The Stuttering Treatment Research Evaluation and Assessment Tool (STREAT): Evaluating Treatment Research as Part of Evidence-Based Practice

Jason H. Davidow
Anne K. Bothe
Robin E. Bramlett
The University of Georgia, Athens

Purpose: This article presents, and explains the issues behind, the Stuttering Treatment Research Evaluation and Assessment Tool (STREAT), an instrument created to assist clinicians, researchers, students, and other readers in the process of critically appraising reports of stuttering treatment research.
Method: The STREAT was developed by combining and reorganizing previously published recommendations about the design and conduct of stuttering treatment research.
Conclusions: If evidence-based practice is to be widely adopted as the basis for stuttering assessment and treatment, procedures must be developed and distributed that will allow students, clinicians, and other readers without specialized knowledge of research design to critically appraise treatment research reports. The STREAT is intended to be such an instrument: It represents the consensus of previous methodological recommendations; it is consistent with and complements existing recommendations in evidence-based medicine and in the broader science of treatment outcome evaluation; and it is formatted into a single instrument for ease of use.

Key Words: stuttering, evidence-based practice, treatment outcomes, critical appraisal

Many recent medical (Sackett, Haynes, Guyatt, & Tugwell, 1991; Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000), allied health (American Speech-Language-Hearing Association [ASHA], 2005; Law, 2002), and even stuttering-specific (Bothe, 2003, 2004; J. C. Ingham, 2003; Onslow, in press) sources recommend a manner of practice known as research-based or evidence-based medicine (EBM) or, more inclusively, evidence-based practice (EBP). EBP encourages clinicians to make decisions that are research-based, client-centered, and outcomes-focused (Bothe, 2004); the triad is classically expressed as requiring research evidence, clinical expertise, and clients' preferences (Haynes, Sackett, Gray, Cook, & Guyatt, 1996). Within this larger framework, however, one of the central defining features of EBP is clearly its emphasis on the individual clinician's identification, evaluation, and thoughtful application of published clinical research.

Even though this basic idea is known to be several centuries old (see Sackett et al., 2000, chap. 1), it is also undeniable that Sackett and colleagues' version of EBM has had substantial influence on the current practice of everything from medicine and allied health to social work, probation, and human resources management (Trinder & Reynolds, 2000). Within ASHA, this shift toward research-based practice has led to the creation of the Joint Coordinating Committee on Evidence-Based Practice and the development of a position statement recommending EBP (ASHA, 2005), as well as several publications highlighting the importance and the contributions of EBP (i.e., Apel & Self, 2003; Dollaghan, 2004; Justice & Fey, 2004; Robey, 2005). A shift toward EBP is also evident in stuttering, as seen in several recent and forthcoming books (Bothe, 2004; Cordes & Ingham, 1998; Onslow, in press) and a recent special section in the *Journal of Fluency Disorders* (Bothe, 2003; Finn, 2003; J. C. Ingham, 2003; Langevin & Kully, 2003; Onslow, 2003). Even the new edition of a widely used textbook previously criticized for not recognizing the importance of research to clinical decision making (e.g., by Onslow, 1999) has now added "a preference for evidence based practice" to its list of "attributes that … make a clinician effective" (Guitar, 2006, p. 278).

As attractive as EBP might be in the abstract, however, application of its principles in real-world clinical settings is

limited by many factors, including clinicians' time and their expertise or comfort level in finding and evaluating published research. If EBP is to be widely adopted for speech-language pathology, as ASHA's (2005) recent position statement suggests it should be, then the discipline and the profession must move well beyond simply recommending EBP. Specifically, systems must be developed to allow practitioners to incorporate EBP into their daily routine, including systems through which research publications can be efficiently appraised and applied to clinical practice. Several relevant structures and systems exist already: The widely accepted Consolidated Standards of Reporting Trials (CONSORT) recommendations (Moher, Schulz, & Altman, 2001), for example, describe how to write the results of a randomized controlled trial (RCT) for publication. There are also relatively standard models for developing what are known as Phase I, Phase II, and Phase III studies of a treatment (see, e.g., Robey, 2004, 2005; Robey & Schultz, 1998; U.S. Department of Health and Human Services [U.S. DHHS], 2004). At the level of a single practitioner's evaluation and application of research, similarly, there are such examples as Sackett et al.'s (2000) "critically appraised topic" forms (see, e.g., Guyatt & Rennie, 2002) and Shaughnessy and colleagues' quite similar "patient oriented evidence that matters" approach (e.g., Shaughnessy, Slawson, & Bennett, 1994; Slawson, Shaughnessy, & Bennett, 1994). Among the most widely recognized of the existing systems for evaluating the quality of treatment research, finally, may be the Oxford Centre for Evidence-Based Medicine's (2005) "Levels of Evidence and Grades of Recommendation" (Guyatt et al., 1995) system for assigning overall grades based on the levels of evidence available. The Oxford system allows clinicians and other readers or decision makers to determine whether the evidence in support of a potential treatment deserves, for example, a grade of "A" because of the existence of "Level 1a" evidence (a good systematic review of multiple homogeneous RCTs) or a grade of "D" because of the existence only of "Level 5" evidence (expert opinion or extrapolation from basic physiological research or principles).

All of these existing systems serve important purposes, but none on its own is ideal for the task of evaluating stuttering treatment research, for several reasons. The CONSORT guidelines, for example, describe the necessary features of the published report about an RCT, primarily from the point of view of the researcher preparing the report. They do not specify the necessary value of those features, however, either for the researcher or for the consumer (e.g., how sample size was determined must be reported, but whether sample size was "big enough" is not dictated). The CONSORT guidelines also assume a starting point of RCTs, but the vast majority of stuttering treatment research uses other designs (Bothe, Davidow, Bramlett, & Ingham, 2005). The distinction between Phase I (testing a new intervention for the first time), Phase II (preliminary testing with a larger number of participants), and Phase III (large-group studies, comparing an intervention to existing standards or to other experimental treatments; U.S. DHHS, 2004) clinical trials, similarly, is an important and useful distinction, but that distinction on its own does not assist readers in determining how well any particular study was conducted and therefore how trustworthy its results might be. Of the existing systems, either the Oxford or Chambless and Hollon's (1998) definitions of "empirically supported therapies" may be the most applicable. Both, however, assume substantial expertise and judgment about everything from calculating and interpreting confidence intervals to whether the heterogeneity of evidence summarized in a systematic review of a given treatment is "worrisome" (Sackett et al., 2000, p. 176). In addition, with respect to evaluating treatment research for any particular disorder, there is the problem that none of the existing general systems for evaluating treatment research takes disorder-specific variables into account. Many of the basics do not change from disorder to disorder, but clinical researchers studying everything from diabetes to orthopedic surgery to stuttering also must consider some very different details and background information.

The goal of the present article, therefore, is to ease the task of critical appraisal of stuttering treatment research by students, clinicians, and other readers by introducing an instrument known as the Stuttering Treatment Research Evaluation and Assessment Tool (STREAT; see Appendix). Originally developed for use in a systematic review of the stuttering treatment literature (Bothe et al., 2005), the STREAT represents a synthesis of seven previous lists of requirements and recommendations for the design and conduct of stuttering treatment research (see below) and is intended to supplement the broader and interdisciplinary tools and systems described above. The differences between the previous sources and the STREAT lie primarily in format; the sources that served as the basis for the STREAT were scholarly discussions of research design requirements, whereas the STREAT was specifically developed as a tool to be used by clinicians and other readers. As described in further detail below, the STREAT is not intended to dictate what might be considered good or bad research, although it does include a set of five basic criteria synthesizing previous recommendations that readers could use as screening or evaluative criteria. The majority of the instrument instead provides a structured format within which readers can gather the information seen as relevant by multiple previous authors with expertise in stuttering treatment evaluation. The STREAT can be used on its own, or it can facilitate readers' use of any of the existing generic or interdisciplinary research-evaluation systems described above.

Construction and Purpose of the STREAT

Seven journal articles and book chapters were identified that discussed the recommended components of a stuttering treatment investigation (Bloodstein, 1981, chap. 11; Conture & Guitar, 1993; Curlee, 1993; J. C. Ingham & Riley, 1998; R. J. Ingham & Cordes, 1997; R. J. Ingham & Costello, 1984; Moscicki, 1993). Six of the seven were written specifically for stuttering treatment but incorporated models and methods from research design and clinical research more generally; Moscicki (1993) described treatment research only in relatively general terms.

All characteristics or requirements of a treatment evaluation investigation listed in any of the seven sources were

identified, resulting in 82 separate recommendations. Repetitive items were collapsed, and all characteristics found in two or more of the seven initial recommendations sources were kept. The four developers of the STREAT instrument itself (Davidow, Bothe, Crowe, & Ingham, 2005) then worked independently and collaboratively to categorize the 82 recommendations and to identify those that could serve as the basis of one consensus document, combining the recommendations made in the seven previous sources into one useful list or evaluation form. A draft version of the STREAT was finalized that included six main categories (Sections II–VII; see Appendix), most with several subcategories. After further discussion and pilot uses of the draft STREAT, four additional main categories were added to create the final version of the STREAT (provided in the Appendix).

The STREAT is intended to be used as a data-gathering instrument, to allow readers to identify and organize all relevant information about a stuttering treatment research report. To be consistent with the suggestion of the Evidence-Based Medicine Working Group (e.g., Haynes et al., 1996; Sackett et al., 2000) that some readers might choose to screen articles for basic methodological soundness before devoting time to reading them thoroughly, the STREAT also includes as its first section five important characteristics of treatment research. Neither Section I nor the STREAT as a whole, however, provides any specific requirements as to which characteristics must be present or which treatment research articles must or must not be used as the basis of stuttering treatment; similarly, the STREAT results in descriptive information about a study, not in a score or total of any sort. These features of the STREAT are intentional, because there is no single criterion and no specifiable number of characteristics that in all situations would necessarily make any treatment research study applicable or inapplicable. The STREAT is intended to serve as a means of organizing relevant information, allowing comparisons across studies, and allowing the reader to decide whether the methodology of any particular article is acceptable for any particular purpose. As described in further detail in the remainder of this article, it is intended to be consistent with the broader intent and recommendations of EBM and EBP, and of existing research evaluation systems, while simultaneously representing the consensus that has developed over the last several decades about stuttering treatment research in particular.

Sections and Items in the STREAT

Section I: Five Basic Characteristics

The acknowledged "gold standard" for treatment research in medicine and allied health is the double-blind RCT (see, e.g., Cochrane, 1972; Cook, Guyatt, Laupacis, Sackett, & Goldberg, 1995; Moscicki, 1993; Oxford Centre for Evidence-Based Medicine, 2005; Sackett et al., 2000). An RCT first identifies one large group of potential participants and then randomly assigns each participant to one of two or more groups, resulting in either a placebo control group or a treatment comparison design. *Double-blind* refers to the fact that data are gathered by experimenters who do not know participants' group assignments or the phase of the study from which the data are being gathered, and participants do not know whether or when they are receiving the experimental treatment; these procedures minimize the influence of placebo effects, expectation, and rater bias (e.g., Rosenthal, 1966; Sackett et al., 2000).

Sackett et al. (2000) suggested that the first question to be asked by a reader facing a new treatment research article is whether the study included random assignment to groups. If it does not, its results are methodologically compromised, and Sackett et al. suggested that readers may choose not to read any further in that article and not to use it as the basis for their clinical practice. This recommendation is problematic for speech-language pathology, however, for several reasons, including in this context the use and the methodological strengths of single-subject experimental designs (Barlow & Hersen, 1984; Kazdin, 1982). Thus, given both the importance of random assignment for group-design research and the strong internal validity of well-designed and well-conducted single-subject experimentation, Item I.a. on the STREAT asks whether the treatment study being reviewed included either of these basic design strategies.

Double-blind techniques are not listed as part of the first item, despite the widespread acknowledgment of the superiority of double-blind RCTs, because double-blind conditions are very difficult to achieve with behavioral or cognitive interventions, or any intervention that is based on extended interactions between a clinician and a client, as opposed to pharmaceutical trials or similar treatments. A more realistic recommendation, adopted for the STREAT, might therefore be that any data gathered by observers should be either gathered by observers who are blind to the participant's group status, instead of being gathered by the treating clinician, or at least gathered by observers who show high interjudge agreement with an independent rater. These ideas create Item I.b.

The final three items listed in Section I are stuttering-specific items that represent a relatively widely held consensus view about minimal requirements for the design of stuttering treatment research (Bloodstein, 1981, 1995; Conture & Guitar, 1993; Curlee, 1993; J. C. Ingham & Riley, 1998; R. J. Ingham & Cordes, 1997; R. J. Ingham & Costello, 1984). As summarized by R. J. Ingham and Cordes (1999), stuttering treatment studies (a) must include repeated evaluations of speech performance before, during, and for a clinically meaningful period of time after treatment, because stuttering is known to vary across time; (b) must evaluate speech performance in beyond-clinic situations, because persons who stutter can be very fluent within the clinic and not transfer that fluency outside of the clinic; and (c) must control for measurement error, changes in speech rate, and unusual speech quality, because all three are known to result in the spurious appearance of improvements in stuttering frequency that cannot be said to constitute an empirically or socially valid improvement in overall speaking or communication abilities. The wording of these elements was made more inclusive for the STREAT, in recognition that some treatment research programs might not be attempting to change stuttering frequency. Regardless of the goal of any stuttering treatment, however, the principles hold: Treatment research is

more difficult to interpret without data from before, during, and after treatment (Item I.c.) and without evidence of beyond-clinic improvement (Item I.d.).

Item I.e. was phrased specifically to emphasize that its concerns apply only when stuttering frequency data are reported or interpreted. Here, as in many other items, readers must use their own judgment in determining whether the absence of speech rate or speech naturalness data is a problem; it is much more of a problem for a study of prolonged speech in adults, for example, than for a study of response-contingent stimulation procedures in preschool children. Nevertheless, for any stuttering treatment research, the presence of speech rate and speech naturalness data allows comparisons to other treatments and addresses concerns often raised not only by researchers but also by speakers who stutter.

Overall, Section I is intended to capture five of the most critical elements in stuttering treatment research, based on widely held views from multiple previous authors. Ideally, any article to be used as the basis for clinical practice in stuttering would meet all five of the criteria listed in this section, unless a reader had a specific and data-based reason to overlook a particular criterion. EBP also recognizes, however, that the literature available in a given discipline at a given point in time is not necessarily ideal (Sackett et al., 2000); thus, the remaining sections of the STREAT can assist readers in gathering and interpreting the best available current evidence.

Section II: Strategy/Participants/Sample Size

As shown in the Appendix, the remaining items on the STREAT are more specific, serving as a means for readers to identify the methodological characteristics emphasized in at least two of the seven recommendations sources and to identify other important details about a stuttering treatment research report. The first subcategory in Section II asks readers to determine the basic research strategy used in a given study, or to determine whether it is a variation on a single-subject design or a variation on a group design (as discussed for Section I).

Item II.c. refers to the manner in which participants were recruited, because the easiest access to research participants can lead to biased samples (Moscicki, 1993; Schiavetti & Metz, 2002). Stuttering self-help groups, for example, can supply researchers with immediate access to large groups of potential participants. However, these individuals may possess such characteristics as motivation or greater knowledge about the disorder that may enable them to be more successful in treatment than other participants might have been. Alternatively, such participants may have sought out the support group because they did not respond to previous treatments; therefore, they may constitute a self-selected group of failures to respond to a given treatment or class of treatments. There is no single correct way to recruit participants, but readers must decide whether the recruitment process biased the reported results.

Inclusion/exclusion criteria (Item II.d.) are also pertinent to evaluation of the study's sample (Moscicki, 1993; Schiavetti & Metz, 2002). Because these criteria will be highly variable depending on various aspects of the study, such as age group and treatment type, an exhaustive list is not practical for the STREAT. Regardless of the actual criteria, the reader should be informed of these in order to judge the appropriateness of the sample. A statement identifying these criteria might be "All participants had not received therapy within the last year, had no other speech or language disorders, and had normal hearing." As is also the case for several other items, a line is provided on the STREAT to note these criteria for future reference.

The subcategory "Sample Size" includes both methodologically necessary items and others that aid in the evaluation process. The authors of the methodological recommendations that served as the basis for the STREAT did not provide specifics regarding sample size, but several, as well as other experimental design texts, mention the need for a sufficient or representative sample of the stuttering population (Bloodstein, 1995; Moscicki, 1993; Sackett et al., 2000; Schiavetti & Metz, 2002). For single-case designs, the STREAT's specification of at least 4 participants (Item II.e.) was based on Barlow and Hersen's (1984) recommendation of "one successful experiment and three successful replications" (p. 346) as the basis for accepting a treatment as successful and beginning a series of systematic replications. Although the success of 4 individuals is hardly enough to warrant a positive prognosis for the stuttering population in general, the logic of successive replications (Sidman, 1960) does not depend on large samples and does support the value of the treatment used, especially for clients with similar characteristics as those tested. As discussed later, the most appropriate size for a group design is related to power (see Section VI), but clearly larger groups are preferable. The very small group size of 10 (Item II.f.) was selected for the STREAT based primarily on the even smaller groups that are common in stuttering treatment research.

Items II.g. and II.h. assist in determining the number of persons who did not complete the study. There are several important issues here; the classic issue is that participants who do not complete a treatment study might differ in important ways from those who do, including leaving the study if they cannot tolerate the side effects or if they are not benefiting from the treatment. Analyses based only on those who do benefit are obviously misleading (Heritier, Gebski, & Keech, 2003; Mahaniah & Rao, 2004). In stuttering research, the long posttreatment times used for some studies raise the almost paradoxical (Onslow, O'Brian, Packman, & Rousseau, 2004) problem that studies can be all but punished for their very thoroughness, as the number of subjects lost to follow-up for any reason is expected to be larger for longer studies (Boberg & Kully, 1994; O'Brian, Onslow, Cream, & Packman, 2003; Onslow, Costa, Andrews, Harrison, & Packman, 1996; Ryan & Ryan, 1995). Nevertheless, readers deserve to be assured that there is a legitimate reason for the smaller groups often reported by the end of a study.

Section III: Dependent Variables

Section III addresses the specific outcome measures used in a study and the frequency and situations in which these outcome measures were gathered. At least half of the methodological recommendations articles that served as the basis for the STREAT included stuttering frequency, speech rate, speech naturalness, stuttering severity, and self-report

(which includes attitude scales, anxiety scales, etc.; Items III.a.–III.e.) as necessary measures. "Other DVs" (Item III.f.) was provided to allow assessment of other, less common, variables (e.g., duration of longest stutter, longest stutter-free segment of speech). We agree with previous assertions that stuttering frequency, speech rate, and speech naturalness data must be gathered to suitably evaluate any stuttering treatment program, but no specific dependent variables are necessary in a larger sense or to all readers; ultimately, the necessity of any specific variable depends on an individual clinician's or client's goals.

Regardless of the variables used in a given study, clear definitions of all dependent variables (Items III.a.1.–III.f.1.) are necessary for the proper evaluation of any treatment outcome (Conture & Guitar, 1993; Moscicki, 1993). Definitions serve the need for replication (Muma, 1993); the treatment can be neither applied by clinicians nor reevaluated by other researchers if its variables cannot be understood. Definitions also must be provided because the exact interpretation of many dependent variables can differ across studies. As an example, several definitions of stuttering may include what other authors would label normal disfluencies (Cordes & Ingham, 1994; Einarsdóttir & Ingham, 2005; Ryan & Ryan, 1995; Yairi & Ambrose, 1992), potentially compromising any interpretations about the actual amount of stuttering. Clear definitions allow readers to decide for themselves whether stutters or normal disfluencies were counted, in order to make a proper decision about the actual outcome of the treatment.

The next subcategory within Section III is "Situations in which speech-related DVs were measured." The importance of obtaining stuttering frequency data beyond the clinic (Item III.g.) is emphasized in most previous sources (e.g., Bloodstein, 1995; J. C. Ingham & Riley, 1998) and is underscored by multiple reports that stuttering frequency may be lower within the clinic during or after treatment (e.g., Hewat, Onslow, Packman, & O'Brian, 2006; Ryan & Ryan, 1995). This issue also clearly relates to the desire for treatments that lead to demonstrable changes in clients' real-world skills or abilities. Within-clinic data (Item III.h.) are also generally accepted to be critical, primarily for treatment planning (Conture & Guitar, 1993; Curlee, 1993; R. J. Ingham & Costello, 1984), although obviously treatments such as Onslow and colleagues' (e.g., Onslow, Andrews, & Lincoln, 1994) Lidcombe Program that are conducted in the client's natural environment may not have or need within-clinic data. Items III.g. and III.h. use the phrase "at least one" because proper evaluation of a treatment requires more than one sample from each condition, but there is no minimum that can reasonably be specified as an absolute requirement for all studies. At the extreme, however, if no beyond-clinic data are reported, readers might be cautious about assuming the real-world effectiveness of the treatment.

The remaining three items in the Situations subcategory pertain to the type of speaking style assessed. At a minimum, a study should specify the style of speech that was gathered (Item III.i.); the results are otherwise uninterpretable. Conversational speech (Item III.j.), rather than oral reading or other less natural tasks, represents the generally accepted ideal (Conture & Guitar, 1993; Curlee, 1993; J. C. Ingham & Riley, 1998; R. J. Ingham & Cordes, 1997; R. J. Ingham & Costello, 1984), because the goal of most persons who stutter is communication in natural interactions with other persons. In addition, to prevent reactive effects, covertly recorded samples (Item III.k.) are sometimes described as desirable (Bloodstein, 1995; R. J. Ingham & Costello, 1984). The data regarding this issue are equivocal, however, with some investigations reporting no meaningful difference between overt and covert speech samples (e.g., Andrews & Tanner, 1982; Ladouceur, Cote, Leblond, & Bouchard, 1982) and others reporting more stuttering during covert assessments (Howie, Woods, & Andrews, 1982; R. J. Ingham, 1975).

The next subcategory in Section III, "Frequency with which all DVs were measured," contains items regarding the phases that should be included in a stuttering treatment investigation and the frequency with which outcomes should be measured. Items III.m. and III.n. refer to the importance of establishing a proper baseline for variables known to change over time (stuttering frequency, speech rate, speech naturalness, and stuttering severity; Conture & Guitar, 1993; J. C. Ingham & Riley, 1998; R. J. Ingham & Costello, 1984). J. C. Ingham and Riley (1998) emphasized that pretreatment measurements should be taken over several months, but that recommendation is conservative as compared with the other sources; a 1-month baseline is specified in the STREAT as a suggested minimum (Item III.n.). It could also be argued that if a group of participants is large enough, the typical day-to-day (and sometimes minute-to-minute) variations in stuttering will cancel each other out with one data point; this is the basic logic of group design research. Because the exact number of participants necessary to ensure the representativeness of one measure is not known, however, most previous sources recommend at least the 1-month baseline selected for this item. The necessity for multiple measurements of self-report or other dependent variables depends on the test–retest reliability of the measure or instrument being used. In the absence of repeated measurements or reliability information about the instrument, the reader may reasonably question any claims made on the basis of self-report data.

The remaining items in this subcategory (Items III.o.–III.bb.) refer to stages of the treatment process. Treatments for many speech and language disorders follow the basic pattern of changing behavior in the clinic (establishment phase), transferring this success to the real world (transfer phase), and then trying to maintain this improvement for a clinically meaningful period of time after treatment has ceased (maintenance phase). In the absence of a data-based reason to structure treatment differently (the Lidcombe Program's use of real-world conditions from the beginning of treatment again provides a good example), stuttering treatment should be no different. The inclusion of all three standard phases in a treatment investigation allows the reader to determine the benefit at each stage of the treatment; thus, information gathered by completing these items can be useful not only to help readers identify the phases used in the study but also to assist readers in identifying any relevant patterns across studies. Items about when data were gathered (III.q., III.t., III.w., III.z.), similarly, are critical for deciding whether that

phase contributed to any reported success. If data were provided only after follow-up for a treatment that included establishment, transfer, and maintenance, the reader would not know the contribution of each phase of treatment and might question whether the treatment itself was responsible for the ultimate changes in dependent variables.

The content of this subcategory should not be interpreted as implying that any single study must include all of the phases mentioned. With young children, in particular, explicit transfer or maintenance programs have been claimed (or shown) to be unnecessary (J. C. Ingham, 1999; Martin, Kuhl, & Haroldson, 1972; Reed & Godden, 1977). For adults, however, treatment studies using prolonged speech, for example, that have reported the best results have included some type of transfer phase or transfer activities (R. J. Ingham et al., 2001; O'Brian et al., 2003; Onslow et al., 1996; Ryan & Ryan, 1995). It is also worth highlighting that the data taken and reported should usually be from times when the treatment is not being actively administered. Data from treatment conditions are necessary to establish the fidelity with which the treatment was administered, a separate issue (see below), but they do not reveal the quality of speech produced in day-to-day conditions beyond treatment.

Definitions of establishment, transfer, maintenance, and follow-up are also necessary to this subcategory. These terms are used in several ways and often can be left to the reader's interpretation. For the sake of completeness, we provide the following definitions, which we use in conjunction with the STREAT:

- *Establishment:* The administration of treatment with the therapist in the clinical environment, not in the client's real-world settings.
- *Transfer:* The administration of treatment in the client's real-world environment, with or without the therapist's literal presence. To be labeled transfer when the therapist is not present, the therapeutic protocol must call for specific active tasks that are known to be completed in the client's natural environment.
- *Establishment and Transfer Activities Performed Simultaneously:* Clinical-environment and real-world tasks are performed in the same daily or weekly time frame, so that no clear division between phases can be recognized.
- *Maintenance:* No active treatment is provided, in the clinical environment or the real world. Measurement of dependent variables is performed, but no treatment is given. Optimally, a contingency schedule is organized in which clinical contact is reduced based on a criterion of success; this type of maintenance phase might therefore be referred to by some authors as an "active" maintenance phase (see R. J. Ingham, 1980, 1999).
- *Follow-up:* No contact from the therapist or researcher beyond discussion of logistical issues such as scheduling.

Many treatment programs do not fit easily into these or any other definitions. Boberg and Kully (1994), for example, wrote of a maintenance program following their intensive establishment and transfer activities that involved several additional treatment activities, which would conflict with the definition of maintenance provided above. The more important point, therefore, is for readers to simply be aware of which phases were actually used in any given study, whether corresponding data are provided, and whether their planned use of any given treatment will use those same phases.

The final subcategory in Section III, "Length of speech samples in which DVs were measured," includes only one item (III.cc.). R. J. Ingham and Costello (1984) suggested speech samples of 3 min, Curlee (1993) suggested 550 words, and Conture and Guitar (1993) suggested 100 to 200 words. R. J. Ingham and Costello (1984) also provided the broader recommendation that beyond-clinic samples should be a length representative of a typical speaking situation: If the client typically speaks for 10 min during a 30-min meal, the sample in that situation should be 10 min. Taking all such opinions into account, and as a summary of previous recommendations, the STREAT recommends at least 3 min or 500 words, which can be roughly equated to 750 syllables.

Section IV: Treatment Fidelity

Items IV.a., IV.c., and IV.d. relate to the accuracy or trustworthiness with which a treatment protocol was followed, allowing the reader to assess whether the data produced are more or less likely to be due to the treatment described by the authors. If a treatment research report does not include these characteristics and reports poor outcome data, there exists the possibility that the poor results are due to improper application of the treatment. Similarly, reductions in fluency coupled with improperly provided treatment may simply be coincidental. In addition, if the treatment is not described well enough that it can reasonably be replicated (Item IV.b.), then it would be a mistake for a clinician to implement it; the critical functional variables may be absent from the published description.

Most articles used in developing the STREAT stressed the importance of treatment fidelity information (Conture & Guitar, 1993; J. C. Ingham & Riley, 1998; Moscicki, 1993). It is worth mentioning here, therefore, that a current treatment review project (Bothe et al., 2005) revealed that only approximately 2% of almost 200 stuttering treatment articles reported treatment fidelity data (e.g., percentage of elements of the experimental protocol completed correctly: Elliot, Miltenberger, Rapp, Long, & McDonald, 1998; accuracy of contingencies provided after a moment of stuttering: James, 1981). Item IV.c., phrased in terms of "any evidence," as opposed to Item IV.d., "treatment fidelity data," recognizes an intermediate level of information, including such statements as that clinicians were observed to ensure the accuracy of treatment administration (when no other data are provided).

Section V: Data Fidelity

Items in this category take into account the accuracy or replicability of the data and the reader's related ability to believe that the data are valid (Conture & Guitar, 1993; Curlee, 1993; J. C. Ingham & Riley, 1998). The importance of intrajudge and interjudge agreement or reliability data (Items V.a., V.b., V.g., V.h.) lies in the many reported difficulties in obtaining agreement for instances of stuttering

(Cordes & Ingham, 1994) and number of syllables produced (Onslow et al., 1994, 1996). Given these known difficulties, data might ideally be presented as averages calculated from several different judges, all of whom showed satisfactory agreement with themselves and with each other. In the absence of this ideal, correlations of .80 or agreement figures of 80%, while imperfect, appear to be the lower bound of recommended levels (Cordes, 1994; Curlee, 1993; Kazdin, 1982); the level of agreement or correlation that must be achieved to ensure treatment effectiveness is clearly a more complicated question that has not been empirically demonstrated.

With respect to judging the reliability of treatment data, it should also be noted that procedures other than calculating reliability coefficients often serve the same purpose more effectively. Onslow et al. (1994), for example, graphically displayed their reliability numbers for stuttering frequency without calculating a coefficient. Such presentations allow readers to see that another judge did not change the general conclusions to be drawn from the data. Such information can be more useful than correlations, especially when correlations are presented without any other data to suggest that observers whose scores are highly correlated are also producing similar data, as opposed to data that differ from each other in consistent (i.e., highly correlated) ways (Cordes, 1994). In all cases of evaluating reliability or agreement data, the ultimate question is the trustworthiness and replicability of the main treatment findings.

Similar issues apply for measures other than stuttering frequency and speech rate, and acceptable reliability for these dependent variables depends on the exact definition of the variable. Speech naturalness for stuttering treatment investigations is typically measured using a 9-point scale (1 = *highly natural*; 9 = *highly unnatural*; see Martin, Haroldson, & Triden, 1984), and deviations of one scale value have generally been accepted as reliable (Finn & Ingham, 1989; R. J. Ingham et al., 2001; Onslow et al., 1996). Stuttering severity, however, has been measured in several different ways, with reliability or replicability not necessarily addressed. These decisions will largely be left to the reader: If the values provided appear to represent reliable data, the relevant items should be marked.

Items V.l. and V.m. also relate to data fidelity. The number of dropouts is identified in a previous portion of the STREAT, but Item V.l. asks the reader to identify whether reasons for dropping out were provided. Knowing that participants withdrew from the investigation for reasons other than unsuccessful treatment provides more confidence in the results when a substantial number of dropouts are reported (Bloodstein, 1995; Moscicki, 1993). In the absence of such explanations, readers should assume that the treatment did not provide a benefit to those who withdrew.

The inclusion of an item regarding the need for an independent observer to take measurements (Item V.m.) is to counteract the possibility of noninteractional observer bias (Rosenthal & Rosnow, 1984), which refers to the observer affecting the recording of the participant's behavior, even in the absence of interactional bias. As discussed above for Section I, this item represents one of the few absolute necessities in evaluating treatment research in stuttering or any other area. If the stuttering frequency judge, to take one example, is aware that a client is in the final stage of treatment, the judge may expect minimal stuttering and decide, probably without explicit awareness of this decision, to count all questionable disfluencies as normal disfluencies, producing data biased toward showing a treatment effect. Ideally, the main data used for all data analyses should be provided by an observer blind as to the participants' diagnoses and treatments; however, as also discussed with respect to Section I, agreement between the primary judge and an appropriately blinded reliability judge may be acceptable.

Section VI: Data Analysis

All items in this section were selected based on their importance to the STREAT authors; the recommendations sources that served as the basis for the other items on the STREAT did not emphasize inclusion of these variables in a treatment outcome study (with the exception that Moscicki, 1993, did mention the need to consider power). Although not imperative, graphic display of data points over time (Item VI.a.), which allows visual inspection of different phases, is preferred for single-case designs (Barlow & Hersen, 1984; Kazdin, 1982). Obviously, if data are clearly explained in the text and can be identified with their respective treatment phase, that may be adequate to determine the outcome of the various phases of treatment.

The other two items in this category relate to group studies, particularly those that do not report individual data. Statistical support for claims of group differences (Item VI.b.) is relevant for both between- and within-group designs, to assure the reader that differences or changes in group means are beyond those expected by chance or through natural variability. Again, an appropriate evaluation in this situation may require some judgment on the part of the reader; if the claimed differences are very large, and if other methodological strengths support the data, then descriptive presentation of data may suffice to support a claim of group differences.

Item VI.c. refers to power, or the probability that a significant difference between groups in a study will be detected when there actually is a significant difference for the conditions that the study has sampled (Cohen, 1988; Jones, Gebski, Onslow, & Packman, 2002). Power is also important for determining with confidence that a study's result of no differences between groups reflects a true lack of differences. Factors affecting power are numerous and beyond the scope of this article (see Cohen, 1988); the most important point is that a finding of no statistically significant difference between groups may be due to a lack of power to find a difference that does exist in the population or populations being studied. Low power in a comparison of groups before treatment may therefore lead to the error of concluding that those groups were equivalent before treatment when in fact they were not, an issue that could create problems in interpreting the claim that any differences found between those groups after treatment can be attributed to the treatment. Similarly, low power may lead to the error of concluding that a truly beneficial treatment is not effective (i.e., the treated and untreated groups may be described as not different in the study when they do represent truly different populations).

Section VII: Results

The majority of the items in this section allow documentation of the final outcome of the treatment. Items VII.a.–VII.f. allow the reader to identify the timing of the final data reported in the study. Items VII.g.–VII.i. then allow the description of the nature and quality of the data provided for up to three different dependent variables; the same items can obviously be repeated for any additional dependent variables in a given study.

Items VII.g.6., VII.h.6., and VII.i.6. ask whether data have been provided from a 1-year follow-up period. All of the recommendations sources used as the basis for the STREAT mention the need for a follow-up period, but the specific duration of that follow-up varied. We are more comfortable with Bloodstein's (1995) recommendation of at least 18 months to 2 years; however, 1 year seems to be the minimally acceptable follow-up period and was therefore selected for the STREAT. This criterion is applicable for all age groups, as there are no data revealing a difference in the most critical relapse period after treatment as a function of age. Beneath these items are spaces to record the actual length of follow-up for future reference and comparison to other studies.

The final subcategory within Section VII, "Length of establishment phase," was included primarily to allow comparisons between treatments, because a treatment that can achieve certain outcomes in 10 hr has obvious advantages over one that requires 30 hr to achieve the same results (R. J. Ingham & Costello, 1984; Ryan & Ryan, 1995). This information is often difficult to determine, a problem that does not detract from its importance.

Additional Issues and Final Considerations

Reliability of the STREAT

As alluded to throughout this article, the STREAT was initially developed to serve as the data-collection instrument for a systematic review of the stuttering treatment literature (Bothe et al., 2005). As part of that review, and as part of providing necessary evidence to support the use of the STREAT, 152 articles were analyzed using the initial version of the STREAT (Bothe et al., 2005). The initial version differed from the current STREAT primarily in cosmetics—formatting, the addition of spaces for recording aspects of the treatment process such as definitions and activities used, changes in a few items, and the absence of the current Section I. The first 90 articles were assessed by all possible groups of three judges from a pool of five judges; the remaining 62 were assessed by two of the five judges. Three of the judges were beginning doctoral students (two of whom had not taken a research design class during their master's studies); the other two were master's students in speech-language pathology. Neither of the master's students had taken a research design course or stuttering course when the projects began, and only one had completed research design and stuttering courses when the project was completed. Therefore, the group of judges can be viewed as comparable to the intended audience for the STREAT (i.e., students, clinicians, and others with little to no specialized training in stuttering or research design).

Each article was first read and evaluated independently by each judge assigned to it. The relevant judges then met to discuss all disagreements, with a goal of complete agreement but with a 5-min limit on circular or unproductive discussion about any one item. As shown in Table 1, interjudge agreement for the items on the STREAT was high for most sections before consensus discussions, but still below 85% for four of the six sections. Agreement was very high for all sections after consensus discussions, both as an average figure for groups of three judges and for the two judges who read the second set of articles (both doctoral students). The lowest agreement, 72.7% of items agreed by all three judges before consensus discussions for the items in Section IV, was due primarily to differences of opinion among some judges as to whether the description of the treatment was "clear enough for replication." Similarly, the relatively low agreement for Section III reflects the readers' difficulties in determining how phases were defined by different authors, a problem compounded by the often vague explanations provided by authors for the activities occurring in each phase. As also shown in Table 1, interjudge agreement was over 95% for all sections of the STREAT after consensus discussions among groups of three judges and was also very high for the independent judgments made later by the two judges who read the 62 additional articles. These data clearly suggest the value of talking with colleagues: Discussions tended to lead to agreements, and judges who had previously worked together tended to produce similar independent judgments when reading later articles.

Intrajudge agreement was assessed by determining the number of judgments that did not change from the initial independent assessments to the consensus discussions, as a measure of how likely it was that any one judgment made by any one judge would be defended and repeated by that judge. As shown in Table 2, intrajudge agreement was high, ranging from 93.7% for the items in Section VII to 98.3% for the items in Section V. The weighted average of the section intrajudge agreement scores shown in Table 2 was 95.4%; that is, over 95% of the over 20,000 individual item judgments represented in Table 2 were repeated by the judge in consensus discussions. The only intrajudge agreement figures below 90% for individual items were 86.9% for Item II.c. ("Clear description of selection procedures"), 87.6% for Item III.f. ("Other dependent variables"), and 85.8% for Item III.u. ("Establishment and transfer activities performed

TABLE 1. Interjudge agreement, expressed as percentage of items agreed by 3 of 3, or by 2 of 2, judges prior to (independent) and following (consensus) consensus discussions, for Sections II–VII of the Stuttering Treatment Research Evaluation and Assessment Tool (STREAT).

	90 articles read by 3 judges		62 articles read by 2 judges	
Section	Independent	Consensus	Independent	Consensus
II	82.5	98.9	96.4	99.8
III	78.6	97.9	91.6	99.1
IV	72.7	96.8	90.8	98.5
V	92.0	98.5	98.4	99.7
VI	90.1	98.0	97.5	99.5
VII	83.8	97.9	92.9	98.8

TABLE 2. Intrajudge agreement, expressed as the number and percentage of independent judgments that did not change during consensus discussions, for Sections II–VII of the STREAT.

Section	Number of judgments	Intrajudge agreement (%)
II	1,804	94.2
III	9,020	94.1
IV	902	97.5
V	4,510	98.3
VI	1,353	97.4
VII	2,706	93.7

simultaneously"). Given the nature of these items, their relatively low intrajudge agreement is understandable; the first requires some judgment and interpretation on the part of the reader, and the second two only come into play for complex studies that may have been particularly difficult to understand from the available information.

Overall, as the data in Tables 1 and 2 show, the evidence available from extended use of the STREAT by multiple relatively inexperienced judges shows that the judgments made about most items are replicable by a single judge and across judges. The exceptions, which readers and users of the STREAT need to be aware of, are for Sections III and IV, where the decisions to be made by the readers are often very difficult, in part because some relevant information is not provided clearly in some stuttering treatment research reports. The solution, as also shown in Table 1, is for readers to discuss with a colleague any items that they cannot confidently answer by themselves, until a consensus can be reached.

Final Issues

In addition to the item-by-item issues discussed in the previous sections, several final issues concerning the use of the STREAT as a whole warrant discussion. As mentioned in several of the previous sections, one of the most important issues is whether any one item on the STREAT must be present for a study to be deemed well designed, of high methodological quality, or otherwise worthy of consideration in an EBP framework. Based on the seven recommendations sources summarized to develop the STREAT, the answer is a qualified "yes"; that is, essentially all of the ideas represented on the STREAT are vital for confidence in the treatment program being reviewed and have been described as such by multiple previous authors. Equally, however, the answer is a qualified "no"; exceptions about the need for any given item are numerous and obvious, and individual readers will draw their own conclusions for any number of reasons. Overall, the strength of an instrument such as the STREAT is that it allows readers to gather necessary information in a standard format. It should also be emphasized that, because of its structure as a synthesis of previous recommendations, the STREAT is not exhaustive of all possibly relevant issues; it does not address, for example, the issue of effect size or the question of whether the researchers providing supportive data about a treatment are those who developed it or are independent of it.

It should also be explicitly mentioned that the STREAT was designed to evaluate one stuttering treatment program or one stuttering treatment investigation. This is, itself, an interesting comment on existing stuttering treatment research; if the standard in our field were RCTs or treatment comparison designs, then the recommendations sources that served as the basis for this instrument would have discussed more than one group, and an instrument such as this would have been designed to accommodate more than one group. As it stands, an instrument designed to focus on one treatment at a time serves the clinical orientation of this project; the goal was a tool that allows readers to assess the evidence about a treatment. Treatment comparison designs, multistage studies, or other variations are assessed using two or more copies of the STREAT, one for each treatment or group.

A related issue arises in considering treatment research that was not intended to address more than one or two of the phases described in Sections III and VII, such as research about the establishment phase alone or about transfer alone. It is important to remember that any one treatment study, on its own, is insufficient; the goal of EBP and the ultimate goal of the STREAT is to accumulate, and to build clinical practice on, all high-quality data about specific treatment programs (e.g., through the completion and use of systematic reviews of the literature). Thus, it is not necessarily problematic if a given study does not present, for example, data from a transfer phase or data from a 1-year follow-up; in some cases, that information will be readily available in other publications, and in other cases readers might be willing to use a treatment for which complete information is not yet available, if that information might be provided in future investigations. Equally, however, if such information is not forthcoming, then readers and clinicians must eventually consider that the continued absence of relevant data, or the continued failure of a treatment's developers or supporters to provide relevant data, might be meaningful in itself.

Overall, the main goal of the STREAT is to assist clinicians with the larger goal of providing the most appropriate and effective stuttering treatment services to their clients, by assisting clinicians with the critical appraisal step of Sackett et al.'s (2000) version of EBP. Based on the review project described above, it appears that even the most complex stuttering treatment journal article can be assessed using the STREAT, with high interjudge and intrajudge agreement, in less than approximately an hour. Multiplied across many articles, this is an important time commitment, but it is not unreasonable, especially considering the alternative of continuing to spend time providing ineffective or otherwise less than ideal treatment. A commitment to research-based or evidence-based practice in speech-language pathology includes a commitment by clinicians, researchers, students, and other readers to assessing the relevant literature, a task that we hope might be eased for stuttering by the use of instruments such as the STREAT in the context of the many larger recommendations made about EBP.

Acknowledgments

The preparation of this article was supported in part by funds from National Institutes of Health Research Grant R01-DC04838 to Anne K. Bothe. Our thanks to Jenna Levy, Kay Taylor, Brad Crowe, and especially Roger Ingham for their assistance at various stages of the related projects that have developed and used the STREAT.

References

American Speech-Language-Hearing Association. (2005). *Evidence-based practice in communication disorders* [Position statement]. Available from http://www.asha.org/members/deskref-journals/deskref/default

Andrews, G., & Tanner, S. (1982). Stuttering treatment: An attempt to replicate the regulated-breathing method. *Journal of Speech and Hearing Disorders, 47,* 138–140.

Apel, K., & Self, T. (2003, September 9). EBP: The marriage of research and clinical services. *The ASHA Leader,* pp. 6–7.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press.

Bloodstein, O. (1981). *A handbook on stuttering* (3rd ed.). Chicago: The National Easter Seal Society.

Bloodstein, O. (1995). *A handbook on stuttering* (5th ed.). San Diego: Singular.

Boberg, E., & Kully, D. (1994). Long-term results of an intensive treatment program for adults and adolescents who stutter. *Journal of Speech and Hearing Research, 37,* 1050–1059.

Bothe, A. K. (2003). Evidence-based treatment of stuttering: V. The art of clinical practice and the future of clinical research. *Journal of Fluency Disorders, 28,* 247–258.

Bothe, A. K. (Ed.). (2004). *Evidence-based treatment of stuttering: Empirical bases and clinical applications*. Mahwah, NJ: Erlbaum.

Bothe, A. K., Davidow, J. H., Bramlett, R. E., & Ingham, R. J. (2005). *Stuttering treatment research, 1970–2005: I. Systematic review incorporating trial quality assessment of behavioral, cognitive, and related approaches.* Unpublished manuscript.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services*. London: Nuffield Provincial Hospitals Trust.

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

Conture, E. G., & Guitar, B. E. (1993). Evaluating efficacy of treatment of stuttering: School-age children. *Journal of Fluency Disorders, 18,* 253–287.

Cook, D. J., Guyatt, G. H., Laupacis, A., Sackett, D. L., & Goldberg, R. J. (1995). Clinical recommendations using levels of evidence for antithrombotic agents. *Chest, 108*(Suppl.), 227S–230S.

Cordes, A. K. (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. K., & Ingham, R. J. (1994). The reliability of observational data: II. Issues in the identification and measurement of stuttering events. *Journal of Speech and Hearing Research, 37,* 279–295.

Cordes, A. K., & Ingham, R. J. (1998). *Treatment efficacy for stuttering: A search for empirical bases.* San Diego, CA: Singular.

Curlee, R. F. (1993). Evaluating treatment efficacy for adults: Assessment of stuttering disability. *Journal of Fluency Disorders, 18,* 319–331.

Davidow, J. H., Bothe, A. K., Crowe, B. T., & Ingham, R. J. (2005). *Stuttering treatment research evaluation and assessment tool (STREAT).* Athens: The University of Georgia, Department of Communication Sciences and Special Education.

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders, 37,* 391–400.

Einarsdóttir, J., & Ingham, R. J. (2005). Have disfluency-type measures contributed to the understanding and treatment of developmental stuttering? *American Journal of Speech-Language Pathology, 14,* 260–273.

Elliot, A. J., Miltenberger, R. G., Rapp, J., Long, E. S., & McDonald, R. (1998). Brief application of simplified habit reversal to treat stuttering in children. *Journal of Behavior Therapy and Experimental Psychiatry, 29,* 289–302.

Finn, P. (2003). Evidence-based treatment of stuttering: II. Clinical significance of behavioral stuttering treatments. *Journal of Fluency Disorders, 28,* 209–218.

Finn, P., & Ingham, R. J. (1989). The selection of "fluent" samples in research on stuttering: Conceptual and methodological considerations. *Journal of Speech and Hearing Research, 32,* 401–418.

Guitar, B. (2006). *Stuttering: An integrated approach to its nature and treatment* (3rd ed.). Baltimore: Lippincott Williams & Wilkins.

Guyatt, G., & Rennie, D. (Eds.). (2002). *Users' guides to the medical literature: Essentials of evidence-based clinical practice*. Chicago: AMA Press.

Guyatt, G., Sackett, D. L., Sinclair, J. C., Hayward, R., Cook, D. J., & Cook, R. J. (1995). Users' guides to the medical literature: IX. A method for grading health care recommendations. *Journal of the American Medical Association, 274,* 1800–1804.

Haynes, R. B., Sackett, D. L., Gray, J. M., Cook, D. J., & Guyatt, G. H. (1996). Transferring evidence from research into practice: I. The role of clinical care research evidence in clinical decisions. *ACP Journal Club, 125*(3), A14–A16.

Heretier, S. R., Gebski, V. J., & Keech, A. C. (2003). Inclusion of patients in clinical trial analysis: The intention-to-treat principle. *Medical Journal of Australia, 179,* 438–440.

Hewat, S., Onslow, M., Packman, A., & O'Brian, S. (2006). A Phase II clinical trial of self-imposed time-out treatment for stuttering in adults and adolescents. *Disability and Rehabilitation, 28,* 33–42.

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. (1999). Behavioral treatment of young children who stutter: An extended length of utterance method. In R. F. Curlee (Ed.), *Stuttering and related disorders of fluency* (2nd ed., pp. 80–109). New York: Thieme.

Ingham, J. C. (2003). Evidence-based treatment of stuttering: I. Definition and application. *Journal of Fluency Disorders, 28,* 197–207.

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. (1975). 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. (1980). Modification of maintenance and generalization during stuttering treatment. *Journal of Speech and Hearing Research, 23,* 732–745.

Ingham, R. J. (1999). Performance-contingent management of stuttering. In R. F. Curlee (Ed.), *Stuttering and related disorders of fluency* (2nd ed., pp. 200–221). New York: Thieme.

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* (pp. 413–437). Boston: Allyn & Bacon.

Ingham, R. J., & Cordes, A. K. (1999). On watching a discipline shoot itself in the foot: Some observations on current trends in

stuttering treatment research. In N. B. Ratner & E. C. Healey (Eds.), *Stuttering research and practice: Bridging the gap* (pp. 211–230). Mahwah, NJ: Erlbaum.
Ingham, R. J., & Costello, J. M. (1984). Stuttering treatment outcome evaluation. In J. M. Costello (Ed.), *Speech disorders in adults: Recent advances* (pp. 189–223). San Diego, CA: College-Hill.
Ingham, R. J., Kilgo, M., Ingham, J. C., Moglia, R., Belknap, H., & Sanchez, T. (2001). Evaluation of a stuttering treatment based on reduction of short phonation intervals. *Journal of Speech, Language, and Hearing Research, 44,* 1229–1244.
James, J. E. (1981). Behavioral self-control of stuttering using time-out from speaking. *Journal of Applied Behavior Analysis, 14,* 25–37.
Jones, M., Gebski, V., Onslow, M., & Packman, A. (2002). Statistical power in stuttering research: A tutorial. *Journal of Speech, Language, and Hearing Research, 45,* 243–255.
Justice, L. M., & Fey, M. E. (2004, September 21). Evidence-based practice in schools: Integrating craft and theory with science and data. *The ASHA Leader,* pp. 4–5, 30–32.
Kazdin, A. E. (1982). *Single-case research designs*. New York: Oxford University Press.
Ladouceur, R., Cote, C., Leblond, G., & Bouchard, L. (1982). Evaluation of regulated-breathing method and awareness training in the treatment of stuttering. *Journal of Speech and Hearing Disorders, 47,* 422–426.
Langevin, M., & Kully, D. (2003). Evidence-based treatment of stuttering: III. Evidence-based practice in a clinical setting. *Journal of Fluency Disorders, 28,* 219–236.
Law, M. (Ed.). (2002). *Evidence-based rehabilitation: A guide to practice*. Thorofare, NJ: Slack.
Mahaniah, K. J., & Rao, G. (2004). Intention-to-treat analysis: Protecting the integrity of randomization. *The Journal of Family Practice, 53,* 644.
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 and Hearing Research, 15,* 743–751.
Moher, D., Schultz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet, 357,* 1191–1194.
Moscicki, E. K. (1993). Fundamental methodological considerations in controlled clinical trials. *Journal of Fluency Disorders, 18,* 183–196.
Muma, J. R. (1993). The need for replication. *Journal of Speech and Hearing Research, 36,* 927–930.
O'Brian, S., Onslow, M., Cream, A., & Packman, A. (2003). The Camperdown program: Outcomes of a new prolonged-speech treatment model. *Journal of Speech, Language, and Hearing Research, 46,* 933–946.
Onslow, M. (1999). [Review of the book *Stuttering: An integrated approach to its nature and treatment* (2nd ed.)]. *Journal of Fluency Disorders, 24,* 319–332.
Onslow, M. (2003). Evidence-based treatment of stuttering: IV. Empowerment through evidence-based treatment practices. *Journal of Fluency Disorders, 28,* 237–245.
Onslow, M. (Ed.). (in press). *Evidence-based clinical practice in stuttering*. Austin, TX: Pro-Ed.
Onslow, M., Andrews, C., & Lincoln, M. (1994). A control/experimental trial of an operant treatment for early stuttering. *Journal of Speech and Hearing Research, 37,* 1244–1259.
Onslow, M., Costa, L., Andrews, C., Harrison, E., & Packman, A. (1996). Speech outcomes of a prolonged-speech treatment for stuttering. *Journal of Speech and Hearing Research, 39,* 734–749.
Onslow, M., O'Brian, S., Packman, A., & Rousseau, R. (2004). Long-term follow up of speech outcomes for a prolonged-speech treatment for stuttering: The effects of paradox on stuttering treatment research. In A. K. Bothe (Ed.), *Evidence-based treatment of stuttering: Empirical bases and clinical applications* (pp. 231–244). Mahwah, NJ: Erlbaum.
Oxford Centre for Evidence-Based Medicine. (2005). *Levels of evidence and grades of recommendation.* Retrieved May 2, 2005, from http://www.cebm.net/levels_of_evidence.asp
Reed, C. G., & Godden, A. L. (1977). An experimental treatment using verbal punishment with two preschool stutterers. *Journal of Fluency Disorders, 2,* 225–233.
Robey, R. R. (2004). A five-phase model for clinical-outcome research. *Journal of Communication Disorders, 37,* 401–411.
Robey, R. R. (2005, May 24). An introduction to clinical trials. *The ASHA Leader,* pp. 22–23.
Robey, R. R., & Schultz, M. C. (1998). A model for conducting clinical-outcome research: An adaptation of the standard protocol for use in aphasiology. *Aphasiology, 12,* 787–810.
Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.
Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research*. New York: McGraw-Hill.
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.
Sackett, D. L., Haynes, R. B., Guyatt, G. H., & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Boston: Little Brown.
Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM* (2nd ed.). New York: Churchill Livingstone.
Schiavetti, N., & Metz, D. E. (2002). *Evaluating research in communicative disorders* (4th ed.). Boston: Allyn & Bacon.
Shaughnessy, A. F., Slawson, D. C., & Bennett, G. H. (1994). Becoming an information master: A guidebook to the medical information jungle. *Journal of Family Practice, 39,* 489–499.
Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.
Slawson, D. C., Shaughnessy, A. F., & Bennett, G. H. (1994). Becoming a medical information master: Feeling good about not knowing everything. *Journal of Family Practice, 38,* 505–513.
Trinder, L., & Reynolds, S. (Eds.). (2000). *Evidence-based practice: A critical appraisal*. Oxford: Blackwell Science.
U.S. Department of Health and Human Services. (2004). *U.S. Department of Health and Human Services Public Health Service Grant Application (PHS 398) PART II: Supplemental Instructions for Preparing the Human Subjects Section of the Research Plan* (OMB No. 0925-0001). Washington, DC: Author.
Yairi, E., & Ambrose, N. (1992). A longitudinal study of stuttering in children: A preliminary report. *Journal of Speech and Hearing Research, 35,* 755–760.

Received May 26, 2005
Revision received October 10, 2005
Accepted January 29, 2006
DOI: 10.1044/1058-0360(2006/013)

Contact author: Anne K. Bothe, Department of Communication Sciences and Special Education, 556 Aderhold Hall, The University of Georgia, Athens, GA 30602.
E-mail: abothe@uga.edu

Appendix (p. 1 of 5)

STREAT: Stuttering Treatment Research Evaluation and Assessment Tool

Name/type of treatment:_____ Age of participants:_____

Date published:_____

Directions: Mark each item that is included in the treatment investigation. Section I is a summary of five important methodological characteristics that can be completed independently or completed as a summary after Sections II–VII have been completed. Asterisks in Sections II–VII identify information summarized in Section I.

I. FIVE BASIC CHARACTERISTICS

a.___ Basic research strategy was either random assignment to groups or a single-subject experimental design with withdrawal or other controls (from II.a.1. and II.b.1.)

b.___ Data requiring observer judgments were gathered by a judge blind as to participant's phase/status/condition (preferable), or original judge's data show 80% or better agreement with data produced by an independent judge (acceptable) (from Section V)

c.___ Data are presented from before, during, and after treatment (from III.l.–III.bb.)

d.___ Data are presented from within and beyond the clinic (from III.g. and III.h.)

e.___ If reduced stuttering frequency is claimed, then the three common potential confounds to interpretation of reduced stuttering frequency are all controlled: speech rate is normal, speech naturalness is normal, interjudge agreement is 80% or better (from Section V and from VII.g.)

II. STRATEGY/PARTICIPANTS/SAMPLE SIZE

Strategy

a.___Single-subject design

1. ___ Design includes withdrawal, multiple baselines, or any feature more complex than case study or AB design*

b.___Group design

1. ___ Random assignment to groups*

Recruitment and Participants

c. ___Clear description of selection procedures

Selection procedures were:_____

d. ___Inclusion/Exclusion criteria listed

Criteria were:_____

Sample Size

e.___At least 4 participants in final data analysis, for a single-subject design

f.___At least 10 participants in final data analysis, for a group design

g.___Number of participants at initiation of study stated

Number of participants:_____

h.___Number of participants at completion of study stated

Number of participants:_____

III. DEPENDENT VARIABLES (DVs)

Selection of Specific DVs

a.___Stuttering frequency

1. ___ Clear definition

Definition was: _____

b.___Speech rate

1. ___ Clear definition

Definition was: _____

c.___Speech naturalness

1. ___ Clear definition

Definition was: _____

d.___Stuttering severity

1. ___ Clear definition

Definition was: _____

e.___Self-report (attitudes, cognitions, etc.)

1. ___ Clear definition

Definition was: _____

f.___Other DVs: _____

1. ___ Clear definition

Definition was: _____

Situations in Which Speech-Related DVs Were Measured

g.___At least one beyond-clinic situation*

h.___At least one within-clinic situation*

i.___Specification of speaking style for assessment (e.g., reading, monologue, or conversation)

1. Style: _____

j.___Conversation sample is included

k.___Covert samples

Frequency With Which All DVs Were Measured*

l.___Pretreatment data were taken and presented

m.___At least 3 data points before treatment for speech dependent variables

1. Speech dependent variables:_____

n.___Pretreatment data taken (at least 2 data points) and presented over at least 1 month for speech dependent variables

1. Speech dependent variables:_____

o.___There was an establishment phase

1. Activities:_____

p.___Data taken and presented at least 1 time during the establishment phase

1. Dependent variables:_____

q.___Data taken and presented immediately after the establishment phase

1. Dependent variables:_____

r.___There was a transfer phase

1. Activities:_____

s.___Data taken and presented at least 1 time during the transfer phase

1. Dependent variables:_____

t.___Data taken and presented immediately after the transfer phase

1. Dependent variables:_____

u.___Establishment and transfer activities performed simultaneously

1. Activities:_____

v.___Data taken and presented at least 1 time during simultaneous establishment and transfer

1. Dependent variables:_____

w.___Establishment and transfer activities were performed simultaneously, and data were taken and presented immediately after these simultaneous phases

1. Dependent variables:_____

x.___There was a maintenance phase

1. Activities:_____

y.___Data taken and presented at least 1 time during the maintenance phase

1. Dependent variables:_____

z.___Data taken and presented immediately after the maintenance phase

1. Dependent variables:_____

aa.___There was a follow-up period

1. Activities:_____

bb.___Data taken after a follow-up period

1. Dependent variables:_____

Length of Speech Samples in Which DVs Were Measured

cc.___Each sample is at least 3 min or 500 words

IV. TREATMENT FIDELITY

Qualified Therapists Doing Treatment

a.___It is stated that the therapists conducting treatment were trained on (or developed) the treatment procedure or the authors performed the procedure

Description of Treatment for Replication

b.___Description clear enough for replication or a reference to another source with a clear explanation

Treatment Administered as Described

c.___Any evidence to suggest that treatment was performed correctly
d.___Treatment fidelity data

V. DATA FIDELITY

Intrajudge Reliability or Agreement Scores of at Least .80 or 80% for:

a.___Stuttering frequency
b.___Speech rate
c.___Speech naturalness
d.___Stuttering severity
e.___Self-report (attitudes, cognitions, etc.)
f.___Other DVs: _____

Interjudge Reliability or Agreement Scores of at Least .80 or 80% for:*

g.___Stuttering frequency
h.___Speech rate
i.___Speech naturalness
j.___Stuttering severity
k.___Other DVs: _____
l.___If participants dropped out, a report of why they dropped out
m.___At least one independent observer taking measurements

Appendix (p. 4 of 5)

STREAT: Stuttering Treatment Research Evaluation and Assessment Tool

VI. DATA ANALYSIS

a.___Graphs, if single-subject design

b.___Statistics to support all claims of group differences (between- and within-group designs)

c.___Power specified

VII. RESULTS

Last Data Reported

a.___After establishment

b.___After transfer

c.___After simultaneous establishment and transfer phases

d.___After maintenance

e.___After follow-up phase:

f. Last treatment phase (besides follow-up):_____

Data Collection*

g. Name of dependent variable #1:_____

1. ___Stable baseline before treatment
2. ___Pretreatment average: _____
3. ___Range of pretreatment scores: _____
4. ___Final average from last data reported: _____
5. ___Range of final scores: _____
6. ___At least 1 year has passed since last treatment phase

   Number of months after last treatment phase that this dependent variable was reported:_____
7. ____Last data reported were collected beyond the clinic
8. ____Last data reported were collected within the clinic

h. Name of dependent variable #2:_____

1. ___Stable baseline before treatment
2. ___Pretreatment average: _____
3. ___Range of pretreatment scores: _____
4. ___Final average from last data reported: _____
5. ___Range of final scores: _____
6. ___At least 1 year has passed since last treatment phase

   Number of months after last treatment phase that this dependent variable was reported:_____
7. ____Last data reported were collected beyond the clinic
8. ____Last data reported were collected within the clinic

i. Name of dependent variable #3:_____

1. ___Stable baseline before treatment
2. ___Pretreatment average: _____
3. ___Range of pretreatment scores: _____
4. ___Final average from last data reported: _____
5. ___Range of final scores: _____
6. ___At least 1 year has passed since last treatment phase

   Number of months after last treatment phase that this dependent variable was reported:_____

Appendix (p. 5 of 5)

STREAT: Stuttering Treatment Research Evaluation and Assessment Tool

7. ____Last data reported were collected beyond the clinic
8. ____Last data reported were collected within the clinic

Length of Establishment Phase

j.___Greater than 30 hr
k.___20 to 30 hr
l.___10 to 20 hr
m.___Less than 10 hr

VIII. REFERENCE AND JUDGMENT INFORMATION

Complete reference of article: _____

_____

Reader's name: _____ Date: _____

Note. From Davidow et al. (2005). Reprinted by permission of the authors.

The Stuttering Treatment Research Evaluation and Assessment Tool (STREAT): Evaluating Treatment Research as Part of Evidence-Based Practice

Jason H. Davidow, Anne K. Bothe, and Robin E. Bramlett
Am J Speech Lang Pathol 2006;15;126-141
DOI: 10.1044/1058-0360(2006/013)