International Journal of Speech-Language Pathology

ISSN: 1754-9507 (Print) 1754-9515 (Online) Journal homepage: informahealthcare.com/journals/iasl20

Assessing methodological quality of randomized and quasi-experimental trials: A summary of stuttering treatment research

Chad Nye & Debbie Hahs-Vaughn

To cite this article: Chad Nye & Debbie Hahs-Vaughn (2011) Assessing methodological quality of randomized and quasi-experimental trials: A summary of stuttering treatment research, International Journal of Speech-Language Pathology, 13:1, 49-60, DOI: 10.3109/17549507.2010.492873

To link to this article: https://doi.org/10.3109/17549507.2010.492873

Published online: 15 Feb 2011.

International Journal of Speech-Language Pathology, 2011; 13(1): 49–60

informa
healthcare

Assessing methodological quality of randomized and quasi-experimental trials: A summary of stuttering treatment research

CHAD NYE & DEBBIE HAHS-VAUGHN

University of Central Florida, Orlando, FL, USA

Abstract
The purpose of this study is to provide a detailed analysis of the methodological quality of experimental and quasi-experimental group designed studies in the area of stuttering intervention. A total of 23 randomized controlled trials (RCT) and quasi-experimental studies of treatment in the area of stuttering were identified and retrieved from an electronic search of nine databases and 13 individual journals. Using the Downs and Black Checklist each study was coded for reporting, external validity, internal validity, and internal validity confounding. Results of the coding indicated that while overall reporting was reasonably complete, the quality of the external and internal validity scores was found to be substantively incomplete. This lack of clarity and completeness of reporting issues related to the external and internal validity makes the interpretation of the findings of individual study results problematic and seriously effects the replicability of the individual study. Implications of these findings are suggested for both researchers and clinicians.

Keywords: *Evidence-based practice (EBP), intervention, validity.*

Introduction

The idea of answering the question of ''what works'' resonates today under the umbrella construct of evidence-based practice (EBP). The effort to demonstrate a scientific basis for clinical decisions permeates the clinical community's cry for defensible programs and methods of treatment for individuals with speech and language disabilities. Enderby and Emerson's (1995) book begs the argument that the professional community in the field of speech-language pathology must face issues such as the ethics of treatment, the quantity and quality of the evidence supporting speech and language therapy, the quality of the clinical services provided by the practicing SLP, the economics of treatment that speaks to the cost effectiveness of treatment, the impact of the professional education and training programs, and the research agenda supporting treatment efficacy and effectiveness.

This special issue dedicated to Dr Pam Enderby is an excellent way to focus attention on the profession that she has been committed to for nearly 40 years. A quick search of major electronic databases will retrieve a lengthy list of Pam's scholarly contributions across a host of disorders, professional issues, treatment effects, research quality, training programs, and in a variety of disciplines including medicine, speech-language pathology, mental health, education, and nursing. Underlying many of these works is a common clinical theme—''what works''! As we approach the 15th anniversary of Pam Enderby and Joyce Emerson's profound yet simple question for the speech-language pathology profession in their book title *Does Speech and Language Therapy Work? A Review of the Literature* (1995), we are reminded of the ultimate goal of the profession—to provide effective speech and language intervention to individuals with communication disorders. The foresight, clarity, and simplicity of that statement is today found in the daily conversations of speech-language pathologists (SLPs) around the world, is being written about in the professional journals and books, is the topic of myriad presentations, posters, symposia, and workshops, and drives a demand for professional practice guidelines. Certainly, Dr Enderby's contribution to the field of speech-language pathology is marked by close scrutiny of the research and scientific bases of the clinical treatment of individuals with speech and language disorders. It is in that spirit and inspiration that this paper will provide an assessment of the quality of research evidence in the area of stuttering intervention as presented in the peer reviewed literature. While we recognize that the quality of research issue extends to numerous other areas of speech-language pathology,

Correspondence: Chad Nye, PhD, Department of Communication Sciences & Disorders, University of Central Florida, Orlando, FL, USA.
E-mail: cnye@mail.ucf.edu

ISSN 1754-9507 print/ISSN 1754-9515 online
Published by Informa UK, Ltd.
DOI: 10.3109/17549507.2010.492873

we hope that this study will at least provide a measure of demonstration of a process for evaluation of research quality as well as summarize the available evidence for a specific area of intervention.

Background

The EBP movement has served to shine a spotlight on the disparity of available research focusing on the assessment of intervention effectiveness and need of clinicians to apply research to clinical decisions (Meline & Paradiso, 2003). In the US, the American Speech-Language Hearing Association established a National Center for Evidence-based Practice (NCEP) (2004), in the UK the Royal College of Speech Language Therapists (RCSLT) developed and published a compendium of the Clinical Guidelines (2005), both designed to serve as a source for practice decisions by SLPs. Both of these enterprises offer a baseline of professional and organizational support to bridge the research–clinic gap. The scientific underpinning of the EBP movement is the original research designed to test the impact of interventions under controlled conditions (Dollaghan, 2007; Hunt, 1998; Reilly, Douglas, & Oates, 2004). One of the tenants of the EBP movement is the use of high quality research as the basis for making policy, programmatic, and practice decisions. Just how the quality of any given study is determined is a matter of some discussion and dispute in professional circles and publications (Hegde, 2007; Reilly, 2004). West, King, Carey, Lohr, McKoy, Sutton, et al. (2002) presented a detailed and comprehensive summary of 40 different models or systems reported in the literature that provided for a scaled judgement of research quality. These 40 models/systems assessed systematic reviews, randomized controlled trials, observational studies, cohort studies, and diagnostic studies that could be graded for quality, quantity, and consistency. West et al. proposed a 10 domain categorization specifically to assess the group studies (e.g., RCT, observational, and cohort):

1) Study question
2) Study population
3) Randomization
4) Blinding
5) Interventions
6) Outcomes
7) Statistical analysis
8) Results
9) Discussion
10) Funding

These systems provide both the producers and consumers of primary and summary research with direction in assessing the nature and quality of the research focusing on intervention effectiveness. However, regardless of the system of analysis or grading, all of the systems seek to identify and evaluate the well established elements of research typically understood as critical to quality assessment of external validity, internal validity, or statistical validity (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). More to the point of this paper, we know of no assessment of the quality of research in the area of stuttering intervention that uses a defined system of analysis and synthesis.

Quality assessment in stuttering research

Several papers have discussed at length the methodological elements needed to construct, implement, and evaluate treatment research in the area of stuttering (e.g., Bothe, Davidow, Bramlett, Franic, & Ingham, 2006; Ingham & Andrews, 1973; Ingham & Lewis, 1978; Moscicki, 1993; Thomas & Howell, 2002). A few meta-analytic analyses have summarized the overall research quality of included studies (Andrews, Howie, & Guitar, 1980; Herder, Howard, Nye, & Vanryckeghem, 2006; Howard, Nye, Vanryckeghem, Schwartz, & Turner, 2006). Recently, Onslow, Jones, O'Brian, Menzies, and Packman (2008) presented a focused tutorial as an orientation and guide for clinicians to assess clinical trials in the area of stuttering to assist in understanding the scientific basis of stuttering interventions.

Given the increased interest in evidence-based practice and the need to address issues of research quality as a basis for making clinical judgements regarding the effects of intervention that lead to a causal conclusion, the purpose of this study is to provide a detailed analysis of the methodological quality of experimental and quasi-experimental group designed studies in the area of stuttering intervention.

Method

The method for this study followed a five-step process of (1) establishing inclusion criteria, (2) retrieving studies from selected journals, (3) evaluating each study for inclusion/exclusion criteria, (4) coding each study, and (5) quantitatively summarizing the findings.

Inclusion criteria

The following inclusion criteria were established prior to initiating the search for studies. In order to be included in the process of a quality evaluation, all studies were required to:

1) Include a group experimental or quasi-experimental research design (QED);
2) Present a comparison of one or more behavioural stuttering treatment programs;

3) Provide at least a post-treatment assessment of the impact of the treatment program on speech fluency; and
4) Be a peer-reviewed manuscript published in the professional literature.

Studies that were excluded included: (1) studies that assessed only attitudinal (e.g., treatment approval), affective (e.g., self-esteem), or personality outcomes (e.g., locus of control), (2) pre-experimental, single subject, longitudinal, or case studies, (3) studies that assessed treatment for cluttering, and (4) studies that reported only pharmacological interventions. If any study included both a pharmacological and behavioural treatment, the reported post-treatment measurements must have presented a separate summary for the behavioural participants in both experimental and control group conditions.

Information retrieval

A complete electronic search without a date of publication restriction was conducted using the nine databases presented in Table I. The electronic search used keywords selected to provide a broad basis for inclusion during the initial information retrieval stage. The following keywords were used to identify potentially appropriate studies for this study: (a) Domain Terms: Stutt*, Stam*; (b) Intervention Terms: Therap*, Treat*, Interven*; (c) Outcome: Stutt*, Fluen*, Dysfl*, Disflu*.

A combination of a hand-search (dating from the earliest issue through 2004) and electronic searches through 2008 was conducted for the following journals:

- American Journal of Speech-Language Pathology (AJSLP)
- Behavior Therapy (BT)
- Folia Phoniatrica et Logopaedica (FPL)
- Journal of Behavior Therapy and Experimental Psychiatry (JBTEP)
- Journal of Communication Disorders (JCD)
- Journal of Fluency Disorders (JFD)
- Journal of Medical Speech-Language Pathology (JMSP)
- Journal of Speech and Hearing Disorders (JSHD)
- Language, Speech, and Hearing Services in the Schools (LSHSS)
- Journal of Speech, Language, and Hearing Research (JSLHR) (includes previous name of Journal of Speech and Hearing Research (JSHR))
- Seminars in Speech and Language (SSL)

In addition, the reference lists of all included studies were scanned for citations of stuttering treatment not found in the electronic or hand search.

Table I. Databases and search engine supplier used to electronically search for stuttering treatment studies.

Database	Supplier
PsychINFO	EBSCO Host
Education Resource Information Center (ERIC)	EBSCO Host
MEDLINE	PubMed
Cumulative Index to Nursing and Allied Health Literature (CINAHL)	EBSCO Host
Cochrane Central Register of Controlled Trials	OVID
Communication & Mass Media	EBSCO Host
Canadian Education Index	CSA
FRANCIS	CSA
British Education Index	

Study characteristics coding

Study characteristics were coded for each study and included: treatment group mean age and age range, total sample size, design type (RCT or QED), research design (multiple treatment groups, etc.), journal source (e.g., JSLHR, AJSLP), and date of publication.

Study quality coding

In order to assess the research quality of the RCT and QED included studies, the 27 item Downs and Black (1998) checklist was selected. West et al. (2002) identified the Downs and Black checklist as being consistent with 18 other recommended quality assessment systems. In addition, Deeks, Dinnes, D'Amico, Sowden, Sakarovitch, Song, et al. (2003) reviewed 60 research design methodology evaluation systems and identified Downs and Black as one of the best evaluation systems available. The Downs and Black checklist provides an overall quality index and four-subscales of quality assessment including reporting, external quality, internal validity-bias, and internal validity-confounding. A rating for power is also provided in the checklist; however, for purposes of this study, the power item was excluded due to inadequate information in most studies that allowed for a power calculation.

Reporting. The report sub-scale includes the first 10 items of the checklist. These items address the completeness and adequacy of the information reported that may minimize potential bias in the interpretation of the results. The items include a clear statement of the purpose or goal of the study, *a priori* inclusion of outcomes to be assessed, participant characteristics, intervention description, recognition of potential confounders, response variability, recognition of adverse effects, accounting for characteristics of participant attrition, and precision of probability values for reported analyses.

External validity. Three items focus on issues related to the potential for generalization to the population at large. These items assess characteristics related to the adequacy of the selected sample to represent the population at both the recruitment and treatment phase of the study and the representativeness of the intervener and environment for purposes of generalization.

Internal validity bias. The seven items measuring the internal validity-bias sub-scale are used to assess research design factors that may account for the degree to which control or planning was exerted for intervention delivery, blinding to condition, fidelity of implementation, and the appropriate selection and measurement of outcomes.

Internal validity confounding. Confounding of results (six items) directs attention to the potential selection bias present in a study. The primary focus is on the degree of experimental control accounting for participant selection, allocation to treatment and control conditions, staff and participant blinding, and accounting for potentially confounding participant and attrition variables at the analysis stage of the study.

Total score. The total score is composed of the sum of the 26 previously mentioned items (i.e., the sum of the sub-scales or the individual sub-sets of quality categories).

Results

Descriptive summary of included studies

As seen in Figure 1, a total of 1348 titles and abstracts were identified in the electronic database and journal hand-search. A total of 973 citations were excluded due to duplicate citations, non-treatment studies, single subject reports, or pharmacological studies which reduced the number of potential studies to 375 for which full text documents were obtained for further analysis. A review of the full texts further reduced the number of studies to a total of 23 studies (see Appendix A) that met all four inclusion criteria.

Coding procedures

The coding for each study was conducted by the authors independently using the Downs and Black checklist. Upon completion of the coding, the reviewers' scoring of each study was compared. Any differences were resolved through discussion until a consensus judgement of the score to be applied to each item was reached. Coding agreement averaged 80.3% across 598 items coded (range = 72–100%).

The results for this study are presented first as a descriptive, general summary of the quality ratings for the individual items as well as the overall quality index and the four sub-scales. Next, differences in the overall quality index and the four sub-scales based on various study characteristics are examined.

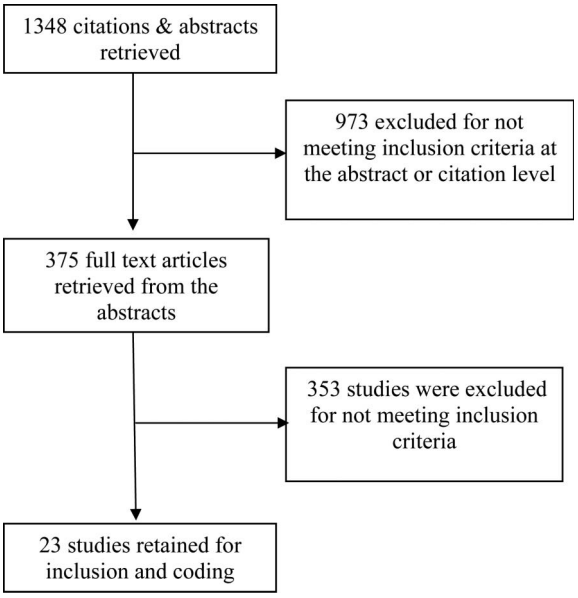

Figure 1. Decision tree for study inclusion criteria.

Methodological quality: General findings

The 23 studies were examined using the four quality sub-scales of the Down and Black (1997) checklist (reporting, external validity, internal validity-bias, and internal validity-confounding) along with the overall quality rating (i.e., all checklist items). For the overall quality index (i.e., all 26 items comprising all sub-scales), a maximum score of 26 was possible. For the 23 studies examined, the average overall quality index was 16.43 (SD = 3.70) with scores ranging from 10–23.

Reporting sub-scale. Table II presents the percentage (and frequency) of studies that met the quality indicator for each item in the reporting sub-scale. The percentage of studies that met the individual reporting sub-scale quality indicators ranged from a high of 100% for three items (main study outcomes, participant characteristics, and intervention and comparison conditions) to a low of 35% for one item (actual probability values for the main outcomes). With a maximum score of 10 possible on the reporting sub-scale, the average score was 8.13 (SD = 1.55; range 5–10). Slightly more than 25% of studies ($n = 6$, 26%) received the maximum score of 10.

External validity sub-scale. The assessment of external validity as a measure of research quality is shown in Table III, with the percentage (and frequency) of studies that met the quality indicator for each item in the external validity sub-scale. Less than one-third of the studies used methods to enrol study participants

Table II. Reporting quality sub-scale scores.

Quality indicator: Does study provide clear description of ...	Percentage (*n*) of studies meeting quality indicator
1. Hypothesis/aims/objectives	91% (21)
2. Main study outcomes	100% (23)
3. Participant characteristics	100% (23)
4. Intervention and comparison conditions	100% (23)
5. Distribution of confounders in each participant group	78% (18)
6. Main findings	83% (19)
7. Estimates of random variability in data for main outcomes	65% (15)
8. Adverse events	87% (20)
9. Characteristics of patients lost to follow-up	74% (17)
10. Actual probability values for the main outcomes	35% (8)

Table III. External validity quality sub-scale scores.

Quality indicator	Percentage (*n*) of studies meeting quality indicator
1. Participants recruited were representative of population	17% (4)
2. Participants who agreed to participate were representative of population	26% (6)
3. Study context (e.g., staff, facilities) representative of population	65% (15)

that ensured representativeness of the population ($n = 6$, 26%) and even fewer studies followed procedures that ensured representativeness of the recruited population ($n = 4$, 17%). With a maximum score of 3 possible on the external validity sub-scale, the average score was 1.09 (SD = .79; range 0–2). There were no studies that received the maximum score. Over one-quarter of the studies received a 0 for this sub-scale score ($n = 6$, 26%).

Internal validity-bias. The percentage (and frequency) of studies that met the quality indicator for each item in the internal validity-bias sub-scale are shown in Table IV. The percentage of study indicators meeting internal validity-bias quality ranged from a low of 4% (blinding of participants to intervention received) to a high of 96% (unplanned analyses were identified). The maximum score on the internal validity-bias sub-scale was 7, with the mean being 4.52 (SD = 1.08; range 2–7). There was one study that received the maximum score of 7 and nearly three-quarters of the studies ($n = 16$, 74%) meeting either four or five internal validity-bias quality indicator criteria.

Internal validity-confounding. Table V presents the percentage (and frequency) of studies that met the quality indicator for each item in the internal validity-confounding sub-scale. The percentage of study indicators meeting internal validity-confounding

Table IV. Internal validity-bias quality sub-scale scores.

Quality indicator	Percentage (*n*) of studies meeting quality indicator
1. Blinding of participants to intervention received	4% (1)
2. Blinding of assessors measuring main study outcomes	48% (11)
3. Unplanned analyses identified (no data dredging)	96% (22)
4. Analyses adjusts for different lengths of follow-ups	87% (20)
5. Statistical tests appropriate for main outcomes	83% (19)
6. Reliable compliance with intervention	44% (10)
7. Scores from main outcome measures were reliable and valid	91% (21)

Table V. Internal validity-confounding quality sub-scale scores.

Quality indicator	Percentage (*n*) of studies meeting quality indicator
1. Participants in different interventions recruited from same population	44% (10)
2. Participants in different interventions recruited during same time period	17% (4)
3. Participants randomly assigned to treatment conditions	65% (15)
4. Random assignment concealed from participants and staff	17% (4)
5. Statistical adjustments for confounding (i.e., intent to treat)	61% (14)
6. Loss of participants (i.e., attrition) addressed in analyses	65% (15)

quality ranged from a low of 17% on two items (participants in different interventions recruited during the same time period and random assignment concealed from participants and staff) to a high of 65% on two items (participants randomly assigned to treatment conditions and loss of participants addressed in analyses). The maximum score on the internal validity-bias sub-scale was 6, with an average score of 2.70 (SD = 1.72; range 0–6). There was one study that received the maximum score of 6 and four studies that received the minimum score of 0 (17%).

Methodological quality: Differences based on study characteristics

In addition to coding for the items in the Downs and Black (1998) checklist, studies were also coded for the following: research design (randomized controlled trial or quasi-experimental design), treatment focus (e.g., treatment vs control or treatment vs modified treatment), age classification (children or adult), year of publication, and journal. Approximately two-thirds of the studies ($n = 16$; 70%) were RCTs with the remaining being QEDs ($n = 7$; 30%). Nearly 40% of studies used a traditional treatment and control focus ($n = 9$, 39%). Slightly more than

one-half of the studies involved children ($n = 13$; 56%). Of the 15 studies that reported mean age of participants, the average age was 18 (SD = 10.95) with a minimum age of 4 and maximum of 33. Most of the studies were published in either the 1980s ($n = 8$, 35%) or in the current decade ($n = 7$, 30%), with the fewest studies published in the 1990s ($n = 3$, 13%). JSLHR ($n = 7$, 30%) and JFD ($n = 5$, 22%) were the journals where studies were most frequently published, followed by JBTEP ($n = 3$, 13%) and International Journal of Language and Communication Disorders (IJLCD) (includes previous name of British Journal of Disorders of Communication (BJDC)) ($n = 3$, 13%). Although most studies were published in JSHR and JSLHR, this reflects less than one-third of the studies that were published in one outlet. A summary of the study characteristics of the 23 studies reviewed are presented in Table VI.

Analyses were then conducted to determine if there were differences in methodological quality (i.e., reporting, external validity, internal validity-bias, internal validity-confounding, and overall quality index) based on various study characteristics (i.e., research design, treatment design, age group, year of publication, and journal). Analysis of the data was conducted using the non-parametric Kolmogorov-Smirnov Z (K-S) test for two group comparisons. Similar to the Mann-Whitney U-test, the K-S test has increased power as compared to the Mann-Whitney when samples sizes are less than 25 per group, as seen in this data. The effect size for the K-S test was r (calculated as $\frac{Z}{\sqrt{N}}$) (Rosenthal, 1991) and is interpreted as the change in the outcome given the predictor.

The non-parametric Kruskal-Wallis was used for multiple group comparisons. Each of the Kruskal-Wallis tests conducted involved four groups, thus each was based on three degrees of freedom. Eta-squared was used as the measure of effect size for the Kruskal-Wallis tests and provides an indication of the percentage of variation in the dependent variable attributed to the independent variable. Given that multiple tests were conducted, the Bonferroni adjustment was made in examining the results to control for the increase chance of a Type I error. Thus, rather than applying an alpha level of .05, results were compared to an alpha level of .01 (.05/5 tests). Although the results of the null hypothesis statistical tests are presented, due to the small sample size and resulting low power, interpretation of the findings emphasize the effect size results.

Methodological quality: Differences by research design

A Kolmogorov-Smirnov test was conducted to determine differences in methodological quality score based on research design (RCT and QED). The K-S test suggests there are no statistically significant differences in methodological quality sub-scale scores based on research design. However, there are statistically significant differences in the overall quality rating based on research design (K-S $Z = 1.75$, $p = .004$, $r = .36$) with RCTs having increased overall quality scores as compared to quasi-experimental designs. Examining the effect size r (calculated as $\frac{Z}{\sqrt{N}}$) and interpreting based on Cohen's (1988) guidelines (i.e., .1 = small effect, .3 = moderate effect, and .5 = large effect), a small to moderate effect was present for reporting internal validity-bias and a moderate effect was evident for internal validity-confounding and overall quality rating. Table VII presents the mean methodological quality ratings by research design and the results of the K-S test.

Table VI. Study characteristics (frequency and percentage).

Study characteristic		Frequency (%)*
Research design	Randomized controlled trial (RCT)	16 (70%)
	Quasi-experimental design (QED)	7 (30%)
Treatment focus	Treatment vs control	9 (39%)
	Two non-traditional treatment comparisons	6 (26%)
	Non-traditional treatment vs traditional treatment	5 (22%)
	Treatment vs modification of that same treatment	3 (13%)
Age classification	Children	13 (56%)
	Adult	10 (44%)
Year of publication	1969–1979	5 (22%)
	1980–1989	8 (35%)
	1990–1999	3 (13%)
	2000–2008	7 (30%)
Journal	Journal of Speech Language Hearing Research (JSLHR)	7 (30%)
	Journal of Fluency Disorders (JFD)	5 (22%)
	Journal of Behavior Therapy and Experimental Psychiatry (JBTEP)	3 (13%)
	International Journal of Speech-Language Pathology (IJSLP)	3 (13%)
	American Journal of Speech-Language Pathology (AJSLP)	1 (4%)
	British Medical Journal (BMJ)	1 (4%)
	Behavior Therapy (BT)	1 (4%)
	Journal of Communication Disorders (JCD)	1 (4%)
	Perceptual and Motor Skills (P&MS)	1 (4%)

*May not round to 100% due to rounding error.

Methodological quality: Differences by age group

A Kolmogorov-Smirnov *Z*-test was conducted to determine differences in methodological quality score based on age group (child vs adult). The results of the K-S test suggest there were no statistically significant differences in methodological quality sub-scale scores or the overall quality index based on age group Examining the effect size *r* (calculated as $\frac{Z}{\sqrt{N}}$) and interpreting based on Cohen's (1988) guidelines, there were generally small effects for the subscales and the overall quality index. Table VIII presents the mean methodological quality ratings by age group.

Methodological quality: Differences by year of publication

A Kruskal Wallis test was conducted to determine differences in quality score based on year of publication. The results of the test suggest there were no statistically significant differences in methodological quality scores for sub-scales or the overall quality index based on research design. Examining the effect size (eta squared) and interpreting using Cohen's (1988) guidelines (.01, small; .06, moderate; .14, large), a large effect is evident for all subscales and the overall quality rating (see Table IX).

Additional analyses were undertaken to examine methodological quality by year of publication based on the point at which evidence-based practices began to be recognized within the profession. Specifically, a Kolmogorov-Smirnov *Z*-test was conducted to determine differences in methodological quality by year of publication by grouping studies into two groups: (a) studies published between 1990–2008, and (b) studies published prior to 1990. The results of the K-S test suggest there were no statistically significant differences in methodological quality sub-scale scores or the overall quality index based on year of publication (reporting, K-S $Z = 1.12$, $p = .17$, $r = .23$; external validity, K-S $Z = .68$, $p = .75$, $r = .14$; internal validity-bias, K-S $Z = .75$, $p = .63$, $r = .16$; internal validity-confounding, K-S $Z = .46$, $p = .99$, $r = .10$; overall quality rating, K-S $Z = .93$, $p = .35$, $r = .19$). Examining the effect size *r* (calculated as $\frac{Z}{\sqrt{N}}$) and interpreting based on Cohen's (1988) guidelines, a small to moderate effect was evident for reporting and the overall quality rating.

Methodological quality: Differences by journal

Table IX presents the mean methodological quality ratings by journal of publication. Due to a number of journals being represented by only one study, comparisons were made only between JFD and JSLHR (JSHR), the two most frequently represented journals in the studies reviewed. The results of the K-S test suggest there were no statistically significant differences in methodological quality sub-scale scores or the overall quality index based year of publication (reporting, K-S $Z = .683$, $p = .74$, $r = .20$; external validity, K-S $Z = .488$, $p = .97$, $r = .14$; internal validity-bias, K-S $Z = .34$, $p = .34$, $r = .10$; internal validity-confounding, K-S $Z = .390$, $p = .998$, $r = .11$; overall quality rating, K-S $Z = .439$, $p = .99$, $r = .13$) (see Table X). Examining the effect size *r* and interpreting based on Cohen's (1988) guidelines, small effect sizes were evident with the exception of a small to moderate effect size for reporting.

Discussion

Interest in the application of clinical research has emerged with growing attention to primary research on issues of both internal and external validity as measures of research quality. Over the years, the components, nature, and role of various research methodological considerations in the development, implementation, and evaluation of clinical research

Table VII. Mean (SD) quality ratings by research design and K-S test results.

	Research design		Results		
Quality category (maximum score)	RCT ($n = 15$)	QED ($n = 7$)	K-S Z	p	r
Reporting (10)	8.69 (1.30)	6.86 (1.35)	1.16	.13	.24
External validity (3)	1.13 (.89)	1.00 (.58)	.65	.79	.14
Internal validity-bias (7)	4.88 (.89)	3.71 (1.11)	.95	.33	.20
Internal validity-confounding (6)	3.44 (1.31)	1.00 (1.29)	1.48	.03	.31
Quality index (26)	18.13 (2.73)	12.57 (2.57)	1.75	.004	.37

Table VIII. Mean (SD) quality ratings by age group and K-S test results.

	Age group		Analysis summary		
Quality category (maximum score)	Child ($n = 13$)	Adult ($n = 10$)	K-S Z	p	r
Reporting (10)	8.38 (1.61)	7.80 (1.48)	.64	.81	.13
External validity (3)	1.23 (.73)	.90 (.88)	.59	.88	.12
Internal validity-bias (7)	4.46 (.88)	4.60 (1.35)	.29	1.00	.06
Internal validity-confounding (6)	2.15 (1.57)	3.40 (1.71)	.81	.54	.17
Quality index (26)	16.23 (3.68)	16.70 (3.92)	.48	.98	.10

Table IX. Mean (SD) quality ratings by year of publication and Kruskal Wallis test results.

	Year of publication				Results		
Quality category (maximum score)	1969–1979 ($n=5$)	1980–1989 ($n=8$)	1990–1999 ($n=3$)	2000–2008 ($n=7$)	χ^2	p	η^2
Reporting (10)	7.40 (1.67)	7.50 (1.61)	8.00 (1.00)	9.43 (.79)	7.64	.05	.35
External validity (3)	.40 (.55)	1.13 (.83)	1.33 (.58)	1.43 (.79)	5.33	.15	.24
Internal validity-bias (7)	4.40 (.55)	4.38 (1.60)	4.00 (1.00)	5.00 (.58)	3.39	.34	.15
Internal validity-confounding (6)	3.20 (1.10)	2.88 (2.17)	.67 (1.15)	3.00 (1.29)	4.94	.18	.22
Quality index (26)	15.40 (1.67)	15.89 (5.11)	14.00 (3.00)	18.86 (1.95)	5.71	.13	.26

Table X. Mean (SD) quality ratings by journal.

	Journal			
Quality category (maximum score)	IJLCD & BJDC** ($n=3$)	JBT&EP** ($n=3$)	JFD ($n=5$)	JSLHR & JSHR** ($n=7$)
Reporting (10)	7.67 (2.08)	8.00 (1.00)	7.60 (2.07)	8.71 (1.11)
External validity (3)	.67 (.58)	.67 (1.15)	1.40 (.55)	1.00 (.82)
Internal validity-bias (7)	4.00 (1.73)	4.00 (0.00)	4.60 (1.14)	4.29 (.76)
Internal validity-confounding (6)	2.67 (2.52)	4.00 (1.00)	1.60 (1.14)	2.14 (1.77)
Quality index (26)	15.00 (5.58)	16.67 (2.08)	15.20 (3.96)	16.14 (3.44)

*Standard deviations are not reported for journals in which only one study is published.
**These journals were merged to form the first listed.

in the area of stuttering have been enumerated, recommended, and supported as important elements of clinical research in the field of speech-language pathology in general and in stuttering intervention specifically. The purpose of the present study was to conduct an assessment of the quality of research in the area of stuttering treatment with an accepted system of evaluation, namely the Downs and Black (1998) checklist. Specifically, we identified 23 studies that met the inclusion criteria for clinically controlled trials using the most rigorous group research designs, RCT and QED. A double coding system was conducted for each study and categories of research quality were summarized. The discussion of the findings and their implications will be organized as follows: Study Characteristics, Reporting Outcomes, External Validity Outcomes, Internal Validity Outcomes, Implications for Research, and Implications for Clinicians.

Study characteristics

While over one-half of the 23 studies retrieved for this paper were RCTs ($n=16$), the total number of studies was surprisingly modest. Given the long history and tradition of stuttering research in the field of speech-language pathology, to find only 23 total controlled clinical trials reporting stuttering treatment efficacy results was unexpected. We would argue that, at least as a general rule, it would seem that the research to practice gap has not been substantially impacted. The one bright spot in the research design characteristics is the number of studies reporting treatment efficacy. A total of six of the 13 child-focused studies employed the Lidcombe Program. Some critics might argue that the majority of these studies originated with researchers directly connected to the Lidcombe Program development and promotion, thus allowing for a potential reporting bias. We would suggest that while more studies generated by independent researchers are certainly desirable, the body of efficacy work using the Lidcombe Program is among the most rigorous and methodologically sound of any of the stuttering efficacy treatment research identified for this study. Further, we would suggest that it is up to other researchers to conduct the research needed to challenge or validate the existing findings of the Lidcombe Program. If treatment is the ultimate goal of the body of evidence in a discipline, it seems that there should be substantially more studies assessing the impact of a treatment than reflected in 23 studies over more than 75+ years of published research in the field of stuttering.

Another point of interest, the production/publication of efficacy research in stuttering treatment, suggests that the influence of evidence-based practice in the past decade has had little impact on clinical treatment production. As seen in Table VI, the number of studies published has remained relatively constant over the past five decades. Additionally the fact that the published work is available in at least 10 different professional journals and with no journal publishing more than $\sim$25% of all the included work suggests that the number of outlets for publication is quite varied, with no single source serving as the primary outlet.

Lastly, when we consider the study characteristics results, the presence and impact of study design provide substantial statements regarding both the quality and feasibility of the efficacy research. The fact that about two-thirds of the clinically controlled trials were RCTs and that the overall methodological quality scores were significantly superior to the QED

studies suggests that it is possible to conduct high quality gold standard research (i.e., RCT) in spite of the often heard complaints of the difficulty and ethics of experimental research.

Reporting outcomes

The results of the quality ratings for reporting suggest that studies were clear in their presentation of the basic participant characteristics, study outcomes, and descriptions of the intervention. This should not be understood to suggest that this level of reporting was sufficient for study replication, but the reporting of these characteristics at least met a minimal level of information in order to be able to assess the clinical application of the findings. For example, all the studies presented basic information on the participants of the study that included age and gender breakdowns for the experimental group. In some cases the information was more specific such as mean ages and/or age ranges of each group while in other studies these data were also broken down by gender or treatment/control group status.

The issues that received the lowest reporting quality scores for the 23 studies were typically those that addressed design quality factors such as participant variability, probability levels, and participant attrition. The finding that one-third of the studies did not report estimates of variability (e.g., standard deviation) coupled with the reporting of actual probability values for the main outcomes by only 32% of the studies means that the reader would be unable to readily verify calculations of group differences or assess the magnitude of treatment impact (i.e., effect size).

The explication of the participant attrition was deemed inadequate in over 25% of the studies. That is, the author(s) did not provide an explanation of (1) the characteristics of participants who withdrew from the study, (2) the attrition rates across the experimental or comparison conditions, or (3) the reasons for participant attrition. While the presence of attrition is not necessarily a fatal research flaw, the absence of reporting and accountability for the reasons and participant characteristics for those dropping out of the study poses a potential problem with respect to the efficacy of the treatment.

External validity

The ability to use the data from these studies for clinical implementation decisions may be significantly impacted by the lack of specificity regarding the recruitment and representativeness of the sample participants to the population from which they were drawn. Certainly it would seem reasonable to know whether or not the participants were reflective of the larger population of participants from which the individual study participants were drawn in order to understand the level of population generalization possible from the outcomes measured.

While one might be willing to accept a less exacting description of the population characteristics represented by the participating individuals, 35% of the studies did not give explicit enough information to confidently note the treatment setting or context. That is, one-third of the studies did not indicate whether participants were, for example, treated in a clinic, school, hospital, or private practice.

Internal validity-bias/confounding

The quality scores of the internal validity-bias subscale indicated that over one-half of the studies did not report assessor blinding or treatment fidelity as part of their study summary. The absence of assessor blinding is recognized as particularly troublesome as a substantial source of study bias. The knowledge of which participants received the treatment of interest provides a potential for the assessor to anticipate responses, encourage/solicit a particular type or range of responses, or suppress certain responses. In any case, without blinding during the measurement of the dependent variable, the potential for assessor bias can be a detractor from the accuracy of the measurement of treatment effects.

A prime method for dealing with potentially confounding factors such as attrition is the use of the intent to treat analysis. That is, all participants allocated to the experimental or comparison condition are included in the post-treatment analysis. In fact over 60% of all studies either used the intent to treat approach or accounted for the loss of participants through some statistical correction. While this should be expected for small sample and short-term treatment studies, the intent to treat analysis may provide a more realistic assessment of the treatment effects when applied to a real world clinical setting.

One of the positive findings from the summary of internal validity bias was the lack of 'data dredging', the use of unplanned post-hoc analyses such as subgroup or moderator analyses. All but one study either reported in the method section that a post-hoc analysis was planned and would be conducted as part of the impact of treatment analysis or no post-test analyses were anticipated or conducted. This absence of data dredging suggests that researchers in the area of stuttering treatment research have considered in advance the potential impact of known independent variables that should be considered as explanatory or moderator variables in assess treatment efficacy.

Implications for research

There are a number of implications for research and research reporting based on the results of this study. These include implications for researchers (and

perhaps more importantly for those who train researchers) as well as publishing outlets.

First, let us address implications for those who train researchers. As noted previously, designing rigorous studies to examine efficacy of stuttering treatments, including randomized controlled trials, *is* feasible. Other elements that suggest a rigorous research design, such as blinding assessors, evaluating treatment fidelity, and recruiting participants from a known population, among others, are also feasible. Feasibility obviously does not equate to implementation in practice. For those who educate researchers, therefore, ensuring that graduate students are prepared to evaluate research to determine its rigour and to conduct rigorous research that employs essential research design features is essential. This includes training for graduate students in how to evaluate and critique research. This also includes training for graduate students in how to design and implement rigorous randomized controlled trials and quasi-experimental designs as well as how to apply appropriate and sophisticated statistical procedures to analyse the resulting data.

Secondly, there are implications for researchers. Designing studies that provide meaningful results that can be applied with some degree of confidence is essential. How this can be insured rests in part on those that train these researchers (as mentioned previously) and those whose outlets the researcher may be attempting to use to disseminate the results of their studies (as discussed in the next paragraph). Another implication for researchers is reporting with transparency and ensuring that results disseminated are presented with clarity and with sufficient detail so that the reader fully understands the implications of the results (e.g., to whom the results can be generalized, threats to internal validity, practical significance, and more). This is especially important for clinicians, who may ultimately be making clinical decisions based on this research.

Third, there were a number of areas of concern cited. Although this suggests there may be a lack of rigour in some of the studies, it may also be the case that there was an oversight in reporting key information. Although recommendations for reporting are provided in, for example, the *Publication Manual of the American Psychological Association* (2010), journal editors and their respective reviewers have an increased responsibility in ensuring that the critical methodological pieces of information are provided in sufficient detail in the manuscripts that they review. This may require additional methodological reporting guidelines imposed by journals. A template for their creation could be based very easily on the items in the Downs and Black (1998) checklist, for example. This may also require some training of journal reviewers to ensure that the reviewers understand the standards imposed by the journal to ensure only rigorous research is published.

Implications for clinicians

There are also a number of implications for clinicians based on the results of this study. First, and in relation to efficacy of treatment for stuttering, there exists a limited number of studies (and even fewer that possess a large number of quality methodological traits) on which decisions about treatment can be made. Second, there are still consistent gaps in application of the most rigorous design methodologies. The consistent absence of such rigour across a body of research makes the usefulness of the results and treatment methods difficult. In other words, there is still room for improvement in translating research to practice, and clinicians must be able to critically evaluate research to determine its usefulness and applicability—thus potentially impacting the quality of the clinical decisions that are based on that research.

Another issue for the clinician is the ability to implement an intervention effectively if there is inadequate reporting of critical information. For example, the data from this study indicated that 56% of the studies did not report sufficient information regarding intervention compliance. If the clinician does not know how the treatment was conducted in the research setting, how will they have confidence that their attempt to implement the same program is in fact the same program? The question is one of fidelity of intervention or intervention compliance. The issue of fidelity of intervention represents a relatively new and important dimension of consideration for the evaluation of quality treatment research. Not only is an adequate description of what was designed and implemented in the study important, but also whether or not the researchers adhered to the description.

In summary, the data presented in this study of research quality in stuttering intervention offers a focused spotlight on issues important to both researchers and clinicians. It is clear to us that these data and their interpretation reflect a growing need for greater attention to treatment research methodology by both researchers and clinicians so that clinical applications and decisions that are based on that research have the strongest possible scientific base of support. In the end, treatment research cannot be effective unless researchers approach the research with the highest level of scientific rigour. Nor can clinicians deliver effective research without understanding the scientific bases of the treatment and related discipline. We would suggest that the evidence-based practice model provides an appropriate environment to bridge the researcher–clinician gap.

Pam Enderby's influence

While this paper is topic specific to the area of stuttering, the experience and driving concept behind

it as mentioned at the outset lies in the prime question "what works?" One of the hallmarks of much if not all of Dr Enderby's professional efforts whether as a clinician, teacher, scholar, or administrator have been focused on answering the question "what works?" Dr Enderby has never suggested that the answer to "what works?" was simple or easily accessible (Enderby, 2004; Enderby & Emerson, 1995), but she has suggested that task is required as a matter of professional integrity and relevancy. Our effort in this paper was to assess one dimension of the important work of research—design quality. We sought to demonstrate one approach to shedding light on the design quality that would have a substantial bearing on the way in which researchers construct and clinicians evaluate research addressing the "what works?" question. Our point is not to offer a platform for criticizing studies or authors for their work, but to draw attention to what some have viewed as critical issues of research design quality that would make the application of the findings meaningful in clinical practice. We believe Pam Enderby has precisely the correct perspective with regard to research in the field of communication disorders when she says, "The real answer is not to conduct more or less of each type of research but to conduct better research" (Enderby & Emerson, 1995, p. 172).

References

American Psychological Association (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author.

Andrews, G., Guitar, B., & Howie, P. (1980). Meta-analysis of the effects of stuttering treatment. *Journal of Speech and Hearing Disorders*, *3*, 287–307.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association*, *276*, 637–639.

Bothe, A., Davidow, J., Bramlett, R., Franic, D., & Ingham, R. (2006). Stuttering treatment research 1970-2005: II. Systematic review incorporating trial quality assessment of pharmacological approaches. *American Journal of Speech-Language Pathology*, *15*, 342–352.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs.* Boston, MA: Houghton Mifflin.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field setting.* Boston, MA: Houghton-Mifflin.

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

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., et al. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, *7*, 1–173.

Dollaghan, C. A. (2007). Handbook for evidence-based practice in communication disorders. Baltimore: Brookes Publishing Company.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomized and non-randomised studies of health care interventions. *British Medical Journal*, *52*, 377–384.

Enderby, P. (2004). Making speech pathology practice evidence-based: Is this enough? *Advances in Speech Language Pathology*, *6*, 125–126.

Enderby, P., & Emerson, J. (1995). *Does speech and language therapy work?.* London: Whurr Publishers.

Hegde, H. N. (2007). A methodological review of randomized clinical trials. *Communicative Disorders Review*, *1*, 15–36.

Herder, C., Howard, C., Nye, C., & Vanryckeghem, M. (2006). Effectiveness of behavioral stuttering treatment: A systematic review and meta-analysis. *Contemporary Issues in Communication Sciences and Disorders*, *33*, 61–73.

Howard, C., Nye, C., Vanryckeghem, M., Schwartz, J.B., & Turner, H. T. (November, 2006). Treatment efficacy for children who stutter: Summarizing single case studies. Paper presented at the American Speech Language Hearing Association Conference, San Diego, CA.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage Foundation.

Ingham, R. J., & Andrews, G. (1973). Behavior therapy and stuttering: A review. *Journal of Speech and Hearing Disorders*, *38*, 405–440.

Ingham, R. J., & Lewis, J. I. (1978). Behavior therapy and stuttering: And the story grows. *Human Communication*, *3*, 125–152.

Meline, T., & Paradiso, T. (2003). Evidence-based practice in schools: Evaluation research and reducing barriers. *Language, Speech, and Hearing Services in the Schools*, *34*, 273–283.

Moscicki, E. (1993). Fundamental methodological considerations. *Journal of Fluency Disorders*, *18*, 183–195.

National Center for Evidence-based Practice (NCEP). (2004). *Evidence-based practice (EBP).* Available online at: http://asha.org/members/ebp/, accessed 12 June 2009.

Reilly, S. (2004). The challenges in making speech language pathology practice evidence based. *Advances in Speech-Language Pathology*, *6*, 113–124.

Reilly, S., Douglas, J., & Oates, J. (2004). The move to evidence-based practice within speech pathology. In S. Reilly, J. Douglas, & J. Oates (Eds.), *Evidence based practice in speech pathology* (pp. 3–17). London: Whurr.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage.

Royal College of Speech Language Therapists Clinical Guidelines. (2005). *Clinical guidelines.* Available online at: http://www.rcslt.org/resources/clinicalguidelines, accessed 12 June 2009.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston, MA: Houghton Mifflin.

Wambaugh, J. L., Duffy, J. R., McNeil, M. R., Robin, D. A., & Rogers, M. A. (2006). Treatment guidelines for acquired apraxia of speech: a synthesis and evaluation of the evidence. *Journal of Medical Speech Language Pathology*, *14*, xv–xxxiii.

West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S.F., et al. (2002). *Systems to rate the strength of scientific evidence.* Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-001). AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality.

Appendix: Included studies

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

Bourdeau, L. A., & Jeffery, C. J. (1973). Stuttering treated by desensitization. *Journal of Behavior Therapy and Psychiatry*, *4*, 209–212.

Burgraff, R. I. (1974). The efficacy of systematic desensitaization via imargery as a theraputic technique with stutterers. *International Journal of Language and Communication Disorders*, *9*, 134–139.

Craig, A., Hancock, K., Chang, E., McCready, C., Shepley, A., McCauls, A., et al. (1996). A controlled clinical trial for stuttering persons aged 9 to 14 years. *Journal of Speech and Language Research*, *39*, 808–829.

Franken, M. J., Kielstra-Van der Schalk, C. J., & Boelens, H. (2005). Experimental treatment of early stuttering: A preliminary study. *Journal of Fluency Disorders*, *30*, 189–199.

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., Onslow, M., & Menzies, R. (2004). Dismantling the Lidcombe Program of early stuttering intervention: Verbal contingencies for stuttering and clinical measurement. *International Journal of Language and Communication Disorders*, *39*, 257–267.

Helps, R., & Dalton, P. (1979). The effectiveness of an intensive group speech therapy programme, for adult stutterers. *British Journal of Disorders of Communication*, *14*, 17–30.

Jones, M., Onslow, M., Packman, A., Williams, S., Ormond, T., Schwarz, I., et al. (2005). Randomised controlled trial of the Lidcombe programme of early stuttering intervention. *British Medical Journal*, *331*, 659–664.

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

Latterman, C., Euler, H. A., & Neumann, K. (2008). A randomized control trial to investigate the impact of the Lidcombe Program on early stuttering in German-speaking preschoolers. *Journal of Fluency Disorders*, *33*, 52–65.

Lewis, C., Packman, A., Onslow, M., Simpson, J. M., & Jones, M. (2008). A phase II trial of telehealth delivery of the Lidcombe Program of early stuttering intervention. *American Journal of Speech-Language Pathology*, *17*, 139–149.

Martin, R. R., & Haroldson, S. K., (1969). The effects of two treatment procedures on stuttering. *Journal of Communication Disorders*, *2*, 115–125.

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.

Peins, M., McGough, W. E., & Lee, B. S. (1972). Evaluation of tape-recorded method of stuttering therapy: improvement in a speaking task. *Journal of Speech and Hearing Research*, *15*, 364–371.

Riley, G. D., & Ingham, J. C. (2000). Acoustic duration changes associated with two types of treatment for children who stutter. *Journal of Speech, Language, and Hearing Research*, *43*, 965–978.

Ryan, B. P., & Ryan, B. V. K. (1983). Programmed stuttering therapy for children: Comparison of four establishment programs. *Journal of Fluency Disorders*, *8*, 291–321.

Ryan, B. P., & Ryan, B. V. K. (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.

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

Stocker, B., & Gerstman, L. J. (1983). A comparison of the probe technique and conventional therapy for young stutterers. *Journal of Fluency Disorders*, *8*, 331–339.

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