Methodological quality of intervention research in speech–language pathology: Analysis of 10 years of group-design studies

Laura M. Justice[1], Chad Nye[2], Jamie Schwarz[2], Anita McGinty[3], & Ana Rivera[2], [1]The Ohio State University, [2]University of Central Florida, and [3]University of Virginia

Abstract

The aim of this work was to critically examine the methodological quality of recent treatment research in speech–language pathology, focusing specifically on group-design studies (randomized, controlled trials and quasi-experimental-design studies). We also considered whether methodological quality differed as a function of a particular study's treatment focus (i.e. literacy, language, speech, fluency, voice), the age group studied (pre-school-age child; school-age child; adult), the year of publication, or the publishing journal. In total, 53 treatment studies published in the last 10 years in three journals of the American Speech–Language–Hearing Association (ASHA) were selected for analysis using Downs and Black's (1998) reliable, valid rating tool for examining methodological quality. The results indicated that the quality of the treatment studies was highly variable in terms of the 25 indicators of quality studied, and that there were few systematic differences in quality attributable to treatment focus, age group studied, year of publication, or the publishing journal. Implications for evidence-based practice and study reporting are discussed.

Keywords: Methodological quality, experimental design, intervention research, evidence-based practice

INTRODUCTION

The press for evidence-based practice (EBP) in the field of speech–language pathology is designed to promote the quality of clinical decision-making for professionals who treat individuals with communication disorders, including speech–language pathologists (SLPs) (Dollaghan, 2007; Fey & Justice, 2007; Schlosser, 2003; Schlosser & Raghavendra, 2004). Presumably, the treatment of communication disorders will become more effective as SLPs integrate the best external evidence from systematic, empirical research with clinical expertise, substantive theoretical knowledge, understanding of clients' values and preferences (American Speech–Language–Hearing Association, 2005), as well as other relevant stakeholder perspectives, including those of family members (Schlosser & Raghavendra, 2004). As a result, clinicians who treat individuals with communication disorders have become increasingly vested in identifying interventions for which there is adequate empirical evidence to justify use in clinical practice. Some clinicians may look towards the systematic review for information about effective intervention. The systematic review may include, but does not require, a meta-analysis of the studies included. In a meta-analysis, researchers aggregate findings from a body of studies featuring similar methods and outcomes using specific statistical techniques (Asher, 1990). However, the strength of findings presented in a systematic review or meta-analysis can

only be as good as the quality of the studies considered therein.

The quality of treatment research can be assessed in several different ways. On the one hand, quality may refer to question significance (Shavelson & Towne, 2002). Question significance refers to the "quality of the question posed" and relies on the ability of scientists to select "insightful questions" that serve to accumulate knowledge within a given discipline (Shavelson & Towne, 2002, p. 55). In treatment research on speech–language pathology, question significance may relate to the number of persons who are affected by the condition under investigation, or the degree to which the condition being studied is considered serious (Gersten et al., 2005). As these points suggest, many scientists who study the treatment of speech-language disorders are engaged in a particular type of science referred to as "use-inspired basic research", which is conducted to simultaneously address the traditional goals of science (e.g. to accumulate knowledge) as well as the needs of society (e.g. to address a pressing societal problem; Stokes, 1997). A significant research question is, therefore, one that furthers knowledge of communication disorders at a theoretical level while simultaneously promotes (indirectly or directly) the quality or effectiveness of treatment for communication disorders.

Another way to address the question of quality is to assess methodological rigor and the appropriate alignment of research methodologies to the specific question or questions being posed (Odom et al., 2005). As was discussed in a recent report by the

National Research Council, scientists must use methods that fit the question and then "competently implement the method" (Shavelson & Towne, 2002). As the previous point makes clear, a key issue is that of competent implementation, as two studies on a given topic can vary immensely in the competency of implementation, which in turn can have significant effects on study findings. For instance, we might have less confidence in the outcomes of a treatment study that failed to measure treatment fidelity (i.e. whether the treatment was implemented as intended), compared with the outcomes of a study in which fidelity was carefully measured at every session (Gresham, MacMillan, Beebe-Frankenberger, & Bocian, 2000; Schlosser, 2002). Although there are a range of justifiable research methodologies from which scientists who study treatments for speech–language pathology may select, questions that seek to make causal inferences between a treatment and an outcome will typically require experimental methods (see Raudenbush, 2005). Causal questions seek to establish cause-and-effect relationships between X and Y variables, and a key aim of experimental treatment research is to apply methods that rule out all other potential causal relationships between X and Y. Of the various research methodologies that permit researchers to explicitly test causal relationships, the randomized, controlled trial (RCT) is considered by many to be the 'gold standard' in group-design studies (see Raudenbush, 2005, for discussion)[1]. When we consider research quality within the treatment literature, we must think about not only question significance but also the use of methodologies that allow researchers to make causal inferences about the relationship between a treatment and a particular outcome.

Hallmarks of the RCT, as used in treatment research, is the active manipulation of treatment as an independent variable and the random assignment of participants to treatment condition (e.g. comparison treatment, no treatment). By using random assignment, the theoretical presumption is that all factors that may impact treatment outcomes (e.g. age, maturation) are distributed equally across conditions. The RCT is a particularly favorable type of experimental design when one seeks to establish causal relations between a treatment and outcome, as the random assignment of participants to treatment conditions presumably reduces rival interpretations of what may have caused the outcome (e.g. maturation, *a priori* skill differences). A well-conducted RCT will feature not only random assignment of participants to conditions, but also random selection of participants from a specific population.

With the use of random selection, the external validity of a study is strengthened.

An example of a recent RCT in our treatment literature is that of Gillam and colleagues (2008). This study represents one of the largest and most rigorously conducted studies of the efficacy of language intervention for school-aged children with primary language impairment. In total, 216 children (6–9 years of age) were randomly assigned to one of four conditions: two featured computer-assisted language intervention, one featured academic enrichment, and one featured one-on-one language intervention with a certified SLP. Children in all groups received intervention for more than 45 h over a 6-week period of daily sessions. Although children showed "clinically significant gains" over time on standardized measures of language ability (children's gains between pre-test and post-test measures were more than half a standard deviation unit, on average), there were no statistically significant differences among the four conditions. While the primary hypothesis tested in this study concerned the use of computer-assisted instruction, the finding of a general null effect for all conditions is quite troubling, as it suggests that one-on-one treatment by a certified SLP using 'best practice' clinical protocols does not accelerate language growth beyond what we see for enrichment or computer-based instruction. This RCT exemplified competent implementation in terms of rigorous controls over rival interpretations (e.g. randomized assignment, blinding of assessors); consequently, its results raise serious questions regarding the efficacy of current approaches to language treatment for school-age children.

The RCT design is not always feasible to employ, particularly when one is working with intact groups of individuals who are already receiving particular forms of treatment. For instance, to study the effects of grouping practices (e.g. one-on-one vs. small-group treatment) on language outcomes for adults with aphasia, the researcher may not have the option of randomly assigning adults to groups. Consequently, the researcher may use a quasi-experimental design (QED), which differs from the RCT in that it is not considered a true experimental design, because the researcher randomly assign participants to groups. Rather, the researcher compares gains or outcomes on key variables for groups that are already formed. Although a range of statistical techniques (e.g. matching, covariance adjustments, see Raudenbush, Martinez, & Spybrook, 2007) can be used to increase one's confidence in the results of QED studies, the lack of randomization of participants to treatment conditions is a major weakness because, without it, rival explanations for any ascertained treatment effects

cannot be unequivocally ruled out. Nonetheless, in order to better understand the totality of the treatment research in an area of interest, QED studies can be an important source of evidence on treatment effects.

Early in this article, we discussed EBP and the increasing emphasis on SLPs' active use of evidence derived from the best available scientific investigations when making clinical decisions about the treatment of speech–language pathology (Dollaghan, 2007). Clearly, for SLPs to engage in EBP, they must have access to a sufficient body of treatment literature that includes both forms of quality discussed earlier in this article: namely, literature that includes a sufficient number of studies exhibiting question significance as well as a sufficient number of studies that have explicitly tested causal relations between treatments and outcomes (i.e. RCTs and QEDs for group-design studies). Although there is little consensus regarding how many studies must show causal relations between a given treatment and a given outcome or set of outcomes for a treatment to be considered 'empirically validated', some experts have proposed that empirically validated treatments are supported by the results of two or more well-conducted experimental studies (see Lonigan, Elber, & Johnson, 1998). Of particular import is the recognition that research quality is critical to understanding the cause-and-effect relationships between treatment and outcome.

For most any type of research design, consensus statements and textbook descriptions are available to provide guidance on essential indicators of quality (e.g. Brantlinger, Jimenez, Klingner, Pugach, & Richardson, 2005; Neuman & McCormick, 1995). For RCTs and QEDs, these quality indicators are particularly well established (see Campbell & Stanley, 1966; Cook & Campbell, 1979; Turpin, 2005) and, for the most part, are agreed upon by organizations and experts (e.g. Brighton, Gebski, & Keech, 2002; Gersten et al., 2005; Phillips et al., 2001); these give guidance to both scientists and practitioners regarding essential characteristics of experimental work from which causal inferences may be drawn. When experimental treatment research is conducted poorly, the impact of that work on advancing knowledge and improving clinical practice are limited; yet, when it is conducted well, the potential benefits are extraordinary.

To date, a critical enquiry into the methodological quality of group-design treatment research in speech–language pathology has not, to our knowledge, been conducted. However, methodological quality, more so than quantity, is directly associated with the confidence we can place in a study's findings and the ability to generalize findings to clinical activity (Troia, 1999). Given the pressure on researchers to engage in more 'use-inspired' research that directly tackles pressing clinical questions, it is important to recognize the strengths and limitations of the extant body of work to guide the research community's efforts in conceptualizing future studies. Likewise, given the pressure on clinicians to use research findings to help guide their clinical activity, we must take stock of the quality of research being consumed by SLPs, including whether the methods of treatment studies allow the making of causal claims.

Thus, the aim of this article is to critically examine the methodological quality of recent treatment research in speech–language pathology, focusing specifically on RCT and QED group-design studies. In addressing this aim, we considered whether methodological quality differed as a function of a particular study's treatment focus (i.e. literacy, language, speech, fluency, voice), the age group studied (pre-school-age child; school-age child; adult), the year of publication, and the publishing journal.

METHODS

The methods for this work comprised two sequential activities: article search and selection and quality assessment.

Article search and selection

The article search and selection procedures were designed to identify a corpus of recent research reports by which to best characterize social and behavioral treatment research in speech–language pathology for a range of possible treatment foci (e.g. language, speech, voice). We did not focus on the treatment of hearing problems, but rather on those skills and abilities typically considered within the scope of practice of SLPs. Articles were considered for inclusion if they were published between 1997 and 2006 in one of three journals of the American Speech–Language–Hearing Association (ASHA): the *American Journal of Speech–Language Pathology* (AJSLP); the *Journal of Speech, Language, and Hearing Research* (JSLHR); and *Language, Speech, and Hearing Services in Schools* (LSHSS). (Please note that, for JSLHR, we only considered articles published within the 'Language and Speech' sections, and excluded those in the 'Hearing' section.) These three peer-reviewed journals are flagship publications within the speech–language pathology discipline and regularly publish treatment research on speech and language disorders. Although there are other high-quality journals from which we could have

sampled, we theorized that these three journals would collectively provide a representative indicator of the current level of methodological quality for treatment research in speech–language pathology.

To identify articles for inclusion in the corpus, a hand search was conducted by two of the authors for the 10 volumes of each journal published between 1997 and 2007, comprising 40 issues each for AJSLP and LSHSS and 60 issues for JSLHR. For a research article to be included in the review corpus, it had to meet three criteria. First, the research study must have reported a RCT or QED designed to test the effects of a particular treatment, using definitions provided in Tuckman (1999). Second, the research study must have involved at least 10 participants. While we recognize that group-design RCTs and QEDs can involve fewer than 10 participants, for power to be sufficient to detect effects, a sample size of 10 (5 per cell) is often the minimum with which any statistically significant effects can be detected (Lenth, 2006). Third, the research must have focused on the outcomes of social, behavioral, or cognitive treatments, rather than on pharmaceutical, surgical, or technological procedures.

No additional exclusion or inclusion criteria were used to restrict the content or focus of the treatment research. Consequently, the corpus included studies examining treatment outcomes for typical and disordered populations for a range of foci (e.g. literacy, language, voice, fluency) and across a range of age groups (e.g. toddlers, adolescents, adults).

To ensure the reliability of the selection and search procedures, 5 journal volumes comprising 22 issues were randomly selected from the 30 volumes (140 issues) studied; this corresponds to 17% of volumes and 16% of issues selected for reliability purposes. Each article in the 5 volumes was independently coded for inclusion or exclusion in the corpus by two authors. Inter-rater agreement was calculated by dividing the total number of agreements over the total number of articles examined ($n = 240$) and multiplied by 100; agreement was 99.2%, showing that the procedures for identifying articles for inclusion were reliable.

Data extraction

Study characteristics. The following information about each study was extracted and entered into a database: (a) research design (RCT or QED); (b) age group of participants (adults, older children or adolescents, infants, toddlers, and preschoolers, other); (c) treatment foci (literacy, language, speech, fluency, voice, multiple foci); and (d) journal (JSLHR, LSHSS, and AJSLP).

Quality assessment. The assessment of study quality was conducted using a modified version of the 27-item methodological quality assessment checklist presented in Downs and Black (1998). The Downs and Black checklist provides a comprehensive categorical framework that can be used for generating a profile of study quality for both experimental and quasi-experimental studies. In the original checklist, 27 items are organized into 4 quality categories: reporting; external validity; internal validity bias; and internal validity confounding. (For the purposes of this article, the checklist was modified by omitting two items—one was considered not relevant because it focused on adverse events of treatment, whereas the other could not be scored for many articles because it involved power calculations.) Each category comprises a set of items designed to answer specific questions regarding that component of study quality, and each item is scored as 'YES = 1' or 'NO = 0'. A composite Quality Index score is obtained by summing the response score across all items.

The items on the Downs and Black checklist are highly consistent with quality indicators identified in other checklists (e.g. Boutron et al., 2005; Turpin, 2005) and those presented in consensus statements by various organizations (e.g. the Council for Exceptional Children, see Gersten et al., 2005; Consolidated Standards of Reporting Trials, see Newcombe, 2000).The Downs and Black checklist was selected because it is of established psychometric quality, including internal consistency, test–retest reliability, inter-rater reliability, and concurrent validity; most of its psychometric indicators are strong (e.g. test–retest reliability of the Quality Index = 0.88) and all were adequate (see Downs and Black, 1998). A recent evaluation of 60 methodology rating checklists (Deeks et al., 2003) identified the Downs and Black checklist as being one of the 14 "best tools" available because of its comprehensive nature, usability, and psychometric soundness. The following summary is a brief overview of each of the four quality categories.

(1) *Reporting* (9 items, maximum score = 9): the nine 'reporting' items examine whether adequate information is provided in the research report to ensure an unbiased assessment of findings. The nine reporting items consider the clarity of information presented with respect to hypotheses, aims, and objectives, measurement of main study outcomes, participant characteristics, treatment and comparison conditions, distribution of principal confounders, main study findings, variability of study outcomes, characteristics of

participants lost to attrition, and probability values associated with statistical tests.

(2) *External validity* (3 items, maximum score = 3): the three 'external validity' items examine whether the findings of the study are representative and whether results can be generalized to the population of interest. The three 'external validity' items assess the adequacy of methods for recruiting and selecting participants from the source population and how representative the staffing and location of treatment was.

(3) *Internal validity bias* (7 items, maximum score = 7): the seven 'internal validity bias' items examine the adequacy of controls exerted over measurement and outcomes. These items consider the adequacy of methods for blinding study participants to the treatment condition received and assessors to the treatment of participants. They also assess whether reported analyses were planned (versus dredged), whether duration of treatment was similar for all participants, whether statistical analyses were appropriate, whether compliance was reliable, and whether the main outcome measures were valid and reliable.

(4) *Internal validity confounding* (6 items, maximum score = 6): the six 'internal validity confounding' items examine the adequacy of controls exerted over the selection of participants. These items consider whether participants in different groups were sampled from the same population and over the same period of time, whether random allocation was used and if it was concealed from participants and staff, and whether confounding variables and attrition were accounted for in analyses.

Scoring procedures and reliability. The 53 studies in the corpus were individually scored for each of the 25 items by two reviewers working independently; thus, the entire corpus was double coded. Reviewers were not blind to any identifying information related to the article, such as author name, source journal, or year of publication. Upon completion of all scoring, a random sample of 13 (25%) of the coded studies was selected to identify the percentage of inter-rater agreement by dividing the total number of agreements by the number of agreements plus disagreements and multiplying this figure by 100. There was

84% agreement for all items across the 13 studies, suggesting that the inter-coder judgments for each item were reasonably consistent.

In order to present a more accurate assessment of the included studies, all coding differences were reconciled to 100% agreement between the individual reviewers for all 53 studies. Thus, the data presented reflect a consensus agreement regarding the coding of each item for all included studies.

RESULTS

Study characteristics

In total, 53 articles were included in the review corpus examined for this study. An overview of these articles, including their research design, participants, and treatment foci, is presented in Table 1, and a list of the individual studies is available in Table 2. As can be seen, the majority of studies featured an RCT ($n = 38$, 72%), and the majority of studies focused on children, with 51% ($n = 27$) involving infants, toddlers, and preschoolers and 28% ($n = 15$) involving older children and adolescents; 15% ($n = 8$) of the studies focused on adults, whereas 6% ($n = 3$) targeted other populations (e.g. parents, teachers, peers). In terms of general areas of communication outcomes, language (53% of studies, $n = 28$) was the most prevalent, followed by literacy (19%, $n = 10$). Relatively few studies focused on fluency (9%, $n = 5$), voice (6%, $n = 3$), or speech (4%, $n = 2$).

Methodological quality: General findings

The 53 studies were examined via the 4 quality categories (reporting, external validity, internal validity-bias, and internal validity-confounding) of the modified Downs and Black (1998) checklist. Table 3 shows the percentage of studies that received a score of 1 (study meets quality indicator) for each reporting item. The percentage of studies that met the individual quality indicators ranged from a low of 81% for two items (study provides a clear description of participant characteristics; study provides a clear description of actual probability values for main outcomes) to a high of 98% (study provides a clear description of main findings). With a maximum score of 9 points for the reporting items, the set of studies received an average reporting score of 8.1 (SD = 1.0; range 5–9) and more than a third of the studies (38%) received the maximum score.

Findings for the three external-validity items are laid out in Table 4. For two of the items, the percentage of studies meeting the quality indicator was low, both concerning the how representative the ascertained sample was of the more general

Table 1. Overview of review corpus.

Study description (n = 53)		n	%
Research design	Randomized controlled trial	38	72
	Quasi-experimental design	15	28
Treatment foci	Literacy	10	19
	Language	28	53
	Speech (articulation, phonology)	2	4
	Fluency	5	9
	Voice	3	6
	Multiple goals	5	9
Age group	Adults (21 years of age and over)	8	15
	Older children or adolescents (5–20 years of age)	15	28
	Infants, toddlers, or preschoolers (0–4 years of age)	27	51
	Other (e.g. caregivers, peers)	3	6
Journal	American Journal of Speech–Language Pathology	13	25
	Journal of Speech, Language and Hearing Research	28	53
	Language, Speech, and Hearing Services in Schools	12	23

population. Only 4% and 6% of studies, respectively, used methods to recruit and to enroll a sample representative of the population being studied. With a maximum score of 3 points for the external validity items, the studies received an average external-validity score of 0.4 (SD = 0.72; range 0–3). Two-thirds of the studies (66%; $n = 35$) scored 0.

Table 5 presents findings for the two internal-validity-quality categories, bias and confounding. The percentage of studies meeting the individual quality indicators for bias ranged from a low of 2% (study blinded participants to intervention assignment) to a high of 94% (study identified all unplanned analyses, did not use data dredging). With a maximum score of 7 points for the bias items, the 53 studies received an average bias score of 4.3 (SD = 1.2; range 2–6). For the confounding items, for which there was a maximum score of 6 points, the 53 studies received an average confounding score of 2.8 (SD = 1.1, range 1–5). The percentage of studies meeting the individual quality indicators for confounding ranged from a low of 11% (study recruited participants during the same time period) to a high of 93% (study addressed attrition of participants within analyses).

In terms of the overall Quality Index, a maximum score of 25 points was possible. The studies received an average score of 15.6 (SD = 3.0) with scores ranging from 10 to 23.

Methodological quality: Differences based on study characteristics

We also considered whether methodological quality differed as a function of certain characteristics of studies. Specifically, we considered whether ratings varied for studies across different treatment foci (i.e. literacy, language, speech, fluency, voice), age groups

studied (pre-school-age child, school-aged child or adolescent, adult), year of publication, and publication journal. For all such analyses, the alpha level of 0.05 was used as a benchmark for statistical significance. In our first set of analyses, we grouped studies by treatment foci, examining differences in average scores for reporting, external validity, internal validity bias, internal validity confounding, and the overall Quality Index. The descriptive data in Table 6 show there to be some variation in quality ratings amongst the study groupings.

Most studies focused on literacy as a primary treatment foci seemed to receive quality ratings higher than those in the other groupings. However, results of the non-parametric Kruskal–Wallis comparison of median values, for which quality ratings served as dependent variables and treatment foci as the independent variable, did not identify any statistically significant differences in ratings for reporting, χ^2 (4, $n = 48$) = 2.148, $p = 0.709$; external validity, χ^2 (4, $n = 48$) = 6.714, $p = 0.152$; internal validity-bias, χ^2 (4, $n = 48$) = 4.70, $p = 0.319$; internal validity-confounding, χ^2 (4, $n = 48$) = 5.187, $p = 0.269$; and Quality Index, χ^2 (4, $n = 48$) = 5.909, $p = 0.206$.

In our second set of analyses, we grouped studies by the age of participants, examining differences in average scores across four age groups (see Table 7). Again, the Kruskal–Wallis non-parametric test was used to compare median quality ratings for treatment studies focused on adults ($n = 8$); older children and adolescents ($n = 15$); infants, toddlers, and preschoolers ($n = 27$); and others (e.g. caregivers, peers; $n = 3$). No statistically significant differences were discovered: reporting, χ^2 (3, $n = 53$) = 3.052, p = 0.384; external validity, χ^2 (3, $n = 53$) = 3.884, $p = 0.274$; internal validity-bias,

Table 2. Studies included in review corpus.

Author	Year	Title	Journal	Volume	Pages
Antia & Kreimeyer	1997	The generalization and maintenance of the peer social behaviors of young children who are deaf or hard of hearing	LSHSS	28	59–69
Berninger et al.	2003	Comparison of three approaches to supplementary reading instruction for low-achieving second-grade readers	LSHSS	34	101–116
Bernstien, Auer, & Tucker	2001	Enhanced speechreading in deaf adults can short-term training/practice close the gap for hearing adults?	JSLHR	44	5–18
Chapman et al.	2004	Effects of cognitive-communication stimulation for Alzheimer's disease patients treated with donepezil	JSLHR	47	1149–1163
Cohen et al.	2005	Effects of computer-based intervention through acoustically modified speech (Fast ForWord) in severe mixed receptive–expressive language impairment: Outcomes from a randomized controlled trial	JSLHR	48	715–729
Cole, Coggins, & Vanderstoep	1999	The influence of language/cognitive profile on discourse intervention outcome	LSHSS	30	61–67
Craig et al.	1996	A controlled clinical trial for stuttering in persons aged 9 to 14 years	JSLHR	39	808–826
Crowe	2003	Comparison of two reading feedback strategies in improving the oral and written language performance of children with language-learning disabilities	AJSLP	12	16–27
Elman & Bernstein-Ellis	1999	The efficacy of group communication treatment in adults with chronic aphasia	JSLHR	42	411–419
Farber & Klein	1999	Classroom-based assessment of a collaborative intervention program with kindergarten and first grade students	LSHSS	30	83–91
Fey, Cleave, & Long	1997	Two models of grammar facilitation in children with language impairments: Phase 2	JSLHR	40	5–19
Fey & Loeb	2002	An evaluation of the facilitative effects of inverted yes–no questions on the acquisition of auxiliary verbs	JSLHR	45	160–174
Fey et al.	2006	Early effects of responsivity education/prelinguistic milieu teaching for children with developmental delays and their parents	JSLHR	49	526
Gillon	2000	The efficacy of phonological awareness intervention for children with spoken language impairment	LSHSS	31	126–141
Girolametto, Pearce, & Weitzman	1996	Interactive focused stimulation for toddlers with expressive vocabulary delays	JSLHR	39	1274–1283
Girolametto, Pearnce, & Weitzman	1997	Effects of lexical intervention on the phonology of late talkers	JSLHR	40	338–348
Girolametto, Weitzman, & Greenberg	2003	Training day care staff to facilitate children's language	AJSLP	3	299–312
Girolametto, Weitzman, & Greenberg	2004	The effects of verbal support strategies on small-group peer interactions	LSHSS	35	254–268
Hadley et al.	2000	Facilitating language development for inner-city children: Experimental evaluation of a collaborative, classroom-based intervention	LSHSS	31	280–295
Hancock et al.	1998	Two- to six-year controlled-trial stuttering outcomes for children and adolescents	JSLHR	41	1242–1252
Justice et al.	2003	Emergent literacy intervention for vulnerable preschoolers: Relative effects of two approaches	AJSLP	12	320–332
Justice & Ezell	2000	Enhancing children's print and word awareness through home based parent intervention	AJSLP	9	257–269
Justice & Ezell	2002	Use of storybook reading to increase print awareness in at-risk children	AJSLP	11	17–29
Justice, Meier, & Walpole	2005	Learning new words from storybooks: An efficacy study with at-risk kindergartners	LSHSS	36	17–32

Table 2. Continued.

Author	Year	Title	Journal	Volume	Pages
Kagan et al.	2001	Training volunteers as conversation partners using "Supported Conversation for Adults With Aphasia" (SCA): A controlled trial	JSLHR	44	624–638
Katz & Wertz	1997	The efficacy of computer-provided reading treatment for chronic aphasic adults	JSLHR	40	493–507
Kouri	2005	Lexical training through modeling and elicitation procedures with late talkers who have specific language impairment and developmental delays	JSLHR	48	157–171
Leonard et al.	2006	Tense and agreement morphemes in the speech of children with specific language impairment during intervention: Phase 2	JSLHR	49	749
Lincoln, Onslow, & Reed	1997	Social validity of the treatment outcomes of an early intervention program for stuttering	AJSLP	6	77–84
Logan et al.	2002	Speaking slowly effects of four self-guided training approaches on adults' speech rate and naturalness	AJSLP	11	163–174
Masterson & Perrey	1999	Training analogical reasoning skills in children with language disorders	AJSLP	8	53–61
Nelson et al.	1996	Effects of imitative and conversational recasting treatment on the acquisition of grammar in children with specific language impairment and younger language-normal children	JSLHR	39	850
Riley & Ingham	2000	Acoustic duration changes associated with two types of treatment for children who stutter	JSLHR	43	965–978
Robertson & Weismer	1997	The influence of peer models on the play scripts of children with specific language impairment	JSLHR	40	49–61
Robertson & Weismer	1999	Effects of treatment on linguistic and social skills in toddlers with delayed language development	JSLHR	42	1234–1248
Roy et al.	2001	An evaluation of the effects of two treatment approaches for teachers with voice disorders: A prospective randomized clinical trial	JSLHR	44	286–296
Roy et al.	2003	Three treatments for teachers with voice disorders: A randomized clinical trial	JSLHR	46	670–688
Roy et al.	2002	Voice amplification versus vocal hygiene instruction for teachers with voice disorders: A treatment outcomes study	JSLHR	45	625–638
Rvachew & Nowak	2001	The effect of target-selection strategy on phonological learning	JSLHR	44	610–623
Rvachew, Nowak, & Cloutier	2004	Effect of phonemic perception training on the speech production and phonological awareness skills of children with expressive phonological delay	AJSLP	13	250–263
Segers & Verhoeven	2004	Computer-supported phonological awareness intervention for kindergarten children with specific language impairment	LSHSS	35	229–239
Swanson, Hodson, & Schommer-Aikins	2005	An examination of phonological awareness treatment outcomes for seventh-grade poor readers from a bilingual community	LSHSS	36	336–345
Throneburg et al.	2000	A comparison of service delivery models: Effects of curricular vocabulary skills in the school setting	AJSLP	9	10–20
Tyler et al.	2002	Efficacy and cross-domain effects of a morphosyntax and a phonology intervention	LSHSS	33	52–66
Tyler et al.	2003	Outcomes of different speech and language goal attack strategies	JSLHR	46	1077–1094
van Kleeck, Gillam, McFadden et al.	1998	A study of classroom-based phonological awareness training for preschoolers with speech and/or language disorders	AJSLP	7	65–76
van Kleeck, Vander Woude, & Hammett	2006	Fostering literal and inferential language skills in Head Start preschoolers with language impairment using scripted book-sharing discussions	AJSLP	15	85–95
Wolfe, Presley, & Mesaris	2003	The importance of sound identification training in phonological intervention	AJSLP	12	282–288

Table 2. Continued.

Author	Year	Title	Journal	Volume	Pages
Yoder & Stone	2006	A randomized comparison of the effect of two prelinguistic communication interventions on the acquisition of spoken communication in preschoolers with ASD	JSLHR	49	698–711
Yoder & Warren	2001	Relative treatment effects of two prelinguistic communication interventions on language development in toddlers with developmental delays vary by maternal characteristics	JSLHR	44	224–237
Yoder & Warren	2002	Effects of prelinguistic milieu teaching and parent responsivity education on dyads involving children with intellectual disabilities	JSLHR	45	1158–1174
Yoder & Warren	1998	Maternal responsivity predicts the prelinguistic communication intervention that facilitates generalized intentional communication	JSLHR	41	1207–1219

Abbreviations: AJSLP, American Journal of Speech–Language Pathology; JSLHR, Journal of Speech, Language and Hearing Research; LSHSS, Language, Speech, and Hearing Services in Schools.

Table 3. Quality scores for 53 group-design treatment studies: Reporting.

Quality indicator: Does study provide clear description of...	% (n) of studies meeting quality indicator
Hypothesis/aims/objectives	96% (51)
Main study outcomes	89% (47)
Participant characteristics	81% (43)
Intervention and comparison conditions	85% (45)
Distribution of confounders in each participant group	91% (48)
Main findings	98% (52)
Estimates of random variability in data for main outcomes	89% (47)
Characteristics of participants lost to attrition	96% (51)
Actual probability values for main outcomes	81% (43)

Table 4. Quality scores for 53 group-design treatment studies: External validity.

Quality indicator	% (n) of studies meeting quality indicator
Participants who are recruited representative of population	4% (2)
Participants who agree to participate representative of population	6% (3)
Study context (e.g. staff, facilities) representative of population	34% (18)

χ^2 (3, $n = 53$) = 5.995, $p = 0.112$; internal validity-confounding, $\chi2$ (3, $n = 53$) = 1.824, $p = 0.610$; and Quality Index, χ^2 (3, $n = 53$) = 4.088, $p = 0.252$.

In our third set of analyses, we grouped studies by the year in which the study was published, using a median split to differentiate early studies (published between 1997 and 2001; $n = 27$) and later studies (published between 2002 and 2006; $n = 26$). Table 8 displays comparisons of quality ratings for early studies and later studies. Results of the non-parametric Mann–Whitney U test

showed that the set of later studies had significantly higher scores for reporting, $z = 2.543$, $p = 0.011$, and a trend for higher scores on the Quality Index, $z = 1.737$, $p = 0.082$. The other ratings were similar for the two sets of studies: external validity, $z = 0.182$, $p = 0.855$; internal validity bias, $z = 0.939$, $p = 0.348$; and internal validity confounding, $z = 1.29$, $p = 0.197$.

Finally, we considered whether the publishing journal was associated with quality ratings. Studies in this corpus were published in the three journals of

Table 5. Quality scores for 53 group-design treatment studies: Internal validity.

Quality indicator		% (n) of studies meeting quality indicator
Bias	Blinding of participants to intervention received	2% (1)
	Blinding of assessors measuring main study outcomes	28% (15)
	Unplanned analyses identified (no data dredging)	94% (50)
	Time between intervention and outcome same for all	85% (45)
	Statistical tests appropriate for main study outcomes	70% (37)
	Compliance with intervention reliable	57% (30)
	Main outcome measures valid and reliable	81% (43)
Confounding	Participants recruited from same population	28% (15)
	Participants recruited during same period of time	11% (6)
	Participants randomized to treatment conditions	59% (31)
	Randomized assignment concealed (to participants and staff)	13% (7)
	Statistical adjustments for confounding (e.g. intent to treat)	81% (43)
	Loss of participants (i.e. attrition) addressed in analyses	93% (49)

Table 6. Mean (SD) quality ratings by treatment focus.

Quality category (maximum score)	Treatment focus					
	Literacy (n = 10)	Language (n = 28)	Speech (n = 2)	Fluency (n = 5)	Voice (n = 3)	Multiple (n = 5)
Reporting (9)	8.4 (0.8)	8.0 (1.0)	8.0 (0.0)	7.6 (1.5)	8.0 (1.7)	8.0 (0.7)
External validity (3)	0.8 (0.9)	0.4 (0.6)	0.0 (0.0)	0.0 (0.0)	0.3 (0.6)	0.8 (1.3)
Internal validity bias (7)	4.6 (1.1)	4.4 (1.3)	4.5 (2.1)	3.2 (1.1)	4.7 (0.6)	4.2 (1.1)
Confounding (6)	3.5 (1.0)	2.7 (1.2)	2.5 (0.7)	2.6 (0.9)	2.3 (1.2)	3.2 (1.3)
Quality index (25)	17.3 (3.2)	15.4 (2.9)	15.0 (2.8)	13.4 (1.9)	15.3 (2.1)	16.2 (3.6)

Table 7. Mean (SD) quality ratings by age group studied.

Quality category (maximum score)	Study population			
	Adults (n = 8)	Older children and adolescents (n = 15)	Young children (n = 21)	Other (n = 3)
Reporting (9)	8.0 (1.3)	7.7 (1.3)	8.2 (0.7)	8.7 (0.6)
External validity (3)	0.6 (0.7)	0.3 (0.8)	0.5 (0.7)	0.0 (0.0)
Internal validity bias (7)	4.1 (1.4)	3.7 (1.1)	4.7 (1.1)	4.3 (1.5)
Confounding (6)	3.1 (1.2)	2.7 (1.0)	2.8 (1.2)	3.3 (0.6)
Quality index (25)	15.9 (3.5)	14.4 (3.1)	16.2 (2.8)	16.3 (1.5)

Table 8. Mean (SD) quality ratings by period of publication.

Quality category (maximum score)	Publication period	
	Early studies: 1997–2001 (n = 27)	Later studies: 2002–2006 (n = 26)
Reporting (9)	7.7 (1.1)	8.4 (0.7)
External validity (3)	0.4 (0.7)	0.5 (0.8)
Internal validity bias (7)	4.2 (1.2)	4.5 (1.2)
Confounding (6)	2.7 (1.1)	3.0 (1.1)
Quality Index (25)	14.9 (2.9)	16.4 (2.9)

Table 9. Mean (SD) quality ratings by publishing journal.

Quality category (maximum score)	Journal		
	AJSLP ($n = 13$)	JSLHR ($n = 28$)	LSHSS ($n = 12$)
Reporting (9)	8.0 (1.5)	8.0 (0.8)	8.2 (0.8)
External validity (3)	0.5 (0.9)	0.3 (0.5)	0.7 (0.9)
Internal validity bias (7)	3.9 (1.3)	4.6 (1.1)	3.9 (1.1)
Confounding (6)	2.8 (1.3)	2.9 (1.1)	2.8 (1.0)
Quality index (25)	15.2 (4.0)	15.9 (2.7)	15.5 (2.5)

Abbreviations: AJSLP, *American Journal of Speech-Language Pathology*; JSLHR, *Journal of Speech, Language, and Hearing Research*; LSHSS, *Language, Speech, and Hearing Services in Schools*.

ASHA: AJSLP ($n = 13$), JSLHR ($n = 28$), and LSHSS ($n = 12$). Table 9 shows the average quality ratings for articles published within each of these journals during the 10-year period studied. These descriptive data show that, on average, JSLHR received the highest quality ratings for internal validity bias and internal validity confounding and for the overall Quality Index, whereas LSHSS received the highest quality ratings for reporting and external validity. However, none of these comparisons were statistically significant according to results of the Kruskal–Wallis non-parametric test, indicating that the corpus of articles from each journal was comparable in quality for reporting, χ^2 (2, $n = 53$) $= 0.669$, $p = 0.716$; external validity, χ^2 (2, $n = 53$) $= 1.791$, $p = 0.408$; internal validity-bias, χ^2 (2, $n = 53$) $= 4.715$, $p = 0.095$; internal validity-confounding, χ^2 (2, $n = 53$) $= 0.061$, $p = 0.970$; and χ^2 (2, $n = 53$) $= 0.669$, $p = 0.705$. The trend for a between-group difference among the journals on interval validity-bias ratings was in favor of JSLHR.

DISCUSSION

In recent years, SLPs have been pressed to engage in EBP, a process of clinical decision-making that involves integrating knowledge of the relevant scientific literature with clinical experience, theoretical knowledge, client preferences (Dollaghan, 2007), and relevant stakeholder perspectives (Schlosser & Raghavendra, 2004). SLPs may seek guidance on research findings from the accumulation and synthesis of evidence in a particular area as reported in a systematic review or meta-analysis, if these are not available from individual studies. However, if systematic reviews and meta-analyses are available, the strength of recommendations provided by these can only be as good as the quality of the research studies included in the work. Similarly, the strength of recommendations offered

from individual studies can only be as sound as the quality of these individual studies. To consider an example, Troia's (1999) examination of 39 RCTs that investigated the effects of phonological-awareness intervention showed that many of the studies had significant methodological flaws that could undermine our confidence in their findings. Nonetheless, this body of research has had tremendous influence in shaping national educational policies (e.g. National Reading Panel, 2000) and continues to be used as a source of evidence for effective practice.

To date, we are aware of no study that has conducted a systematic examination of the quality of group-design treatment studies in speech–language pathology. On occasion, reviews of treatment effectiveness have included an assessment of methodological quality (e.g. Law, Garrett, & Nye, 2004); however, these quality assessments are narrowly focused on a particular treatment focus and, therefore, do not provide an all-encompassing examination of communication disorders treatment research. The present study conducted a broad survey of the speech–language pathology treatment literature to index the methodological quality of 53 group-design research studies published over a 10-year period. Using a psychometrically sound checklist designed to investigate the methodological quality of experimental studies, this body of work was evaluated in terms of quality of reporting, external validity, internal validity bias, and internal validity confounding. The potential outcomes of such an examination are twofold. First, we are hopeful that the results may influence the quality of future treatment research in speech–language pathology, by identifying areas in which methodology is typically weak and that require greater attention in design of future studies. Second, we are also hopeful that the results may influence the way in which treatment research is translated for use in the field by consumers; that is, we hope that SLPs and other professionals will use the information in this report to identify whether

studies are of sufficient quality to inform their clinical decision making.

Turning to the major findings of this work, the first finding of interest was that quality of the treatment studies was highly variable using the 25 quality indicators studied. As many as 96% of studies met some indicators (e.g. study included a clear description of the hypothesis/aim/objective), whereas only 2% of studies met other indicators (e.g. participants were blinded to the intervention condition received). As a body of work, this finding suggests that speech–language pathology treatment research is not of uniformly high quality. Considering the percentage of studies exhibiting the quality indicators across the four categories examined using the Downs and Black checklist, it seems that this body of work is relatively strong in reporting clarity and relatively weak in external validity.

The second finding of interest, related to the first, is that many studies showed significant methodological weaknesses. Troia (1999) identified criteria for RCTs that are "deemed imperative for eliminating rival interpretations of causality and for permitting generalizations of findings to other populations and settings" (p. 32; see also Gersten et al., 2005). Failure to meet these criteria, which relate to both internal validity and external validity, may be considered a 'fatal flaw' in treatment research; they include, for instance, use of random assignment, measurement of intervention fidelity or adherence, and blinding of assessors to treatment conditions. A substantial percentage of studies in the present corpus exhibited methodological shortcomings in these areas: 41% did not use random assignment, 43% did not provide adequate details on fidelity/adherence, and 72% did not use blinded assessors. These shortcomings allow uncertainty to enter into the interpretation of findings regarding the causal relations between treatment and outcomes.

A third finding of interest is that methodological quality was not found to systematically vary when studies were on different foci. Whether studies were focused on literacy, language, speech, fluency, or voice outcomes, study quality was generally similar—although the data suggested that studies of literacy received slightly higher scores across all quality indicators compared with other substantive foci. Quality did not seem to vary when considering publication venue or studies focused on different age groups. That is, studies were of generally the same level of quality whether they focused on adults or on young children, school-aged children and adolescents; likewise, studies published in AJSLP, JSLHR, and LSHSS were of similar quality. The only study characteristic that was associated with methodological quality was the timing of the study: studies published between 2002 and 2006 had significantly higher ratings for reporting and a trend towards higher ratings on the overall Quality Index, compared with studies published between 1997 and 2001. This finding suggests, perhaps, that researchers are becoming more aware of quality indicators for group-design studies, or that journal editors and reviewers are requiring higher levels of quality for publication.

IMPLICATIONS

This study suggests that study findings (and findings presented in systematic reviews and meta-analyses) must be carefully considered with respect to the quality of the research, particularly when research findings are used to guide treatment decisions. As this review indicates, some studies contain methodological weaknesses of consequence, such as the failure to blind assessors to the treatment received by participants and the failure to monitor or report intervention compliance and fidelity. As has been described elsewhere (see Troia, 1999), methodological weaknesses such as these can and should raise questions about the confidence we can place in a study's findings. Given the incidence of some methodological weaknesses in the 53 studies reviewed, our results suggest that, at the very least, the ASHA publications board should consider establishing minimum reporting standards for research into treatment in speech–language pathology. This board guides publication standards for the journals studied in this work (AJSLP, JSLHR, and LSHSS). And, although it is not possible for us to generalize our findings to other journals that publish treatment research on communication disorders (e.g. Journal of Fluency Disorders, Communication Disorders Quarterly, Evidence-Based Communication Assessment and Intervention), it seems reasonable to suggest that the editorial board of these and other journals should also establish minimum standards. Specifically, the description of the treatment protocol, fidelity of treatment implementation, and presentation of the magnitude of effect (e.g. effect size) should be reported for all treatment research studies. We view these characteristics as particularly important to clinicians who seek to use research evidence when making treatment decisions.

Description of treatment protocol

In order for the practicing clinician to be able to assess the appropriateness of an intervention, it is imperative that the treatment protocol, as intended to be implemented, is presented in sufficient detail

that replication of the intervention procedure is possible. The description must include, at a minimum, the specific objectives, techniques, and materials used to deliver the treatment, the overall length of the treatment program as well as number of treatment sessions, and the duration of the individual treatment sessions. The studies included in the present corpus were variable in the extent to which this information was provided in enough detail.

Fidelity of treatment implementation

Even for those studies in which the treatment protocol is fully described, it is imperative that the authors also identify any variation in protocol that occurred during the study. The clinician can then make a decision as to whether or not these variations are important in either interpreting the original researchers findings or in the replication process if the treatment is applied to an individual in the a clinical setting. In effect, we are arguing for a transparent description of exactly what happens when the treatment is delivered to human participants. While it may be the case that some variations are of no substantive consequence, the true impact of the variations cannot be understood (and tested) if they are not known. In the present review, we found that just over half of the studies (57%) included data on compliance or fidelity to intervention.

Presentation of the magnitude of effect

In the present research, we did not document reporting of effect-size estimates, as many studies did not report these indices and we were unable to reliably calculate these by hand, given the data available. Moreover, the purpose of this review was not to provide estimates of treatment effects. However, as we consider quality of treatment research in future research reports, it warrants note that effect-size estimates ought to be included and interpreted. This is essential for understanding the practical impacts of a study's findings and for interpreting specific findings in the context of treatment effects available within an accumulated literature. Providing sufficient information within a report (e.g. group means, standard deviations, and confidence intervals for effect size estimates) for the quality of the effect-size indices reported to be ascertained is also necessary (American Psychological Association, 2001).

While the standards presented here, which we suggest should serve as minimum guidelines for reporting results from RCT and QED treatment studies, are not as comprehensive as those instituted by other organizations (e.g. American Educational Research Association, Council for Exceptional

Children), they do provide a basic standard for treatment research reporting that is noticeably absent in much of the current research in the field of communication disorders. We believe that by implementing these, further discussion can build upon these principles to address all published research in the field of communication disorders and, thus, provide a stronger and more consistent scientific basis for clinicians engaged in the treatment of communication disorders.

NOTE

1. Researchers who conduct treatment research on relatively low-incidence conditions, such as apraxia or autism, may not have adequate access to large enough participant samples to conduct RCTs. Single-subject experimental designs that feature rigorous experimental controls may serve as an alternative 'gold standard'; see Horner et al. (2005) for a discussion of high-quality single-subject research designs.

REFERENCES

American Speech–Language–Hearing Association (2005). *Evidence-based practice in communication disorders: Position statement*. Retrieved December 22, 2007, from http://www.asha.org/about/publications/leader-online/archives/2005/050524/f050524a4.htm

Asher, W. (1990). Educational psychology, research methodology, and meta-analysis. *Educational Psychologist, 25,* 143–158.

Boutron, I., Moher, D., Tugwell, P., Giraudeau, B., Poiraudeau, S., Nizard, R., et al. (2005). A checklist to evaluate a report of a nonpharmacological trial (CLEAR NPT) was developed using consensus. *Journal of Clinical Epidemiology, 58,* 1233–1240.

Brantlinger, E., Jimenez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children, 71,* 195–208.

Brighton, J.K., Gebski, V.J., & Keech, A.C. (2002). Specifying interventions in a clinical trial. *Medical Journal of Australia, 176,* 281–282.

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

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

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

Dollaghan, C.A. (2007). *The handbook for evidence-based practice in communication disorders.* Baltimore, MD: Paul H. Brookes Publishing Co.

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

Fey, M.E., & Justice, L.M. (2007). Evidence-based decision making in communication intervention. In R. Paul, & P.W. Cascella (Eds.), *Introduction to clinical methods in communication disorders,* (2nd ed.)Baltimore, MA: Paul. H. Brookes Publishing Co.

Gersten, R., Fuchs, L.S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M.S. (2005). Quality indicators for group experimental and quasi-experimental research. *Exceptional Children, 71,* 137–149.

Gillam, R., Loeb, D.F., Hoffman, L., Bohman, T., Champlin, C., Thibodeau, L., et al. (2008). The efficacy of Fast ForWord language intervention in school-age children with language impairment: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research, 51*, 97–119.

Gresham, F.M., MacMillan, D.L., Beebe-Frankenberger, M.E., & Bocian, K.M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research and Practice, 15*, 198–205.

Horner, R.H., Carr, E.G., Halle, J., McGee, G., Odom, S.L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179.

Law, J., Garrett, Z., & Nye, C. (2004). The efficacy of treatment for children with developmental speech and language delay/disorder: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 47*, 924–943.

Lenth, R. V. (2006). Java applets for power and sample size (Version 1.63) [Computer software]. Retrieved November 25, 2007. Available from http://www.stat.uiowa.edu/~rlenth/Power

Lonigan, C., Elber, J., & Johnson, S (1998). Empirically supported interventions for children: An overview. *Journal of Clinical Child Psychology, 27*, 138–145.

*Nancollis, A., Lawrie, B.A., & Dodd, B. (2005). Phonological awareness intervention and the acquisition of literacy skills in children from deprived social backgrounds. *Language, Speech, and Hearing Services in Schools, 36*, 325–335.

National Reading Panel (2000). Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Neuman, S.B., & McCormick, S. (1995). *Single-subject experimental research: Applications for literacy.* International Reading Association. Newark, DE: International Reading Association.

Newcombe, R.G. (2000). Reporting of clinical trials in the JO—The CONSORT Guidelines. *Journal of Orthodontics, 27*, 69–70.

Odom, S.L., Brantlinger, E., Gernsten, R., Horner, R.H., Thompson, B., & Harris, K. (2005). Research in special education: Scientific methods and evidence-based practices. *Council for Exceptional Children, 71*, 137–148.

Phillips, B., Ball, C., Sackett, D., Badenoch, D., Straus, S., Haynes, B., et al. (2001). *Oxford Centre for Evidence-Based Medicine levels of evidence.* Oxford, England: Center for Evidence-Based Medicine.

Raudenbush, S.W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*, 25–31.

Raudenbush, S.W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*, 5.

Schlosser, R.W. (2002). On the importance of being earnest about treatment integrity. *Augmentative and Alternative Communication, 18*, 36–44.

Schlosser, R.W. (2003). *The efficacy of augmentative and alternative communication: Toward evidence-based practice.* San Diego, CA: Academic Press.

Schlosser, R.W., & Raghavendra, P. (2004). Evidence-based practice in augmentative and alternative communication. *Augmentative and Alternative Communication, 20*, 1–21.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Stokes, D.E. (1997). *Pasteur's quadrant.* Washington, DC: Brookings Institution Press.

Troia, G.A. (1999). Phonological awareness intervention research: A critical review of the experimental methodology. *Reading Research Quarterly, 34*, 28–52.

Tuckman, B. (1999). Conducting educational research, Fifth ed. Orlando, FL: Harcourt Brace & Company.

Turpin, D.L. (2005). CONSORT and QUOROM guidelines for reporting randomized clinical trials and systematic reviews. *American Journal of Orthodontics and Dentofacial Orthopedics, 128*, 681–685.