



Assessing efficacy of stuttering treatments[☆]

Carys Thomas, Peter Howell*

Department of Psychology, University College London, Gower Street, London WC1E 6BT, UK

Received 20 September 2000; received in revised form 27 February 2001; accepted 15 June 2001

Abstract

Efficacy has been defined as the extent to which a specific intervention, procedure, regimen, or service produces a beneficial result under ideally controlled conditions when administered or monitored by experts. Studies on efficacy can be divided into those that study methods of conducting treatment (i.e., treatment process research) and those that are concerned with the effects of treatments (i.e., treatment outcome research). This review covers both areas, emphasizes the former, and considers such key determinants of efficacy as measurement, treatment integrity, and design issues. A set of criteria is given and a meta-analysis of whether studies published since 1993 meet these criteria is reported (incorporating some pragmatic and ethical considerations). The review ends by considering directions that warrant further investigation in the future.

Educational objectives: The reader will learn about and be able to describe (1) measurements appropriate for evaluating treatment efficacy studies; (2) how to evaluate reports of stuttering treatment programs; and (3) different designs used in treatment efficacy studies. © 2001 Elsevier Science Inc. All rights reserved.

Keywords: Efficacy; Treatment; Stuttering

1. Introduction

The treatment of stuttering has been described as a controversial and perplexing issue for speech language pathologists (Ingham & Riley, 1998), and

[☆] This research was supported by a grant from the Wellcome Trust. Authorship is shared equally.

* Corresponding author. Tel.: +44-207-679-7566; fax: +44-207-436-4276.

E-mail address: p.howell@ucl.ac.uk (P. Howell).

recent concerns have been expressed about the absence of rigorous documentation regarding the efficacy of particular interventions (Ansel, 1993; Conture, 1996; Conture & Guitar, 1993; Cordes & Ingham, 1998; Starkweather, 1993). It has even been asserted that the state of stuttering treatment research, at least up to early 1996, was abysmal and that some leaders in the field appear to have abandoned basic scientific principles that are at the heart of any attempt to establish treatment efficacy (Cordes, 1998). Efficacy is the extent to which a specific intervention, procedure, regimen, or service produces a beneficial result under ideally controlled conditions when administered or monitored by experts (Last, 1983). In contrast, treatment effectiveness is the extent to which an intervention or treatment employed in the field does what it is intended to do for a specific population (Last, 1983). Treatment efficacy research can be characterized as an investigative tool for examining the effects of environmental variables (i.e., treatment) on organismic variables (i.e., communication behaviors). Moreover, it has been suggested that the beauty of efficacy research is its ability to address both theoretical and clinical questions simultaneously (Ols-wang, 1993).

The dawn of a behavioral orientation to stuttering treatment in the 1960s introduced a set of principles and practices for determining treatment efficacy. This model was based primarily upon the quantification of the target of treatments, plus systematic evaluations of relevant behaviors across clinically important settings for meaningful periods of time (Ingham & Andrews, 1973; Kazdin, 1978) and did, to some extent, transcend theoretical orientations (Bloodstein, 1987). However, Schwartz (1976) published his account of “solving stuttering,” with its accompanying claim that the disorder had been treated with an 89% success rate. Reaction throughout the field was “principally directed at a glaring absence of data-based therapy evaluation” (Ingham, 1993, p. 134).

Then, a second catalyst occurred in 1987, when Cooper claimed that “at least two out of every five adolescent and adult abnormally disfluent individuals are incurable stutterers” (Cooper, 1987, p. 381), which was attacked on similar grounds. It has since been suggested that the procedures recommended for evaluating the efficacy of stuttering treatment have become overwhelmingly complex, while at the same time, prevailing notions about the nature of stuttering have become increasingly biological (Ingham & Cordes, 1997). As a result, Ingham and Cordes (1997) claimed that even the most recent studies of stuttering treatment seem to have been conducted without evaluation procedures and that treatments are now being recommended with little or no empirical support. In support of this claim, Cordes (1998) reviewed 88 selected publications and reported that treatments that were most often recommended were not treatments that had been the most comprehensively researched.

Systematic assessments of the efficacy of treatments utilized by a profession are “essential to the maintenance of the clinical integrity of any profession” (Curlee, 1993, p. 328). Hence, the purpose of this article is to identify some of the fundamental issues that should form the bases of evaluating treatment efficacy for

stuttering. As Purser (1987) noted, evaluations of treatment efficacy involve both treatment process research (i.e., study of methods of conducting treatment) and treatment outcome research (i.e., study of the effects of treatments). This review, therefore, examines both aspects, with particular emphasis on the former, as well as the issues that need to be resolved or addressed before treatment efficacy can be assessed.

Our review endeavors to provide an impartial look at studies of efficacy across treatments of stuttering that will complement recent discussions in the literature in this area (Yaruss, 1998) and in wider aspects of health care services (Kazdin & Kendall, 1998). We do not intend to promote any particular theory or to evaluate in detail programs of treatment. These objectives rule out, respectively, our consideration of Steps 2, 3, 6, and 7 of Kazdin and Kendall's (1998) list of "Steps Toward Developing Effective Treatments." This review, instead, critically addresses: (i) measurement issues, (ii) treatment integrity, and (iii) design issues, that correspond to Kazdin and Kendall's conceptualization of dysfunction (Step 1), specification of treatment (Step 4), and tests of treatment outcome (Step 5). Yaruss (1998) provided a framework for describing the etiology and range of problems that a speaker may experience using language that other health professionals employ. However, this does not readily lend itself to assessment of the activities of professionals engaged in delivering treatments, which fall under topics (ii) and (iii).

2. Clinical issues in the measurement of stuttering

It was once thought that stuttering was a comparatively simple disorder to measure (Ingham & Andrews, 1973), and the counting of moments of stuttering, which began in the 1930s, not only operationalized the measurement of stuttering but brought with it the rigor of scientific inquiry. There is general agreement in the literature, as well as considerable content validity, supporting the notion that reductions in stuttering frequency and severity are associated with effective treatment outcome (Conture & Guitar, 1993; Ingham & Costello, 1984; Yairi, 1993, 1997). However, the literature also expressed growing concern about the reliability and validity of clinic-based perceptual measures of stuttering, which stemmed primarily from problems that independent observers had in agreeing satisfactorily on the loci of stutters, thereby threatening the validity of a study's results (see Cooper, 1990; Ingham, 1990). This problem was not resolved even when observers were given a definition of stuttering (Curlee, 1981; Martin & Haroldson, 1981; Young, 1975a), were required to repeat their judgments over several sessions (Cordes, Ingham, Frank, & Ingham, 1992; Young, 1975b), or listened to slowed nondistorted recordings of stuttered speech (Kroll & O'Keefe, 1985). The implications of such findings have been described as undermining measurement of treatment success with stuttering (Bloodstein, 1990; Cooper, 1990; Ingham, 1990).

In addition, Kully and Boberg (1988) had 10 audio recordings of speech, eight of stuttered, and two of nonstuttered speech, evaluated at 10 clinics throughout the world, including Australia, Canada, England, and the US. They reported not only large interclinic discrepancies (e.g., the percentage of syllables stuttered ranged from 3.80% SS to 13.70% SS), but large discrepancies in % SS between samples of the two normal speakers as well. Such high variability in stuttering measures questioned not only the value of data that depict outcome evaluations, but the fact that some clinics rated fluent speakers with relatively high % SS (e.g., one clinic scored a fluent speaker with 4.79% SS), which raised further concerns about the clinical significance of posttreatment measures of stuttering reduction. These findings were later replicated by Ingham and Cordes (1992), and a study by Ham (1989) also reported large differences in the ways that clinical researchers quantify stuttering events.

The identification of stuttering behaviors has an established literature of its own (e.g., see Costello & Ingham, 1984; Smith, 1990; Starkweather, 1987 for further discussion). It is apparent that the reliability of clinical measures of stuttering provides a shaky foundation for evaluating treatment efficacy (Ingham, 1990). Nevertheless, some authors have proposed that such measures do not necessarily pose a serious threat to the assessment of treatment efficacy (e.g., Starkweather, 1993). Kully and Boberg clearly showed that different clinicians employ different protocols in counting stutters, but the purpose of collecting such speech measures in clinical practice is to document trends in clients' speech performance before, during, and after treatment (Onslow, 1996). Kully and Boberg's data showed disparate percentages of stuttered syllables; however, their data also indicated that clinicians generally identify the same trends, which suggests that several measurement issues may play a more critical role in evaluations of treatment efficacy.

Even though interjudge reliability is an important issue, intrajudge reliability may be more important in clinical practice because measures of stuttering need to be internally consistent from session to session across extended periods of time. Packman, Ingham, and Onslow (1993) examined seven clinicians who worked in the same clinic and found that the number of stutters counted in various speech samples differed when the clinicians re-counted the same samples at a later time; however, the relatively high levels of intrajudge agreement indicated that the clinicians were capable of making consistent clinical measures. As might be expected, the most valid measures of stuttering are those based on perceptual judgments of reliable observers who are well acquainted with the clinical signs of stuttering and that systematic comparisons of such measures across situations and time are appropriate for evaluating treatment efficacy (Curlee, 1993). In addition, Cordes et al. (1992) reported that a judge's amount of experience with stuttering and ability to review repeatedly recorded speech samples play an important part in determining levels of agreement.

Due to the well-documented variability of stuttering within subjects, speech samples should ideally be obtained under multiple conditions and on multiple

occasions (Conture, 1997; Conture & Guitar, 1993; Yaruss, 1997). This can be particularly important for young children, as stuttering has been reported to fluctuate greatly over time and sometimes cease entirely (Ingham & Riley, 1998). Druce, Debney, and Byrt's (1997) study of 6- to 8-year-old children obtained two pretreatment measures during conversational interactions with a family member and with an unknown person to represent low- and high-stress speaking situations, respectively. Boberg and Sawyer (1977) reported that follow-up measures are likely to be biased if collected in the same clinical environment in which treatment was administered because clients were observed to stutter more when conversing outside the clinical setting than when speaking with a stranger in the familiar surroundings of the clinic.

It has also been recommended that speech measures be collected without clients' knowledge that their speech is being evaluated, so that they do not react to being assessed and try to create a favorable outcome. Ingham (1972) compared covert and overt assessments and found that stutterers generally did speak more fluently when they were aware that their speech was being evaluated. However, these findings have been greatly contested by studies (Andrews & Craig, 1982; Howie, Tanner, & Andrews, 1981; Howie, Woods, & Andrews, 1982) which have found that such behavior occurs only in some individuals. Perhaps more important is that speech samples be obtained in clients' natural environments, as ecological validity is uncertain when only measures obtained in the clinic are used (Conture & Guitar, 1993; Costello & Ingham, 1984; Starkweather, 1993). In an experimental trial of an operant treatment of early stuttering, Onslow, Andrews, and Lincoln (1994) collected (a) home recordings of children speaking to family members in the family home; (b) away-from-home recordings of children speaking to non-family members in the homes of family friends and relatives; (c) covert recordings of children speaking to family members at home without the children's knowledge; and (d) recordings of children conversing with an investigator. Without a broad collection of measures in various settings, investigators cannot be certain that clinic-collected data can be generalized to many outside-of-clinic speaking situations. These authors were able to use the mean/median posttreatment speech measures obtained in everyday speaking situations of these children to show significant decreases from those collected pretreatment in the same situations.

Many types of dependent variables have been measured, in addition to the percentage of stutter and syllable counts, and have played an increasingly important role in evaluating treatment efficacy. Profiles of treatment outcomes should include stuttering severity, speech rate, and speech naturalness. Various methods of assessing stuttering severity have been described (Conture, 1997; Costello & Ingham 1984), and severity can be influenced by a number of factors such as parental concerns, teasing from peers, and a child's own frustrations. Furthermore, if a treatment employs techniques that aim to alter speech rate, (e.g., prolonged speech treatments), speech rate measures are critical in evaluating outcomes because such treatments may result in unnaturally slow or monotone speech.

Some researchers believe that assessments of treatment efficacy should also include measures of speech naturalness (Costello & Ingham, 1984; Martin, Haroldson, & Triden, 1984; Starkweather, 1993), which can be affected by speech rate, inflection, articulation, resonance, loudness, vocabulary, and sentence structure (Ingham & Riley, 1998). Because so many treatments incorporate changes in clients' speech patterns to induce fluency, measures of speech naturalness allow clinicians to ensure that spontaneous, normal-sounding speech has not been sacrificed or overlooked by efforts to objectify the efficacy of a given treatment. Druce et al. (1997) collected measures of % SS, speech rate, speech naturalness, and subjective ratings of stuttering severity in evaluating an intensive treatment program for 15 young children who stuttered. This combination of measures enabled them to conclude that the reduction in % SS was consistent with improvements in both severity and naturalness ratings, thereby offering complementary support of the children's improved fluency.

Issues of reliability are important in treatment efficacy studies because stuttering frequency, severity, speech rate, and speech naturalness are filtered through what Ingham and Riley (1998) described as "the complexities of human perception," which are subject more often to error and bias than are measures collected with objective instruments. This again raises the issue of adequate inter- or intrajudge reliability and the need to report on a number of such key variables as level of training and qualifications of clinicians and the procedures used to ensure independent judgments (Cordes, 1994; Cordes et al., 1992; Lewis, 1994) if the efficacy of treatments is to be critically examined.

One variable, which was noted as underinvestigated elsewhere (Woods, Fuqua, & Waltz, 1997), is clients' use of avoidance behaviors to diminish stuttering severity, which may be particularly relevant to some operant-based treatments. For example, clients may undergo avoidance conditioning whenever they perform certain behaviors to avoid stuttering (Bandura, 1969). Woods et al. (1997) reported that a 6-year-old boy's frequent replies of "I don't know" were used to avoid more extensive answers to conversational questions. Starkweather (1993) cited such behaviors as stalling for time, forcing words out, changing words altogether, or losing eye contact as ways to minimize the aversive experience of stuttering. If such avoidance behaviors displace stuttering, they may limit or prevent effective application of treatment as well as limit assessment of therapeutic efficacy. Future research should include assessment of avoidance behaviors through functional analyses of avoidance responses, which can be accomplished by monitoring the occurrence of various avoidance responses in stuttering-prone situations and during treatment (Woods et al., 1997) so that avoidance responses can be identified and procedures for their elimination introduced.

Conture and Guitar (1993) argued that neither the short-term, medium-term, nor long-term efficacy of therapy can be documented without objective measures. Subjective measures of experimenters, people who stutter, family, or friends are inadequate on their own and may often be colored by expectations, hope, etc. Bloodstein (1995) stated that variables which can affect subjective judgments

may be subtle and difficult to predict. For example, Lanyon, Lanyon, and Goldsworthy (1979) found that clinician's predictions of how successful a person who stutters would be in mastering a biofeedback treatment procedure were significantly related to their eagerness for clients to present themselves in a favorable light, as indicated by *k* scores on the MMPI. Findings also revealed that the *k* scores of persons who stutter had little relationship to objective measures of their progress. Substantial individual variation is often present at onset and in the development of stuttering, which encourages the development of individualized treatment programs. However, this may mislead clinicians to believe that a client has achieved more than he/she has. Starkweather (1993) claimed that the only way to control this problem is to maintain clear identification and measurement of treatment goals. Research has been unable to provide a conclusive demonstration that one method of counting stuttering is more valid than another (Ingham & Riley, 1998). Therefore, studies of treatment efficacy must describe how stutters were identified and recorded, the clinicians who made them, their experience, and the reliability of their counts.

3. Treatment integrity

Another important issue that needs to be addressed is whether a clinician administers an intervention correctly because this determines if the treatment that was administered is the same as that for which efficacy data are reported (Conture & Guitar, 1993). The call for high-quality, detailed descriptions in published reports of the specific treatments, procedures, and measures employed (Ingham & Riley, 1998) to support the reliability and validity of efficacy data is the hallmark of science that would ensure that others applying the same procedures would obtain comparable treatment effects. Siegel (1990) has argued that replication is the most stringent test of reliability and for extending the external validity of research findings (Venty & Schiavetti, 1986). Similarly, Meyers (1990, p.178) emphasized the importance of replication in noting that "Replication is worth 1000 *t* tests."

Onslow (1992) critically examined the literature on stuttering intervention and found that only a few of the many studies conducted have been replicated even once, over a period of 20 years. Muma's (1993) survey of one fourth of the studies published in the *Journal of Speech and Hearing Disorders* and *Journal of Speech and Hearing Research* over a decade (1979–1989) led him to claim that Types I and II errors likely account for 50–250 false findings in the 1712 studies surveyed and could be misdirecting treatment in the field. Only 12 direct replications were found, in contrast to the unknown number of false findings estimated by Muma, revealing an urgent need for more replications in this area. Replications that yield conflicting results raise considerable questions about which study is to be believed, but Muma pointed out that different outcomes could be obtained because of different subject samples, different performance samples, invalid measures, variations in procedures, and inappropriate data

analyses. For example, Onslow, Adams, and Ingham (1992) may have failed to replicate the results of Martin et al. (1984) for a number of these reasons. In contrast, if replications yield comparable results, such as Vihman and Greenlee's (1987) replication of Grunwell's (1981) work on children's phonological development, the substantive base of the field is extended. Researchers' replications of the work of others not only increases their own knowledge but may expand their own research interests and test new hypotheses as well.

The implications for treatment efficacy research are highly significant, replications would ascertain the likelihood of false findings, or would extend the generalizability of studies with relatively few participants, which applies to the large proportion of stuttering treatment research that involved single-subject research designs. A number of researchers have commented that replication is ignored too often in treatment efficacy research (e.g., Attanasio, 1994; Meline & Schmidt, 1997; Muma, 1993; Onslow, 1992). Moreover, a critical review of children's treatment success by Craig, Chang, and Hancock (1992) found that a considerable number of treatment studies, ranging from response contingent to prolonged speech methods, lacked replication, thereby reducing confidence in the success of such treatments.

Another important issue has been described as the surety of treatment fidelity (Yeaton & Sechrest, 1981). Ingham and Riley (1998) maintained that information on the training and supervision of clinicians is important but that empirical evidence that a treatment was, in fact, administered correctly is much more powerful by the focus on the reliability of administration of independent variables. In recent years, parent's presence in therapy sessions, as observers or active agents in home programs, has significantly increased with the aim of facilitating generalization and maintenance of treatment effects (Felsenfeld, 1997). As a result, several findings have been identified that may have implications for treatment efficacy. Studies of fathers' and mothers' paralinguistic behaviors have shown that fathers tend to talk more during parent-child interactions (Kelly, 1993; Kelly & Conture, 1992; Schulze 1991) and that reductions in parents' speech rates are correlated ($r=.47$) with reductions in children's disfluencies (Starkweather & Gottwald, 1993). Several other variables have also been found to correlate with children's disfluencies, including interruptions (Rustin & Cook, 1983), methods of discipline (Prins, 1983), negativity, and excessive questioning (Fosnot, 1993). A key variable that can affect studies of the efficacy of various treatment approaches is the accuracy with which clinicians are delivering prescribed treatments. Ryan and Van Kirk Ryan (1983) described several discrepancies in clinicians' behavior, even after training, which included failing to teach slow, prolonged speech patterns during initial stages of treatment, and undercounting or failing to count stutters in later stages. As Ingham and Riley (1998) pointed out, if a treatment efficacy study does not include evidence of treatment fidelity in all aspects of administration, readers of the study cannot be sure that the results reported are the product of the treatment applied.

4. Design issues

Treatment efficacy research requires designs which can establish that the treatment effects observed are clearly the product of the treatments applied. The advantages and disadvantages of the different designs used in stuttering treatment research have an extensive literature base (e.g., Barlow & Hersen, 1984; Ingham & Riley, 1998; Schiavettii & Metz, 1997); therefore, this discussion focuses on factors or components within designs that are critical for evaluating treatment efficacy. It should be noted, however, that single-subject experiments are suitable for demonstrating overall effects of a given treatment and differential effects of individual components of a treatment (e.g., Costello, 1975). Group studies are best done once single-subject studies have convincingly demonstrated that a treatment produces desirable clinical effects (Ingham & Riley, 1998). Although effect-size statistics (Meline & Schmidt, 1997; Young, 1993), which estimate the magnitude of statistically significant differences, have improved the clinical relevance of group data analyses, this is still a less sensitive demonstration of treatment efficacy than are findings from single-subject studies.

An efficient way of illustrating “good” and “bad” designs is to examine key criticisms that have been directed toward various therapy approaches. Traditional approaches to treating stuttering, especially for children who stutter, involve parent counseling. These approaches attempt to enhance fluency indirectly through manipulations aimed at improving parent-child relationships (Bloodstein, 1987). Such manipulations may include play therapy, reducing anxiety, increasing confidence, or parental modeling of communication behaviors (Gregory & Hill, 1980). Most reports of this work are anecdotal or present only case history evidence of a treatment’s success, and the lack of objective evidence severely limits conclusions that can be drawn about the efficacy of such treatment. Response contingent methods, which are viewed as scientifically rigorous methods, have attracted their share of criticisms on the conclusions drawn about treatment success. For example, Manning, Trutna, and Shaw (1976) attempted to determine if tangible forms of rewards were more effective than verbal rewards in reducing stuttering. Both forms of reward appeared to be successful, but their unique contributions could not be measured because treatment involved a number of therapy procedures. As a result, the efficacy of individual treatment components could not be isolated and assessed. Response-contingent procedures have also received substantial criticism for small sample sizes, lack of control groups (Martin, Kuhl, & Haroldson, 1972; Reed & Godden, 1977), poor external validity and lack of replication (Costello, 1975), and failure to monitor speech rate and progress in the long term (Costello, 1975; Onslow, Costa, & Rue, 1990).

The Gradual Increase in Length and Complexity of Utterance (GILCU) program is another treatment reported to have significant clinical success (Ryan, 1971). Ryan and Van Kirk Ryan (1983) compared the effectiveness of four treatment programs (i.e., programmed traditional, delayed auditory feedback,

time-out contingency, and GILCU) with 16 school age children who stuttered. All four treatments were reported to be effective in reducing stuttering, but only 4 of 16 children completed follow-up stages. The study's large dropout rate and lack of a control group, which avoided raising ethical concerns, raised questions about the success of these treatments. However, this study does illustrate some good design principles. For example, treatment effects were assessed not only in the clinic but also in the children's classroom across the four treatment groups. In addition, replications of the GILCU program have shown that it can be a successful treatment for stuttering (Rustin, Ryan, & Ryan, 1987).

Treatment variants of prolonged speech such as smooth speech have also produced high success rates, but many studies failed to obtain long-term outcome data (e.g., Debney & Druce, 1987), failed to use case history reporting (e.g., Casteel & McMahon, 1978), and lacked matched control groups and replication support (e.g., Cooper, 1987; Culp, 1984). Regulated breathing techniques have also come under criticism for using a number of treatments in combination. Azrin and Nunn (1974) used a mixture of relaxation therapies, self-awareness of stuttering, regulated breathing techniques, together with parent and social support. Although almost total elimination of stuttering was reported, there were no control groups and only subjective measures of outcome. Also Ladouceur and Martineau (1982) reported that regulated breathing was more effective in a combination treatment program than as a single treatment procedure. Several other approaches to stuttering therapy have fallen into similar design traps, which has led to questions about the treatments' efficacy (see Craig et al., 1992).

Whatever treatment is applied, long-term monitoring of clients, especially children who stutter, should not be ignored. Because of the high proportion of children who evidence spontaneous remissions of stuttering, plus the high dropout rates in studies, treatment success is difficult to measure. Furthermore, Andrews and Harvey (1981) claimed that regression to the mean invalidates findings of intervention studies that do not establish stable pretreatment baselines or use adequate randomly assigned or matched untreated control groups. Moreover, Hanna and Owen (1977) reported that stutterers and their parents have a tendency to seek help when stuttering appears to be at its worst. After treatment is sought and before it begins, there appears to be a spontaneous return of symptoms to their average level, which may be nonspecific benefits of having sought help. Andrews and Harvey (1981) described six studies that assessed stuttering 2 months prior to treatment. All six showed a trend (i.e., differences were not statistically significant) to less stuttering on the second assessment. Nevertheless, the authors concluded that study designs should allow time for regression of stuttering to mean severity levels in pre-posttreatment outcome designs to avoid inflating the magnitude of treatment effects and confounding estimates of improvement that are due to therapy. They suggested collecting time series data until a stable baseline is achieved or holding people who stutter on a waiting list for at least 3 months. As noted earlier, speech

samples assessing treatment outcomes should be obtained in nontreatment settings (Ingham & Riley, 1998).

Another issue is how to control for spontaneous recoveries in young children, which occurs in a relatively large proportion of them within the first year of onset and presents a serious complication for treatment efficacy studies of preschool age children (Andrews & Harris, 1964; Bloodstein, 1995; Yairi & Ambrose, 1992). This undermines confidence of the findings reported by treatment efficacy studies, as stuttering may disappear irrespective of any treatment received. Curlee and Yairi (1997) argued that a control group is essential in such cases so that a treatment group would have to “beat the odds” to provide convincing evidence of treatment effects (Ingham & Riley, 1998). It has been further suggested that groups should be matched on a number of key variables that may be predictive of spontaneous recovery, such as duration of stuttering, age at onset, gender, and family history of stuttering recoveries (Curlee & Yairi, 1997; Yairi, Ambrose, Paden, & Throneburg, 1996). There is, however, an ethical question that needs to be addressed — is it unfair to assign young people who stutter into untreated control groups, denying them treatment? Some parents whose children were in control groups opted for their treatment (Onslow et al., 1994).

Given the nature and number of therapy approaches employed for stuttering, it is not surprising that no one design has emerged as the best for assessing treatment efficacy. It is unfortunate that few firm conclusions can be drawn about most treatments because there has been little attention paid to assessing long-term outcomes, a reliance on single-subject designs without replications or larger numbers, and group research lacking adequate controls, all of which may create false impressions and beliefs. Although incorporating adequate controls, pre-posttreatment measures, and many features noted earlier in this review, appear to be critical for improving studies of treatment efficacy; investigators face a number of challenges, not the least of which is balancing ethical with methodological issues. Nevertheless, long-term assessments are essential (Starkweather, 1993). Onslow (1996) sets the minimum post-treatment interval at one year, with some advocating at least a 2- to 5-year follow-up assessments before confidence can be placed in a treatment’s success (Bloodstein, 1995; Craig et al., 1992).

5. Criteria for assessing efficacy studies: comparison with Moscicki (1993) and a meta-analysis of studies from that date

The issue of balancing ethical and methodological considerations is picked up in this section. A previous review on efficacy research in the *Journal of Fluency Disorders* (Moscicki, 1993) presented a list of criteria that studies should meet:

- (a) Careful attention should be paid to the selection and representativeness of participants, and estimates of the reliability and validity of all measures should be provided;

- (b) Sample sizes must be sufficiently large to show statistically significant differences between experimental and control groups;
- (c) Studies should have clear operational definitions of outcomes, with careful attention given to the instrumentation used in assessing and classifying outcomes;
- (d) Treatment must be provided using a standard treatment protocol by clinicians who have received standardized training in administering the procedures;
- (e) Administration of treatment should include procedures for monitoring clinicians' adherence to the treatment protocol and by study participants;
- (f) Conducting pilot studies is essential to resolving design difficulties and other issues prior to initiation of full-scale experimental studies;
- (g) Follow-up assessments must be of sufficient duration to determine the differential effects of treatment;
- (h) Potential sources of bias need to be anticipated and accounted for in the study's design and by analytic models; and
- (i) Data analysis must be appropriate to the study's design.

There are some differences of emphasis between Moscicki's methodological criteria and those suggested in this review as a consequence of differences in focus. First, Moscicki's remarks are directed specifically at randomized control designs. Second, her criteria are not divided into those pertaining to measurement, integrity, and design. Third, we believed it would be unfair to use her criteria in evaluating published efficacy studies. After 8 years, however, it did seem appropriate to see if some efficacy studies had met at least some of these criteria. An evaluation of such studies will employ the criteria we propose, which include both pragmatic and ethical factors but still overlap those of Moscicki to a large degree. As Robinson Crusoe learned, there is no point in building something, a boat in his case, which is of such quality that prevents it from being used in practice (i.e., the weight of Robinson's boat prevented it being moved into the sea). On the other hand, a boat that is too flimsy is a danger to its passengers. The real question is how to get the right balance between ideal and practice.

As will be seen, there is agreement with Moscicki on theoretical perspective as well as broad agreement on criteria pertaining to efficacy–practice, too. However, some of Moscicki's criteria involve requirements that are rarely met in empirical studies, which can sometimes be justified on pragmatic or ethical grounds. An example is “representativeness” in criterion (a), which is often ignored when speakers' frequency of stuttering must be high enough to allow the effects of a treatment to be shown (e.g., Ryan & Van Kirk Ryan, 1995). Few studies involve control groups because of concerns about the ethics of denying treatment to certain speakers who stutter. Although Hancock et al.'s (1998) study included a control group, control subjects were not prohibited from treatment in the long term, which prevented the study from obtaining long-term control data.

Efficacy studies are costly to perform, and alternative ways can obtain the information that pilot studies provide, such as single-subject experimental designs (criterion f). Our criteria are not specific to randomized control designs, attempt to recognize the ethical and pragmatic constraints, and are organized as three categories: measurement, treatment integrity, or design. Correspondences with Moscicki's criteria are designated at the end of the following criteria by the letter from her list:

Measurement issues

- (1) Clear identification and measurement of treatment goals must be provided, for example, by checks on the administration of the treatment and/or detailed descriptions of specific treatments. Other researchers performing the treatments can then check and see whether they obtain comparable effects. Effects of treatment outcome should at least assess reduction in stuttering frequency and preferably the influence of treatment on severity [c].
- (2) Measures must be reliable and valid. Methods for assessing reliability and validity need to be evaluated separately and reported in studies [a].
- (3) Sampling and satisfactory analysis of speech measures (which is essential for reliability and consistency). Although it is often recommended that a range of samples from different contexts be obtained (see number 13, below), this may be impractical at times, and fewer data analysed properly are more valuable than a large data set analysed badly [h, i].
- (4) Accurate data analysis is essential. Valid data have little or no value when analysed inappropriately [h, i].

Treatment integrity

- (5) Adequate supervision and training should be provided to clinicians to ensure the reliability of administration of the treatment [d].
- (6) Empirical evidence should be collected to establish that a treatment was administered properly, as diverse administrations of treatment can affect outcome. This criterion assumes that treatment procedures were properly described [e]. (Note: Subjects' adherence, as Moscicki mentions, can be checked only if subjects are self-administering treatments.)

Design

- (7) Adequate sample size is necessary. Few current studies include effect–size statistics, which needs to be remedied in the near future. Single-subject experiments may be appropriate for individualized treatments. Group studies may best be done after single-subject experiments demonstrate that a procedure produces desirable effects [b].
- (8) Control groups need to be appropriately matched on key variables (e.g., factors known to be associated with spontaneous recovery) [b].

- (9) Potential confounding factors in the treatment design must be identified, monitored, and reduced or eliminated when possible. Such factors can gravely affect the outcome of results. Treatments programs need to be carefully examined in terms of the impact of different rewards and punishments, avoidance behaviors, impact of attrition on outcome, and effects of individual treatments in combined treatment programs [d, e, h, i].
- (10) Follow-up assessments should be conducted for at least a year after treatment, especially for children. If treatment is unsuccessful, however, it may be unethical to hinder a client from trying other treatments in the long term (Hancock et al., 1998) [g].
- (11) Stable, pretreatment baselines or matched or randomly assigned control group data are needed. If the need for treatment is believed to be urgent, collecting baseline data may not be ethical.

Desirable but not essential features

- (12) Measures of intrajudge reliability should be collected, which should show higher reliability than interjudge measures. On occasions, it may not be possible or practical to have the same judge, for example, when there are larger samples. Single-subject studies should employ intrajudge assessments, unless more than one judge participates in the study. If interjudge measures are used, it is important to focus on the trends or patterns of measures, rather than specific numbers.
- (13) Speech assessments should be obtained under multiple conditions and on various occasions. This may not be feasible if particular equipment can be used only in laboratory settings or if parents are not able to bring a child to the clinic, necessitating clinicians to visit a child at home.

This review supports the need for all but one of Moscicki's criteria (i.e., pilot studies are not warranted). Also, different emphases are placed on some criteria, such as the need to document adherence to treatments or to examine avoidance behaviors. Some criteria are always met to the letter but not necessarily the spirit of the criterion. For example, describing how treatment is provided and its outcomes is covered in all studies, but only a few invested effort to fully meet these criteria. Reliability varies substantially across studies with some reporting reliability statistics for overall number of stutters and others individual stuttering events. Relatively little work has been done on intrajudge rather than interjudge reliability, which may be problematic because a high proportion of studies report judgements from a single judge, suggesting that intrajudge measures are more appropriate, but may also report interjudge reliability on subsets of samples. Similar problems occur in reports of the statistics that most, but not all studies, include at present. For instance, it is not always clear if researchers have checked their data to see that the assumptions for parametric tests are met. Studies can be

divided into those that use designs that are appropriate for single-subject or groups. For reasons that will be apparent later, we agree with Moscicki's (1993) preference for group designs. Some of the criteria may inter-relate. For example, ethical considerations may encourage early reports about promising investigations on small samples, which may outweigh statistical criteria.

Table 1

Summary of whether eight studies since 1993 (identifiable by the letters along the top) met the criteria along the side (y) or not (n) when the criterion is appropriate for evaluating a study (NA indicates that a criterion was not appropriate)

Study	a	b	c	d	e	f	g	h
(1) Define treatment outcome	y	y	y	y	y	y	y	y
(2) Reliability/validity of measure	n	y	y	y ^a	y	n ^b	y	y
(3) No source of bias in analysis	y	y ^c	y	y	y	n ^b	y	y
(4) Data analysis appropriate	– ^d	y	y ^d	y	y	– ^{b,d}	y	y
(5) Describe treatment	y	y	y	y	y	y	y	y
(6) Check adherence	y	y	NA	y	y	NA	y	y
(7) Sample size	5	15	2	16	97	3	24	12
	1 case		1 case			1 case		
(8) Control data ^e	n	y?	y	y	y ^f	y	y?	y
(9) No source of bias in design	y	y	y	y	y ^f	y	y	y
(10) Follow up	y	y	y	y	y	y	y	y
(11) Baseline	y	y	y	y	y	y	y	y

The studies are (a) Eichstadt et al. (1998), (b) Druce et al. (1997), (c) Craig and Kearns (1995), (d) Stager, Ludlow, Gordon, Cotelingam, and Rapoport (1995), (e) Hancock et al. (1998), (f) Ingham et al. (1997),

(g) Ryan and Ryan (1995), and (h) Onslow et al. (1996).

^a A criterion was set for two judges to meet. If they did not, the speech was reassessed.

^b The time interval measure used in study (f) does not provide a valid estimate of stuttering (Howell, Staveley, Sackin, & Rustin, 1998).

^c Study (b) included a high number of bilinguals, though analysis indicated that this had little effect.

^d Reporting no statistics in single case studies may be an overstrict criterion, so “–” is added to studies (a) and (f). Craig and Kearns (1995) (study c), however, did perform statistical analyses. They found significant effects of treatment that they dismissed (raising the possibility of a Type II error).

^e “Control data” means different things in different designs — Moscicki regards this as coming from nontreated individuals as she considered randomized control designs. (This criterion was not met by Hancock et al., 1998 [study e] and then only in the initial phase — see footnote f.) In the Stager et al. (1995) drug study (d), control was appropriate (a placebo phase was followed by a phase in which one of two drugs was administered, an active experimental drug and nonactive control drug, and the experimenters were blind as to which drug a subject was receiving). In single case studies (c) and (f), the control condition was the period when no treatment was administered between periods where treatment was given (control was not satisfactory in study a). In studies (b), (f), and (h), a control derived from a single baseline measure (e.g., as in study h) may not be satisfactory, as people who stutter may seek help when their problem is worst. Nevertheless, at this time, a single baseline control was categorised adequate. When the length of time between diagnosis and treatment is not specified, this concern is signified by a “?” after y.

^f The Hancock et al. study was a 4-year follow-up study. This made for some ethical problems about what to do with the controls, as the authors recognized. They did not deny subjects who had initially served as controls subsequent treatment in the follow-up period. This seems justifiable from a pragmatic perspective (bearing in mind the length of the follow-up period).

Having voiced these concerns, there is no reason to avoid the issue of assessing how adequately studies published since 1993 have met the revamped criteria. Table 1 summarises assessment of eight studies with respect to the first 11 of our criteria that were listed above. The studies were obtained from the *Journal of Fluency Disorders* and *Journal of Speech, Language, and Hearing Research* between 1993 and 2000. A total of nine were found, one of which was a follow-up of a study that appeared earlier during this period, so the earlier one was not included (i.e., Hancock et al., 1998, was used instead of Craig et al., 1996). The studies are: (a) Eichstadt, Watt, and Gibson (1998); (b) Druce et al. (1997); (c) Craig and Kearns (1995); (d) Stager et al. (1995); (e) Hancock et al. (1998); (f) Ingham, Moglia, Frank, Ingham, and Cordes (1997); (g) Ryan and Van Kirk Ryan (1995); (h) Onslow, Costa, Andrews, Harrison, and Packman (1996). Note that the countries in which these studies were performed are limited to Australia, South Africa, and the US.. Evaluating them with the criteria we proposed is somewhat subjective for the reasons discussed above. A minimum criterion of acceptability was used for all criteria. For instance, any indication that judges' reliability was measured was deemed satisfactory, implicit definitions of treatment outcomes were considered acceptable, and if statistics were reported, they were considered appropriate even if there were no indications that there had been a check that the assumptions of the test had or had not been met. Clearly, there is an urgent need to tighten the acceptability criterion before the next review appears. In Table 1, we present a breakdown of whether each study met (y) or did not meet (n) each criterion. If a criterion was not appropriate for assessing a study (e.g., sample size in a single subject design, NA indicates that the criterion was not applicable for assessing the study). In single case studies (i.e., a and f), statistics were not reported (criterion 4), but descriptive data were given. Statistical analysis is possible in single-subject studies (e.g., time series analyses if there are enough data points or nonparametric randomization tests if there are fewer data). Few studies of conditioning treatments do this, so rather than indicating that such studies had not met the criterion, a “–” was entered. The principal finding of this analysis is that a high number of studies met the minimal criteria. So, with the proviso of minimum criteria being used, the standards of efficacy research published since Moscicki's (1993) review seems high. One other factor, not apparent in the table, is that the Ryan and Van Kirk Ryan (1995) was a replication of a previous report of their own work.

6. Future directions

Few disorders pose greater challenges to the assessment of treatment efficacy than does stuttering (Curlee, 1993), and it is important to note that this review has addressed only some of the key determinants for efficacy evaluations. For example, there is little discussion of sampling speech measures, data, or power analyses (Murphy & Myors, 1999), parents as treatment components or admin-

isters of treatment, self-administration of treatment, or the influence of nontreatment variables on treatment effectiveness (Yairi, 1997). It is evident that methodological considerations are fundamental for documenting treatment efficacy. The issues discussed illustrate a significant point, however: A considerable amount of work is needed to refine, document, and assess treatment procedures and their efficacy. In most cases, treatment research has not provided rigorous experimental evidence (Zebrowski & Conture, 1998). Furthermore, as Onslow (1996) noted, posttreatment periods of at least 1 year are the minimum for adequate outcome evaluation and few studies met this methodological standard for examining long-term treatment effects. In addition, Conture and Wolk (1990) noted that long-term treatment effectiveness research is long overdue, especially studies of the relationship of specific behavioral and attitude changes to the changes in stuttering.

Baer (1990) suggested that treatment research would be improved if clinical researchers discerned what clients' main complaints or concerns are. It is not until this is done that variables that need changing can be identified (e.g., children may want to reduce feelings of self-consciousness). A number of measures (i.e., stuttering frequency, speech rate, abnormal speech quality) may not evidence significant differences between stuttered speech and normally fluent speech of stutterers and nonstutterers, respectively (Ingham & Cordes, 1997). Thus, assessment of stuttering might benefit from a renewed focus on self-judged measures. On the whole, however, the need for well-controlled research aimed at developing and investigating the success of stuttering treatment is evident, including variables not directly involved in treatment being considered in parallel with treatment variables.

Many treatment programs employ several treatment approaches, such as the Monterey Fluency Program (Ryan & Van Kirk Ryan, 1995) and Stuttering Intervention Program (Pindzola, 1999). Therefore, when evaluating their efficacy, the contribution of each treatment component should be determined in order to evaluate their combined effects. Although systematic replications are missing from the stuttering literature on treatment efficacy, Attanasio (1994) suggested that this reflects, in part, the reluctance of some journals to publish replication studies and an unwillingness to give journal space to unsuccessful or negative treatment findings. It would be a positive move for such publication practices to modify such attitudes as a means of encouraging clinicians to focus on sound methodological treatment efficacy issues in their work. There has been little systematic study of treatment failures, which could, in fact, provide crucial information for improving the efficacy of stuttering treatments. If a treatment is not effective, there is a need to know if the fault lies with treatment procedures, clinician's application of those procedures, client's practice habits, motivation, or other variables.

Initiation of a systematic program of treatment efficacy research of stuttering treatments presents an important challenge and is one that the field must embrace in order to advance (Moscicki, 1993). It is critical that standards are

adopted for evaluating treatment outcomes so that significant variations in outcomes can be reliably identified. On the whole, studies should focus on identifying specific therapy procedures that contribute the most to successful treatment outcomes as well as variables that are responsible for treatment failures. Children and adults who stutter deserve nothing less than rigorously tested and empirically supported treatments (Cordes, 1998); therefore, future research needs to incorporate these basic principles in their designs and to identify new critical variables for study if sounder treatment efficacy evaluations are to become available.

References

- Andrews, G., & Craig, A. (1982). Stuttering: overt and covert measurement of the speech of treated subjects. *Journal of Speech and Hearing Disorders*, 47, 96–99.
- Andrews, G., & Harris, M. (1964). *The syndrome of stuttering*. London: Heinemann.
- Andrews, G., & Harvey, R. (1981). Regression to the mean in pre-treatment measures of stuttering. *Journal of Speech and Hearing Disorders*, 46, 204–207.
- Ansel, B. (1993). Treatment efficacy research in stuttering. *Journal of Fluency Disorders*, 18, 121–123.
- Attanasio, J. S. (1994). Inferential statistics and treatment efficacy studies in communication disorders. *Journal of Speech and Hearing Research*, 37, 755–759.
- Azrin, N. H., & Nunn, R. G. (1974). A rapid method of eliminating stuttering by a regulated breathing approach. *Behaviour Research and Therapy*, 12, 279–286.
- Baer, D. M. (1990). The critical issue in treatment efficacy is knowing why treatment was applied. In: L. B. Olswang, C. K. Thompson, S. F. Warren, & N. J. Minghetti (Eds.), *Treatment efficacy research in communication disorders* (pp. 31–39) (Rockville, MD).
- Bandura, A. (1969). *Principles of behavior modification*. New York: Holt, Rhinehart, and Winston.
- Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs*. New York: Pergamon.
- Bloodstein, O. (1987). *A handbook on stuttering* (4th ed.). Chicago: National Easter Seal Society.
- Bloodstein, O. (1990). On pluttering, skivvering and floggering: a commentary. *Journal of Speech and Hearing Disorders*, 55, 392–393.
- Bloodstein, O. (1995). *A handbook on stuttering* (5th ed.). Chicago: National Easter Seal Society.
- Boberg, E., & Sawyer, L. (1977). The maintenance of fluency following intensive therapy. *Human Communication*, 2, 21–28.
- Casteel, R. L., & McMahon, R. L. (1978). The modification of stuttering in a public school setting. *Journal of Childhood Communication Disorders*, 2, 6–17.
- Conture, E. G. (1996). Treatment efficacy: stuttering. *Journal of Speech and Hearing Research*, 39, S18–S26.
- Conture, E. G. (1997). Evaluating childhood stuttering. In: R. F. Curlee, & G. M. Siegel (Eds.), *Nature and treatment of stuttering: new directions* (2nd ed., pp. 239–256). Needham Heights, MA: Allyn and Bacon.
- Conture, E. G., & Guitar, B. E. (1993). Evaluating efficacy of treatment of stuttering: school-age children. *Journal of Fluency Disorders*, 18, 253–287.
- Conture, E., & Wolk, L. (1990). Efficacy of intervention by speech–language pathologists: stuttering. *Seminars in Speech and Language*, 11, 200–211.
- Cooper, E. B. (1987). The chronic perservatives stuttering syndrome; incurable stuttering. *Journal of Fluency Disorders*, 12, 381–388.
- Cooper, J. A. (1990). Research needs in stuttering: roadblocks and future directions. *ASHA report 18*. Rockville, MD: ASHA.

- Cordes, A. (1994). The reliability of observational data: I. Theories and methods for speech–language pathology. *Journal of Speech and Hearing Research*, 37, 264–278.
- Cordes, A. (1998). Current status of the stuttering treatment literature. In: A. K. Cordes, & R. J. Ingham (Eds.), *Treatment efficacy for stuttering: a search for empirical bases*. San Diego: Singular Publishing Press.
- Cordes, A. K., & Ingham, R. J. (1998). *Treatment efficacy for stuttering: a search for empirical bases*. San Diego: Singular Publishing Group.
- Cordes, A. K., Ingham, R. J., Frank, P., & Ingham, J. C. (1992). Time interval analysis of interjudge and intrajudge agreement for stuttering event judgements. *Journal of Speech and Hearing Research*, 35, 483–494.
- Costello, J. M. (1975). Time-out procedures for the modification of stuttering: three case studies. *Journal of Speech and Hearing Disorders*, 40, 216–231.
- Costello, J. M., & Ingham, R. J. (1984). Assessment strategies for stuttering. In: R. Curlee, & W. H. Perkins (Eds.), *Nature and treatment of stuttering: new directions* (pp. 303–333). San Diego: College-Hill Press.
- Craig, A., Chang, E., & Hancock, K. (1992). Treatment success for children who stutter: a critical review. *Australian Journal of Human Communication Disorders*, 20, 81–92.
- Craig, A., Hancock, K., Chang, E., McCreedy, 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 and Hearing Research*, 39, 808–826.
- Craig, A. R., & Kearns, M. (1995). Results of a traditional acupuncture intervention for stuttering. *Journal of Speech and Hearing Research*, 38, 572–578.
- Culp, D. (1984). The preschool fluency development program: assessment and treatment. In: M. Peins (Ed.), *Contemporary approaches to stuttering therapy* (pp. 39–71). Boston: Little, Brown, and Company.
- Curlee, R. (1981). Observer agreement on stuttering and disfluency. *Journal of Speech and Hearing Research*, 24, 595–600.
- Curlee, R. F. (1993). Evaluating treatment efficacy for adults: assessment of stuttering disability. *Journal of Fluency Disorders*, 18, 319–331.
- Curlee, R. F., & Yairi, E. (1997). Early intervention with early childhood stuttering: a critical examination of the data. *American Journal of Speech–Language Pathology*, 6, 8–18.
- Debney, S., & Druce, T. (1987). *Intensive fluency programme long-term follow-up*. Poster presentation ASHA conference.
- Druce, T., Debney, S., & Byrt, T. (1997). Evaluation of an intensive treatment program for stuttering in young children. *Journal of Fluency Disorders*, 22, 169–186.
- Eichstadt, A., Watt, N., & Gibson, J. (1998). Evaluation of the efficacy of a stutter modification program with particular reference to two new measures of secondary behaviours and control of stuttering. *Journal of Fluency Disorders*, 23, 231–246.
- Felsenfeld, S. (1997). Epidemiology and genetics of stuttering. In: R. F. Curlee, & G. M. Siegel (Eds.), *Nature and treatment of stuttering: new directions* (2nd ed., pp. 3–23). Needham Heights, MA: Allyn and Bacon.
- Fosnot, S. M. (1993). Research design for examining treatment efficacy in fluency disorders. *Journal of Fluency Disorders*, 18, 221–251.
- Gregory, H. H., & Hill, D. (1980). Stuttering therapy for children. *Seminars in Speech, Language and Hearing*, 1, 351–364.
- Grunwell, P. (1981). The development of phonology. *First Language*, 2, 161–191.
- Ham, R. E. (1989). What are we measuring? *Journal of Fluency Disorders*, 14, 231–243.
- Hancock, K., Craig, A., McCreedy, C., McCaul, A., Costello, D., Campbell, K., & Gilmore, G. (1998). Two- to six-year controlled-trial stuttering outcomes for children and adolescents. *Journal of Speech, Language and Hearing Research*, 41, 1242–1252.
- Hanna, R., & Owen, N. (1977). Facilitating transfer and maintenance of fluency in stuttering therapy. *Journal of Speech and Hearing Disorders*, 42, 65–76.

- Howell, P., Staveley, A., Sackin, S., & Rustin, L. (1998). Methods of interval selection, presence of noise and their effects on detectability of repetitions and prolongations. *Journal of the Acoustical Society of America*, *104*, 3558–3567.
- Howie, P. M., Tanner, S., & Andrews, G. (1981). Short- and long-term outcome in an intensive treatment program for adult stutterers. *Journal of Speech and Hearing Disorders*, *46*, 104–109.
- Howie, P. M., Woods, C. L., & Andrews, G. (1982). Relationship between covert and overt speech measures immediately before and immediately after stuttering treatment. *Journal of Speech and Hearing Disorders*, *47*, 419–422.
- Ingham, J. C., & Riley, G. (1998). Guidelines for documentation of treatment efficacy for young children who stutter. *Journal of Speech, Language and Hearing Research*, *41*, 753–770.
- Ingham, R. J. (1972). A comparison of covert and overt assessment procedures in stuttering therapy outcome evaluation. *Journal of Speech and Hearing Research*, *18*, 346–354.
- Ingham, R. J. (1990). Stuttering. In: A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (pp. 599–631). New York: Plenum.
- Ingham, R. J. (1993). Stuttering treatment efficacy: paradigm-dependent or -independent. *Journal of Fluency Disorders*, *18*, 133–149.
- Ingham, R. J., & Andrews, G. (1973). Behavior therapy and stuttering: a review. *Journal of Speech and Hearing Disorders*, *38*, 405–441.
- Ingham, R. J., & Cordes, A. K. (1992). Interclinic differences in stuttering-event counts. *Journal of Fluency Disorders*, *17*, 171–176.
- Ingham, R. J., & Cordes, A. K. (1997). Self-measurement and evaluating stuttering treatment efficacy. In: R. F. Curlee, & G. M. Siegel (Eds.), *Nature and treatment of stuttering: new directions* (2nd ed., pp. 413–438). Needham Heights, MA: Allyn and Bacon.
- Ingham, R. J., & Costello, J. M. (1984). Stuttering treatment outcome evaluation. In: J. M. Costello (Ed.), *Speech disorders in children: recent advances* (pp. 313–346). San Diego: College-Hill Press.
- Ingham, R. J., Moglia, R. A., Frank, P., Ingham, J. C., & Cordes, A. K. (1997). Experimental investigation of the effects of frequency-altered auditory feedback on the speech of adults who stutter. *Journal of Speech, Language and Hearing Research*, *40*, 361–372.
- Kazdin, A. E. (1978). *History of behaviour modification*. Baltimore: University Park Press.
- Kazdin, A. E., & Kendall, P. C. (1998). Current progress and future plans for developing effective treatments: comments and perspectives. *Journal of Clinical Child Psychology*, *27*, 217–226.
- Kelly, E. M. (1993). Speech rates and turn-taking behaviors of children who stutter and their parents. *Seminars in Speech and Language*, *14*, 203–214.
- Kelly, E. M., & Conture, E. G. (1992). Speaking rates, response time latencies, and interrupting behaviours of young stutterers, nonstutterers, and their mothers. *Journal of Speech and Hearing Research*, *35*, 1256–1267.
- Kroll, R., & O'Keefe, B. (1985). Molecular self-analyses of stuttered speech via speech time expansion. *Journal of Fluency Disorders*, *10*, 93–105.
- Kully, D., & Boberg, E. (1988). An investigation of interclinic agreement in the identification of fluent and stuttered syllables. *Journal of Fluency Disorders*, *13*, 309–318.
- 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.
- Lanyon, R. I., Lanyon, B. P., & Goldsworthy, R. J. (1979). Outcome predictors in the behavioural treatment of stuttering. *Journal of Fluency Disorders*, *4*, 131–139.
- Last, J. M. (1983). *A dictionary of epidemiology*. New York: Oxford University Press.
- Lewis, K. E. (1994). Reporting observer agreement on stuttering event judgments: a survey and evaluation of current practice. *Journal of Fluency Disorders*, *19*, 269–284.
- Manning, W. H., Trutna, P. A., & Shaw, C. K. (1976). Verbal versus tangible reward for children who stutter. *Journal of Speech and Hearing Disorders*, *41*, 52–62.

- Martin, R. R., & Haroldson, S. K. (1981). Stuttering identification: standard definition and moment of stuttering. *Journal of Speech and Hearing Research, 24*, 59–63.
- Martin, R. R., Haroldson, S. K., & Triden, K. A. (1984). Stuttering and speech naturalness. *Journal of Speech and Hearing Disorders, 49*, 53–58.
- Martin, R. R., Kuhl, P., & Haroldson, S. (1972). An experimental treatment with two preschool stuttering children. *Journal of Speech, Language and Hearing Research, 15*, 743–752.
- Meline, T., & Schmidt, J. F. (1997). Case studies for evaluating statistical significance in group designs. *American Journal of Speech–Language Pathology, 6*, 33–41.
- Meyers, S. (1990). Tempest is a *t* test? A reply to Finn and Gow. *Journal of Speech and Hearing Disorders, 55*, 173–174.
- Moscicki, E. K. (1993). Fundamental methodological consideration in controlled clinical trials. *Journals of Fluency Disorders, 18*, 183–196.
- Muma, J. R. (1993). The need for replication. *Journal of Speech and Hearing Research, 36*, 927–930.
- Murphy, K. R., & Myers, B. (1999). Testing the hypothesis that treatments have negligible effects: minimum-effect tests in the general linear model. *Journal of Applied Psychology, 84*, 234–248.
- Olswang, L. B. (1993). Treatment efficacy research: a paradigm for investigating clinical practice and theory. *Journal of Fluency Disorders, 18*, 125–131.
- Onslow, M. (1992). Choosing a treatment procedure for early stuttering: issues and future directions. *Journal of Speech and Hearing Research, 35*, 983–993.
- Onslow, M. (1996). *Behavioral management of stuttering*. London: Singular Publishing Group.
- Onslow, M., Adams, R., & Ingham, R. (1992). Reliability of speech naturalness ratings of stuttered speech during treatment. *Journal of Speech and Hearing Disorders, 35*, 994–1001.
- Onslow, M., Andrews, C., & Lincoln, M. (1994). A control/experimental trial of operant treatment for early stuttering. *Journal of Speech and Hearing Research, 37*, 1244–1259.
- Onslow, M., Costa, L., Andrews, C., Harrison, E., & 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. *Journal of Speech and Hearing Disorders, 55*, 405–416.
- Packman, A., Ingham, R. J., & Onslow, M. (1993). *Reliability of listeners' stuttering counts: the effect of instructions designed to reduce ambiguity*. Submitted for publication.
- Pinzola, R. (1999). The stuttering intervention program. In: M. Onslow, & A. Packman (Eds.), *The handbook of early stuttering intervention*. San Diego: Singular Publishing Group.
- Prins, D. (1983). Continuity, fragmentation and tension: hypothesis applied to evaluation and intervention with pre-school disfluent children. In: D. Prins, & R. J. Ingham (Eds.), *Treatment of stuttering in early childhood: methods and issues* (pp. 21–42). San Diego: College-Hill Press.
- Purser, H. (1987). The psychology of treatment evaluation studies. In: L. Rustin, H. Purser, & D. Rowley (Eds.), *Progress in the treatment of fluency disorders* (pp. 258–273). London: Taylor and Francis.
- Reed, C. G., & Godden, A. L. (1977). An experimental treatment using verbal punishment with two preschool stutters. *Journal of Fluency Disorders, 2*, 225–233.
- Rustin, L., & Cook, F. (1983). Intervention procedures for the disfluent child. In: P. Dalton (Ed.), *Approaches to the treatment of stuttering* (pp. 47–75). London: Croom Helm.
- Rustin, L., Ryan, B., & Ryan, B. (1987). Use of the Monterey programmed stuttering treatment in Great Britain. *British Journal of Disorders of Communication, 22*, 151–162.
- Ryan, B. (1971). Operant procedures applied to stuttering treatment for children. *Journal of Speech and Hearing Disorders, 36*, 264–280.
- Ryan, B. P., & Van Kirk Ryan, B. (1983). Programmed stuttered therapy for children: comparison of four established programs. *Journal of Fluency Disorders, 8*, 291–321.
- Ryan, B. P., & Van Kirk Ryan, B. (1995). Programmed stuttering treatment for children: comparison of two establishment programs through transfer, maintenance, and follow-up. *Journal of Speech and Hearing Research, 38*, 61–75.

- Schiavettii, N., & Metz, D. (1997). *Evaluating research in speech pathology and audiology* (3rd ed.). Boston: Allyn and Bacon.
- Schulze, H. (1991). Time pressure variables in the verbal parent–child interaction patterns of fathers and mothers of stuttering, phonologically disordered and normal preschool children. In: H. F. M. Peters, W. Hulstijn, & C. W. Starkweather (Eds.), *Speech motor control and stuttering* (pp. 441–451). New York: Elsevier.
- Schwartz, M. F. (1976). *Stuttering solved*. New York: McGraw-Hill.
- Siegel, G. M. (1990). Concluding remarks. In: L. B. Olswang, C. K. Thompson, S. S. Warren, & N. J. Mingetti (Eds.), *Treatment efficacy research in communication disorders*. Rockville, MD: American Speech–Language–Hearing Foundation.
- Smith, A. (1990). Toward a comprehensive theory of stuttering: a commentary. *Journal of Speech and Hearing Disorders, 55*, 398–401.
- Stager, S. V., Ludlow, C. L., Gordon, C. T., Cotelingam, M., & Rapoport, J. L. (1995). Fluency changes in persons who stutter following a double blind trial of clomipramine and desipramine. *Journal of Speech and Hearing Research, 38*, 516–525.
- Starkweather, C. W. (1987). *Fluency and stuttering*. Englewood Cliffs, NJ: Prentice-Hall.
- Starkweather, C. W. (1993). Issues in the efficacy of treatment for fluency disorders. *Journal of Fluency Disorders, 18*, 151–168.
- Starkweather, C. W., & Gottwald, S. R. (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.
- Venty, I. M., & Schiavetti, N. (1986). *Evaluating research in speech pathology and audiology* (2nd ed.). New York: Macmillan.
- Vihman, M., & Greenlee, M. (1987). Individual differences in phonological development: ages one to three years. *Journal of Speech and Hearing Research, 30*, 503–521.
- Woods, D. W., Fuqua, R. W., & Waltz, T. J. (1997). Evaluation and elimination of an avoidance response in a child who stutters: a case study. *Journal of Fluency Disorders, 22*, 287–297.
- Yairi, E. (1993). Epidemiological and other considerations in treatment efficacy research with preschool children who stutter. *Journal of Fluency Disorders, 18*, 197–219.
- Yairi, E. (1997). Disfluency characteristics of childhood stuttering. In: R. F. Siegel, & G. M. Siegel (Eds.), *Nature and treatment of stuttering: new directions* (2nd ed., pp. 49–78). London: Allyn and Bacon.
- Yairi, E., & Ambrose, N. (1992). A longitudinal study of stuttering in children: a preliminary report. *Journal of Speech and Hearing Research, 36*, 521–528.
- 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. S. (1997). Clinical implications of situational variability in preschool children who stutter. *Journal of Fluency Disorders, 22*, 187–203.
- Yaruss, J. S. (1998). Describing the consequences of disorders: stuttering and the international classification of impairments, disabilities and handicaps. *Journal of Speech, Language and Hearing Research, 41*, 249–257.
- Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity and effectiveness. *Journal of Consulting and Clinical Psychology, 49*, 156–176.
- Young, M. A. (1975a). Onset, prevalence, and recovery from stuttering. *Journal of Speech and Hearing Research, 40*, 49–58.
- Young, M. A. (1975b). Observer agreement for marking moments of stuttering. *Journal of Speech and Hearing Research, 18*, 530–540.
- Young, M. A. (1993). Supplemental tests of statistical significance: variation accounted for. *Journal of Speech and Hearing Research, 36*, 644–656.
- Zebrowski, P. M., & Conture, E. G. (1998). Influence of nontreatment variables on treatment effectiveness for school-age children who stutter. In: A. K. Cordes, & R. J. Ingham (Eds.), *Treatment efficacy for stuttering: a search for empirical bases*. San Diego: Singular Publishing Press.

CONTINUING EDUCATION

Assessing efficacy of stuttering treatments

QUESTIONS

1. Lanyon, Lanyon & Goldsworthy's (1979) study suggested that:
 - a. Biofeedback was an effective form of treatment for stuttering
 - b. Patients should be given the MMPI prior to enrollment in fluency therapy
 - c. Objective outcome measures are necessary because subjective judgements of therapy outcomes may be controlled by a number of variables
 - d. The MMPI is not a valid instrument.
 - e. MMPI K scores can be used to measure therapeutic progress
2. What was Muma's estimate of Types I and II statistical errors in studies published in the decade 1979–1989?
 - a. as many as 1 in 2
 - b. as many as 1 in 3
 - c. as many as 1 in 4
 - d. as many as 1 in 5
 - e. as many as 1 in 6
3. Which of these studies is a replication of an earlier one?
 - a. Conture and Guitar (1993)
 - b. Druce et al. (1997)
 - c. Eichstadt et al. (1998)
 - d. Martin et al. (1984)
 - e. Grunwell (1981)
4. What is the minimum suggested period advocated for following-up clients after they have received stuttering treatment?
 - a. 6 months
 - b. 1 year
 - c. 2 years
 - d. 3 years
 - e. 4 years
5. Which of the following treatments was not investigated by Ryan and Van Kirk Ryan in their 1983 study?
 - a. Programmed traditional
 - b. Delayed auditory feedback
 - c. Controlled breathing
 - d. Time-out contingency
 - e. GILCU