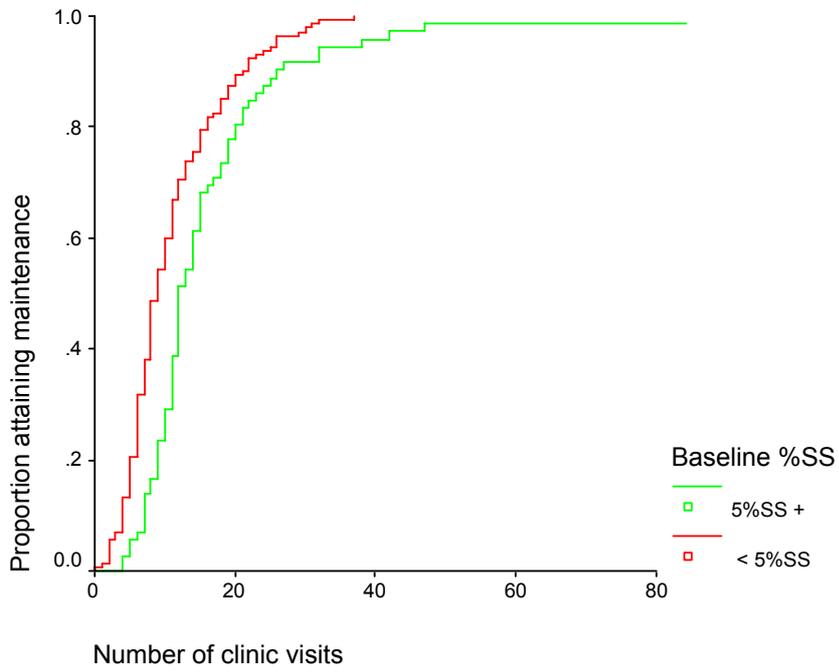


Figure 4.4: Number of clinic sessions required to complete Stage 1 by age

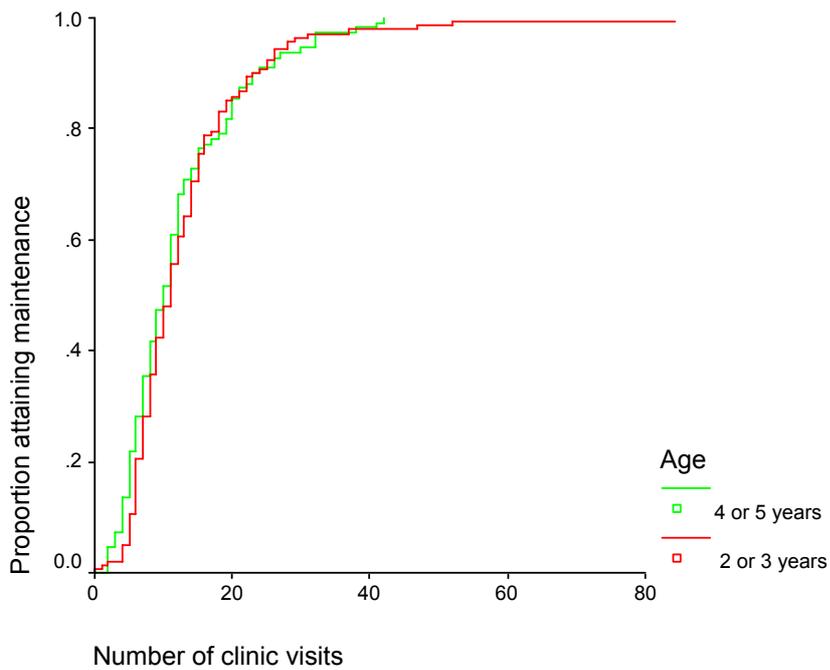
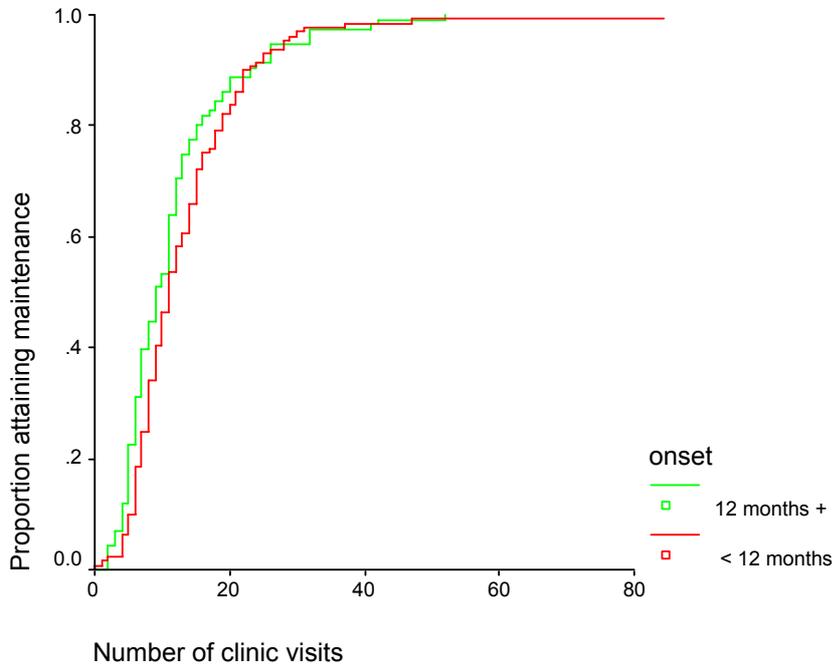


Figure 4.5: Number of clinic sessions required to complete Stage 1 by time since onset of stuttering



In summary, the Kaplan-Meier plots and log-rank analyses support the regression analyses in confirming a highly significant difference in treatment time between more severe and less severe children, and that the difference in treatment time for children with shorter and longer onset-to-treatment interval approaches significance.

Discussion

This was a study of children who began treatment for stuttering prior to their sixth birthday. In a clinical study of this age group, it is possible that some of the children may have recovered without the formal treatment, and it is also possible that

for some of the children a process of natural recovery may have occurred concurrently with the treatment.

Nonetheless, it is of interest that the present data suggest that a longer period since onset is associated with quicker treatment. In other words, for a child treated with the Lidcombe Program, the chance of recovery increases with the time since onset. This is diametrically opposed to what occurs with natural recovery, where the chance of recovery decreases as the time since onset increases (Yairi, Ambrose, Paden, & Throneburg, 1996). These contrasting effects suggest the presence of a fundamental difference underpinning the recoveries in the present clinical study and those in the studies of Yairi and colleagues in the natural recovery literature.

Because natural recovery may occur for any child who begins to stutter, the issue of exactly how soon after onset intervention should occur is a critical one for clinicians. The benefits of immediate intervention need to be weighed up against the benefits of delaying treatment to increase the chance of natural recovery (Onslow & Packman, 1999b). Using a non-clinical population, the Illinois Early Childhood Stuttering Project is generating information that clinicians can incorporate into that decision-making process. At present, for example, it is known that gender and family history may influence the chance of natural recovery but that severity does not (for example, Ambrose et al., 1997). The present study identified variables that may influence the time required for treatment recovery. In that regard, the most decisive finding of the present study is that stuttering severity at the first treatment session is a predictor of the time required for treatment; medians of 9 and 12 clinic sessions were required for less and more severe stuttering respectively. This finding is in agreement with Starkweather and Gottwald's (1993) finding that children with more severe stuttering take longer to treat. However, the present findings are not in agreement

with Starkweather and Gottwald's finding that older children require more treatment time.

The present data provide evidence that delaying treatment during the preschool years will not worsen the tractability of the condition, at least in terms of the treatment time required. However, there are three important caveats to these data. First, these findings relate to the Stage 1 of the Lidcombe Program only, and say nothing about treatment outcome in the long term. Second, the findings pertain only to children in the first few years after onset of stuttering and cannot be generalised to later childhood or early adolescence.

Third, the present findings pertain to one treatment method only. The Lidcombe Program requires active participation from the child and this may explain why children who had been stuttering longer tended to take fewer clinic visits than children closer to onset. When the author discussed this finding with the Lidcombe clinicians, they found it to be consistent with their clinical experiences. They commented that when children are very young, shortly after stuttering onset, they often do not manage the demands of the treatment as well as children who are older. This may not be an important factor when treatment is indirect, such as with the treatment investigated by Starkweather and Gottwald. Such treatments, which are based primarily on changes to children's everyday environment and to parental communication styles, require less participation from the child and do not focus as much on stuttering per se (for overviews see Guitar, 1998; Onslow & Packman, 1999a). Consequently, there is a need for a replication of the present study with different treatment methods.

One limitation of the present study is its retrospective nature. On the other hand, the data can be regarded as more generalisable than data from outcome studies

as they were generated during the everyday administration of a health service in speech clinics. Although the clinicians were experienced in treating stuttering, both clinics are located in a culturally and ethnically diverse middle to low socioeconomic area of Sydney, and there were no selection criteria. The fact that the median treatment time in this study is identical to that of a previous outcome study (Onslow, Andrews, & Lincoln, 1994) suggests that the present findings are valid: a clinical caseload of 250 children required a median of 11 clinic visits to meet stringent criteria for controlling stuttering.

The findings of this study have implications for the randomised clinical trial of the Lidcombe Program presented in Chapter 7. Providing that family circumstances do not change throughout the trial making attendance at the treatment clinic difficult, and that the child does not develop additional problems, then the vast majority of children should complete the first stage of Lidcombe Program treatment and hence attain fluency within 1 year of beginning treatment. In fact, based on the present data it would be expected that around 90% of children who remain in therapy would complete Stage 1 within 6 months. Therefore the duration of follow-up required to ensure most children in the Lidcombe Program treatment group are fluent or close to fluent at the final follow-up point would be between 6 months and 1 year. For a no-treatment control group, delaying treatment for at least a year after onset of stuttering should not jeopardise their responsiveness to treatment, if that is required at the end of follow-up. Because more severe children tend to require more clinic visits to complete Stage 1, %SS should be included as a stratification variable in the randomisation procedure of the trial.

Chapter 5: Predictor Study Two

Introduction

This chapter presents a second retrospective study of the Lidcombe Program. Considering the clinical importance of the findings of the first file audit study (Study 1) presented in the previous chapter, the aim of the present study was to replicate them on a group of children who had received the Lidcombe Program in Britain. Of enormous interest was determining whether the results of a second study would be consistent with the original or whether differences exist and, if so, for what reasons. The current study also used a meta-analysis of the new British data and the extant Australian data to explore, in particular, the non-significant result presented in the previous chapter indicating that children who have been stuttering longer may take a shorter time to complete Stage 1 of the program.

Method

Participants and Treatment

Participants were 78 children who had attended a specialist stuttering National Health Service clinic in Norwich, Norfolk, between 1997 and 2001. The children were identified as stuttering by speech pathologists with a method similar to that described in Study 1, using a consensus procedure described by Onslow (1992). For a child to receive treatment, it was necessary for one or both parents to believe the child to be stuttering and for two speech pathologists with extensive experience in stuttering management to agree that the child was stuttering.

All children began treatment before their sixth birthday. Of these children, 12 (15.3%) did not complete Stage 1. Table 5.1 presents their age, gender, onset-to-treatment interval, and percent syllables stuttered (%SS) at first clinic visit, as well as the reasons why treatment was not completed. The reasons for failing to complete the

Lidcombe Program were similar to those for Study 1 and are unremarkable for a speech clinic. These 12 children were not included in the analyses. The remaining 66 children all successfully completed Stage 1 by achieving the program criteria, as described in the previous chapter. Mean age at first Stage 1 clinic visit was 52 months, range 32-71 months. There were 20 girls and 46 boys. The children were treated in the order that they presented to the clinic.

The treatment was conducted according to the Lidcombe Program manual by five experienced speech pathologists, one of whom had had extensive experience in the Lidcombe Program and regularly conducts workshops on the method in the Britain. This clinician supervised the other four.

Dependent Variable

The dependent variable in this study was the same as that in Study 1, namely number of clinic visits required to complete Stage 1. Again, the dependent variable was dichotomised to fewer than 10 clinic visits (short treatment period) and 10 or more clinic visits (long treatment period). This procedure was used because, as was the case for Study 1, treatment time as a continuous variable did not meet requisite assumptions for least squares regression. The categorisation resulted in 26 children with a short treatment period and 40 children with a long treatment period.

Predictor Variables

Predictor variables were the same as those in Study 1. Again, they were ascertained from client records: gender, age at the first Stage 1 clinic visit, onset-to-treatment interval (time from reported onset of stuttering to the first Stage 1 clinic visit) and %SS at the first Stage 1 clinic visit. These data are obtained routinely by

Lidcombe Program clinicians; however, in this study the investigators could not retrieve three %SS measures and three onset dates from the records.

Table 5.1: Details of the 12 children who did not complete Stage 1

Particip- -ant number	Gender	Age	Time since onset	%SS at first clinic visit	Reason for not completing the Lidcombe Program
1	M	61	19	1.4	Mother struggled with delivering treatment
2	F	61	19	17.5	Treatment interrupted due to challenging behaviours, eg. Aggression, making herself ill
3	F	55	13	1.6	Treatment interrupted due to aggressive behaviour, referred to specialist clinic.
4	M	43	16	13.8	Family moved.
5	M	43	8	10.2	Family abandoned treatment as there were significant problems with younger sibling.
6	M	51	15	5.7	Father withdrew child from treatment as the family had broken up.
7	F	47	13	5.3	Treatment interrupted twice due to family relocation.
8	M	41	11	1.6	Therapy suspended several times due to family problems. Consistently poor attendance. Treatment finally abandoned by family.
9	M	59	23	3.2	Mother requested discharge due to transport problems and marriage break-up.
10	M	52	10	11.2	Mother found treatment in structured conversations and identifying stutters difficult. She failed to attend for further appointments.
11	M	43	13	23.4	Two unsuccessful attempts at treatment. Treatment was terminated firstly due to extreme domestic problems and subsequently due to transport difficulties
12	F	55	15	12	Treatment abandoned due to behaviour issues. Child was monitored for 15 months.

The predictor variables were categorised before the regression analysis was conducted in order to avoid the undesirable assumption that any relation between those variables and treatment time would be linear. Also, in the case of onset-to-treatment interval, categorisation improved the reliability of this variable by reducing its dependence on accurate parent recall. The a priori categorisation of predictor variables was done identically to Study 1. Onset-to-treatment interval was categorised as shorter than 12 months (23 children) and longer than or equal to 12 months (40 children). Age at first Stage 1 clinic visit was categorised into 2-3 years (24 children) and 4-5 years (42 children). Stuttering rate at first Stage 1 visit was categorised as low being below 5.0 %SS (32 children) and high being above 5.0 %SS (31 children).

Results

Descriptive Statistics on the British Data

Means, standard deviations, medians and ranges for each predictor variable are presented in Table 5.2. There were three children for whom %SS at first clinic visit was unable to be obtained and three children for whom onset-to-treatment interval was unable to be obtained. The median %SS score at the first Stage 1 clinic visit was 4.5 %SS. However, the cohort included 12 children with %SS scores of less than 1.0 at this time, despite the fact that they met the above criteria for stuttering. Like Study 1, this is a retrospective study and so they were not removed. These 12 children had a median treatment time of 6 sessions with a range from 3 to 17. Hence the children were clearly stuttering and needed treatment.

Table 5.2: Descriptive statistics for the 66 children

Variable	Number of Measures	Mean	SD	Median	Min	Max
%SS at first clinic visit	63	5.8	5.4	4.5	0	31
age at first clinic visit (months)	66	52	10	51	32	71
onset-to-treatment interval (months)	63	17	9.7	13	2	49
number of clinic visits	66	12	5.7	11	3	32

Logistic Regression on the British Data

Univariate logistic analyses were conducted for each predictor variable to determine which variables were statistically significant ($p < 0.1$). A conservative cut-off for the p -value was used so that potentially predictive variables would not be eliminated. The significant variables were then combined in a multivariate analysis, and backwards elimination conducted to determine which variables were independently associated with the dependent variable.

Results of the multivariate logistic regression are in line with those obtained for the Australian cohort in Study 1, and are presented in Table 5.3. Percent syllables stuttered was found to be a significant predictor of time to complete Stage 1 ($p = 0.029$), with an odds ratio of 3.8. There was a non-significant result suggesting that onset-to-treatment interval may be related to treatment time (odds ratio 0.33, $p = 0.084$).

Table 5.3: Results of multivariate logistic regression of treatment time
(British data only)

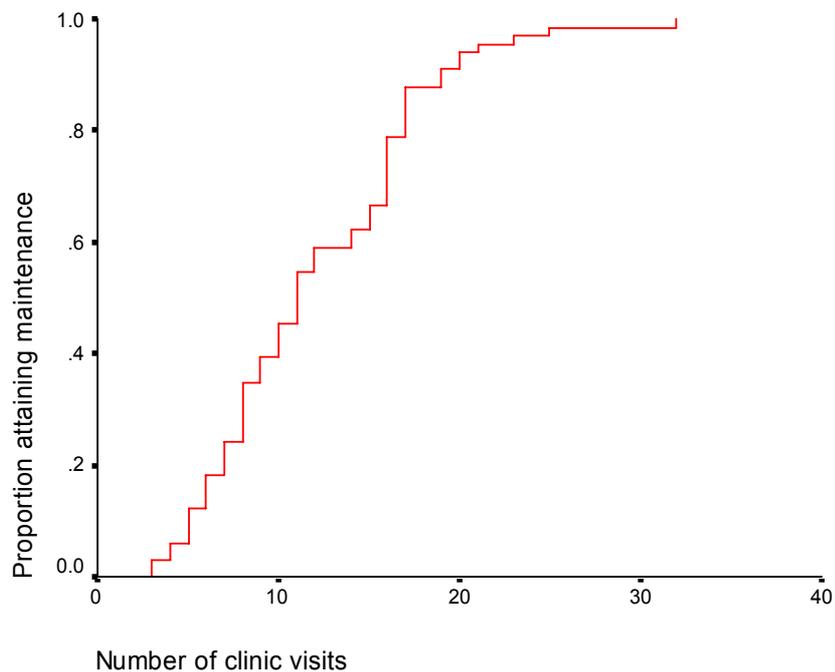
Variable		Odds Ratio	95% Confidence Interval	<i>p</i>-value
%SS at first clinic visit	<5 %SS	1.0*		
	>= 5 %SS	3.8	1.1 - 12	0.029
onset-to-treatment interval	<12 mths	1.0*		
	>= 12 mths	0.33	0.09 – 1.2	0.084

* = reference category

Kaplan-Meier Plots and Log-rank Tests

Figure 5.1 is a Kaplan Meier recovery plot (Kaplan & Meier, 1958), of number of clinic visits required to complete Stage 1. The number of clinic visits required for these children to complete Stage 1 ranged from 3 to 32 with a median of 11 visits. Figure 5.1 also shows that 95% of the children completed Stage 1 within 21 clinic visits. This is consistent with the results of Study 1. Log-rank tests showed no significant difference for age at first clinic visit, onset-to-treatment interval, or gender. However, %SS at first clinic visit was statistically significant ($p = 0.032$), confirming the result from the multivariate logistic regression analysis that showed that children whose stuttering was more severe took longer to complete Stage 1 than the children whose stuttering was less severe.

Figure 5.1: Kaplan-Meier plot of number of clinic visits to complete Stage 1



Meta-analysis

Because the methods of data collection were identical for the Australian and British cohorts (Study 1 and Study 2 respectively), meta-analyses of the combined data were conducted. Logistic regression was used to model the 316 participants after testing for heterogeneity between the two cohorts using interaction terms in logistic regression models. There was no evidence of heterogeneity, so a logistic regression adjusting for the fixed effect of whether participants originated from the British or Australian cohorts was performed.

Figures 5.2-5.5 show for the individual and combined cohorts, odds ratios and confidence intervals from a logistic regression model for each of the predictor variables. The odds ratios presented for the individual studies are based on univariate models. Statistical significance is implied when a 95% confidence interval does not cross the line where the odds ratio equals one. For the combined cohort (Total), both %SS and onset-to-treatment interval are statistically significant.

The result of a multivariate meta-analysis is shown in Table 5.4. A similar method to that described under the heading “Logistic Regression on the British Data” (see above) was used to determine which variables were independently associated with the dependent variable for the combined data. The increased statistical power of the meta-analysis resulted in a significant association between treatment time and both %SS at the first clinic visit ($p < 0.0001$) and onset-to-treatment interval ($p = 0.013$). The odds ratio of 3.5 for %SS at first clinic visit suggests that the odds of children whose stuttering was more severe were three and a half times greater than the odds of children whose stuttering was less severe to take longer to complete Stage 1. The odds ratio of 0.52 for onset-to-treatment interval suggests that the odds of children who had been stuttering for longer were half as great as the odds for children who had been stuttering for less time to take longer to complete Stage 1.

Figure 5.2: Meta-analysis of time to complete Stage 1 by stuttering severity at first clinic visit

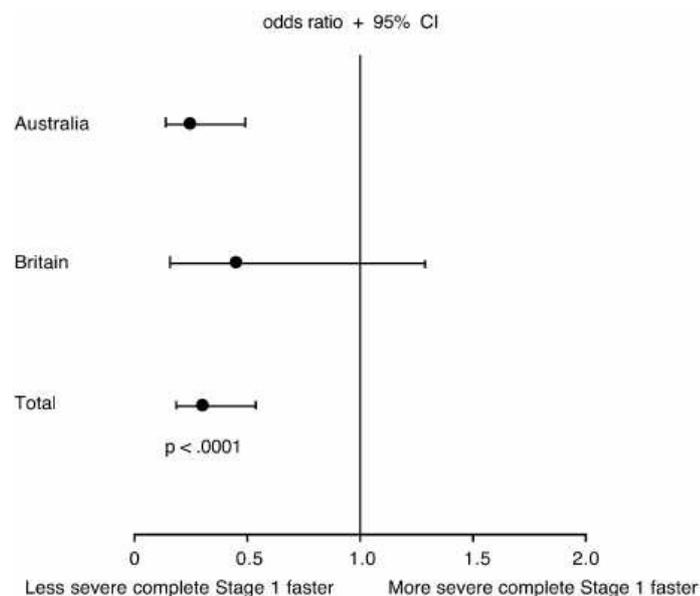


Figure 5.3: Meta-analysis of time to complete Stage 1 by time from onset of stuttering to treatment

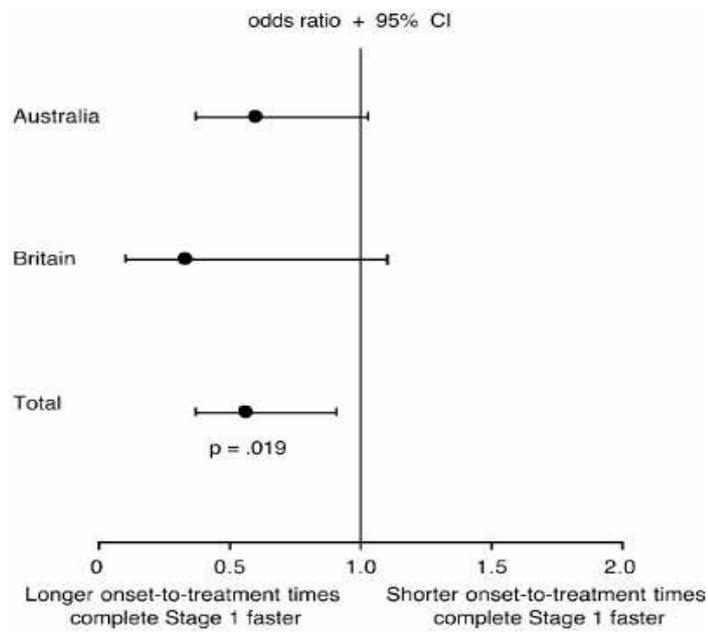


Figure 5.4: Meta-analysis of time to complete Stage 1 by gender

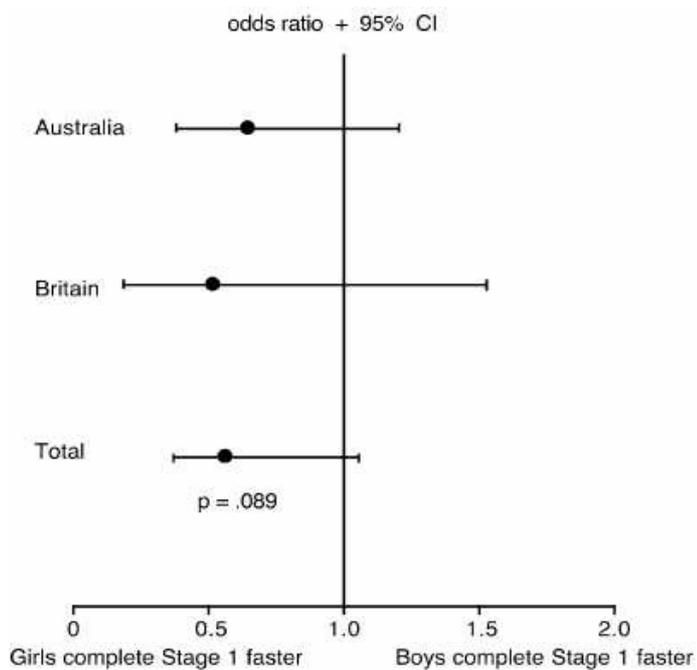


Figure 5.5: Meta-analysis of time to complete Stage 1 by age

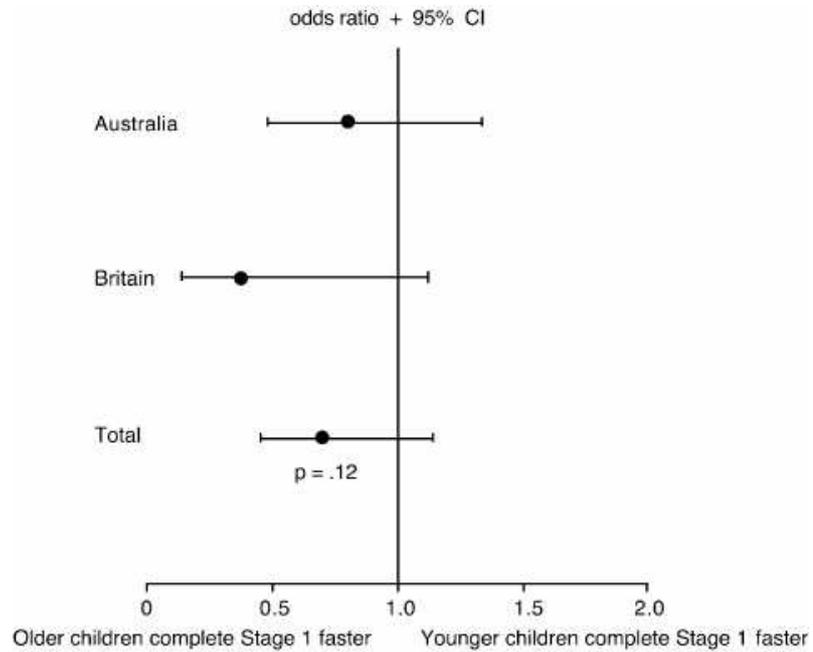


Table 5.4: Results of multivariate logistic regression of treatment time
(British and Australian data combined)

Variable		Odds Ratio	95% Confidence Interval	p-value
%SS at first clinic visit	<5 %SS	1.0*		
	>= 5 %SS	3.5	2.0 – 6.2	<0.0001
onset-to-treatment interval	<12 mths	1.0*		
	>= 12 mths	0.52	0.31 – 0.87	0.013

* = reference category

Discussion

The present results replicate those of Study 1. In a cohort of British children treated with the Lidcombe Program, the median number of clinic visits to complete Stage 1 was 11, with 95% of children completing Stage 1 within 21 clinic visits. This replication is particularly significant, as the treatment was conducted in a different country from where the Lidcombe Program originated. It should be stressed that there was no communication about treatment issues between the Australian group and the British clinicians during the treatment period. The British results, then, were obtained completely independently.

The combined data indicate that stuttering frequency at the first clinic visit is a highly significant predictor of number of clinic visits taken to complete Stage 1 of the Lidcombe Program. Gender and age are not significant predictors, although data suggest that girls and older children may require less treatment time. Hence, Starkweather and Gottwald's (1993) finding of a relation between stuttering severity and requisite treatment time in their treatment program has now been replicated twice with children receiving the Lidcombe Program. In a clinical caseload, children with more severe stuttering will be over three times more likely than children with less severe stuttering to require more than the median number of clinic visits to complete Stage 1 of the Lidcombe Program. These findings are, of course, intuitive. It seems reasonable to expect that treatment would take longer for children whose stuttering is more severe. However, confirmation that this is in fact the case is valuable information for clinicians who are deciding whether to intervene immediately with a stuttering child or whether to wait for a period to give natural recovery a chance to occur. Armed with the knowledge that there is a relationship between severity and treatment time, clinicians may not be prepared to wait as long

with a child whose stuttering is severe as they might with a child whose stuttering is mild.

However, Starkweather and Gottwald's finding that children who had been stuttering longer took longer to treat was not replicated in this study. In fact, the meta-analysis showed the opposite to be the case. Although neither the Australian or British cohorts alone produced a statistically significant effect in this regard, the data in each cohort was suggestive of such a relationship. This relationship also held when other factors such as severity were adjusted for in a multivariate analysis. The meta-analysis combined the two cohorts, thereby increasing the size of the sample and consequently the statistical power of the analysis. In the combined analysis, onset-to-treatment interval was significantly associated with treatment time (odds ratio 0.52, $p = 0.013$, after adjustment for severity and cohort). Although this effect is not as strong as the effect of severity on treatment time, it can still be considered clinically significant. A liberal interpretation of this finding is that children who have been stuttering for longer are likely to take less time to complete Stage 1 of the program. A more conservative interpretation is that it is very unlikely that delaying treatment for up to 12 months after onset of stuttering will result in increased treatment time.

Unlike the result for severity, the reasons for this finding are not apparent, and it is not clear why it differs so markedly from that of Starkweather and Gottwald (1993). The post-treatment stuttering data in Starkweather and Gottwald's study are not available, so the outcome of the two studies cannot be compared. However, as discussed in Chapter 4 the Multi-process Approach and the Lidcombe Program are very different treatments. The Multi-process Approach is primarily indirect while the Lidcombe Program relies on direct operant procedures. It is possible that this may

account for the differences between them in the effect of time since onset on treatment responsiveness.

There are a number of caveats to this study as in Study 1. First, as discussed above, the findings relate only to children treated with the Lidcombe Program. Second, none of the findings can be extrapolated beyond the preschool years. Third, the findings pertain only to completion of Stage 1 of the Lidcombe Program, and not to long-term outcomes. Finally, the data were obtained retrospectively from clinic files, and so should be interpreted with some caution.

Nonetheless, it seems clear—from the two clinical file audits reported in this thesis at least—that delaying intervention with the Lidcombe Program for a short period after onset is unlikely to decrease a child’s responsiveness to the treatment. Naturally this result should not be taken to mean that treatment should necessarily be delayed for 1 year after onset. It has been suggested that clinicians take many factors into account when deciding when to implement treatment for early stuttering, such as severity of stuttering (see above), whether stuttering severity is decreasing, age of the child, time since onset, the reaction of parents and child to the stuttering, and whether parents wish treatment to proceed (Packman & Lincoln, 1996; Packman & Onslow, 1998; Packman, Onslow & Attanasio, 2003; Onslow & Packman, 1999b; Zebrowski, 1997). In the end, the decision of when to begin treatment will most likely always remain a clinical judgement informed by science rather than a purely scientific, data-based judgement (Packman et al., 1999). The present results can be seen as contributing to the evidence that clinicians can use to inform that judgment, at least as far as intervention with the Lidcombe Program is concerned.

The implication of these results on a prospective clinical trial of the Lidcombe Program treatment additional to those stated in the discussion section of

the previous chapter would be the following. For a no-treatment control group, it is confirmed that delaying treatment for at least a year after onset of stuttering should not reduce their responsiveness to the Lidcombe Program, if treatment is still required at the end of that period. In fact evidence suggests that treatment time with the Lidcombe Program is less for those children with longer onset-to-treatment times, although it is uncertain how much delay can be afforded. In addition to including baseline severity of stuttering as a stratification factor, it would also be advisable to include time from onset as a stratification factor as the results of the meta-analysis indicate this variable is associated with recovery duration.

Unfortunately, while the results of Study 1 were available at the start of the randomised controlled trial reported in Chapter 7, the results of the present study were not hence onset-to-treatment time was not included as a stratification factor. However, age was included as a stratification factor (as it was thought at the time that this variable may impact upon treatment outcome) and this correlates strongly with onset-to-treatment time.

**PART THREE: RANDOMISED CONTROLLED TRIAL OF THE
LIDCOMBE PROGRAM**

**Chapter 6: A Simulated Study of Analysis of the Primary
Outcome Measure**

Introduction

Prior to embarking on the analysis of the randomised controlled trial of the Lidcombe Program presented in the following chapter, a simulation study was undertaken to determine the most appropriate method of analysis. The main outcome for the trial is percentage of stuttered syllables (%SS), which is a proportion that has a positively skewed distribution. Although many simulation studies have investigated the analysis of non-normal data, none has dealt specifically with %SS. Additionally as the sample size for the randomised controlled trial was 54 participants in total, the impact of small sample size on analysis of %SS was also investigated.

Measures of Percent Syllables Stuttered

An objective, behavioural measure of stuttering makes an essential contribution to assessment of the severity of the disorder at one time and over a period. Severity measures that are based on counts of stuttering are informative because they reflect the rate at which perceived moments of speech breakdown occur. Although there are reliability issues associated with stuttering-count measures (Cordes & Ingham, 1994), researchers seem generally able to publish stuttering-count data with adequate inter-judge and intra-judge reliability. Further, in the sense that replication of findings is a sure test of reliability (Sidman, 1960) stuttering count measures have conveyed a range of replicated effects (Siegel, 1990). Such measures are valid to the extent that they are generally interchangeable with a perceptual measure of stuttering severity (Bloodstein, 1995), with the exception of cases where stuttering behaviour contains prominent portions of either repeated movements or fixed postures (O'Brian, Packman, Onslow, & O'Brian, 2004).

The stuttering count measure that is most commonly incorporated in current research is percentage of syllables stuttered (%SS). Dual button-press counting devices are available to assist the observer in collecting this measure (Lincoln & Packman, 2003). As stated in Chapter 1, percentage of syllables stuttered provides a straightforward calculation involving the total number of syllables associated with unambiguous stuttering and the total number of syllables in the speech sample. The calculation is simply:

$$\%SS = 100 \times (\text{total number of stuttered syllables} / \text{total number of syllables}).$$

Although the datum %SS is commonly referred to as a stuttering rate or stuttering frequency measure, in the arithmetical sense it is a proportion. Hence, %SS scores will always be a positive number between zero and 100.

Assumptions in Statistical Analysis of %SS

The mathematically related procedures of t-test, least squares regression, and analysis of variance (ANOVA) are common methods of statistical inference in stuttering research that involves %SS scores. These procedures rely on fundamental assumptions as follows:

1. Randomly selected samples
2. Independence of samples and observations
3. Normal distribution
4. Homogeneous variance between groups

The two assumptions, randomness of the samples selected, and independence of samples and observations, should be met in most cases providing samples are randomly chosen from the clinical or non-clinical populations they are intended to represent and one measurement from each participant contributes to the statistical inference. If more than one measurement is taken from each participant, more

complicated analyses would normally be undertaken, which in turn may require further assumptions.

However, the two other assumptions underlying ANOVA, least squares regression, and t-test analyses warrant consideration in the context of analysis of %SS measures. These assumptions are that outcomes are normally distributed, so that the sampling errors also follow a normal distribution when modelling the outcomes, and homogeneity of population variances between the groups.

The possible violation of these assumptions is an important issue because the results of the analysis may be incorrect or misleading. For example if the assumption of normality is violated then a standard analysis such as ANOVA, least squared regression, or t-test may not be the most powerful available and this could mean the difference between detecting a true difference or not. A more serious problem would be an increased Type I error resulting in an increased probability of concluding a true difference when in reality one does not exist. Often, the effect of an assumption violation on the result depends on the extent of the violation such as how unequal the population variances are, or how skewed one or the other population distribution is. Some small violations may have little practical effect on the analysis, while other violations may render the result incorrect or uninterpretable.

In reviewing a number of randomised studies evaluating treatments for stuttering (see Chapter 2), it was found that the majority of methods used to analyse stuttering data were standard techniques such as ANOVA, t-test and least squares regression. Often, in recognition of the fact that these data are skewed and in potential violation of assumptions, transformations were performed or non-parametric methods were used. However whether it is undesirable to analyse skewed data or whether transforming them or using non-parametric methods improves the

situation is not clear in this area of research. In this simulated study these matters are pursued and an attempt to develop some guidelines for the analysis of stuttering data is made.

In stuttering research, recruitment of participants to study is problematic because the incidence of stuttering is low and many do not consider their stuttering to be a problem. Additionally, it is common for participants to agree initially but then withdraw from study. In many cases obtaining follow up samples of speech can be difficult, if not impossible. Hence, achieved sample sizes are often low and can be quite different for different groups under study. In this regard, the impact of small sample sizes and differing sample sizes will be investigated in this simulated study. It is also quite usual for different groups under study to have quite different levels of stuttering. Thus, in this study the effects of differing levels of stuttering will also be compared. The impact of these factors will be evaluated for the various analytical techniques under study.

The Distribution of %SS

Percentage stuttered syllables is a proportion, hence it will always be a positive number between zero and 100%. It is assumed that for non-stutterers the proportion will be zero, or at least close to zero. However, for stutterers, the proportion will typically be greater than zero. How much greater than zero is illustrated for a number of cohorts of stutterers presented in Figures 6.1-6.4. The distributions of pre-treatment %SS for young children are from the two cohorts presented in Chapter 4 (Australian children) and Chapter 5 (British children). Additionally, the distributions of %SS for a cohort of adult stutterers from Melbourne, Australia are described before and after treatment (Block, Onslow, Packman, Gray, & Dacakis, 2005).

The three cohorts represent a broad range of the stuttering population. They represent adults and children, from different continents, before and after treatment. For these samples, %SS is skewed to the right with the majority of participants having small proportions of stuttered syllables but some having larger proportions and occasionally much larger proportions. Mean %SS for the four samples differ, as do their standard deviations. However, the distributions have two similarities, one being the shape and the other, the relationship between the mean and standard deviation.

For distributions with a smaller mean, there is a corresponding smaller standard deviation. In fact, the standard deviation appears to be approximately equal to the mean for all the distributions. This skew and relationship between the mean and standard deviation appears to be consistent for all the distributions of %SS observed.

Figure 6.1: Distribution of %SS in a clinical caseload of English children

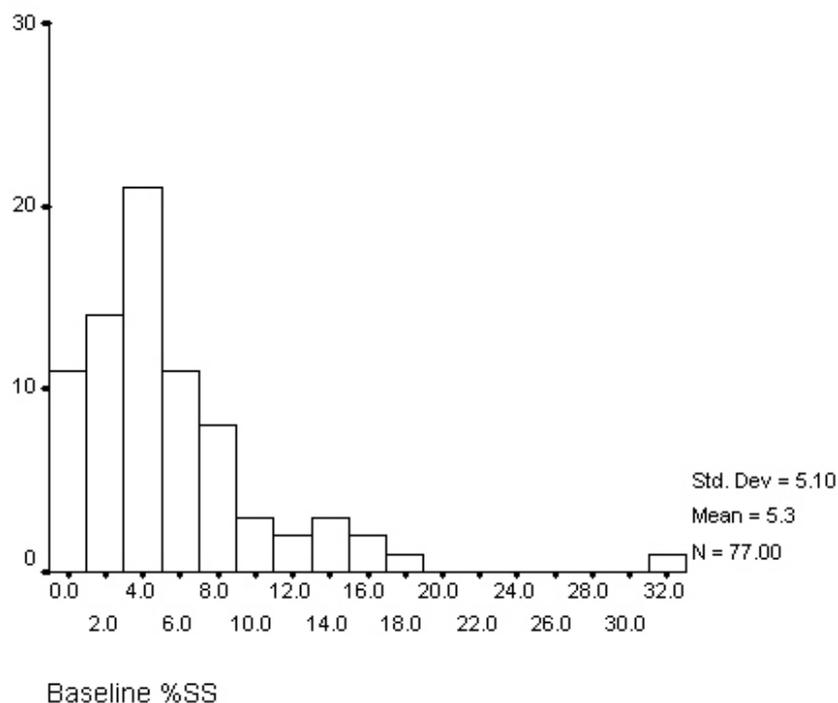


Figure 6.2: Distribution of %SS in a clinical caseload of Australian children

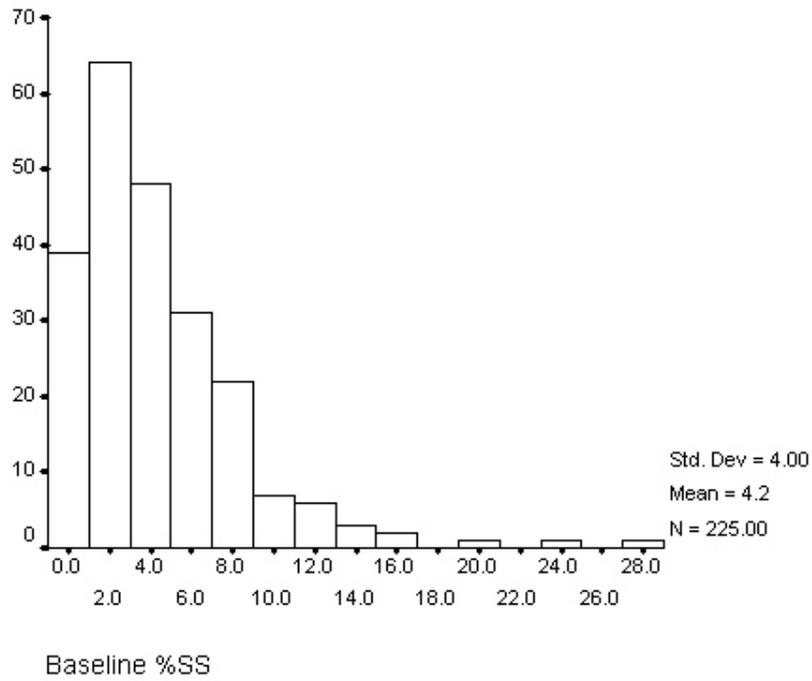


Figure 6.3: Distribution of %SS in a cohort of adult stutterers before treatment

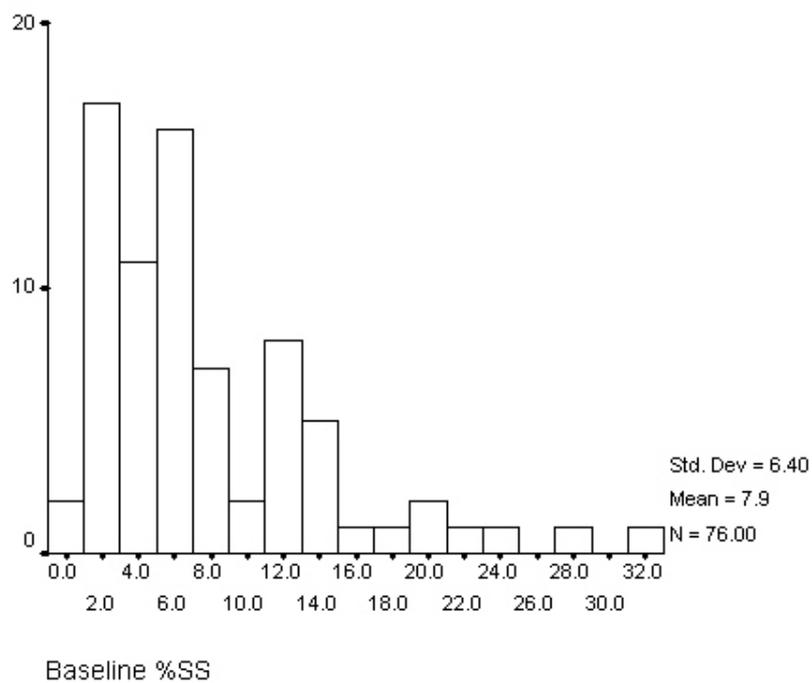
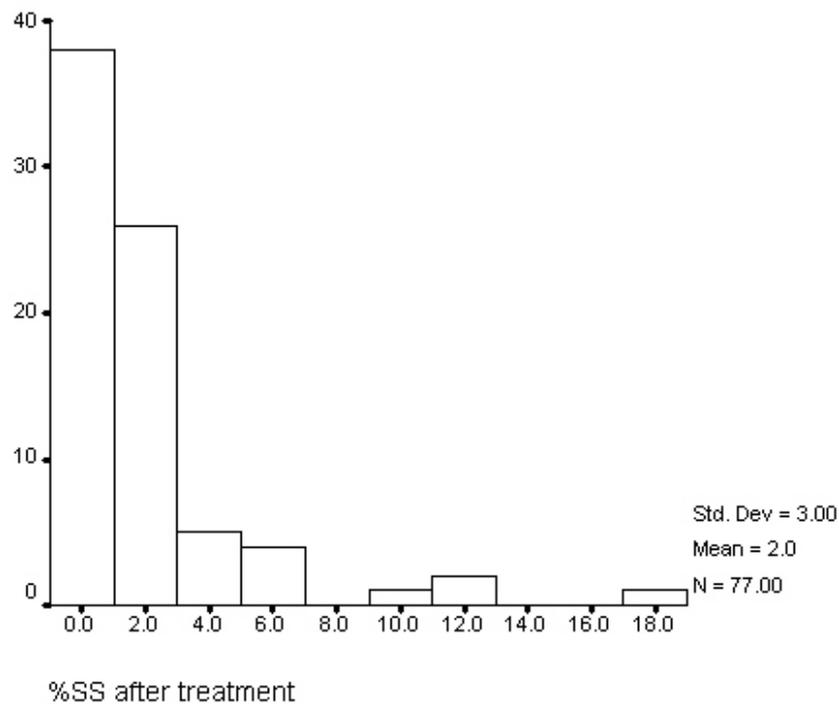


Figure 6.4: Distribution of %SS in a cohort of adult stutterers after treatment



The data shown in these plots have an obvious non-normal distribution. How can this distribution be modelled? Whilst there are many statistical distributions that may provide an adequate fit to the data, a common distribution in statistical science is that based on the gamma family.

The gamma distribution arises naturally in processes that involve waiting times between certain types of events such as traffic accidents or the number of stoppages on a production line. Most statistical distributions including the gamma distribution can be characterised by their parameters. For example, the normal distribution has two such parameters: the mean, which is a location parameter, and the standard deviation, which is a scale parameter. The location parameter, in this case, determines where along the x-axis the distribution is located. A mean of zero implies the centre of the distribution is located at $x = 0$, while a mean of 10 implies the centre of the distribution is at $x = 10$. The effect of the scale parameter is to

stretch out the distribution. Hence a standard deviation of 3 implies a much wider distribution than a standard deviation of 1.

The gamma distribution can be characterised by three parameters: a shape parameter, a scale parameter and a location parameter (sometimes referred to as the threshold parameter). The location or threshold parameter determines the minimum value possible for the distribution. For example, a location parameter of zero indicates that all values are greater than zero. The shape parameter, as the name suggests, determines the shape of the distribution. The shape of the gamma distribution can vary as gamma is not a single distribution, like the normal distribution, but is rather a family of distributions. These distributions are particularly useful in modelling applications since they are flexible enough to model a variety of data sets. As is the case for the normal distribution, the scale parameter for the gamma distribution determines the width of the distribution. Hence, a scale parameter of 10 indicates a much wider or more variable distribution than a scale parameter of, say, 5 (Jambunathan, 1954).

Just as there are statistical tests for normality, a test to determine whether data are adequately described by other known distributions, such as the gamma distribution, are also available. The Statistical Analysis System (SAS) (SAS Institute, Cary, NC) includes the Kolmogorov-Smirnov goodness-of-fit test (D'Agostino & Stephens, 1986) for the gamma distribution in the Proc Univariate procedure and when this is performed on the data described in the four figures above, it indicates adequate goodness-of-fit ($p > 0.1$ for all distributions). It was reassuring to note that the gamma distribution appeared to provide adequate goodness-of-fit for both pre-treatment and post-treatment data. Figures 6.5 and 6.6 present the normal and gamma Q-Q plots of %SS for the British cohort of children. Q-Q plots plot the quantiles of a

variables distribution against the quantiles of any of a number of test distributions. These plots are generally used to determine if the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points cluster around a straight line. It can be seen from Figures 6.5 and 6.6 that the gamma distribution appears to fit the data better than the normal distribution. Similar results are obtained for the other cohorts.

The SAS program also generates a theoretical distribution based on the type of distribution specified and the data from the actual distribution. SAS generated theoretical distributions based on gamma and the four actual distributions, described above, of location parameter zero, shape parameter of approximately one, and with varying scale parameters relating to the different mean and standard deviations of the four actual distributions.

Figure 6.5: Normal Q-Q plot of %SS

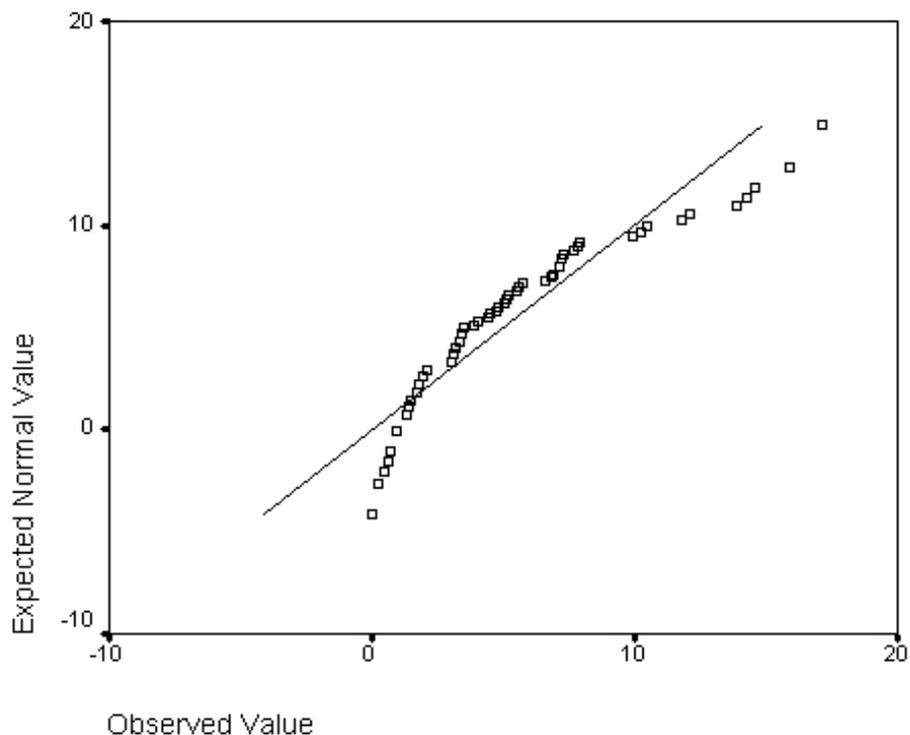


Figure 6.6: Gamma Q-Q plot of %SS

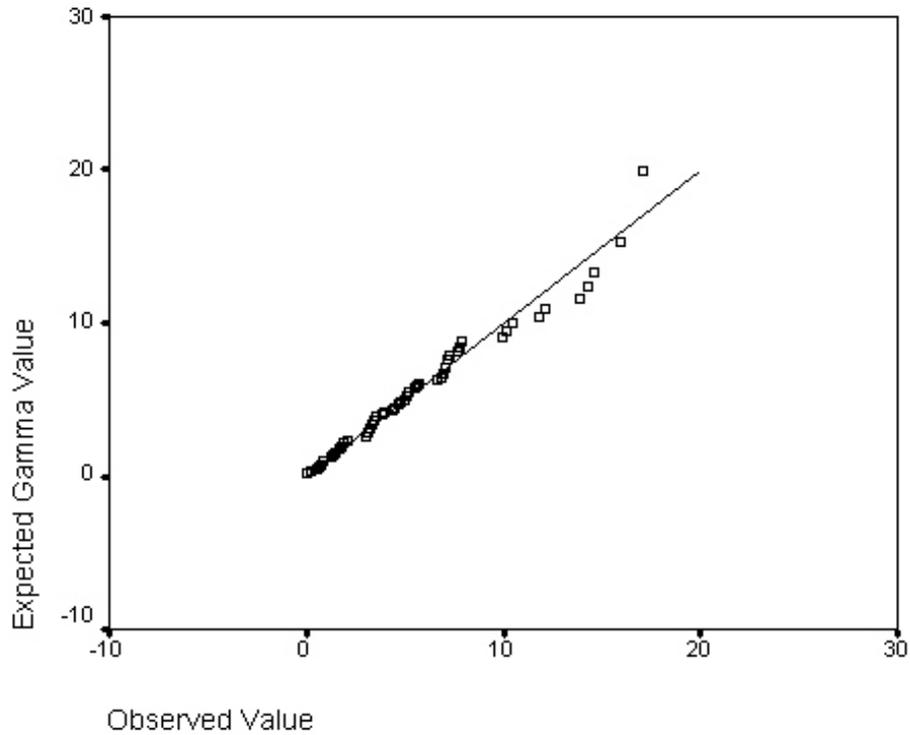
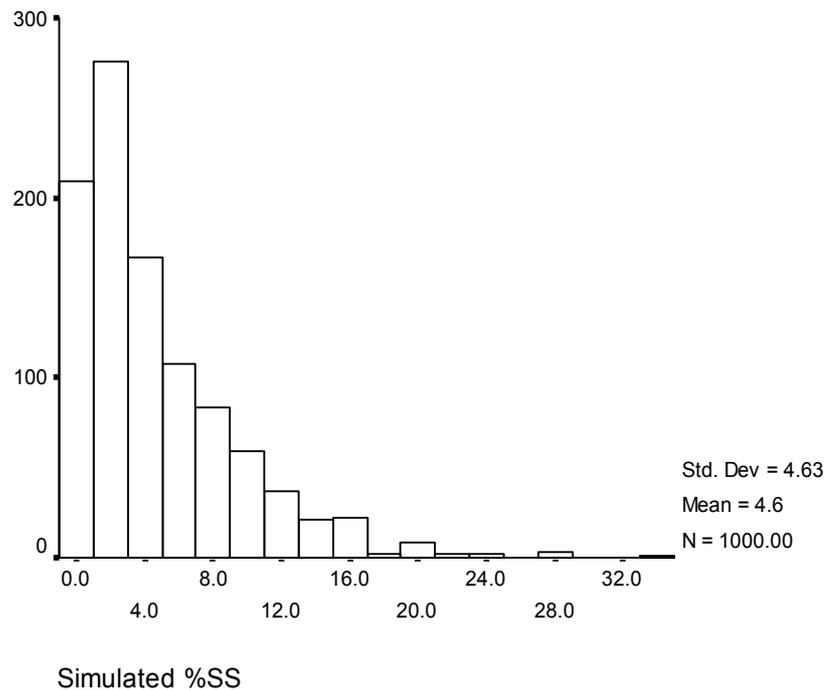


Figure 6.7 shows a gamma distribution that has a shape parameter of one, location parameter of zero and scale parameter of five, simulated to resemble a typical distribution for %SS. A shape parameter of one implies that the expected mean and standard deviation of the distribution will be equal. A shape parameter of greater than one would give rise to a less skewed distribution where the expected standard deviation is less than the expected mean whereas a shape parameter of less than one would give rise to a more skewed distribution where the expected standard deviation is greater than the expected mean. A location parameter of zero implies that all values will be greater than zero. A scale parameter of five, in this case, ensures that the expected mean and standard deviation will be five.

Figure 6.7: Simulated gamma distribution



In summary, the gamma distribution was chosen to approximate the distribution of %SS for the following reasons. First the shape of the gamma distribution closely resembles that of the shape of the distribution of %SS scores, both pre-treatment and post-treatment. Second goodness-of-fit tests suggest the gamma distribution adequately describes the distribution of %SS, both pre-treatment and post-treatment, and this is confirmed with a Q-Q plot. Third a location parameter of zero can be chosen so that all simulated scores are greater than zero, as for %SS. In addition the maximum score is extremely unlikely to be greater than 100, the theoretical maximum for %SS.

Appropriateness of using Standard Techniques for Analysis of %SS

Given that the distribution of %SS is not normal but can be adequately approximated by a gamma distribution, how appropriate are standard techniques such

as ANOVA for analysis of %SS scores? From the central limit theorem, a well-known statistical theory, as the size of a sample increases, the distribution of the mean of the sample tends to normality even when the distribution of the population is non-normal. Hence, standard analysis techniques such as ANOVA that compare means should perform adequately with data that are non-normal providing the sample size is large enough. How large the sample size needs to be depends on how non-normal the data are. In most cases, a sample size as low as 10 is sufficient (Miller, 1986). Additionally when two samples of non-normal data are compared, the problems associated with the non-normality tend to cancel out if the distributions and sizes of the two samples are similar.

With regard to homogeneity of variances, if the sample sizes for the two groups are similar then standard analytic techniques, such as ANOVA and t-test, are quite robust to quite large differences in the variances of the two groups. Even ratios of the variances of up to four and even higher can be tolerated if the sample sizes are nearly equal (Miller, 1986). However this is not true when the sample sizes of the two groups are also not similar. If sample sizes and variances are both substantially different for the two groups, then standard analytical techniques may not be appropriate. More specifically if the group with the smaller variance also has the larger sample size this will weight the pooled variance estimate toward a smaller value, thereby inflating power. However if the larger sample size occurs with the group with larger variance, the pooled variance estimate will result in a test that has less than optimal power.

A number of researchers have investigated the performance of standard analysis techniques on non-normal data. Stonehouse and Forrester (1998) investigated the performance of standard techniques as well as non-parametric

techniques on normal and non-normal data. However only Type 1 error was assessed; power was not investigated. This research showed that standard techniques performed adequately in most cases. Two situations when this was not the case were when skew was severe and when the groups compared differed in sample size and in their spread, or variance. Welch's (1938) approximate t-test for unequal variances addressed the latter problem for normal data but it was sensitive to departures from normality.

Kingman and Zion (1994) investigated the effect of transformations on non-normal data with respect to power and Type 1 error. Their investigation included gamma distributed data and transformation by logs and square root. Conclusions stated that neither transforming gamma distributed data prior to performing t-test nor performing t-test on untransformed gamma data had much effect on the rate of Type 1 error. However, compared to the universally most powerful test for the gamma distribution, which is based on the ratio of sample means, transformation by square root prior to performing t-test resulted in a small loss of power (3-6%), whereas transformation by logs resulted in a larger loss of power (5-10%). Performing t-test on the untransformed gamma data resulted in similar loss of power as transformation by square root.

Boneau (1960) violated a number of assumptions underlying the t-test and found through simulation that generally such violations produce a minimal effect on the distribution of the t values and that the t-test is a remarkably robust test. One exception was when both the sample sizes and variances were different in which case procedures that deal with unequal variances should be used such as those proposed by Satterthwaite (1946) or Welch (1938). However when the two distributions are skewed and of different shape a lack of robustness can occur especially when sample

sizes are small. Boneau only investigated violation of assumptions on the Type 1 error; power was again not investigated.

In the area of stuttering research, it is common for groups with small sample sizes or even different sample sizes to be compared. It is also quite possible that groups with different variation could be compared. And finally, it is likely that standard techniques such as ANOVA would be used to perform the analysis. This leads to the main questions addressed here. How appropriate are standard techniques for the analysis of %SS scores? Does a transformation of the data or do non-parametric methods produce better results? And, finally, under what circumstances do these techniques perform adequately and under what circumstances do they not?

In summary, ANOVA, t-test and least squares regression are commonly used to analyse %SS data, however assumptions of normality and homogeneity of variances are unlikely to hold for this kind of data. Transformation of %SS and the use of non-parametric methods are often used in the belief that these analytic techniques are more appropriate than the standard methods described above. However it is unknown which of this collection of analytic techniques is optimal for the analysis of %SS and under what conditions. The overall aim of this study is to develop some guidelines for the appropriate analysis of stuttering data under the various conditions that are likely to occur in practice. In particular, the study was conducted to determine the most appropriate analysis for the Randomised Controlled Trial reported in Chapter 7.

Methods

To investigate the most appropriate method for analysis of %SS data, gamma distributed data were simulated with location parameter zero, shape parameter one and varying scale parameters to imitate data likely to be obtained in practice. Four

methods for analysing %SS are included in the comparison: ANOVA performed on raw %SS data, ANOVA performed on transformed data, the Wilcoxon-Mann-Whitney non-parametric test and the t-test proposed by Satterthwaite (1946) for unequal variances. Transformation is a common response to skewed data such as %SS, and two transformations commonly used for moderately right-skewed data (square root and log) were compared. Consistent with the results of Kingman and Zion (1994), square root transformation generally provided greater power (15% on average) than transforming using logs, hence the results obtained after log transformation are not presented.

The Wilcoxon-Mann-Whitney test was chosen above the many other non-parametric tests for comparing two continuous distributions because it is commonly used and also commonly available in the majority of statistical software. Although non-parametric tests do not require data to be from a normal distribution they are not necessarily assumption-free tests. In fact one of the assumptions for the Wilcoxon-Mann-Whitney test is that the populations, from which the two samples have been drawn, have the same variance. Therefore this test will not solve the problem of heterogeneous variance however the t-test proposed by Satterthwaite (1946) deals specifically with unequal variances and that is why it has been included.

Three different sample sizes were investigated in various combinations. That is, samples of the same size were compared and samples of different sizes were compared. These three sizes were chosen to represent a range of sample sizes likely to be obtained in practice. The sample sizes were 5, 20 and 50. For each combination of sample sizes, two different treatment effects were investigated. One of these treatment effects was zero or no treatment effect, included to assess the performance of the models when there is no treatment effect in the data. This analysis represents

the investigation of Type 1 error and is described more fully below. The other treatment effect was chosen based on a sample size calculation for normally distributed data and approximately 80% power. More details are provided in the results section.

The simulation method used for investigating power (the importance of which is discussed in Chapter 3) involved taking random samples from a Gamma distribution of varying sample sizes and treatment effects. The random number generator used was that provided in the RAND function in base SAS (Fishman, 1978) with the seed generated from the system clock (see Appendix A for a sample SAS program used to generate the simulations). Analysis by the various methods proposed was then performed. This procedure was repeated 2,000 times for each method of analysis and the proportion of times a statistically significant result ($p < 0.05$) was obtained was calculated for each method. This analysis gives an estimate of the power for detecting a true difference between the two simulated treatment groups. The estimated level of precision for the power obtained with this analysis is an exact binomial 95% confidence interval around 80% of 78.2 - 81.7% (Clopper & Pearson, 1934).

A similar method was used to compare the rate of Type 1 error for the two models except that the two random samples were simulated from the same distribution. Analysis was then performed and this procedure was repeated 2,000 times for each method of analysis. The proportion of times a statistically significant result ($2p < 0.05$) was obtained was used to estimate the rate of Type 1 error. Because the two samples are generated from the same distribution, it is expected that there should be no difference between them. Obtaining a statistically significant difference in this situation is a Type 1 error and ideally should not occur more than

5% of the time. The estimated level of precision for the Type I error obtained with this analysis is an exact binomial 95% confidence interval around 5% of 4.09-6.05%. The number of simulations of 2,000 was chosen so that the 95% confidence interval around an estimate of Type 1 error of approximately 5% would be approximately +/- 1%.

SAS, version for Windows 8.2 (SAS Institute, Cary, NC) or SPSS, version for Windows 11.5 (SPSS Inc., Chicago, Illinois) were used for all analyses.

Results

Each of the five tables of results presented (see Tables 6.1-6.5) deals with a different sample size scenario with varying means and standard deviations. For the first table of results (Table 6.1), the sample size is 50 for each group and for the power analysis the scale parameters are 3.6 for Group 1 and 2 for Group 2. These scale parameters imply that the expected means and standard deviations for the two groups would be 3.6 and 2, respectively, which equates to a 44% reduction in %SS for Group 2 compared to Group 1 (Cohen's [1988] measure of effect size $d = [\text{mean}_1 - \text{mean}_2] / \sqrt{[(\sigma_1^2 + \sigma_2^2) / 2]} = 0.55$). The values for these parameters have been chosen based on a sample size calculation for normally distributed data and approximately 80% power. For the Type 1 error analyses, the scale parameters are 3 for both groups, which equates to a 0% reduction in %SS. For the power comparisons, the power of each method relative to the other methods is of greatest importance, however the absolute power is also of interest as this can be compared to 80%, which is what would be expected if the data were normally distributed and standard analysis performed. For the Type 1 error analyses, the comparison of Type 1 error is less important, as of most interest is that the frequency of Type 1 error does not exceed 5%.

Table 6.1: Power and Type 1 error for sample size of 50 in each group, with expected mean and standard deviations 3.6 for Group 1 and 2.0 for Group 2

Method of analysis	Type I error (n = 50)	Power (n = 50, Cohen's d = 0.55)
ANOVA (no transform)	5.5%	80.9%
ANOVA (sqrt transform)	5.0%	79.7%
Wilcoxon-Mann-Whitney	4.7%	69.6%
Satterthwaite's t-test	5.6%	81.1%

In Table 6.2, the sample size is 20 for Group 1 and 50 for Group 2 (Power 1) and 50 for Group 1 and 20 for Group 2 (Power 2). The expected means and standard deviations were 4.4 for Group 1 and 2.0 for Group 2, corresponding to 55% difference in %SS scores (Cohen's $d = 0.70$). In Table 6.3, the sample size is 20 for each group and for the power analysis the expected means and standard deviations were 5.4 for Group 1 and 2.0 for Group 2, representing a 63% difference in %SS scores (Cohen's $d = 0.84$).

Table 6.2: Power and Type 1 error for sample size of 20 and 50 in each group, with expected mean and standard deviations 4.4 for Group 1 and 2.0 for Group 2.

Method of analysis	Type 1 error (n = 20, 50)	Power 1 (n = 20, 50 Cohen's d = 0.70)	Power 2 (n = 50, 20, Cohen's d = 0.70)
ANOVA (no transform)	5.1%	84.0%	72.2%
ANOVA (sqrt transform)	5.2%	80.8%	75.2%
Wilcoxon-Mann-Whitney	5.0%	70.5%	72.0%
Satterthwaite's t-test	6.6%	68.6%	87.4%

Table 6.3: Power and Type 1 error for sample size of 20 in each group with expected means and standard deviations 5.4 for Group 1 and 2.0 for Group 2

Method of analysis	Type 1 error (n = 20)	Power (n = 20, Cohen's d = 0.84)
ANOVA (no transform)	6.0%	81.4%
ANOVA (sqrt transform)	5.0%	81.9%
Wilcoxon-Mann-Whitney	5.9%	72.7%
Satterthwaite's t-test	5.3%	80.9%

Table 6.4 presents results obtained with a sample size of 5 for Group 1 and 20 for Group 2 (Power 1) and 20 for Group 1 and 5 for Group 2 (Power 2). For the power analysis the expected means and standard deviations were 13.0 for Group 1 and 2.0 for Group 2, representing an 85% difference in %SS (Cohen's $d = 1.18$).

Table 6.4: Power and Type 1 error for sample size of 20 and 5 in each group, with expected mean and standard deviations 13.0 for Group 1 and 2.0 for Group 2

Method of analysis	Type 1 error (n = 5, 20)	Power 1 (n = 5, 20, Cohen's d = 1.18)	Power 2 (n = 20, 5, Cohen's d = 1.18)
ANOVA (no transform)	4.9%	95.6%	40.7%
ANOVA (sqrt transform)	5.2%	94.3%	79.5%
Wilcoxon-Mann-Whitney	4.3%	76.3%	83.0%
Satterthwaite's t-test	7.9%	23.7%	97.8%

In Table 6.5, the sample size was 5 for each group, and for the power analysis the expected means and standard deviations were 20 for Group 1 and 2 for Group 2, a 90% difference (Cohen's $d = 1.27$). These parameters were chosen based on a

sample size calculation for normally distributed data and approximately 70% power, because 80% power was not able to be obtained with such a small sample size. It was noted that with a sample size of 10 in each group and an 80% reduction in %SS (Cohen's $d = 1.10$), power was 78.1% for ANOVA (no transform), 85.5% for ANOVA (square root transform), 77.8% for Wilcoxon-Mann-Whitney and 72.8% for Satterthwaite's t-test. With a sample size of 15 in each group and a 70% reduction in %SS (Cohen's $d = 0.94$), power was 80.5% for ANOVA (no transform), 82.6% for ANOVA (square root transform), 74.6% for Wilcoxon-Mann-Whitney and 76.7% for Satterthwaite's t-test.

Table 6.5: Power and Type 1 error for sample size of 5 in each group, with mean and standard deviations 20.0 for Group 1 and 2.0 for Group 2.

Method of analysis	Type 1 error (n = 5)	Power (n = 5, Cohen's d = 1.27)
ANOVA (no transform)	3.9%	52.4%
ANOVA (sqrt transform)	4.2%	71.1%
Wilcoxon-Mann-Whitney	2.9%	61.0%
Satterthwaite's t-test	2.6%	31.2%

Discussion

When the sample size was 50 in both groups, the results for all methods were reasonably consistent, except that power was lower for the Wilcoxon-Mann-Whitney test. The power obtained for both ANOVA methods as well as Satterthwaite's t-test was approximately 80% and the Type 1 error rate was approximately 5% for all four methods. When the sample size for the first group was reduced to 20, both the

Wilcoxon-Mann-Whitney test and Satterthwaite's t-test had lower power than the other methods. Both ANOVA methods achieved slightly more than 80% power. Satterthwaite's t-test was also associated with an increased Type 1 error (6.6%, 95% confidence interval: 5.5-7.7%) while the other three methods had approximately 5% Type 1 errors. When the sample size of the second group was reduced to 20 (and the first group remained at 50), power was very similar for three of the methods of analysis but lower than 80%, however power was substantially higher for Satterthwaite's t-test and higher than 80%.

In Table 6.3, the sample size for both groups was 20 and the performance of the Wilcoxon-Mann-Whitney was once again below the other three methods with regard to power. Power for both ANOVA methods as well as Satterthwaite's t-test was just over 80% and the rate of Type 1 error was 5% or slightly greater for all four methods. When the sample size for the first group was reduced to 5, the Wilcoxon-Mann-Whitney test and especially Satterthwaite's t-test had lower power than both ANOVA methods but power for those two methods was approximately 95%. The rate of Type 1 errors was high for Satterthwaite's t-test (7.9%, 95% confidence interval: 0.67-0.91%) but approximately 5% for the other three methods. When the sample sizes for Group 1 and Group 2 were 20 and 5, respectively, the results were very different. ANOVA performed on the raw data had much lower power than the other three methods and the power for Satterthwaite's t-test was almost 98%. Power for the other two methods was around 80%.

In Table 6.5, the Wilcoxon-Mann-Whitney test, ANOVA performed on raw data and Satterthwaite's t-test had reduced power compared to ANOVA performed on square root transformed data. ANOVA performed on square root transformed data had approximately 70% power and all three methods had rates of Type 1 error of

around 4% or even lower. Additional analyses showed that when there were 10 participants in each group, ANOVA performed on square root transformed data had greater power than ANOVA performed on raw data but when sample size was increased to 15 in each group, power became quite similar although was slightly lower for ANOVA on the raw data. Power was lower for Satterthwaite's t-test and the Wilcoxon-Mann-Whitney test compared to ANOVA performed on square root transformed data under both additional sample size scenarios.

For this experiment, an adequate analysis technique would provide less than or equal to 5% Type I error and 80% (70% for the final table) power under all scenarios tested. All analysis techniques tested performed adequately with respect to Type I error except for Satterthwaite's t-test which had greater than 5% Type I error when the sample sizes were different in the two groups. ANOVA performed on square root transformed data performed adequately with respect to power under all scenarios except perhaps one where only 75% power was achieved. The Wilcoxon-Mann-Whitney test generally had less power than the other methods and less than 80% power barring one exception where 83% power was achieved. ANOVA performed on the raw data and Satterthwaite's t-test generally performed adequately with respect to power but there were two (different) occasions when the estimated power for these tests was much lower than 80% (70% for Table 6.5). So why were these results obtained and what do they mean in practice for analysis of %SS scores?

As discussed in the Introduction section of this chapter, statistical theory states that as the size of a sample increases, the distribution of the mean of the sample tends to normality even when the distribution of the population is non-normal. This theoretical effect can be seen quite clearly with the results of the simulations. Providing the sample sizes of both groups are at least 20 and similar,

then the performance of ANOVA without transformation and Satterthwaite's t-test are at least as good as the other two methods. However once the sample size is below 20, this is no longer the case. The performance of ANOVA without transformation and Satterthwaite's t-test fall away quite rapidly so that when sample size is just 5 in each group, power is much lower than for the other two methods. However, the rate of Type 1 error is maintained at around 5%, or even lower, quite consistently, even when the sample size is below 20 for both groups.

As was also discussed in the Introduction section of this chapter, if samples sizes and variances are both substantially different for the two groups, then standard analytical techniques may not be appropriate. Again, the results of the simulations reflect what would have been expected given the theoretical considerations explained above. Quite large differences in the variances of the two groups are well tolerated in the power scenarios in Tables 6.1 and 6.3 where sample sizes are the same for both groups. This is not true for the power scenario in Table 6.5 but that is more likely due to the non-normality rather than the discrepancy in the variances. However, when the sample sizes of the two groups are different, the difference in the variances appears to have an effect. This is evident for the power scenarios in Table 6.2 but more so in Table 6.4 where the ratio of the sample sizes and variances of the two groups is largest. In these two tables, power is increased for ANOVA when the smaller group has the larger variance but reduced when the smaller group also has the smaller variance. In contrast power is decreased for Satterthwaite's t-test when the smaller group has the larger variance but increased when the smaller group also has the smaller variance. In addition Type 1 error is increased for Satterthwaite's t-test when the sample sizes of the two groups are different despite the fact that variances are similar in the two groups.

The results obtained here for %SS are consistent with the previous research of Stonehouse and Forrester (1998), Kingman and Zion (1994) and Boneau (1960) summarised in the introduction section. Previous research suggests that analysis of log transformed gamma distributed data results in less power compared to analysis of untransformed data or square root transformed data. This was also true for the current study in most cases. For gamma distributed data, it was previously found that the rate of Type I error was largely unaffected by the type of analysis performed. T-test performed on untransformed data, log transformed data, or square root transformed data all produced similar acceptable levels of Type I error. This was also true for the present study. However when sample sizes in the two groups were different Satterthwaite's t-test was associated with increased Type I error. Previous research also suggests data that is of different spread and sample size in the two groups is associated with an increased Type I error for standard analysis techniques. The present study found that groups of widely differing sample size and variance resulted in a loss of power when standard analysis techniques are applied. Using Satterthwaite's t-test specifically for groups with unequal variances did not appear to improve the situation presumably because the distributions of the two populations compared were non-normal. The findings associated with Satterthwaite's t-test are consistent with the results of the Stonehouse and Forrester (1998) study which found the similar Welch's approximate t-test to be sensitive to departures from normality.

The results of the simulations suggest that standard techniques like ANOVA can be used with confidence under most scenarios. Two occasions where this is not true is when sample size is small (less than 20) or when sample sizes and variances of the groups are markedly different. This latter problem can be avoided by designing balanced experiments with similar sized groups. If sample sizes are small

or if sample sizes and variances are markedly different, then a suitable transformation should be used prior to analysis or a non-parametric method such as the Wilcoxon-Mann-Whitney test used. Providing a suitable transformation can be found, this would normally be preferred over the Wilcoxon-Mann-Whitney test as the results of the simulations suggest power is lower for this test under most circumstances. Using Satterthwaite's t-test proved to be of no additional value compared to ANOVA when variances were different in the two groups and is not recommended.

Assuming the distribution of stuttering data is similar to that shown in this study, a square root transformation appears to be appropriate. However if the distribution is not similar, then it would be important to find a suitable transformation. Simulations performed for this study (but not shown) based on a log transformation generally gave inferior results with respect to power compared to those performed after square root transformation. However, if the distribution is, for example, more highly skewed due to the presence of outliers, then a log or some other transformation may be preferred. If sample sizes are small or variances and sample sizes are substantially different, and the distribution is not similar to that shown in this paper, then it would be important to investigate various transformations until a suitable transformation could be determined prior to analysis.

A suitable transformation is one that removes the skewness of the original distribution resulting in a symmetrical transformed distribution. A plot of the resulting data after transformation will determine how successful the transformation was at removing the skewness. Tests for normality and Q-Q normal plots will determine whether the resulting data are approximately normally distributed. In choosing from the many possible transformations the ladder of powers is a useful

organising concept (Velleman & Hoaglin, 1981). The ladder of powers is a list of power transformations that can be visualised as a continuous series. For positively skewed data the recommended procedure would be to move down the ladder of powers, that is, use square root, then cube root, logs, and so on until a suitable transformation is found.

It could be argued that a transformation should always be performed prior to using a standard analysis, as power should always be at least as high as performing standard analysis on untransformed data. Based on the results of the current research, this would seem reasonable. However there are two arguments against this recommendation. The first is that transforming data introduces a further complication into the analysis. Interpretation of results becomes problematic, as the original scale of the data has been lost due to the transformation. Estimates produced from the analysis may not have any intuitive meaning. In the case of log transformed data, the estimate obtained from the analysis can be back transformed to obtain an estimate with intuitive meaning however this cannot be done for square root transformed data. The second argument is that it is not always straightforward to find an appropriate transformation for skewed data. In fact it may not be possible.

Conclusions

Standard analytical techniques such as ANOVA, t-test and least squares regression are appropriate for analysis of %SS scores when sample size is at least 20 and similar in the two groups. If sample size is less than 20 in each group then a suitable transformation should be performed prior to analysis. If sample sizes and variances of the groups are markedly different then transformation should also be undertaken prior to analysis. The Wilcoxon-Mann-Whitney non-parametric method is not recommended, as power is reduced compared to standard techniques, under

most circumstances. Satterthwaite's t-test for unequal variances is also not recommended as it did not achieve improved results compared to ANOVA when variances were different in the two groups. With regard to the randomised controlled trial of the Lidcombe Program presented in the following chapter, a standard analysis (ANOVA, t-test or linear regression) is appropriate as there are 20 or more participants in each group and the sizes of the groups are reasonably similar.

Chapter 7: The Randomised Controlled Trial

Introduction

As was outlined in detail in the preceding chapters of this thesis, considerable research has already been undertaken on the Lidcombe Program for early stuttering. The logical next step in this program of research was to conduct an adequately powered randomised controlled trial (RCT) with long-term follow up. Because the Lidcombe Program was already well established in Australia, a RCT of this treatment with a 6-12 month follow up would have been problematic if conducted in Australia. Indeed, the majority of participants in a no-treatment control group of an earlier trial conducted in Sydney withdrew from the study and demanded treatment for their child (Onslow, Andrews, & Lincoln, 1994). Hence the RCT comparing the Lidcombe Program with standard treatment reported in this chapter was undertaken in New Zealand. New Zealand was an obvious choice of venue because it is geographically close to Australia, the Lidcombe Program was not being used in New Zealand at the time the trial commenced, and the two countries are culturally quite similar.

The retrospective study outlined in Chapter 4 provided valuable information for designing the RCT. It was found that the median number of weekly clinic visits required to complete Stage 1 of the Lidcombe Program was 11, with around 90% of children completing Stage 1 in 22 visits. As clinic visits are weekly, this would equate to a 6-month period providing the children missed fewer than 5 appointments. In addition, almost all children (> 99%) had completed Stage 1 of the program within 1 year. Therefore, a follow up period of between 6 and 12 months would be required for a prospective study so that the vast majority of children in the Lidcombe Program group would have completed Stage 1 and hence have reached a minimal level of stuttering within the follow up period. When the trial began the duration of follow up

was 12 months however it was apparent within the first few months of the trial that the control group parents were reluctant to wait 12 months before they could receive treatment for their child. Hence follow up was reduced to the compromise duration of 9 months.

Stuttering severity predicted treatment duration in the studies presented in Chapters 4 and 5, with children whose stuttering was more severe requiring longer treatment than children whose stuttering was less severe. There was also a non-significant result suggesting that time-since-onset of stuttering may be associated with treatment duration, with children who had been stuttering for more than 1 year seemingly requiring shorter treatment than children who had been stuttering for less than 1 year. These findings were confirmed with the meta-analysis reported in Chapter 5. This finding meant that children in a control group who were still stuttering after 9 months would not have their ultimate treatment response jeopardised by this.

The primary and secondary hypotheses for the trial were as follows:

Primary Hypothesis

After 9 months, stuttering severity in percent syllables stuttered (%SS) will be lower by at least an average of 1.0 for children randomised to the Lidcombe Program than for children randomised to the control group.

Secondary Hypothesis

The children randomised to the Lidcombe Program will take less time to attain minimal stuttering ($\%SS < 1.0$) compared to children randomised to the control group.

Methods

A trial protocol was written and ethical approval was obtained from the University of Canterbury in Christchurch, New Zealand and the University of Sydney in early 1999. The information sheet and informed consent used for the trial are presented in Appendix B and Appendix C. The following were the inclusion and exclusion criteria for the trial:

Inclusion Criteria

1. Age at enrolment between 3 and 6 years inclusive.
2. Stuttering identified using standard procedures (Onslow, 1996).
3. Children and parents proficient in English.

Exclusion criteria

1. Severity of stuttering less than 2.0 %SS.
2. Child treated in the previous 12 months for stuttering.
3. Onset of stuttering during 6 months prior to date of enrolment.

Children who were stuttering at less than 2.0 %SS were excluded because the definition of treatment success in the Lidcombe Program is 1.0 %SS or below.

Hence, children with stuttering severity below 2.0 %SS would have only required slight improvement to be deemed recovered. Children with onset of stuttering during the previous 6 months were excluded because clinicians typically would not intervene with these children. Children treated in the past 12 months were excluded so that any changes in stuttering could be attributed to the study treatment or natural recovery rather than a previous treatment. Three years was chosen as the lower age limit because of suggestions that children younger than 3 years may not have sufficient cognitive development for the treatment to be maximally effective. These

suggestions arose from data from the retrospective studies outlined in Chapters 4 and 5 suggesting that children with a short onset-to-treatment interval require a longer time for treatment than children with a long onset-to-treatment interval. As reported in Chapter 4, the treating clinicians in the first retrospective study suggested that this finding was intuitive from the perspective of the cognitive development of the children. Six years was used as the upper age limit because of data that suggest that the effectiveness of the Lidcombe Program starts to decrease at around this age (Lincoln, Onslow, Wilson, & Lewis, 1996).

Treatment Plans

Once eligibility had been assessed and a participant met all criteria and informed consent had been obtained, a participant was allocated to one of two treatment arms. The control treatment for the trial involved the child and at least one parent having an initial consultation with the clinician where the trial was explained and pre-treatment speech measures and demographic characteristics collected prior to treatment allocation. After the initial consultation, the parents were to provide recordings of their child's speech at the times outlined in the study protocol and document any treatment for stuttering given to the child during the trial. Standard treatment was the treatment the child would normally receive assuming that the Lidcombe Program was unavailable.

At the time the trial began, the Lidcombe Program was not considered standard practice for early stuttering in New Zealand and was not routinely presented to children who began to stutter. Preliminary investigations determined that there was no generally accepted treatment format for stuttering preschool children in New Zealand, with some clinicians using variants of Goldiamond's (1965) prolonged-speech procedure adapted for children, and family counselling based on Rustin's

procedures (Rustin, Botterill, & Kelman, 1996). However, at the time the trial started, the majority of preschool children in New Zealand were not treated. Hence the standard practice control group for the trial was, generally, no formal treatment, with some cases being treated as outlined above.

The Trial Management Committee considered it was ethical to have a control group that would probably not receive treatment for 9 months for the following reasons. First children who are experiencing early stuttering are quite likely to recover from stuttering without formal treatment. Second, as mentioned above in this chapter, the retrospective studies presented in Chapters 4 and 5 showed that a delay of at least 1 year after onset of stuttering does not jeopardise a child's response to Lidcombe Program treatment and in fact there is evidence to suggest successful treatment occurs more quickly in older preschool children. It was a policy in this trial that control children would be given access to Lidcombe Program treatment, should they require it, 9 months after randomisation.

The children allocated to the experimental Lidcombe Program group attended 1-hour weekly sessions in the clinic with a clinician and at least one parent. The parents were trained to administer the treatment described in the clinician's guide (Onslow, Packman, & Harrison, 2003) and in Chapter 1 of this thesis.

Randomisation

Eligible and consenting children were allocated to experimental or control group independently by way of a centralised telephone randomisation service provided by the Australian National Health and Medical Research Council (NHMRC) Clinical Trials Centre. The allocation process involved the treatment clinician contacting the NHMRC Clinical Trials Centre by telephone, fax or email. Treatment assignment was conducted independently of the study personnel and

reported back to the treatment clinician. The method of treatment allocation used was dynamically balanced randomisation (Signorini, Leung, Simes, Beller, & Gebski, 1993), where maximum treatment imbalances are pre-specified for each stratification factor, as well as for the trial overall. Treatment allocation is always random providing the maximum imbalances are not exceeded. If a maximum imbalance will be exceeded by a random allocation, then the allocation will be forced to the treatment group with the smaller treatment allocation to ensure that balance is maintained. For example, if the maximum imbalance is set at six for the trial overall and the current distribution is 30 participants in the Lidcombe Program arm and 24 participants in the control arm, then the next allocation will be control. This may occur randomly if the "coin toss" is in favour of control, or it may be forced if the coin toss is in favour of Lidcombe Program.

Stratification

Stratification factors for the trial were:

1. Institution
2. Severity of stuttering (less than 5.0 %SS versus 5.0 %SS or more)
3. Age at enrolment (younger than 5 years versus 5 years and older)
4. Gender
5. Family history (at least one first degree relative reported by parents to have recovered naturally from stuttering)

There are several research findings that support the choice of factors used for stratification. Pre-treatment severity of stuttering is associated with time to reach Stage 2 of the Lidcombe Program as outlined in Chapters 4 and 5. Further, age, family history of recovery, and gender are likely to be related to natural recovery

(Yairi & Ambrose, 1992a; Packman & Onslow, 1999). Hence it is possible that those factors are related also to treatment recovery. Although there was only one institution initially in the study, another was added later. Stratification by institution was necessary because different institutions may have different treatment practices that would be used with control children, as well as other differences such as clinician experience, which would be pertinent to the experimental children.

Clinician Training

As clinician experience could be an important factor with regard to treatment efficacy in this type of study, clinicians who delivered the Lidcombe Program treatment were required to meet the following criteria to be accepted into the trial:

1. Successful completion of a 2-week intensive training course in delivering the Lidcombe Program treatment.
2. Successful treatment of at least two children. Successful treatment meant the children had reached Stage 2.

Outcome Measures

The primary outcome measure for the trial was %SS (see Chapter 6). This was estimated for both treatment groups from audio tape recordings made on occasions outside the treatment clinic in everyday speaking situations. Participating parents provided recordings of their child's speech 2 weeks and 1 week pre-randomisation, and 1 week, 1 month, 2, 3, 4, 6, and 9 months post-randomisation. On each occasion, the children were audio-tape recorded by their parents in three beyond-clinic speaking situations: (1) speaking to a family member at home, (2) speaking to a non-family member at home, and (3) speaking to a non-family member away from home.

Every effort was made to ensure that the tape recordings were made at approximately the same time of day and that the time of recording was documented. Measures of %SS were gathered from the recordings by two experienced speech pathologists, who were blinded to the identity of the children and the group to which they belonged. One rated all the Auckland tapes and the other rated all the Christchurch tapes. The measures of %SS were made from samples of speech of at least 300 syllables. In order to establish reliability, 5% of the total recordings for each child were re-rated by the same speech pathologist and also by the second speech pathologist so that estimates of intra-judge and inter-judge agreement could be made.

Sample Size

The sample size required for the study was calculated using the following assumptions using STPLAN (Brown, 1993):

1. 2-sided test
2. One primary outcome of %SS
3. 80% power
4. Level of significance 0.05
5. Exponentially distributed data (see Jones, Onslow, Harrison, & Packman, 2000)

It was also assumed that %SS = 1.0 for the treated group 9 months after commencing treatment and that the minimum clinically worthwhile difference is 1.0 %SS (control group %SS = 2.0). This was based on a difference of 1.0 %SS being perceptually distinguishable. A sample size of 34 in each group was sufficient for this difference but drop-ins and drop-outs were expected. A drop-in rate of 10% and

a dropout rate of 10% were assumed, resulting in a sample size of 55 in each group (see Chapter 3 for more details on estimation of sample size).

Interim Analysis

An interim analysis was proposed at the mid-point of the trial when half of the proposed sample size (55 participants) had been followed for 9 months. A data monitoring committee (DMC) was setup to independently review the interim data. This committee consisted of a speech pathologist, a clinical biostatistician and a medical doctor who is also an epidemiologist. The DMC responsibilities included appropriate recommendation to the trial management committee that could include continuation of the trial without amendment, protocol modification, or early stopping of the trial if one of the treatment arms was found to be highly statistically significantly superior to the other treatment arm with respect to the primary outcome ($p < 0.003$). Safety of the Lidcombe Program treatment had already been established in previous studies as well as in routine clinical practice so it was highly unlikely that safety would be an important issue in the interim analysis.

Statistical Analysis

Analyses were by intention to treat using SAS, version for Windows 8.02 (SAS Institute, Cary, NC) and Stata/SE 8.0 for Windows (Stata Corporation, College Station, TX). The primary comparison of difference in average %SS at 9-months post-randomisation was performed using a 2-sample t-test (see Chapter 6). The frequency of stuttering analysed for each child was an average of their 9 month samples. Treatment effect was estimated with and without adjustment for baseline characteristics and within important subgroups using least squares regression. Tests

of heterogeneity were performed using interaction terms in multiple regression models. Last observation carried forward was used for two participants without follow up tapes at 9 months.

See Appendix D for a diagrammatic depiction of the design of the trial.

Conduct of the Trial

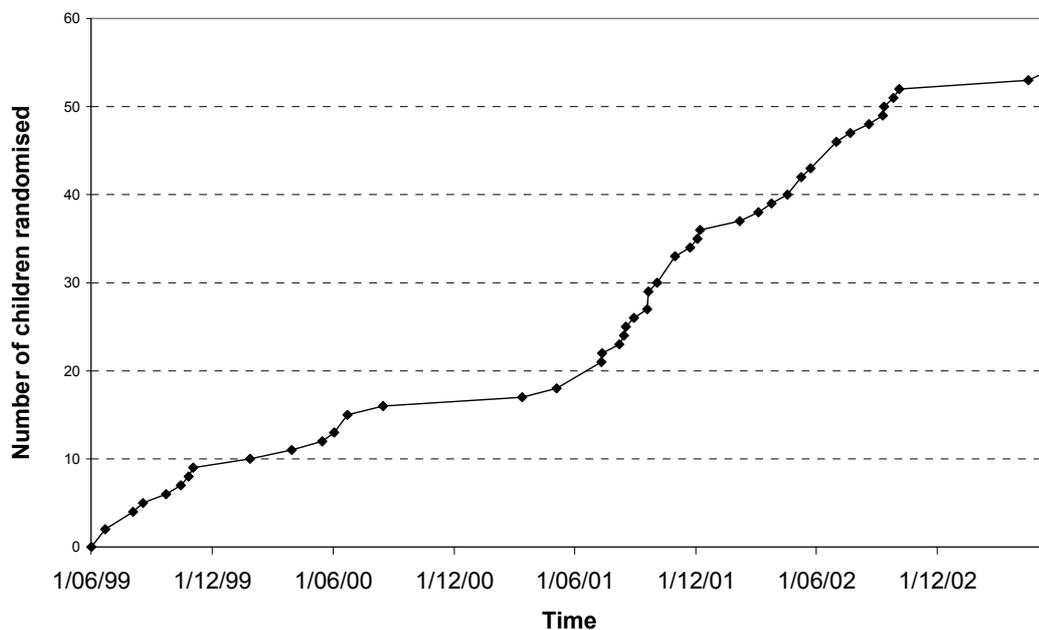
Recruitment

The first participant was randomised to the trial in June 1999 at the University of Canterbury in Christchurch, New Zealand. Christchurch has a population of approximately 350,000. While it was hoped that one child would be randomised to the study every week and that recruitment would be finished in two years, this was not the case. Recruitment was a lot slower than hoped with approximately one child being randomised per month for the first 15 months and then the recruitment stopped for no apparent reason. At that time, the Trial Management Committee decided it was critical that another centre was recruited to the study. In early 2001 the ‘Stuttering and Research Trust’ in Auckland, New Zealand received ethical approval and joined the collaboration, and in March of that year, the first participant in Auckland was randomised to the study. Auckland has a population of approximately 1.2 million.

Recruitment increased once Auckland joined the study and between March 2001 and October 2002, a further 36 participants were randomised. After a year of not randomising any new participants to the trial (from August 2000 to August 2001) Christchurch began recruiting again. However, in 2002 it became increasingly difficult to retain children in the Christchurch arm of the study. In fact, data were obtained for 24 of the first 25 children recruited to the Christchurch arm of the study

but only for one of the following six children recruited. The reasons for this are not clear but it was decided to stop recruitment to the trial in Christchurch in September 2002. Recruitment continued in Auckland but the rate was also lower than expected. From October 2002 to June 2004, only two additional participants were randomised to the study in Auckland (see Figure 7.1).

Figure 7.1: Recruitment over time



The Auckland site joined the trial on 13 March 2001 and the Christchurch site stopped recruiting to the trial on 9 September 2002.

Based on this lack of accrual, it was decided to stop the trial before the proposed sample size had been obtained. The reasons for this lack of accrual are unknown however a possible explanation was that as the trial progressed, parents in Christchurch and Auckland became increasingly aware of the Lidcombe Program and of the opportunity it presented for stuttering to be treated in the preschool years. Therefore they may have been unwilling to accept the possibility of not treating their child for 9 months.

Control Group

Another issue that became apparent early on in the trial in Christchurch was that parents of the children randomised to the control group were often unhappy that their child would receive no treatment for the next 9 months. This was despite the fact that they probably would not have received treatment had they not joined the study anyway. The speech pathologist assured the parents that delaying treatment would not harm their child and that treatment with Lidcombe Program would be offered should their child require treatment after 9 months. However, some parents were not convinced and withdrew their children from the control group of the study. Three children in Christchurch ended up receiving some Lidcombe Program treatment. One control participant in Auckland also received some Lidcombe Program treatment.

Speech Tapes

An additional problem encountered during the study was that a number of the parents did not provide all the tapes requested and this was especially true for the parents of the control children. As a result of the problems with retaining the control group and obtaining follow-up tapes, the design of the trial was modified in consultation with the Trial Management Committee. The first modification was that the control group would be monitored more often at 3-monthly intervals. This ensured that the speech pathologist saw the control children at regular intervals during the 9-month follow up period. These additional follow-up sessions allowed the speech pathologist to reassess the child and reassure the parents that there was nothing to worry about with respect to the speech of their child. However, if it was obvious that the speech of the child had begun to cause significant stress for the child

and/or parent, then beginning treatment earlier than planned was an option for the control children.

A second modification was the decision to concentrate on obtaining the critical tapes for performing the primary analysis: the pre-treatment and 9-month post-randomisation tapes. In a few cases, the 9-month tape was obtained by a speech pathology student visiting the family home and tape recording the child's speech. A second tier of tapes was also considered important: those to be collected at the 3 and 6 month post-randomisation points. This was when it was planned that the control children and parents would attend the treatment clinic for a follow-up interview. The post-randomisation tapes scheduled for 1 week, 1, 2 and 4 months were all considered to be the least important and least likely to be provided, and obtaining these recordings was de-prioritised.

A further issue related to the problem of obtaining tapes from the parents of the control children was identified in Auckland. The parent of one of the children in the study had promised to provide a pre-treatment tape of the child's speech, however that tape was not forthcoming and the child was withdrawn from the study soon afterwards. There was no indication that this was likely to occur at the time the parent and child agreed to be in the study. In response to this, it was decided not to randomise any children to the study until parents had provided their pre-treatment tape. Provision of the tape was regarded as an indication that the parent was likely to provide follow up tapes of the child's speech as requested. This was similar to a run-in phase of a RCT where participants who do not adhere to the trial protocol during the run-in phase are consequently not randomised to the study.

Results

Recruitment began in June 1999 and concluded in May 2003 with a total of 54 participants randomised, 31 at the University of Canterbury and 23 at the Stuttering Treatment and Research Trust in Auckland.

Participant Flow and Follow up

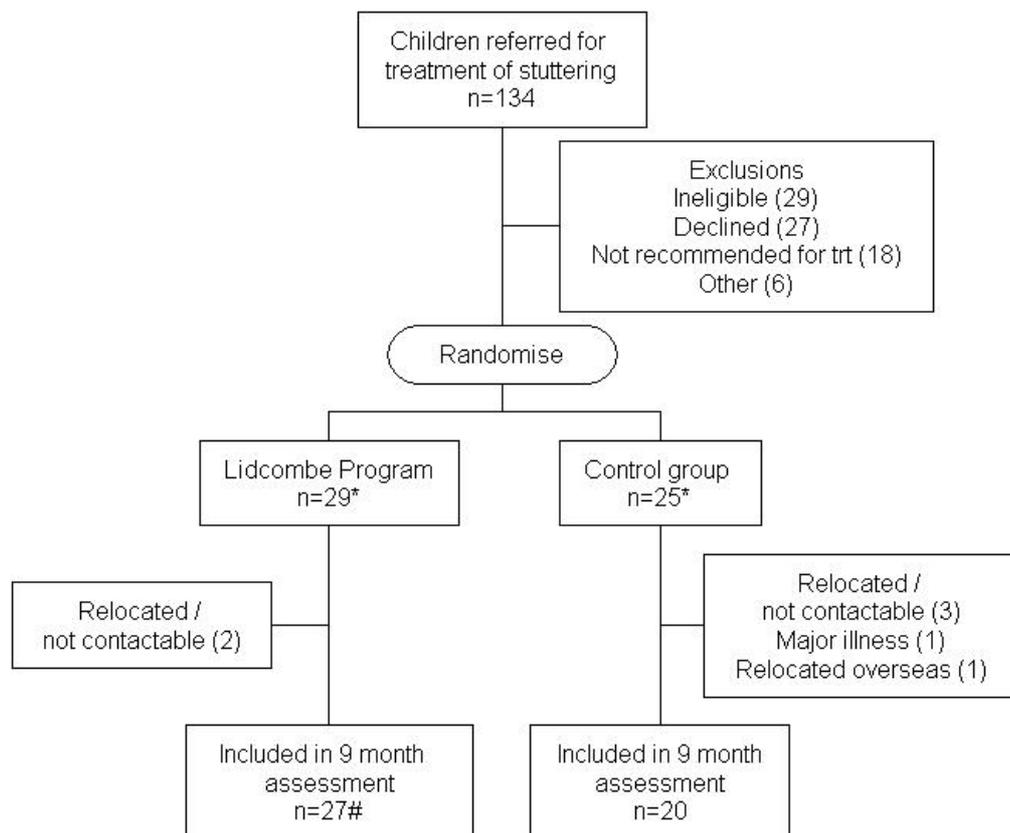
Seven (13%) of the 54 randomised participants did not complete the trial hence final data were not obtained, with all analyses being performed on 47 participants. The reasons for these children being lost to follow up appear to be typical for young children attending a speech clinic (Table 7.1; Figure 7.2).

A comparison of the pre-treatment characteristics of the participants who were lost to follow up with the others showed no statistically significant differences in severity, gender, treatment group assigned and family history of recovery from stuttering (Table 7.2). However there was a statistically significant difference in age ($p = 0.015$, 2-sample t-test). The mean age and range for the seven participants who were lost to follow up was 4.7 (3.9 - 5.8) years. In contrast, the participants followed up had a mean and range of 3.9 (2.8 - 6.0) years resulting in an average difference of 9 months (95% confidence interval: 2-16 months).

One participant had only 6-month follow up data available and another had only 7-month follow up data available for the primary analysis. Both of these participants were allocated to the Lidcombe Program group. Final tapes of speech were due at 9 months post-randomisation but were often recorded sometime after that. The median time from randomisation to final follow up in the Lidcombe Program group was 9 months (inter-quartile range: 9-12 months) compared to 11 months (inter-quartile range: 10-17 months) in the control group.

There were protocol violations with four control children receiving some Lidcombe Program treatment and one child randomised to the Lidcombe Program only receiving 3 weeks of treatment (Table 7.3; Figure 7.2). Three control children received treatment other than the Lidcombe Program. Two participants received Easy-does-it (Roseman & Johnson, 1998) and one received components of the Lidcombe Program on an ad-hoc basis. The components of the Lidcombe Program received were praise for smooth speech, acknowledgement for bumpy speech, and self monitoring.

Figure 7.2: Flowchart of participants



* One participant randomised to Lidcombe Program only received Lidcombe Program for 3 weeks and four participants randomised to control received some Lidcombe Program treatment

6-month follow up data were used for one participant and 7-month for another, as 9-month follow up data were not obtained

Table 7.1: Participants lost to follow up

Participant	Treatment group	Site	Reason
CT	Control	Auckland	Not contactable soon after randomisation
BD	Control	Christchurch	Family relocated, not contactable
EH	Control	Christchurch	Became sick and was admitted to hospital
LD	Lidcombe Program	Christchurch	Failed to attend after randomisation and did not return calls
GH	Control	Christchurch	Not contactable (phone cut off, mail returned)
II	Control	Christchurch	Family relocated overseas
CH	Lidcombe Program	Christchurch	Not contactable (phone cut off, mail returned) soon after beginning treatment

Table 7.2: Baseline characteristics of participants lost to follow up and participants remaining on study

	Lost to follow up	Remaining on study	Fisher's exact test
	N (%)	N (%)	<i>p</i>-value
%SS: 2-5	3 (43)	15 (32)	0.7
>5	4 (57)	32 (68)	
Female	2 (29)	10 (21)	0.6
Family history of recovery	3 (43)	21 (45)	0.9
Treatment group:			
Lidcombe Program	2 (29)	27 (57)	0.2
Control	5 (71)	20 (43)	
Age in years <i>Mean (SD)</i>	4.7 (0.7)	3.9 (0.6)	0.015 (2-sample t-test)

Table 7.3: Treatment received

	Lidcombe Program	Control	Total
	N (%)	N (%)	N (%)
Lidcombe Program	*27 (50)	*4 (7)	31 (57)
No treatment	2 (4)	18 (33)	21 (37)
Other treatment	0 (0)	3 (6)	3 (6)
Total	29 (54)	25 (46)	54 (100)

* Four controls received some Lidcombe Program treatment throughout the study period and one participant randomised to Lidcombe Program only received 3 weeks of treatment

Speech Samples

Of the 517 speech samples collected for analysis, 239 (46%) were recorded with a family member in the family home, 118 (23%) were recorded with a family member away from the family home, 108 (21%) involved the child speaking to a non-family member in almost all cases within the family home, and 11 (2%) were samples recorded by a speech pathologist within the clinic or in the family home as recordings by parents were not forthcoming. On 41 (8%) occasions it was unclear where the speech sample was recorded. For each child at each time-point there was an average of 2.3 (range 1-6) speech samples available for analysis with a mean duration of 433 syllables and standard deviation of 12. Intra-judge and inter-judge agreement was found to be satisfactory with intra-class correlations of $r = 0.99$ for both.

Analysis

Of the 54 participants recruited to the study, 29 were allocated to the Lidcombe Program group and 25 to the control group. A comparison of pre-randomisation characteristics by treatment group shows the groups to be similar (Table 7.4). Stuttering severity measured by %SS was similar in the two groups at pre-randomisation, however, a t-test comparison between the two groups at 9-month post-randomisation showed a highly statistically significant improvement in the Lidcombe Program group compared to the control group (Table 7.5 and 7.6). The mean %SS at 9 months post-randomisation was 1.5 (SD = 1.4) for the Lidcombe group compared to 3.9 (SD = 3.5) for the controls resulting in a treatment effect of 2.3%SS (95% confidence interval: 0.8-3.9). Stuttering severity in the two groups at pre-randomisation and 9-months post-randomisation by treatment site is presented in Table 7.7.

The time taken for children to attain minimal stuttering (less than 1.0 %SS) was unable to be tested appropriately due to most parents not collecting the speech recordings strictly at the times outlined in the trial protocol. However, the proportion of children with less than 1.0% SS at 9-months post-randomisation was higher in the Lidcombe Program group compared to the control group when adjusted for pre-treatment severity in a logistic regression model (Table 7.8).

Estimates of treatment effect (reduction in %SS) within subgroups are provided in Table 7.9. Whilst it is acknowledged that sample size is small and power is low for a test of heterogeneity, this analysis was performed to determine whether the treatment effect appeared to be consistent amongst important subgroups. Results suggest there is a larger effect of treatment for those children without a family history of recovery from stuttering compared to children with a history. A closer

inspection of the data shows this result is mainly due to the controls with a family history of recovery doing particularly well (mean %SS = 2.4, SD = 1.6) compared to the controls with no such history (mean %SS = 5.1, SD = 4.2). This could be explained by previous research that has shown children with a family history of recovery from stuttering are more likely to naturally recover than children without a family history of recovery (Yairi et al, 1996). Treatment effect appeared to be consistent within all other important subgroups with the possible exception of treatment site. However, the control group doing better in Auckland compared to Christchurch could explain this possible difference (Table 7.7).

Table 7.4: Baseline characteristics of all participants

	Lidcombe Program	Control	Total
	N (%)	N (%)	N (%)
Age: 3 to 4 years	17 (59)	12 (48)	29 (54)
4 to 5 years	9 (31)	12 (48)	21 (39)
5 to 6 years	3 (10)	1 (4)	4 (7)
%SS > 5.0	20 / 29 (69%)	16 / 25 (64%)	36 / 54 (67%)
Female	7 / 29 (24%)	5 / 25 (20%)	12 / 54 (22%)
Family history of recovery from stuttering	13 / 29 (45%)	11 / 25 (44%)	24 / 54 (44%)

Table 7.5: Stuttering severity (%SS) at pre-treatment and 9-months post-randomisation by treatment group

	Lidcombe Program	Control
	Mean (SD)	Mean (SD)
Pre-treatment	6.4 (4.3)	6.8 (4.9)
At 9-months post-randomisation [#]	1.5 (1.4)	3.9 (3.5)

[#] $p = 0.003$ (2-sample t-test)

Table 7.6: Comparison of stuttering severity (%SS) by treatment group at 9-months post-randomisation

	Estimated difference	<i>p</i>-value
	(95% CI)	
Unadjusted	2.3 (0.5-3.9)	0.003
Adjusted*	2.2 (0.5-3.6)	0.003

* Adjusted for baseline severity, treatment site, gender, age, family history of recovery

Table 7.7: Stuttering severity (%SS) at baseline and 9-months by treatment site

	Auckland		Christchurch	
	Lidcombe P.	Control	Lidcombe P.	Control
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Baseline	5.1 (4.9)	6.1 (3.8)	7.7 (3.2)	7.3 (5.8)
At 9-months	1.5 (1.6)	2.6 (2.2)	1.6 (1.2)	4.9 (4.2)

Table 7.8: Number of participants achieving %SS of less than 1.0 at 9-months post-randomisation

	Lidcombe Program	Control
	N (%)	N (%)
Less than 1.0 %SS	14 (52)	3 (15)
1.0 %SS or more	13 (48)	17 (85)

Odds ratio (adjusted for baseline severity): 0.13 (95% CI: 0.03-0.63), $p = 0.011$

Table 7.9: Estimates of treatment effect within subgroups

Subgroup	Number of participants	Difference in %SS	95% confidence interval	<i>p</i>-value*
Treatment site:				
Auckland	22	1.1	-0.6 – 2.8	0.15
Christchurch	25	3.3	0.9 – 5.8	
Gender:				
Boys	37	2.4	0.6 – 4.2	0.8
Girls	10	2.0	-1.6 – 5.5	
Age:				
Less than 4 years	28	2.4	0.1 – 4.7	0.9
Four years or more	19	2.3	0.4 – 4.2	
Family history of recovery:				
No	26	3.8	1.5 – 6.2	0.027
Yes	21	0.5	-1.6 – 2.7	
Baseline severity				
Less than 5%SS	19	2.1	0.2 – 4.0	0.6
5%SS or more	28	2.7	0.5 – 4.9	

* Test for heterogeneity

Discussion

The present study is the first randomised controlled trial of a treatment for early stuttering, demonstrating that the Lidcombe Program is more effective than natural recovery over a 9 month period. This finding is underscored by the apparent larger treatment effect for children with no family history of natural recovery than children with such a family history because this suggests that natural recovery did not significantly aid the Lidcombe Program treatment effect. The estimated treatment effect of 2.3 %SS is more than double the minimum clinically worthwhile difference as specified in the trial protocol. In the present study only 15% of children in the control arm attained a minimal level of stuttering as defined in the trial protocol in a median follow up of 1 year.

Study Limitations

The present study has a number of limitations that should be acknowledged, but these are not sufficient to alter the main conclusion. The achieved sample size was only half that proposed, and seven of the 54 randomised participants did not complete the trial. These participants were older than the participants who completed the trial by an average of 9 months. However this difference in age is unlikely to have a significant effect on the results of the trial as the number of participants lost to follow up is small and there is no evidence to suggest that age is correlated with %SS at 9 months for the participants who completed the trial.

Four control participants received some Lidcombe Program treatment and one participant randomised to Lidcombe Program only received 3 weeks of treatment. These protocol violations would tend to dilute the treatment effect. In addition, for two participants randomised to the Lidcombe Program, data were only available at 6 and 7 months respectively. This, too, would tend to dilute the treatment

effect, as it would be expected that children receiving treatment would tend to have reduced levels of stuttering as time goes on. Time from randomisation to final follow up was often longer than 9 months and slightly longer for the control group compared to the Lidcombe Program group. Again, this would tend to dilute the treatment effect as longer follow up would allow the control group more time to naturally recover whereas shorter follow up in the Lidcombe Program group would allow less time for the treatment to have an effect.

Another limitation is that the duration of follow up for this study was 9 months. Ideally the follow-up period should have been longer than 9 months so that the children in the treatment group could have completed maintenance stage of the program (Stage 2) and the control group children could have had more time to recover naturally. However, it is very difficult to retain a control group for longer than 9 months, as parents are often reluctant to allow their child to go untreated for a long period especially if their stuttering is severe or appears to be worsening. Although the data suggest that a number of children in the Lidcombe Program arm had not attained a minimal level of stuttering at 9 months post-randomisation, previous research indicates they will attain zero or near-zero levels of stuttering and maintain these very low levels in the long-term (Lincoln & Onslow, 1997).

PART FOUR: SUMMARY

Chapter 8: Summary and Conclusions

Summary of the Thesis

The topic of this thesis is an evaluation of the effectiveness of the Lidcombe Program of Early Stuttering Intervention, culminating in the presentation of the first randomised controlled trial of a treatment for early stuttering. Chapter 1 began with background information on stuttering with particular emphasis on early stuttering and natural recovery from early stuttering. Treatments in current use for early stuttering were summarised, followed by an outline of the Lidcombe Program. It was concluded that the Lidcombe Program is unique because it is supported by a systematic program of research. In contrast, other treatment programs for early stuttering appear to be supported by very little objective data.

A systematic review of randomised studies of treatment for stuttering was presented in Chapter 2. In all there were 27 studies included in the review. The review showed that there is only one randomised study of treatment for early stuttering. In this study the Lidcombe Program reduced stuttering rates beyond that of natural recovery over a 3-month period in 23 subjects.

A review of statistical power of studies undertaken in the area of stuttering research in the late 1990s was presented in Chapter 3. It concluded that sample sizes are generally small hence studies are of insufficient power to detect even large effects in most cases. None of the 26 studies included in the review stated that they had performed an a priori sample size calculation.

As shown in Chapters 4 and 5, when treated with the Lidcombe Program, children require a median of 11 clinician hours to attain the minimal stuttering criteria, with 90% of children attaining this in 22 hours. As children are seen weekly in the clinic for 1-hour sessions, the overwhelming majority of children will complete Stage 1 of the Lidcombe Program within 6 months and enter Stage 2

(maintenance) which normally lasts for at least 1 year. More severe stuttering prior to treatment is associated with a greater number of clinician hours to attain the minimal stuttering criterion while a delay in treatment for at least 1 year after onset of stuttering is associated with fewer clinician hours. This latter finding is especially important when deciding when to treat a child for early stuttering. Delays of at least 1 year post onset do not appear to affect responsiveness to treatment with the Lidcombe Program.

A simulation study presented in Chapter 6 showed that standard analytical techniques such as ANOVA, t-test and least squares regression are appropriate for analysis of %SS scores when sample size is at least 20 and similar in the two groups. If sample size is less than 20 in each group then a suitable transformation should be performed prior to analysis. If sample sizes and variances of the groups are markedly different, then transformation should also be undertaken prior to analysis. The results of the simulation study guided the analysis of the RCT and have wider implications for research that measures stuttering with %SS.

The first randomised study of the Lidcombe Program was Harris et al. (2002), and a second randomised study of the Lidcombe Program was presented in Chapter 7 of this thesis. This randomised controlled trial showed that the Lidcombe Program reduced stuttering rates compared to a no-treatment control group over a follow up period of 9 months. At 9 months the control group had reduced their frequency of stuttering by an average of 43% presumably from a combination of natural recovery and the ad-hoc treatment given to some of the participants. In contrast the Lidcombe Program group had reduced their stuttering by 77% resulting in a mean frequency of 1.5 %SS. The majority of the children in the Lidcombe Program group were still in Stage 2 of the program at the 9 month follow up point.

An a priori sample size calculation resulted in a proposed sample size of 110 subjects for the RCT presented in Chapter 7. Even with 4 years of recruitment from two treatment sites, only 54 subjects were entered into the study. This study suggested a reason for the findings of Chapter 3 that sample size is generally low in stuttering research: it is in fact very difficult to recruit even moderate numbers of stuttering participants in research.

The randomised controlled trial of the Lidcombe Program as well as the previous Harris et al. randomised study provides strong evidence that this intervention for early stuttering is more effective than natural recovery. When these two studies are combined in a meta-analysis of individual participant data, the fixed effect of treatment in a linear regression model after adjustment for treatment site and baseline level of stuttering is estimated to be 2.5 %SS (95% confidence interval: 1.2 to 3.8; $p < 0.001$). As well as being statistically significant, this treatment effect is also clinically significant as the 95% confidence interval rules out a difference of less than 1.0 %SS, the minimum clinically worthwhile difference as specified in the protocol for the randomised controlled trial.

Previous research has shown that a family history of recovery from stuttering is predictive of natural recovery from early stuttering (Yairi, Ambrose, Paden & Throneburg, 1996). Subgroup analysis performed in the randomised controlled trial of the Lidcombe Program confirmed this. There was a significantly larger treatment effect in those children without a family history of recovery from stuttering. A closer investigation of the data indicated this difference was mainly due to the control children with a family history of recovery from stuttering having a much lower level of stuttering at 9-months post-randomisation compared to the control children who

did not have such a history. This subgroup analysis result adds weight to the conclusion that the Lidcombe Program is more effective than natural recovery.

Conclusions

Although a number of children presenting at a clinic with early stuttering may recover without treatment, identifying these children in advance is not possible. It is not acceptable to wait for an extended period to see if natural recovery occurs because it appears that the Lidcombe Program is less effective once children move into the school years (Lincoln, Onslow, Wilson & Lewis, 1996). In addition, delaying treatment until the school years is not a viable option because of the negative social and cognitive consequences of the disorder at this age (see Chapter 1). If stuttering persists into the school years, then a child is exposed to unacceptable risk of experiencing the disabling effects of chronic and intractable stuttering through the lifespan. It is arguable that, with the few limitations of the clinical trial reported in this thesis, the first “gold standard” randomised controlled trial evidence of the effectiveness of an early stuttering intervention has been attained. Therefore, the fundamental conclusion from the research in this thesis is that, when treatment for stuttering is required during the preschool years, the Lidcombe Program is the only treatment program that has been shown to be more effective than natural recovery.

A Replication

Given that ultimate evidence for an intervention for a disorder is a meta-analysis of randomised controlled trials (National Health and Medical Research Council, 1999) the most desirable line of future research subsequent to that reported here should be a replication of the present findings. Ideally, such replications should be independent of the group that conducted the present trial. It is encouraging that

such a replication is under way. During the course of the conduct of the studies reported in this paper, the author became aware of another clinical trial of the Lidcombe Program being conducted as part of a doctoral program of research in Germany by Ms Christina Latterman. This trial is being conducted under the supervision of Dr. Katrin Neumann at the Klinik für Phoniatrie und Pädaudiologie, Johann-Wolfgang Goethe Universität, in Frankfurt, and Dr. Harald Eurler at the Universität Kassel, in Kassel. At the time of writing, 46 children had been randomised to this study, experimental treatments had been concluded, and data analysis had begun (Personal Communication from C. Latterman to the author's supervisor, M. Onslow, 28 February 2005).

This thesis opened with mention of the history of the diagnosogenic theory and the historical Prins and Ingham (1983) speculation about the possibility of effective, direct, early stuttering intervention. Perhaps the present trial will be the first of many that provides a compelling data set that makes such speculation a reality.

References

Adams, M. (1984). The differential assessment and direct treatment of stuttering. In J. Costello (Ed.) *Speech disorders in children*. San Diego, CA: College-Hill Press.

Althaus, M., Vink, H., Minderaa, R., Goorhuis-Brouwer, S., & Oosterhoff, M. (1995). Lack of effect of clonidine on stuttering in children. *American Journal of Psychiatry*, *152*, 1087-1089.

Altman D., Schulz K., Moher D., Egger M., Davidoff F., Elbourne D., Gøtzsche P., & Lang T., for the CONSORT group. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine*; *134*, 663-694.

Ambrose, N., Cox, N., & Yairi, E. (1997). The genetic basis of persistence and recovery in stuttering. *Journal of Speech, Language, and Hearing Research*, *40*, 567-580.

Andrews, G. (1984). Epidemiology of stuttering. In R. Curlee & W. Perkins (Eds.), *Nature and treatment of stuttering: New directions*. San Diego, CA: College-Hill Press.

Andrews, G., Craig, A., Feyer, A., Hoddinott, S., Howie, P., & Nielson, M. (1983). Stuttering: A review of research findings and theories circa 1982. *Journal of Speech and Hearing Disorders*, *48*, 226-46.

Andrews, G., & Dozsa, M. (1977). Haloperidol and the treatment of stuttering. *Journal of Fluency Disorders*, *2*; 214-24.

Andrews, G., Guitar, B., & Howie, P. (1980). Meta-analysis of the effects of stuttering treatment. *Journal of Speech and Hearing Disorders*, *45*, 287-307.

Andrews, G., & Harris, M. (1964). The Syndrome of Stuttering. *Clinics in Developmental Medicine, No. 17*. London: Spastics Society Medical Education and Information Unit in association with Wm. Heinemann Medical Books.

Australian Stuttering Research Centre. (2001). *Lidcombe Program manual*. Retrieved from the Australian Stuttering Research Centre site on World Wide Web: [http:// www.fhs.usyd.edu.au/ASRC](http://www.fhs.usyd.edu.au/ASRC)

Azrin, N., & Nunn, R. (1974). A rapid method of eliminating stuttering by a regulated breathing approach. *Behaviour Research & Therapy, 12*, 279-286.

Azrin N., Nunn, R., & Frantz, S. (1979). Comparison of regulated-breathing verses abbreviated desensitization on reported stuttering episodes. *Journal of Speech & Hearing Disorders, 44*, 331-339.

Bernstein Ratner, N. (1997). Leaving Las Vegas: Clinical odds and individual outcomes. *American Journal of Speech-Language Pathology, 6*, 29-33.

Bishop, J., Gebiski, V., Simes, J., Dewar, J., Snyder, R., Byrne, M., & Forbes, J. (1999). Sequential randomized studies in metastatic breast cancer: CMFP vs doxorubicin/cyclophosphamide, mitoxantrone or paclitaxel. *Oncology, 13*, 1-6.

Block, S., Onslow, M., Packman, A., Gray, B., & Dacakis, G. (2005). Treatment of chronic stuttering: outcomes from a student training. *International Journal of Language and Communication Disorders*. In press.

Blood, I., Wertz, H., Blood, G., Bennett, S., & Simpson, K. (1997). The effects of life stressors and daily stressors on stuttering. *Journal of Speech, Language and Hearing Research 40*, 134-143.

Bloodstein, O. (1995). *A handbook on stuttering*. San Diego, California: Singular Publishing Group.

Bluemel, C. (1932). Primary and secondary stammering. *Quarterly Journal of Speech, 18*, 187–200.

Boneau, C. (1960). The effects of violations of assumptions underlying the t-test. *Psychological Bulletin, 57*, 49-64.

Bonelli, P., Dixon, M., Bernstein Ratner, N., & Onslow, M. (2000). Child and parent speech and language and the Lidcombe Program of Early Stuttering Intervention. *Clinical Linguistics and Phonetics, 14*, 427-446.

Bosshardt, H. & Fransen, H. (1996). Online sentence processing in adults who stutter and adults who do not stutter. *Journal of Speech, Language and Hearing Research, 39*, 785-797.

Brady, J. (1991). The pharmacology of stuttering: A critical review. *American Journal of Psychiatry, 148*, 1309-1316.

Brady, J., Price, T., McAllister, T., & Dietrich, K. (1989). A trial of verapamil in the treatment of stuttering in adults. *Biological Psychiatry, 25*, 630-633.

Brown, B. (1993). *STPLAN (Version 4.0): Calculations for Sample Sizes and Related Problems*. University of Texas, Dept. of Biomathematics.

Brumfitt, S. & Peake, M. (1988). A double-blind study of verapamil in the treatment of stuttering. *British Journal of Disorders of Communication, 23*, 35-40.

Büchel, C., & Sommer, M. (2004). What causes stuttering? *PloS Biology [serial on-line]; 2*(2). Retrieved Jan 27, 2005, from <http://www.plosbiology.org>.

Burns, D., Brady, J., & Kuruvilla, K. (1978). The acute effect of haloperidol and apomorphine on the severity of stuttering. *Biological Psychiatry, 13*, 255-264.

Clopper, C., & Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.

Conture, E. (1990). *Stuttering (2nd Ed.)*. Englewood Cliffs, N.J.: Prentice-Hall.

Conture, E., & Caruso, A. (1987). Assessment and diagnosis of childhood disfluency. In L. Rustin, H. Purser, & D. Rowley (Eds.), *Progress in the treatment of fluency disorders*. London, UK: Taylor and Francis.

Conture, E., & Melnick, K. (1999). Parent-Child Group Approach to Stuttering in Preschool Children. In M. Onslow & A. Packman (Eds.) *The Handbook of early stuttering intervention*. San Diego, CA: Singular Publishing Group.

Cordes, A., & Ingham, R. (1994). The reliability of observational data .2. Issues in the identification and measurement of stuttering events. *Journal of Speech and Hearing Research*, 37, 279-294.

Costa, A., & Kroll, R. (1995). Sertraline in stuttering. *Journal of Clinical Psychopharmacology*, 15, 443-444 (letter).

Costello, J. (1983). Current behavioural treatments for children. In D. Prins & R. Ingham (Eds.), *Treatment of stuttering in early childhood: Methods and issues*. San Diego, CA: College-Hill.

Craig, A., & Calver, P. (1991). Following up on treated stutterers: Studies of perceptions of fluency and job status. *Journal of Speech & Hearing Research, 34*, 279-284.

Craig, A., Hancock, K., Chang, E., McCready, C., Shepley, A., McCaul, A., Costello, D., Harding, S., Kehren, R., Masel, C., & Reilly, K. (1996). A controlled clinical trial for stuttering in persons aged 9 to 14 years. *Journal of Speech, Language and Hearing, 39*, 808-826.

Curlee, R. (1984). Stuttering disorders: An overview. In J. Costello (Ed.) *Speech disorders in children*. San Diego, CA: College-Hill Press.

Curlee, R. (1993). Identification and management of beginning stuttering. In R. F. Curlee (Ed.). *Stuttering and related disorders of fluency*. New York: Thieme Medical Publishers.

Curlee, R., & Yairi, E. (1997). Early intervention with early childhood stuttering: A critical examination of the data. *American Journal of Speech-Language Pathology, 6*, 8-18.

D'Agostino, R. and Stephens, M., (Eds). (1986). *Goodness-of-fit Techniques*. New York: Dekker.

Davidow, J., Crowe, B., & Bothe, A. (2004). GILCU and ELU treatment programs for stuttering: assessing the implications of strong but limited evidence. In A. Bothe (Ed.) *Evidence based treatment of stuttering: Empirical bases and clinical applications*. Mahwah, New Jersey: Laurence Erlbaum Associates, Inc.

Davis, S., Howell, P., & Cooke, F. (2002). Sociodynamic relationships between children who stutter and their non-stuttering classmates. *Journal of Child Psychology and Psychiatry, 43*, 939-947.

DeNil, L., & Brutton, G. (1991). Speech-associated attitudes of stuttering and nonstuttering children. *Journal of Speech and Hearing Research, 34*, 60-66.

Dupont, W., & Plummer, W. (1990). Power and sample size calculations: A review and computer program. *Controlled Clinical Trials, 11*, 116-128.

Edgren, B., Leanderson, R., & Levi, L. (1970). A research programme on stuttering and stress. *Acta Oto-Laryngologica - Supplement, 263*, 113-118.

Ellis, A. (1961). *Reason and emotion in psychotherapy*. New York: Lyle Stuart.

Eversham, M., & Fransella, F. (1985). Stuttering relapse: the effect of a combined speech and psychological reconstruction programme. *British Journal of Disorders of Communication, 20*, 237-248.

Faul, F., & Erdfelder, E. (1992). *GPOWER: A priori, post-hoc, and compromise power analyses for MS-DOS* [Computer program]. Bonn, FRG: Bonn University, Dept. of Psychology.

Felsenfeld, S. (1997) Epidemiology and genetics of stuttering. In R. Curlee and G. Siegel (Eds.) *Nature and treatment of stuttering: New directions (2nd ed.)* Boston, MA: Allyn & Bacon.

Finn, P. (1996). Establishing the validity of recovery from stuttering without formal treatment. *Journal of Speech, Language and Hearing Research, 39*, 1171-1181.

Finn, P. (1997). Adults recovered from stuttering without formal treatment: Perceptual assessment of speech normalcy. *Journal of Speech, Hearing, and Language Research, 40*, 821-831.

Finn, P., Ingham, R., Ambrose, N., & Yairi, E. (1997). Children recovered from stuttering without formal treatment: Perceptual assessment of speech normalcy. *Journal of Speech, Hearing, and Language Research, 40*, 867-876.

Fishman, G. (1978). *Principles of Discrete Event Simulation*. New York: John Wiley & Sons, Inc.

Franklin, Riley, Maguire, et al. (2002). *Olanzapine in the treatment of adult developmental stuttering: A double-blind, placebo-controlled trial*. Poster presentation at the 42nd Annual National Institute of Mental Health New Clinical Drug Evaluation Unit meeting.

Goldiamond, I. (1965). Stuttering and fluency as manipulatable operant response classes. In L. Krasner & L. Ullman (Eds.), *Research in behavior modification*. New York: Holt, Rinehart & Winston.

Gordon, C., Cotelingam, G., Stager, S., Ludlow, C., Hamburger, S., & Rapoport, J. (1995). A double-blind comparison of clomipramine and desipramine in the treatment of developmental stuttering. *Journal of Clinical Psychiatry, 56*, 238-242.

Gottwald, S., & Starkweather, C. (1999). Stuttering prevention and early intervention: A multiprocess approach. In M. Onslow and A. Packman (Eds.) *The Handbook of early stuttering intervention*. San Diego, CA: Singular Publishing Group.

Gregory, H. (1999). Developmental Intervention: Differential Strategies. In M. Onslow and A. Packman (Eds.) *The Handbook of early stuttering intervention*. San Diego, CA: Singular Publishing Group.

Gregory, H., & Hill, D. (1998). Differential evaluation - differential therapy for stuttering. In R. Curlee (Ed.) *Stuttering and related disorders of fluency*. New York: Thieme Medical Publishers.

Guitar, B. (1998). *Stuttering: An Integrated Approach to Its Nature and Treatment (2nd. Ed)*. Baltimore: Williams and Wilkins.

Harris, E., & Albert, A. (1991). *Survivorship Analysis for Clinical Studies*, New York: Marcel Dekker.

Harris, V., Onslow, M., Packman, A., Harrison, E., & Menzies, R. (2002). An experimental investigation of the impact of the Lidcombe Program on early stuttering. *Journal of Fluency Disorders*, 27, 203-214.

Harrison, E. (2002). *The Lidcombe Program of early stuttering intervention: Current status and future development needs*. Doctoral thesis.

Hayhow R., Cray A., & Enderby P. (2002). Stammering and therapy: views of people who stammer. *Journal of Fluency Disorders*, 27, 1-17.

Hays, P. (1987). Bethanecol chloride in treatment of stuttering. *Lancet*, 329, 271 (letter).

Hedges, O., Umar, F., Mellon, C., Herrick, L., Hanson, M., & Wahl, M. (1995). Direct comparison of family history method and the family study method using a large stuttering pedigree. *Journal of Fluency Disorders*, 20, 25-34.

Hoenig, J. & Heisey, D. (2001). The abuse of power: The persuasive fallacy of power calculations for data analysis. *American Statistical Association*, 55, 19-24.

Ingham, J. (1994). Therapy of the stuttering child. In G. Blanken, J. Dittman, H. Grimm, J. Marshall, & C. Wallesch (Eds.), *Linguistic Disorders and pathologies*. Berlin, Germany: Walter De Gruyter.

Ingham, J., & Riley, G (1998). Guidelines for documentation of treatment efficacy for young children who stutter. *Journal of Speech Language & Hearing Research, 41*, 753-770.

Ingham, R. (1980). Modification of maintenance and generalization during stuttering treatment. *Journal of Speech and Hearing Research, 23*, 732-745.

Ingham, R. (1981). Evaluation and maintenance in stuttering treatment: A search for ecstasy with nothing but agony. In E. Boberg (Ed.), *Maintenance of fluency*. New York: Elsevier.

Ingham, R. (1984). *Stuttering and behavior therapy : current status and experimental foundations*. San Diego, CA: College-Hill Press.

Ingham, R., & Cordes, A. (1997). Identifying the authoritative judgements of stuttering: comparisons of self-judgements and observer judgements. *Journal of Speech, Language and Hearing Research, 40*, 581-594.

Ingham, R., & Cordes, A. (1998). Treatment decisions for young children who stutter: Further concerns and complexities. *American Journal of Speech-Language Pathology, 7*, 10-19.

Ingham, R., Fox, P., Ingham, J., Zamarripa, F., Martin, C., Jerabek, P., & Cotton, J. (1996). Functional-lesion investigation of developmental stuttering with positron emission tomography. *Journal of Speech, Language and Hearing Research, 39*, 1208-1227.

- Ingham, R. & Onslow, M., (1984). *Stuttering treatment evaluation manual*.
Cumberland College of Health Sciences, Sydney.
- Jambunathan, M. (1954). Some Properties of Beta and Gamma Distributions.
Annals of Mathematical Statistics, 25, 401-405.
- James J. (1983). Parameters of the influence of self-initiated time-out from
speaking on stuttering. *Journal of Communication Disorders*, 16, 123-32.
- Janssen, P., Kloth, S., Kraaimaat, F., & Brutten, G. (1996). Genetic factors in
stuttering: a replication of Ambrose, Yairi, and Coxís (1993) study with adult
probands. *Journal of Fluency Disorders*, 21, 105-108.
- Johnson, W., et al. (1942). A study of the onset and development of
stuttering. *Journal of Speech Disorders*, 7, 251–257.
- Johnson, W., et al. (1959). *The onset of stuttering*. Minneapolis: University
Minneapolis Press.
- Jones, M., Gebski, V., Onslow, M., & Packman, A. (2000). *Rationale,
Methods, and Design of a Randomised Controlled Trial of the Lidcombe Program of
Early Stuttering Intervention*. Poster session presented at the Society for Clinical
Trials meeting, Toronto, Canada.
- Jones, M., Gebski, V., Onslow, M., & Packman, A. (2001). Design of
randomized controlled trials: Principles and methods applied to a treatment for early
stuttering. *Journal of Fluency Disorders*, 26, 247-267.
- Jones, M., Onslow, M., Harrison, E., & Packman, A. (2000). Treating
stuttering in young children: Predicting treatment time in the Lidcombe Program.
Journal of Speech Language and Hearing Research, 43, 1440-1450.

Kampman, K., & Brady, J. (1993). Bethanechol in the treatment of stuttering. *Journal of Clinical Psychopharmacology*, *13*, 284-285.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistics Association*, *53*, 457-481.

Kingman, A., & Zion, G. (1994). Some power considerations when deciding to use transformations. *Statistics in Medicine*, *13*, 769-783.

Kingston, M., Huber, A., Onslow, M., Jones, M., & Packman, A. (2003). Predicting treatment time with the Lidcombe Program: Replication and meta-analysis. *International Journal of Language and Communication Disorders*, *38*, 165-77.

Kirby, A., Gebski, V., & Keech, A.C. (2002). Determining the sample size in a clinical trial, *Medical Journal of Australia*, *177*, 256-257.

Kloth, S., Janssen, P., Kraaimaat, F., and Brtten, G. (1995). Communicative behavior of mothers of stuttering and non-stuttering high-risk children prior the onset of stuttering. *Journal of Fluency Disorders*, *20*, 365-377.

Kraemer, H., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

Kunz, R., & Oxman, A. (1998). The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal*, *317*, 1185-1190.

Ladouceur, R., Boudreau, L., & Theberge, S. (1981). Awareness training and regulated-breathing method in modification of stuttering. *Perceptual & Motor Skills*, *53*, 187-194.

Ladouceur, R., & Martineau, G. (1982). Evaluation of regulated-breathing method with and without parental assistance in the treatment of child stutterers. *Journal of Behaviour Therapy and Experimental Psychiatry*, 13, 301-306.

Lang, T., & Secic, M. (1997). *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers (Medical Writing and Communication)*. Philadelphia: American College of Physicians.

Langevin, M., Bortnick, K., Hammer, T., & Wiebe, E. (1998). Teasing/bullying experienced by children who stutter: Toward development of a questionnaire. *Contemporary Issues in Communication Science and Disorders*, 25, 12-24.

Langevin, M., & Hagler, P. (2003). Development of a scale to measure peer attitudes toward children who stutter. In A. Bothe (Ed.), *Evidence-based treatment of stuttering: Empirical issues and clinical implications*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Lemeshow, S., Hosmer Jr, D. W., Klar, J., & Lwanga, S. K. (1990). *Adequacy of sample size in health studies*. Chichester, UK: John Wiley & Sons.

Lincoln, M., & Onslow, M. (1997). Long-term outcome of an early intervention for stuttering. *American Journal of Speech-Language Pathology*, 6, 51-58.

Lincoln, M., Onslow, M., & Reed, V. (1997). Social validity of an early intervention for stuttering: The Lidcombe Program. *American Journal of Speech-Language Pathology*, 6, 77-84.

Lincoln, M., Onslow, M., Wilson, L., & Lewis, C. (1996). A clinical trial of an operant treatment for school-age stuttering children. *American Journal of Speech-Language Pathology, 5*, 73-85.

Lincoln, M., & Packman, A. (2003). Measuring stuttering. In M. Onslow, A. Packman, & E. Harrison (Eds.) *The Lidcombe Program of early stuttering intervention: A clinician's guide*. Austin, TX: Pro-Ed.

Lipsey, M. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Logan, K. & Conture, E. (1997). Selected temporal, grammatical, and phonological characteristics of conversational utterances produced by children who stutter. *Journal of Speech, Language and Hearing Research, 40*, 107-120.

Machin, D., Campbell, M., Fayers, P., & Pinol A. (1997). *Sample size tables for clinical studies*. Oxford, UK: Blackwell Science.

Mackay, L., Finn, P., & Ingham, R. (1997). Effect of speech dialect on speech naturalness ratings: a systematic replication of Martin, Haroldson, and Triden (1984). *Journal of Speech, Language and Hearing Research, 40*, 349-360.

Maguire, G., Riley, G., Franklin, D., & Gottschalk, L. (2000). Risperidone for the treatment of stuttering. *Journal of Clinical Psychopharmacology, 20*, 479-482.

Mansson, H. (2000). Identification of stuttering in preschool children - A multifactorial, longitudinal study in development. *Journal of Fluency Disorders, 25*, 241-241.

Martin, R., Kuhl, P., & Haroldson, S. (1972). An experimental treatment with two preschool stuttering children. *Journal of Speech and Hearing Research, 15*, 743-752.

MathSoft Inc. [Computer Software]. (2000). *S-Plus Professional Release 1*.
Seattle, Washington: MathSoft Inc., StatSci Division.

Max, L., Caruso, A., & Vandevenne, A. (1997). Decreased stuttering frequency during repeated readings: a motor learning perspective. *Journal of Fluency Disorders*, 22, 17-33.

McClellan, M. (1996). Lip-muscle reflexes during speech movement preparation in stutterers. *Journal of Fluency Disorders*, 21, 49-60.

Miller, R. (1986). *Beyond ANOVA, basics of applied statistics*. New York, USA: John Wiley & Sons.

Moher, D., Schulz, K., & Altman, D. for the CONSORT Group. (2001). Consort statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. *Lancet*, 357, 1191-1194.

Moleski, R., & Tosi, D. (1976). Comparative psychotherapy: rational-emotive therapy versus systematic desensitization in the treatment of stuttering. *Journal of Consulting Clinical Psychology*, 44, 309-311.

Moore, W., & Ritterman, S. (1973). The effects of response contingent reinforcement and response contingent punishment upon the frequency of stuttered verbal behavior. *Behaviour Research & Therapy*, 11, 43-48.

Murray, T., Kelly, P., Campbell, L., & Stefanik, K. (1977). Haloperidol in the treatment of stuttering. *British Journal of Psychiatry*, 130, 370-373.

National Health and Medical Research Council. (1999). *A guide to the development, implementation and evaluation of clinical practice guidelines*. Commonwealth of Australia, Canberra.

O'Brian, S., Onslow, M., Cream, A., & Packman, A. (2003). The Camperdown Program: Outcomes of a new prolonged-speech treatment model. *Journal of Speech, Language, and Hearing Research, 46*, 933-946.

O'Brian, S., Packman, A., Onslow, M., & O'Brian, N. (2004). Measurement of stuttering in adults: Comparison of stuttering rate and severity scaling methods. *Journal of Speech, Language, and Hearing Research, 47*, 1081-1087.

Onslow, M. (1992). Identification of early stuttering: Issues and suggested strategies. *American Journal of Speech-Language Pathology, 1*, 21-27.

Onslow, M. (1996). *Behavioral Management of Stuttering*. San Diego, CA: Singular Publishing Group.

Onslow, M. (2003). From laboratory to living room: The origins and development of the Lidcombe Program. In Onslow, M., Packman, A. & Harrison, E. (Eds.) *The Lidcombe Program of early stuttering intervention: A clinician's guide*. Austin, TX: Pro-Ed.

Onslow, M. (2005). (in preparation) Overview of stuttering. In M. Onslow (Ed.) *Evidence based clinical management of stuttering*. Austin, TX: Pro-Ed.

Onslow, M., Andrews, C., & Costa, L. (1990). Parental severity scaling of early stuttered speech: four case studies. *Australian Journal of Human Communication Disorders, 18*, 47-61.

Onslow, M., Andrews, C., & Lincoln, M. (1994). A control/experimental trial of an operant treatment for early stuttering. *Journal of Speech and Hearing Research, 37*, 1244-1259.

Onslow, M., Costa, L., Andrews, C., Harrison, L., & Packman, A. (1996). Speech outcomes of a prolonged-speech treatment for stuttering. *Journal of Speech and Hearing Research, 39*, 734-749.

Onslow, M., Costa, L., & Rue, S. (1990). Direct early intervention with stuttering: Some preliminary data. *Journal of Speech and Hearing Disorders, 55*, 405-416.

Onslow, M., Gardner, K., Bryant, K. Stuckings, C., & Knight, T. (1992). Stuttered and normal speech events in early childhood: The validity of a behavioral data language. *Journal of Speech and Hearing Research, 35*, 79-87.

Onslow, M., Harrison, E., & Jones, M. (1993). *Early stuttering: Onset, treatment, and recovery*. Paper presented at the Annual Conference of the Australian Association of Speech and Hearing, Darwin, Northern Territory.

Onslow, M., & Menzies, R., & Packman, A. (2001). The Lidcombe Program: Development of a parent-conducted operant early intervention for stuttering. *Behavior Modification, 25*, 116-139.

Onslow, M., & Packman, A (1999a). (Eds.) *The handbook of early stuttering intervention*. San Diego, CA: Singular Publishing Group.

Onslow, M., & Packman, A. (1999b). The Lidcombe Program and natural recovery: Potential choices of initial management strategies for early stuttering. *Advances in Speech Language Pathology, 1*, 113-121.

Onslow, M., Packman, A. & Harrison, E. (Eds.) (2003). *The Lidcombe Program of early stuttering intervention: A clinician's guide*. Austin, TX: Pro-Ed.

Packman, A., & Attanasio, J. (2004). *Theoretical issues in stuttering*. London: Psychology Press.

- Packman, A., & Lincoln, M. (1996). Early stuttering and the Vmodel. *Australian Journal of Human Communication Disorders, 24*, 45-54
- Packman, A., & Onslow, M. (1998). What is the take-home message from Curlee and Yairi? *American Journal of Speech-Language Pathology, 7*, 5-9.
- Packman, A., & Onslow, M. (1999). Issues in the treatment of early stuttering. In M. Onslow & A. Packman (Eds.) *The handbook of early stuttering intervention*. San Diego, CA. Singular Publishing Group.
- Packman, A., Onslow, M., & Attanasio, J. (2003). The timing of early intervention with the Lidcombe Program. In M. Onslow, A. Packman, & E. Harrison (Eds.) *The Lidcombe Program of early stuttering intervention: A clinician's guide*. Austin, TX: Pro-Ed.
- Paden, E., & Yairi, E. (1996). Phonological characteristics of children whose stuttering persisted or recovered. *Journal of Speech, Language and Hearing Research, 39*, 981-990.
- Peins, M., McGough, W., & Lee, B. (1972). Evaluation of tape-recorded method of stuttering therapy: improvement in a speaking task. *Journal of Speech and Hearing Research, 15*, 364-371.
- Petitti, D. (1994). *Meta-analysis, Decision Analysis and Cost-Effectiveness Analysis*. Oxford, UK: Oxford University Press.
- Piantadosi, S (1997). *Clinical trials, a methodological perspective*. New York: John Wiley & Sons.
- Pindzola, R. (1987). *Stuttering prevention program: age 3 to grade 3*. Austin, TX: Pro-Ed.

Pindzola, R. (1999). The stuttering intervention program. In M. Onslow and A. Packman (Eds.) *The Handbook of early stuttering intervention*. San Diego, CA: Singular Publishing Group.

Prins, D. (1983). Continuity, fragmentation, and tension: Hypotheses applied to evaluation and intervention with preschool disfluent children. In D. Prins & R. Ingham (Eds.) *Treatment of stuttering in early childhood: Methods and issues*. San Diego, CA: College-Hill Press.

Prins, D., & Ingham, R. (1983). *Treatment of stuttering in early childhood: Methods and issues*. San Diego: College Hill Press.

Prins, D., Main, V., & Wampler, S. (1997). Lexicalization in adults who stutter. *Journal of Speech, Language and Hearing Research, 40*, 373-384.

Prins, D., Mandelkorn, T., & Cerf, F. (1980). Principal and differential effects of haloperidol and placebo treatments upon speech dysfluencies in stutterers. *Journal of Speech and Hearing Research, 23*, 614-629.

Rantala, S., & Petri-Larmi, M. (1976). Haloperidol (Serenase) in the treatment of stuttering. *Folia Phoniatica, 28*, 354-361.

Reed, C., & Godden, A. (1977). An experimental treatment using verbal punishment with two preschool stutterers. *Journal of Fluency Disorders, 2*, 225-233.

Riley, G., & Riley, J. (1983) Evaluation as a basis for intervention. In D. Prins & R. Ingham (Eds.), *Treatment of stuttering in early childhood: Methods and issues*. San Diego, CA: College-Hill Press.

Riley, J., & Riley, G. (1999). Speech motor training. In M. Onslow and A. Packman (Eds.) *The Handbook of early stuttering intervention*. San Diego, CA: Singular.

Robb, M., & Blomgren, M. (1997). Analysis of F2 transitions in the speech of stutterers and non-stutterers. *Journal of Fluency Disorders*, 22, 1-16.

Roseman, B., & Johnson, K. (1998). *Easy does it for fluency: preschool/primary*. East Moline, IL: LinguiSystems, Inc.

Rosenberger, P., Wheelden, J., & Kalotkin, M. (1976). The effect of haloperidol on stuttering. *American Journal of Psychiatry*, 133, 331-334.

Runyan, C., & Runyan, S. (1993). Therapy for school-aged stutterers: An update on the Fluency Rules Program. In R. Curlee (Ed.), *Stuttering and related disorders of fluency*. New York: Thieme Medical Publishers.

Runyan, C., & Runyan, S. (1999). The fluency rules program. In M. Onslow and A. Packman (Eds.) *The Handbook of early stuttering intervention*. San Diego, CA: Singular Publishing Group.

Rustin, L., Botterill, W., & Kelman, E. (1996). *Assessment and therapy for young disfluent children*. London, England: Whurr Publishers.

Rustin, L., Kuhr, A., Cook, P., & James, I. (1981). Controlled trial of speech therapy versus oxprenolol for stammering. *British Medical Journal Clinical Research Education*, 283, 517-519.

Ryan, B., & Ryan, B. (1999). The Monterey fluency program. In M. Onslow and A. Packman (Eds.) *The Handbook of early stuttering intervention*. San Diego, CA: Singular Publishing Group.

Ryan, B., & Van Kirk, B. (1978). *Monterey Fluency Program*. Monterey, CA: Monterey Learning Systems.

Saint Laurent, L., & Ladouceur, R. (1987). Massed versus distributed application of the regulated-breathing method for stutterers and its long-term effect. *Behaviour Therapy, 18*, 38-50.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110–114.

Schulz, K., Chalmers, I., Hayes, R., & Altman, D. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association, 273*, 408-412.

Sidman, M. (1960). *Tactics of Scientific Research*. New York: Basic Books.

Siegel, G. (1990). Conference on treatment efficacy: Concluding remarks. In Olswang, L., Thompson, C., Warren, S., & Minghetti (Eds.) *Treatment Efficacy Research in Communication Disorders*. Rockville, MD: American Speech-Language-Hearing Foundation.

Signorini, D., Leung, O., Simes, R., Beller, E., & Gebski, V. (1993) Dynamic balanced randomization for clinical trials. *Statistics in Medicine, 12*, 2343-2350.

Silverman, S., & Bernstein Ratner, N. (1997). Syntactic complexity, fluency, and accuracy of sentence imitation in adolescents. *Journal of Speech, Language and Hearing Research, 40*, 95-106.

Stager, S., Calis, K., Grothe, D., Bloch, M., Turcasso, N., Ludlow, C., & Braun, A. (1997). A double-blind trial of pimozine and paroxetine for stuttering. In Hulstijn, W., Peters, H., & van Lieshout, P. (Eds.) *Speech production: motor control, brain research and fluency disorders*. Amsterdam, The Netherlands: Elsevier.

Stager, S., Denman, D., & Ludlow, C. (1997). Modifications in aerodynamic variables by persons who stutter under fluency-envoking conditions. *Journal of Speech, Language and Hearing Research, 40*, 832-847.

Starkweather, C. (1987). *Fluency and stuttering*. Englewood Cliffs, N.J.: Prentice-Hall.

Starkweather, C. (1990). Current trends in therapy for stuttering children and suggestions for future research. *ASHA Reports, 18*, 82-90.

Starkweather, C., & Gottwald, S. (1993). A pilot study of relations among specific measures obtained at intake and discharge in a program of prevention and early intervention for stuttering. *American Journal of Speech-Language Pathology, 2*, 51-58.

Starkweather, C., Gottwald, S., & Halfond, M. (1990). *Stuttering prevention: a clinical method*. Englewood Cliffs, N.J.: Prentice-Hall.

Stein M., Baird A., & Walker J. (1996). Social phobia in adults with stuttering. *American Journal of Psychiatry, 153*, 278-280.

Stonehouse, J., & Forrester, G. (1998). Robustness of the t and U tests under combined assumption violations. *Journal of Applied Statistics, 25*, 63-74.

Stuart, A., Kalinowski, J., Armson, J., Stenstrom, R., & Jones, K. (1996). Fluency effect of frequency alterations of plus/minus one-half and one-quarter octave shifts in auditory feedback of people who stutter. *Journal of Speech, Language and Hearing Research, 39*, 396-401.

Swift, W., Swift, E., & Arellano, M. (1975). Haloperidol as a treatment for adult stuttering. *Comprehensive Psychiatry, 16*, 61-67.

Taves, D. R. (1974). Minimisation: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics*, *15*, 443-453.

Teesson, K., Packman, A., & Onslow, M. (2003). The Lidcombe behavioral data language of stuttering. *Journal of Speech, Language, and Hearing Research*, *46*, 1009-1015.

Van Lieshout, P., Hulstijn, W., & Peters, H. (1996a). Speech production in people who stutter: testing the motor plan assembly hypothesis. *Journal of Speech, Language and Hearing Research*, *39*, 76-92.

Van Lieshout, P., Hulstijn, W., & Peters, H. (1996b). From planning to articulation in speech production: what differentiates a person who stutters from a person who does not stutter? *Journal of Speech, Language and Hearing Research*, *39*, 546-564.

Van Riper, C. (1963). *Speech correction: Principles and methods (4th Ed.)*. Englewood Cliffs, N.J.: Prentice-Hall.

Van Riper, C. (1982). *The Nature of Stuttering (2nd ed.)*. Englewood Cliffs, NJ: Prentice-Hall.

Vanryckeghem, M., & Brutten, G. (1996). The relationship between communication attitude and fluency failure of stuttering and non-stuttering children. *Journal of Fluency Disorders*, *21*, 109-118.

Vanryckeghem, M., & Brutten, G. (1997). The speech-associated attitude of children who do and do not stutter and the differential effect of age. *American Journal of Speech-Language Pathology*, *6*, 67-73.

Velleman, P., & Hoaglin, D. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston, Mass.: Duxbury Press.

Waterloo, K., & Gotestam, K. (1988). The regulated-breathing method for stuttering: an experimental evaluation. *Journal of Behaviour Therapy & Experimental Psychiatry*, 19, 11-19.

Welch, B. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.

Wells, P., & Malcolm, M. (1971). Controlled trial of the treatment of 36 stutterers. *British Journal of Psychiatry*, 119, 603-604.

Wenker, R., Wegener, J., & Hart, K. (1996). The impact of presentation mode and disfluency on judgments about speakers. *Journal of Fluency Disorders*, 21, 147-159.

Wille, A. (1999). Bioresonance therapy (biophysical information therapy) in stuttering children. *Forschende Komplementarmedizin*, 6, 50-52.

Wolpe, J. (1958). *Psychotherapy for reciprocal inhibition*. Palo Alto, CA: Stanford University Press.

Woods, D., Twohig, M., Fuqua, R., & Hanley, J. (2000). Treatment of stuttering with regulated breathing: Strengths, limitations and future directions. *Behaviour Therapy*, 31, 547-68.

Woods, S., Shearsby, J., Onslow, M., & Burnham, D. (2002). The psychological impact of the Lidcombe Program of early stuttering intervention: Eight case studies. *International Journal of Language and Communication Disorders*, 37, 31-40.

Yairi, E. (1981). Disfluencies of normal speaking two-year-old-children. *Journal of Speech and Hearing Research, 24*, 490-495.

Yairi, E. (1997). Disfluency characteristics of childhood stuttering. In R. Curlee & G. Siegel (Eds.), *Nature and treatment of stuttering: New directions*. Needham Heights, MA: Allyn & Bacon.

Yairi, E., & Ambrose, N. (1992a). A longitudinal study of stuttering in children: A preliminary report. *Journal of Speech and Hearing Research, 35*, 755-760.

Yairi, E., & Ambrose, N. (1992b). Onset of stuttering in preschool children: Selected factors. *Journal of Speech and Hearing Research, 35*, 782-788.

Yairi, E., & Ambrose, N. (1999). Early childhood stuttering I: Persistency and recovery rates. *Journal of Speech, Language, & Hearing Research, 42*, 1097-1112.

Yairi, E., Ambrose, N., & Cox (1996). Genetics of stuttering: A critical review. *Journal of Speech and Hearing Research, 39*, 771-784.

Yairi, E., Ambrose, N., Paden, E., & Throneburg, R. (1996). Predictive factors of persistence and recovery: Pathways of childhood stuttering. *Journal of Communication Disorders, 29*, 51-77.

Yaruss, J. (1997). Clinical implications of situational variability in pre-school children who stutter. *Journal of Fluency Disorders, 22*, 187-203.

Yaruss, J., & Conture, E. (1996). Stuttering and phonological disorders in children: examination of the covert repair hypothesis. *Journal of Speech, Language and Hearing Research, 39*, 349-364.

Young, M. (1994). Evaluating differences between stuttering and non-stuttering speakers: the group difference design. *Journal of Speech and Hearing Research, 37*, 522-534.

Zebrowski, P. (1997). Assisting young children who stutter and their families: Defining the role of the speech-language pathologist. *American Journal of Speech-Language Pathology, 6*, 19-28.

Appendices

Appendix A. Sample SAS Program for Running Simulations

```
*type 1 error simulations ;

%macro error (howmany) ;

%do i = 1 %to &howmany;

data sim1 ;
  do i = 1 to 5;
    x = rand ('gamma',1) ;
    pss = (3*x) ;
    group = 1 ;
    output;
  end;

data sim2 ;
  do i = 1 to 5;
    x = rand ('gamma',1) ;
    pss = (3*x) ;
    group = 2 ;
    output;
  end;

data sim12 ;
set sim1 sim2 ;
run ;

filename routed 'test';

proc printto print = routed new;
run ;

proc ttest data = sim12 ;
class group ;
var pss ;
run ;

proc printto print = print ;
run;

data pval ;
infile routed ;
input x word $ @;
if word = 'Satterth' then do ;
  input x1 x2 x3 pval $ ;
  keep pval ;
  output ;
end ;
```

```
end ;  
run;  
  
data count ;  
set pval ;  
if pval = 'of' then delete ;  
else if pval < 0.05 then count = 1 ;  
else count = 0 ;  
run ;  
  
proc append base = error data = count ;  
  
%end ;  
  
%mend ;  
  
%error (2000) ;  
  
proc freq data = error ;  
tables count ;  
run ;
```

Appendix B. Trial Consent Form

1. I have read the patient information sheet and understand its content.
2. I have had opportunity to discuss this study and have my questions answered by the investigators.
3. I have received satisfactory answers to my questions.
4. I have received enough information about the study.
5. I understand that my participation is voluntary.
6. I understand that I am free to withdraw from the study:
 - at any time
 - without having to give reasons
 - without affecting my child's future treatment
7. Sections of my child's treatment records relating to his/her participation in the study may be inspected by individuals involved in the study. All personal details will be treated as strictly confidential. I give my permission for these individuals to have access to my child's records.
8. I have had sufficient time to come to my decision.
9. I agree for my child to participate in this study.

Child's Parent:

Signature: _____ Name: _____ Date: _____

Witness:

Signature: _____ Name: _____ Date: _____

Investigator:

Signature: _____ Name: _____ Date: _____

Appendix C. Trial Information Sheet

NAME OF THE PROJECT: LIPS Trial: A randomised controlled trial of the Lidcombe Program for early stuttering

INVESTIGATORS: *Mark Jones, Professor Mark Onslow, Dr Ann Packman, Associate Professor Val Gebski, University of Sydney; Tika Ormond, University of Canterbury; Shelley Williams, Stuttering and Treatment Research Trust (START).*

SUBJECT INFORMATION SHEET

The aim of this research project is to investigate the effectiveness of a treatment for young children who stutter. The treatment, known as the Lidcombe Program, is used in some parts of Australia and the UK; however its overall effectiveness is not known, because it is believed that about four out of five children who start to stutter get better without any treatment at all. The findings of this research will make a significant contribution to our understanding and treatment of early stuttering.

If you decide to participate in this project, you will be allocated to either a *Treatment* or a *Non-Treatment* group. Parents in both groups will be required to audiotape-record their child's speech in their normal environment every few months, for 9 months. This enables the Investigators to measure improvement in stuttering. Parents will be able to borrow tape recorders if necessary. There is a significant chance that children in the *Non-Treatment* group will recover from their stuttering without treatment. However, any child who is still stuttering after one year will be offered the treatment, if it has been shown to be effective. Previous research in Australia indicates that waiting a year does not decrease the effectiveness of the treatment.

The treatment is safe and enjoyable for children, and there are no risks or pain involved. The treatment involves the parent and the child attending a speech clinic once a week and during these visits the parent is shown how to do the program at home. The home treatment involves the parent and child doing speech-related activities together for about 15 minutes each day. The average time for the treatment is around 12 weeks, with subsequent follow-up appointments, but it may take longer than this.

Participation in this research is voluntary. If you decide to participate, you and your child are free to withdraw from the project at any time without penalty. Withdrawal will not jeopardise any future relationship you may have with the University of Canterbury or START, in any way.

The results of the project may be published, but you may be assured of the complete confidentiality of data gathered in this investigation. To ensure anonymity and confidentiality only the clinician's treating the children will have access to their name and all data obtained will be stored under strict security.

This research is being conducted as a requirement for a PhD by Mark Jones under the supervision of Associate Professor Mark Onslow and Dr Ann Packman at the University of Sydney. The project has been reviewed by the University of Canterbury and the University of Sydney. Any person with concerns or complaints about the conduct of this research study can contact the Manager of Ethics and Biosafety Administration, University of Sydney, on +61 2 9351 4811. For further information or if you have any concerns about participation in the project, please contact Professor Mark Onslow / Dr Ann Packman (telephone: +61 2 9351 9061).

Appendix D. Design of Randomised Controlled Trial

