Neuropsychol Rev (2006) 16:161–169
DOI 10.1007/s11065-006-9013-7

ORIGINAL PAPER

Evaluating Single-Subject Treatment Research: Lessons Learned from the Aphasia Literature

Pélagie M. Beeson · Randall R. Robey

Received: 2 November 2006 / Accepted: 5 November 2006 / Published online: 7 December 2006

Abstract The mandate for evidence-based practice has prompted careful consideration of the weight of the scientific evidence regarding the therapeutic value of various clinical treatments. In the field of aphasia, a large number of single-subject research studies have been conducted, providing clinical outcome data that are potentially useful for clinicians and researchers; however, it has been difficult to discern the relative potency of these treatments in a standardized manner. In this paper we describe an approach to quantify treatment outcomes for single-subject research studies using effect sizes. These values provide a means to compare treatment outcomes within and between individuals, as well as to compare the relative strength of various treatments. Effect sizes also can be aggregated in order to conduct meta-analyses of specific treatment approaches. Consideration is given to optimizing research designs and providing adequate data so that the value of treatment research is maximized.

Keywords Effect size · Treatment · Rehabilitation · Outcomes · Evidence based practice · Stroke · Meta-analysis

P. M. Beeson (✉)
Department of Speech, Language, and Hearing Sciences,
Department of Neurology,
The University of Arizona, 1131 E. Second Street,
Tucson, AZ 85721-0071, USA
e-mail: pelagie@u.arizona.edu

R. R. Robey
Communication Disorders Program, University of Virginia,
Charlottesville, Virginia, USA

Introduction

A prominent question in the field of aphasiology, as in all clinical disciplines, is *What are the empirically supported treatment approaches?* There is a relatively rich literature on the treatment of aphasia and related disorders, such as acquired alexia and agraphia, so this question would appear to be one with a ready answer. In fact, a review of the published literature yields over 600 articles spanning about five decades that specify and examine treatment approaches for aphasia and related disorders (see ANCDS aphasia treatment website http://www.u.arizona.edu/∼pelagie/ancds/index.html). At last count, 288 articles employed group designs to examine treatment effects, and the remaining 332 involved single-subject research, either in the form of case reports (80) or single-subject experimental studies (252). Although this large body of treatment literature serves to inform clinicians and researchers, it is difficult to discern the relative potency of the various treatments and to synthesize the findings in a meaningful manner.

Several researchers have conducted systematic reviews of the aphasia treatment literature and provided useful meta-analyses of the outcomes from group studies (Whurr, Lorch, & Nye, 1992; Robey, 1994; Robey, 1998). In general, these meta-analyses offer converging evidence to suggest that aphasia treatment brings about meaningful, positive change in language performance relative to untreated controls. By necessity, the meta-analyses combined the outcomes from an array of heterogeneous treatment approaches described with varying levels of specificity. Because a large number of aphasia treatment studies employ single-subject methodology, Robey, Schultz, Crawford, and Sinner (1999) set out to complement the meta-analyses of group studies with a synthesis of the data from single-subject treatment research. Of 63 studies meeting criteria for inclusion, only 12 provided adequate data to allow the quantification of effect sizes, and those studies were directed toward a diverse set of language behaviors, so a meta-analysis was not deemed appropriate. However, the calculation and reporting of effect sizes from this cohort of studies provided a starting point from which to evaluate subsequent research in aphasia treatment. In summary, the endeavors to synthesize the aphasia treatment literature offered support for the therapeutic value of behavioral treatments in general, but also served to guide ensuing efforts to increase the evidence base for specific treatment approaches.

Among the issues that have come into better focus in the past decade is the fact that treatment outcome research is best conducted in phases, so there is a logical, principled progression in rehabilitation research that encompasses single-subject as well as group research designs (Robey & Schultz, 1998; World Health Organization, 1975). New treatments should first be examined with a small number of individuals to test the therapeutic effect (Phase 1), followed by additional studies to optimize procedures, discern the most appropriate candidates for treatment, and further explore the potential efficacy of the treatment (Phase 2). Positive results from these pre-efficacy studies should prompt well-controlled group designs that test the efficacy of the treatment under ideal conditions (Phase 3). In other words, large-scale research efforts should be reserved for techniques that have positive outcomes from Phase 1 and 2 studies (Garrett & Thomas, 2006; Robey & Schultz, 1998). Only after a treatment has been shown to be efficacious in Phase 3 studies should research ensue to examine the potency of treatment under typical conditions of service delivery (i.e., Phase 4 effectiveness studies). Finally, practical issues such as cost-benefit analysis can be addressed (Phase 5).

The fact of the matter is that the majority of aphasia treatment research was completed before the five-phase model was broadly recognized. Not surprisingly then, some appraisals of the body of literature have been rather harsh when assessment criteria were grounded in the five-phase system. For example, the Cochrane Review of "speech and language therapy for aphasia following stroke" restricted its selection criteria to include randomized control trials, which yielded only 12 studies for consideration at that time (Greener, Enderby, & Whurr, 1999). Such restrictive criteria limit the extent to which the existing literature can be used to address questions regarding the value of aphasia treatment. On this point, we appreciate Tim Pring's (2004) appraisal of this situation in his essay entitled "Ask a silly question: Two decades of troublesome trials." We acknowledge that under ideal circumstances the best available evidence would consist of meta-analyses of high-quality randomized control trials. However, in the absence of such evidence, the highest quality and most relevant studies must suffice in constituting the best current evidence. It is our goal to promote better use of the existing body of aphasia treatment literature and to enhance the potential contributions of future research.

In this paper, we focus on the evaluation of single-subject research because such experimental designs have played (and continue to play) a foundational role in the development and testing of aphasia treatments, yet they are often neglected in attempts to evaluate and synthesize the literature. Our perspective comes from ongoing systematic review of the aphasia treatment literature as we seek to synthesize the outcomes and formulate evidence-based practice guidelines for aphasia.[1] Carrying out this process with single-subject research forces a decision on how best to assess the reported data in a consistent manner. Extending the synthesis process established for group studies, we chose to calculate effect sizes as a means of quantifying the outcomes for individual participants. We describe our approach here and provide the rationale for various decisions. This description is intended to assist other researchers, clinicians, and consumers of research in the quantification of single-subject treatment outcomes. Although the information presented here is in the context of aphasia treatment, it has application to a broad range of neuropsychological research directed toward the remediation of behavioral impairments.

Why concern ourselves with effect sizes for single-subject data?

Early applications of single-subject treatment designs were evaluated by visual analysis of graphical data. However, unless differences in performance between measurement periods are very large, a reliable method of quantification is needed to detect treatment effects (Johnston, Ottenbacher, & Reichardt, 1995). That is, a visual analysis can be flawed, and the impression of a positive treatment effect may be false and lead to Type 1 error (i.e., concluding an effect is present when none exists). For example, Matyas and Greenwood (1990) found that Type I error rates for visual analyses ranged from 16% to 84%. Applying inferential statistics that make assumptions about the distributional properties of the parent population is also problematic because single-subject data are inherently autocorrelated. In other words, repeated measures within the same subject are clearly not independent of one another, thus limiting the choice of appropriate statistical analyses (Kromrey & Foster-Johnson, 1996; Robey, Schultz, Crawford, & Sinner, 1999).

An alternative to visual inspection and the use of inferential statistics is the calculation of a standardized effect size as a means for assessing change. An effect size is simply

[1] See the Academy of Neurologic Communication Disorders and Sciences (ANCDS) website http://ancds.org for information regarding evidence-based practice guidelines for neurogenic communication disorders.

a quantity that characterizes the degree of departure from the null state, which, in this case, is the degree to which a treatment outcome differs from zero. In other words, an effect size contrasting pre-treatment and post-treatment levels of performance provides a measure of change observed through some variable of interest. Because effect sizes are quantified in standard deviation units, they can be compared across studies and combined in meta-analyses.

The reporting of effect sizes is not a new idea. Over the past several decades, many statisticians have advocated this practice (Bezeau & Graves, 2001; Cohen, 1988; Fan, 2001; Hyde, 2001; Kirk, 1996, 2001; Nix & Barnette, 1998; Shaver, 1993; Thompson, 1998, 2002; Vacha-Haase, 2001). In fact, the fifth edition of the *Publication Manual of the American Psychological Association* (2001) stated, "For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section. You can estimate the magnitude of effect or the strength of the relationship with a number of common effect size estimates . . ." The manual further states that authors should "provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship" (pp. 25–26). Implementation of this practice is increasing for group studies, but very few researchers conducting single-subject research report effect sizes.

When effect sizes are included in published reports, they allow clinicians and researchers to develop a sense of the relative strength of the specific treatments. In addition, when a line of research produces multiple estimates of an effect size parameter in the context of independent experiments, it is possible to pool (i.e., average) them. Finding an average effect size for a certain body of research literature is the basis for meta-analysis. From an evidence-based practice perspective, a meta-analysis of highest quality research on a certain intervention protocol constitutes the most persuasive form of clinical scientific evidence (Harbour & Miller, 2001). Of course, the accumulation of evidence for specific treatments is an ongoing process, so new evidence must be considered as it becomes available. Consistent reporting of effect sizes by researchers would facilitate the ease with which this is accomplished.

What studies should be considered?

When examining the literature for relevant studies that evaluate a treatment approach of interest, the researcher's goal is to glean information provided by fundamentally sound scientific evidence regarding the magnitude of treatment effects. Potential studies are those that state testable hypotheses regarding treatment outcomes. Single-subject treatment studies typically examine pre-treatment versus post-treatment performance within a given participant, or treatment versus no treatment conditions across individuals. The studies must provide quantities necessary for calculating effect size (e.g., raw data in tables or graphs, exact probabilities, descriptive statistics). Studies with poor experimental validities that would render an effect meaningless must be excluded (Wortman, 1994).

What are the dependent variables of interest?

Whereas many of the group studies of aphasia treatment quantify change using an index of overall language performance, the single-subject studies typically describe the effects of treatment directed toward relatively specific language processes. Therefore, these studies can be logically grouped according to the nature of the dependent variable. In our review of the literature, we have found that essentially all single-subject aphasia treatment studies are directed toward one or more of the following language processes: lexical retrieval; syntactic production or comprehension; speech production or fluency; auditory comprehension; and reading, writing, or some form of alternative communication, such as gesture.

The majority of aphasia treatment studies focus on the measurement of direct treatment effects, that is, changes in the targeted behaviors or items that are trained. Additional measures of generalization are reported with varying consistency, including performance on untrained items or the use of specific linguistic variables in connected speech. Because these measures sample different levels of performance, it is best to consider them separately. Therefore, in our reviews, three different outcomes are gleaned from treatment studies: direct treatment effects, generalization to untrained items, and generalization to connected speech. In general, these can be considered to reflect increasing levels of difficulty.

How to calculate effect sizes for single-subject research

Single-subject treatment studies typically are designed to measure change in performance on a variable (or variables) of interest in response to the described treatment. Such studies commonly include behavioral measures before treatment (the first *A* phase, A_1), during treatment (the *B* phase), and after treatment (the second *A* phase, A_2). The ABA design is ideal for calculating effect size from single-subject research, and its value is enhanced when the ABA phases are repeated across several sets of stimuli or behaviors (Backman, Harris, Chisholm, & Monette, 1997; McReynolds & Thompson, 1986; Sidman, 1960). This type of multiple baseline design is illustrated in Fig. 1. The value of this design lies in the repeated measures of the behaviors of interest over time and the staggered initiation of treatment that allows

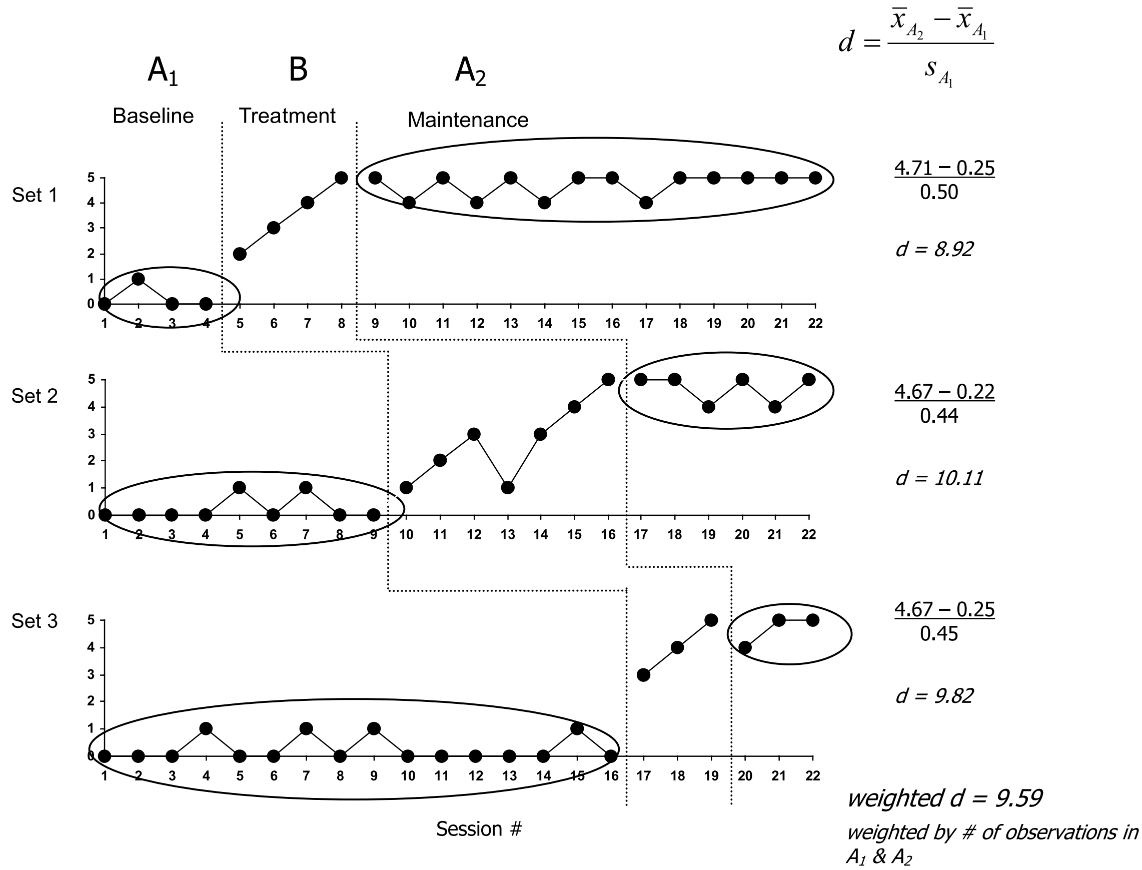

Fig. 1 Example of data from single-subject multiple baseline design. The circled data points are those that contribute to the calculation of *d* statistic to index changes in level of performance (see raw data in Table 1)

one to examine the specificity of treatment effect relative to a control condition.

In the example data provided in Fig. 1, three sets (or groups) of stimuli were targeted for treatment, and each set comprised five items (or words). In this example, the participant might be an individual with acquired agraphia who was retrained on the spelling of three sets of five words. Using a multiple baseline design, spelling performance was probed on items during the first four sessions, and then treatment was initiated for the five items comprising Set 1. A criterion for mastery was established *a priori*, so the items were to be treated until performance was 80% (4 out of 5 words) or better over two sessions. When criterion for Set 1 was achieved, treatment for Set 2 was initiated, and so forth for Set 3. In this example, performance was probed on all items during each session, allowing for consistent sampling throughout ABA phases. The second *A* phase traditionally is referred to as the withdrawal phase. In treatment research, however, it is not uncommon for some maintenance-promoting activities, such as homework, to continue during the A_2 phase, so it is better characterized as the post-treatment or maintenance phase.

In order to measure the degree of change in the behavior of interest from pre-treatment to post-treatment, the level of performance from the first *A* phase is compared to that of the second *A* phase. For the example in Fig. 1, the baseline performance was near zero, but it might be the case that pre-treatment performance on the dependent variable of interest is at chance or some intermediate level of performance. Regardless, the comparison of interest is the change in the level of performance from pre- to post-treatment. The null hypothesis H_0 is that pre-treatment levels will be equal to or greater than post-treatment levels.

$$H_0 : \beta_{\text{level}_{A1}} \geq \beta_{\text{level}_{A2}}$$

The research hypothesis for H_0 asserts that the overall level of performance increases from pre-treatment to post-treatment. In applications where benefit is realized through a decrease in the target behavior, the sign is reversed. When the rate of change, or the profile of change, is of interest, it can be assessed through a different null hypothesis, contrasting the slope of the A_1 period with the slope of the *B* period. For the data depicted in Fig. 1, changes in level from A_1 to A_2 and changes in slope from A_1 to B are both evident. In our current review of aphasia treatments, we focus on determining *how much* change can be effected by the treatment, and we are less concerned with how fast the change is made or how variable the performance is during the treatment phase itself. For that reason, we calculate the effect size based on changes in level of performance, rather than changes in slope. In later phases of outcome research, after treatment efficacy has been established, information provided by examination of slope may be of particular interest. For example, if two treatments have been shown to be effective, it may be useful to compare the relative efficiency of each by examining differences in the acceleration and duration of slopes in the learning curve (i.e., the slope of B compared to the slope of A_1).

In order to quantify the magnitude of the change in level of performance, we use a variation of Cohen's (1988) d statistic as calculated by Busk and Serlin (1992, pp. 197–198):

$$d_1 = \frac{\bar{x}_{A_2} - \bar{x}_{A_1}}{S_{A_1}}$$

where A_2 and A_1 designate post-treatment and pre-treatment periods, respectively, $\bar{x}_A$ is the mean of the data collected in a period, and S_A is the corresponding standard deviation. This statistic was empirically selected from seven potential estimators for change, on the basis of a Monte Carlo simulation. The estimators included Percent Non-Overlapping Data (PND) (Scruggs & Mastropieri, 1998), f^2 (Kromrey & Foster-Johnson, 1996), and several different equations for calculating a d statistic: d_1 (Busk & Serlin, 1992), d_2 (Busk & Serlin, 1992; White, Rusch, Kazdin, 1989), d overall (Faith, Allison, & Gorman, 1997), d level-only (Faith et al., 1997), and d (Center, Skiba, & Casey, 1985–1986). Although our evaluation of these statistics is ongoing, empirical assessment to date indicates that the first of Busk and Serlin's d statistics (d_1) is the most reliable estimator of the effect size when the pre-treatment variance is a non-zero value (see discussion below).

In multiple baseline designs, effect sizes can be calculated for each series of data points and then averaged to represent the treatment effect for a single individual. As illustrated in Tables 1 and Table 2, a d statistic is calculated for Set 1 on the basis of 4 pre-treatment probes and 14 post-treatment probes, for a total of 18 observations. Because each baseline can comprise a different number of observations, averaging over baselines is best accomplished using a weighted mean. In this case, the first d statistic of 8.92 is weighted for the 18 observations (8.92 × 18). It is then added to the corresponding weighted values from Set 2 (10.11 × 15) and Set 3 (9.82 × 19). The sum is then divided by the total number of observations (18 + 15 + 19). As shown in the figure and tables, the weighted average d statistic for the treatment effect for this participant was 9.59. It should be evident from this example that calculation of the d_1 statistic is accomplished through basic mathematics. Thus, practicing clinicians could calculate effect sizes to quantify the magnitude of change demonstrated by their patients in response to treatment.

The one circumstance under which the Busk and Serlin d_1 statistic cannot be calculated is when there is no variance during the A_1 phase. In other words, if each pre-treatment probe has the same value (such as zero), then the A_1 variance equals zero, so the calculation for d becomes impossible. In such cases, some other estimate of variance must be used.

Table 1 Raw data that are plotted in Fig. 1

	Set 1		Set 2		Set 3	
Session	Phase	Value	Phase	Value	Phase	Value
1	A_1	0	A_1	0	A_1	0
2	A_1	1	A_1	0	A_1	0
3	A_1	0	A_1	0	A_1	0
4	A_1	0	A_1	0	A_1	1
5	B	2	A_1	1	A_1	0
6	B	3	A_1	0	A_1	0
7	B	4	A_1	1	A_1	1
8	B	5	A_1	0	A_1	0
9	A_2	5	A_1	0	A_1	1
10	A_2	4	B	1	A_1	0
11	A_2	5	B	2	A_1	0
12	A_2	4	B	3	A_1	0
13	A_2	5	B	1	A_1	0
14	A_2	4	B	3	A_1	0
15	A_2	5	B	4	A_1	1
16	A_2	5	B	5	A_1	0
17	A_2	4	A_2	5	B	3
18	A_2	5	A_2	5	B	4
19	A_2	5	A_2	4	B	5
20	A_2	5	A_2	5	A_2	4
21	A_2	5	A_2	4	A_2	5
22	A_2	5	A_2	5	A_2	5

Note. A_1 = pre-treatment phase; B = treatment phase, A_2 = post-treatment phase. To calculate the d statistic, the values from A_1 and A_2 phases are evaluated.

Table 2 Analysis of data presented in Table 1 (and Fig. 1)

	Set 1	Set 2	Set 3	Sum
Mean A_1	0.25	0.22	0.25	–
Mean A_2	4.71	4.67	4.67	–
Mean A_2 – Mean A_1	4.46	4.45	4.42	–
SD A_1	0.50	0.44	0.45	–
d	8.92	10.11	9.82	–
# observations $A_1 + A_2$	18.00	15.00	19.00	52.00
Weighted d	160.56	151.70	186.62	498.88
Weighted d for all data	–	–	–	9.59

Note. A_1 = pre-treatment phase; A_2 = post-treatment phase; *SD* = standard deviation.

Busk and Serlin (1992) and White et al. (1989) addressed this issue by pooling the variance from A_1 and A_2 for the calculation of a different d statistic (Busk and Serlin's equation for d_2). However, as Busk and Serlin point out, the pooling of the A_1 and A_2 variances assumes that they are relatively homogeneous, a condition that is often violated in the single-subject data that we have reviewed. Another option is to replace the zero-variance value with the mean variance of the other A_1 phase data for the same individual. Discerning the best resolution of this issue is a focus of our current research. In the meantime, we use Busk and Serlin's d_2 equation:

$$d_2 = \frac{\bar{x}_{A_2} - \bar{x}_{A_1}}{s_{\text{pooled}}}$$

where A_2 and A_1 designate post-treatment and pre-treatment periods, respectively, $\bar{x}_A$ is the mean of the data collected in a period, and s_{pooled} is the square root of the weighted average of the variances for A_1 and A_2.

Many researchers conducting single-subject experiments collect the necessary information to calculate effect sizes, but the data are not included in the published manuscript. In order to calculate the effect sizes, it is necessary to determine the individual values for the pre-treatment and post-treatment phases for each set of trained items. The values may be available in tables, or they may be retrievable from graphs. The values are easy to retrieve from graphs when plotted in whole number increments. For example, when there are 5 items to be trained in a particular set, and the probe data are presented as the total number correct, one can easily discern the value of each data point. When the units are fractional, or the resolution of the plots is insufficient for determining exact values of the data points, it may be easiest to enlarge the size of the graph and use a ruler to line up the Y-axis values with data points. If uncertainty remains regarding the value of the data points, an alternative approach is to measure the vertical distances from abscissa to the center of plotted points, and to substitute distance for the original values (Faith et al., 1997). This is simply a linear transformation of the original values, in which $x' = xc$, where x' is distance, x is the original score, and c is a constant. Calculation of the d statistic can be accomplished using the set of distances in place of the original values. When measuring these values, particular care should be taken to ensure reliability. We suggest taping a photocopy of the graph to a table top with the page adjusted so the abscissa is parallel to the drawing arm of a T-square. A vertical line is drawn from the center of each data point through the abscissa. Distances are then measured using a digital caliper, with one arm of the caliper set in the center of a data point and the other set at the intersection of the corresponding vertical line and abscissa. Measurement increments should be consistent, for example, at 0.001 inch.

It is important to establish reliability in the measurement of effect sizes, in particular when the values are gleaned from other papers and the measurements are made by hand. Inter-rater reliability should be established by re-measurement of a subset of plots (0.10–0.20). Levels of inter-rater reliability should be greater than 0.90.

Addressing some challenging questions when calculating effect sizes

Despite efforts to provide guidelines for the evaluation of research data, the actual process of calculating effect sizes is fraught with questions regarding how to handle data that are not ideal. While there may be some latitude in establishing rules for the gray areas, it is important to document the decision-making process so that uniform rules are applied to all studies. Below we address some of the challenging questions that we have faced.

What is the minimum number of pre-treatment baseline probes that will still allow calculation of effect size? Mathematically, two observations in the baseline period are necessary to solve for d_I. As in all forms of research, the greater the number of valid observations, the more accurate is the result of any analysis. However, as a practical matter in clinical-treatment research, the first baseline period is often brief so that treatment can begin. Since a crucial estimate arising out of the initial A period is the standard deviation, we suggest three observations as a minimum. When the data are available, d_I is averaged across multiple baselines, thus providing an opportunity to combine short A_1 periods in early baselines with longer durations in later baselines.

What is the minimum number of post-treatment probes that will still allow calculation of effect size? Mathematically, only one observation in the A_2 period is necessary for the calculation of d_I. Once again, a greater number provides a better estimate and, more importantly, offers information regarding the durability of the treatment effect over time. We prefer three probes, but two are considered allowable. As noted above, there is benefit from averaging d_I across multiple baselines so that shorter A_2 periods in later baselines are combined with long A_2 periods from early data sets.

Can an effect size be calculated from data in an AB design, that is, when there is no A_2 phase? The problem with such data is that they only provide information about the slope in phases A and B, and do not sample the performance after treatment is completed (A_2). Therefore, they do not provide adequate information for the calculation of effect size.

How are the data used for items that are never treated? The data from performance on untreated items do not contribute to calculation of effect sizes for direct treatment. However, these data may be examined as an index of experimental control to confirm the treatment effect is specific to the trained items. Alternatively, improved performance on untrained items may provide evidence of generalization. Change in the level of performance on the untrained items could be estimated by comparing a series of representative data points collected at the beginning of the treatment protocol (to approximate an A_1 phase) with a series of representative data points collected at the end of the maintenance period (to approximate an A_2 phase).

How should data be analyzed for behaviors probed infrequently or at highly irregular intervals? These data should not be included in calculation of effect sizes.

How should follow-up data be analyzed? There is marked variability in research reports regarding the inclusion of follow-up probes obtained well after the cessation of treatment, so it is difficult to include such data in a meta-analysis. However, it is mathematically feasible to calculate an effect size to measure the change in level of performance from A_1 to the mean performance at follow-up. The design can be conceived as $A_1BA_2A_3$, where A_3 represents the extended follow-up probes. The effect size obtained from the A_1 to A_3 phase can be compared to the effect size from the A_1 to A_2 phase, thus providing an index of the durability of the trained behavior.

How should data be considered in more complex designs, such as, $A_1BA_2CA_3$? When B and C are two complementary treatments offered in sequence (e.g., phonetic and semantic modules in a protocol combining phonetic and semantic cueing), the calculated effect size should be centered on the A_1 and A_3 periods. However, if B and C are completely different treatments, the effect for B should be calculated for the initial A_1BA_2 sequence, and the effect for C should be calculated from the A_2CA_3 sequence. To control for order effects in such cases, some participants should undergo treatment using $A_1CA_2BA_3$ sequence. Ultimately, however, if the data from such designs are to be summarized in a single meta-analysis, the comparison of A_1 and A_3 can be used which reflects the overall effects of sequential treatment.

How are effect sizes calculated if the treatment is designed to decrease an undesirable behavior rather than increase a desirable behavior? In such cases, absolute values of calculated effect sizes are used.

How to interpret the magnitude of the effect size

The interpretation of the magnitude of the effect size is not an easy task. It requires an informed means of developing benchmarks to discern the magnitude of small, medium, and large effect sizes for a particular treatment. Ideally, this is an empirical endeavor. In the absence of such data, the benchmarks set forth by Cohen (1988) for the d statistic based on between-group designs are often cited, with 0.2, 0.5, and 0.8 as benchmarks for small-, medium-, and large-sized effects, respectively. However, Cohen (1988) makes it clear that these benchmarks are based on certain applications in psychology, and the referents may have very little utility in other contexts. Even limited exposure to the magnitude of effect sizes in single-subject research makes it clear that Cohen's benchmarks are inappropriate in this context. A more reasoned approach is to examine the available effect sizes from an array of single-subject studies directed toward similar behavior. In the area of aphasia treatment, a starting point is offered by the effect sizes reported in the Robey et al. (1999) review of single-subject research in aphasia. With one extreme outlier removed from the effect sizes derived from 12 studies, the first, second, and third quartiles for the d statistic were 2.6, 3.9, and 5.8, corresponding to small-, medium-, and large-sized effects. These values offered initial benchmarks for the interpretation of the data in several recent single-subject studies in acquired alexia and agraphia (Beeson & Egnor, 2006; Beeson, Magloire, & Robey, 2005).

Greater understanding of the magnitude of effect sizes is emerging as meta-analyses are conducted for treatments directed toward specific language processes. We recently provided tentative benchmarks for single-subject effect sizes for syntactic production treatment, based on a review of 14 studies with retrievable effect sizes for the direct treatment effect (Robey & Beeson, 2005). Using rounded values from the 95% confidence intervals, small, medium, and large effect sizes yielded the following benchmarks: 6.0, 12.0, and 18.0. Similarly, a review of 12 studies with retrievable effect sizes for lexical retrieval treatments yielded benchmarks of 4.0, 7.0, and 10.1 for small, medium, and large effect sizes (Robey & Beeson, 2005).

Averaging effect sizes across studies: Conducting a meta-analysis

A meta-analysis can be conducted when a group of effect sizes is available that is relevant to a common question. The application of meta-analysis to single-subject research in aphasia is discussed in detail in Robey et al. (1999), so we provide only an overview here. In effect, the meta-analysis provides an average effect size derived from all valid and relevant evidence. The primary studies in the meta-analysis should meet inclusionary criteria with regard to relevance and validity (Moher et al., 1999). Each study contributes no more than one effect size to any averaging process in order to avoid excess influence from any one study. However, some studies contribute estimates of effect size for direct treatment effects (e.g., probes of treated tokens) as well as generalization effects (e.g., probes of untreated tokens). All averages are weighted for the number of observations corresponding to each effect. In the case of multiple baseline across-subjects designs, average effect sizes can be calculated for each subject, and a weighted average of the effect sizes for all subjects represents the treatment effect for the entire single study. When implemented appropriately, a meta-analysis objectively and fairly combines the results of many independent studies into a single and coherent conclusion. Ideally, separate meta-analyses will be conducted for each of the relevant dependent variables of interest.

Conclusions and advice to researchers and clinicians

In this paper we have presented our current approach to the analysis of single-subject data so that treatment outcomes can be quantified in a standard manner and synthesized using meta-analysis. The advantage of this approach is that it offers a means to evaluate new treatment approaches relative to existing approaches, and it helps to shape an emerging expectation regarding what constitutes a potent treatment. This enterprise will advance more rapidly as researchers routinely report effect sizes or at least provide the necessary data required for their calculation.

To increase the likelihood that a given study will contribute to the treatment outcomes research in a significant manner, researchers should clearly state the intention of the study so that the testable hypothesis regarding treatment outcomes is evident. For example, information should be provided regarding the following questions: What is the phase of the research? What are the dependent variables of interest? Does this study examine direct treatment effects only or are there measures of generalization? With regard to sampling performance over time, at least three pre-treatment and post-treatment probes should be obtained on the dependent variable(s) of interest. The addition of follow-up probes at some time after the cessation of treatment will enhance the value of the study by providing an index of durable changes in level of performance over time. Graphic display and/or tabled data should clearly provide the values obtained for the probes during pre-treatment, treatment, post-treatment (withdrawal or maintenance), and follow-up phases. The appropriate effect sizes should be reported along with the raw data and the equation(s) used, so that subsequent researchers can verify the quantification of the outcome. Finally, the discussion of the results should interpret the findings relative to other aphasia treatment outcomes, both from quantitative and qualitative perspectives, and it should offer comments regarding the direction of future research that will advance evidence-based practice.

In closing, we appreciate the mandate to provide empirical support for treatments implemented in clinical practice has prompted critical review of the existing aphasia treatment literature. This endeavor inevitably engenders a certain amount of regret regarding the weaknesses of previous studies, but it also provides insight regarding the direction and rigor necessary for future research. We anticipate that increased understanding of single-subject treatment outcomes and routine calculation of effect sizes will help to promote evidence-based practice in aphasia and other areas of neurorehabilitation.

Acknowledgement The authors thank Susan Carnahan and Ronald Lazar for their helpful comments and suggestions regarding this paper. This work was supported by the Academy of Neurologic Communication Disorders and Sciences (ANCDS), the American Speech-Language-Hearing Association (ASHA), and Division 2: Neuropsysiology and Neurogenic Communication Disorders of ASHA. The first author is also supported in part by RO1DC007646 and RO1DC008286 from the National Institute on Deafness and Other Communication Disorders.

References

Backman, C. L., Harris, S. R., Chisholm, J. M., & Monette, A. D. (1997). Single-subject research in rehabilitation: A review of studies using AB, withdrawal, multiple baseline, and alternate treatments designs. *Archives of Physical Medicine and Rehabilitation*, *78,* 1145–1153.

Beeson, P. M., & Egnor, H. (2006). Combining treatment for written and spoken naming. *Journal of the International Neuropsychological Society, 12,* 816–827.

Beeson, P. M., Magloire, J., & Robey, R. R. (2005). Letter-by-letter reading: Natural recovery and response to treatment. *Behavioural Neurology*, *16,* 191–202.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of neuropsychology research. *Journal of Clinical and Experimental Neuropsychology*, *23,* 399–406.

Busk, P. L., & Serlin, R. (1992). Meta-analysis for single case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, *19,* 387–400.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, Second edition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1997). Meta-analysis of single-case research. In D. R. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fan, X. (2001). Statistical significance and effect size in educational research: two sides of a coin. *Journal of Educational Research*, *94,* 275–282.

Fink, R. B., Brecher, A., Schwartz, M. F., & Robey, R. R. (2002). A computer-implement protocol for treatment of naming disorders: Evaluation of clinician-guided and partially self-guided instruction. *Aphasiology*, *16,* 1061–1086.

Garrett, Z., & Thomas, J. (2006). *International Journal of Language and Communication Disorders*, *43,* 95–105.

Greener, J., Enderby, P., & Whurr, R. (1999). Speech and language therapy for aphasia following stroke (Review). *The Cochrane Library*, *2,* 1–62.

Harbour, R., & Miller, J. (2001). A new system for grading recommendations in evidence based guidelines. *British Medical Journal*, *323,* 334–336.

Hyde, J. S. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. *Educational and Psychological Measurement*, *61,* 225–228.

Johnston, M. V., Ottenbacher, K. J., & Reichardt, C. S. (1995). Strong quasi-experimental designs for research on the effectiveness of rehabilitation. *American Journal of Physical Medicine and Rehabilitation*, *74,* 383–392.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56,* 746–759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61,* 213–218.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The Journal of Experimental Education*, *65,* 73–93.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23,* 341–351.

McReynolds, L. V., & Thompson, C. K. (1986). Flexibility of single-subject experimental designs. Part I: Review of the basics of single-subject designs. *Journal of Speech and Hearing Disorders*, *51,* 194–203.

Moher, D., Cook, D. J., Eastwood, S., Okin, I., Rennie, D., Stroup, D. F., the QUOROM Group. (1999). Improving the quality of reports of meta-analyses of randomized controlled trials: The QUOROM statement. *The Lancet*, *354,* 1896–1900.

Nix, T. W., & Barnette, J. J. (1998). A review of hypothesis testing revisited: Rejoinder to Thompson, Knapp, and Levin. *Research in Schools*, *5,* 55–57.

Publication Manual of the American Psychological Association (5th ed.). (2001). Washington, DC: American Psychological Association.

Pring, T. (2004). Ask a silly question: two decades of troublesome trials. *International Journal of Language and Communication Disorders*, *39,* 285–302.

Robey, R. R. (1994). The efficacy of treatment for aphasic persons: A meta-analysis. *Brain and Language*, *47,* 582–608.

Robey, R. R. (1998). A meta-analysis of clinical outcomes in the treatment of aphasia. *Journal of Speech, Language and Hearing Research*, *41,* 172–187.

Robey, R. R., & Beeson, P. M. (2005). Aphasia treatment: Examining the evidence. Presentation at the American Speech-Language-Hearing Association Annual Convention. San Diego, CA.

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.

Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Single-subject clinical-outcome research: designs, data, effect sizes, and analyses. *Aphasiology*, *13,* 445–473.

Scruggs, T. E., Mastropieri, M. A. (1998). Summarizing single-subject research. *Behaviour Modification*, *22,* 221–242.

Shaver, J. P. (1993). What statistical testing is, and what it is not? *Journal of Experimental Education*, *61,* 293–316.

Sidman, M. (1960). *Tactics of scientific research*. New York: Basic Books, Inc.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in Schools*, *5,* 33–38.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher (3)*, *31,* 25–32.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement*, *61,* 219–224.

White, D. M., Rusch, F. R., & Kazdin, A. E. (1989). Applications of meta-analysis in individual-subject research. *Behavioral Assessment*, *11*(3), 281–296.

Whurr, R., Lorch, M., & Nye, C. (1992). A meta-analysis of studies carried out between 1946 and 1988 concerned with the efficacy of speech and language therapy treatment for aphasic patients. *European Journal of Disorders of Communication*, *27,* 1–17.

World Health Organization (1975). WHO Scientific Group on Guidelines for Evaluation of Drugs for Use in Man (Geneva: World Health Organization).

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis*. New York: Russell Sage Foundation.