Research Article

Assessment Fidelity in Aphasia Research

Jessica D. Richardson,^a Sarah Grace Hudspeth Dalton,^a
Jennifer Shafer,^b and Janet Patterson^c

Purpose: In aphasia treatment literature, scarce attention is paid to factors that may reduce a study's validity, including adherence to assessment and treatment procedures (i.e., fidelity). Although guidelines have been established for evaluating and reporting treatment fidelity, none exist for assessment fidelity.

Method: We reviewed treatment fidelity guidelines and related literature to identify assessment fidelity components. We then examined 88 aphasia treatment studies published between 2010 and 2015 and report the frequency with which researchers provide information regarding the following assessment fidelity components: assessment instruments, assessor qualifications, assessor or rater

training, assessment delivery, assessor or rater reliability, and assessor blinding.

Results: We found that 4.5% of studies reported information regarding assessment instruments, 35.2% reported information regarding assessor qualifications, 6.85% reported information regarding assessor or rater training, 37.5% reported information regarding assessor or rater reliability, 27.3% reported on assessor blinding, and no studies reported information regarding assessment delivery.

Conclusions: There is a paucity of assessment fidelity information reported in aphasia treatment research. The authors propose a set of guidelines to ensure readers will be able to evaluate assessment fidelity, and thus study validity.

Individuals with aphasia who participate in treatment demonstrate greater improvements than those who do not (Brady, Kelly, Godwin, & Enderby, 2012; Holland, Fromm, DeRuyter, & Stein, 1996; Robey, 1994, 1998), although response to treatment varies greatly. There are many potential sources of variance in aphasia treatment research (e.g., aphasia severity and type, personal motivation, treatment adherence, scoring errors, etc.) that reduce the power to detect effects, making it difficult to answer central questions about how assessment and treatment characteristics interact to engender a positive treatment outcome. A contributor of unwanted, and potentially avoidable, variance could be suboptimal methodological quality, specifically with regards to scientific validity. A discussion of scientific validity, or the truthfulness of inferences (Shadish, Cook, & Campbell, 2002), in aphasia treatment research is warranted in the absence of clear answers about treatment outcomes after more than a century's worth of systematic aphasia treatment investigations (Academy of Neurologic

Communication Disorders and Sciences [ANCDS] Aphasia Treatment Website, <http://aphasiatx.arizona.edu/>; speechBITE, <http://speechbite.com/>).

Two types of validity, statistical conclusion validity and internal validity, are particularly important in intervention research. *Statistical conclusion validity* refers to accuracy of inference about the presence and strength of the relationship between two variables. It can be threatened by measurement limits (e.g., restricted range), measurement error (e.g., unreliability), incorrectly applied statistics, unreliable treatment implementation, and other sources of variance introduced into the experimental setting, as well as by low statistical power (which can also be a result of the random variance introduced by the forenamed threats; Shadish et al., 2002). These threats can increase the chance of Type I or Type II error, or the additional error (called Type III error by some) of concluding significance or non-significance when in fact the assessments or the treatments were not correctly implemented (Bell et al., 2004; Hinckley & Douglas, 2013; Nigg, Allegrante, & Ory, 2002). *Internal validity* refers to whether or not causation can be inferred from the statistical conclusions. It can be threatened by those same threats to statistical conclusion validity and other assessment factors, such as change in instrumentation over time or test exposure (Shadish et al., 2002). Although this list of threats is not exhaustive, it is clear that many are related to implementation of clinical procedures, and

^aUniversity of New Mexico, Albuquerque

^binVentiv Health Clinical, Princeton, NJ

^cVA Northern California Health Care System

Correspondence to Jessica D. Richardson: jdichslp@gmail.com

Editor: Anastasia Raymer

Associate Editor: William Hula

Received September 16, 2015

Revision received March 15, 2016

Accepted June 13, 2016

DOI: 10.1044/2016_AJSLP-15-0146

Disclosure: The authors have declared that no competing interests existed at the time of publication.

it follows that a contributor to the historically mixed results in aphasia treatment research could be poor validity. *Fidelity*, or the adherence to and consistent implementation of study procedures, may alleviate some threats to validity. Without establishing and monitoring fidelity, investigators cannot confidently determine whether results (significant or non-significant) were caused by the targeted independent variable or by other random factors introduced because the clinician “drifted” from or “contaminated” the assessment or treatment protocol by adding or omitting elements (Bellg et al., 2004).¹

Although this article addresses the importance of fidelity in assessment procedures, treatment fidelity has so far been the focus of most clinical implementation discussions. *Treatment fidelity* addresses whether the essential elements of a treatment are delivered as intended and are distinguishable from comparison conditions (Bellg et al., 2004; Gearing et al., 2011; Hinckley & Douglas, 2013). In general psychology research, several meta-analyses demonstrated that studies that took steps to ensure treatment fidelity had larger effect sizes (two to three times higher) for treatment outcomes than those studies that did not (Durlak & DuPre, 2008). School psychology researchers showed that as the degree of treatment fidelity increases, the rates of positive outcomes also increase (Gresham, Gansle, Noell, Cohen, & Rosenblum, 1993). In addition, treatment fidelity, also defined as *program integrity*, has been clearly identified as a moderator of outcome variables in substance abuse prevention research (Dusenbury, Brannigan, Falco, & Hansen, 2003; Hansen, Graham, Wolkenstein, & Rohrbach, 1991). Thus, treatment implementation can be a determinant of success.

The National Institutes of Health’s Behavior Change Consortium (Bellg et al., 2004) established a work group to define treatment fidelity and develop guidelines to ensure treatment fidelity. According to this group, establishing treatment fidelity requires attention to processes in the following components: (a) study design—adequate to test hypotheses and produce valid inferences; (b) training—manuals and procedures that are standardized across clinicians; (c) treatment delivery—monitoring to assure delivery as intended; (d) treatment receipt—assure the patient understands the treatment procedures and demonstrates understanding and use within experimental sessions; and (e) treatment enactment—use of behaviors targeted in treatment in real-world settings.

In addition to the guidelines for establishing treatment fidelity, there are also reporting guidelines that allow readers to evaluate the quality of the research. Many of these guidelines are available through the Enhancing the Quality and Transparency of Health Research Network (<http://www.equator-network.org/>). Guidelines are available for randomized controlled trials (Consolidated Standards of Reporting

Trials [CONSORT]; Schulz, Altman, & Moher, 2010), observational studies (Strengthening the Reporting of Observational Studies in Epidemiology [STROBE]; von Elm et al., 2007), systematic reviews (Preferred Reporting Items for Systematic Reviews and Meta-Analyses [PRISMA]; Moher, Liberati, Tetzlaff, & Altman, 2009), case reports (Case Reporting [CARE]; Gagnier et al., 2013), diagnostic/prognostic studies (Standards for Reporting of Diagnostic Accuracy Studies [STARD]; Bossuyt et al., 2015), and for reporting study protocols (Standard Protocol Items: Recommendations for Interventional Trials [SPIRIT]; Chan et al., 2013), among others. An additional set of guidelines has been developed that focuses specifically on describing the research intervention, to ensure that readers would be able to replicate the procedures (Template for Intervention Description and Replication [TIDieR]; Hoffmann et al., 2014).

Despite the availability of numerous guidelines for establishing and reporting treatment fidelity, it is not widely reported within the aphasia literature. Hinckley and Douglas (2013) examined treatment fidelity reporting in 149 aphasia treatment studies published between 2002 and 2011. Their examination revealed that although almost half of the studies reported treatment methods that would enable replication, only 21 out of 149 (14%) of the studies specifically described their method for establishing or monitoring treatment fidelity. Twenty of those studies utilized a single treatment fidelity method—monitoring of adherence to treatment protocol, supervising treatment sessions, utilizing a training manual, or preimplementation role-playing. Only one study utilized more than one treatment fidelity method, combining the use of a training manual with monitoring of adherence to the treatment protocol.

There is a need to expand our focus to assessment fidelity, for just as there can be clinician drift from treatment procedures (e.g., cueing hierarchy not delivered as designed) and contamination of treatment procedures (e.g., clinician incorporates cues from Treatment B into Treatment A), assessor and/or rater drift and contamination are just as likely to occur. This is especially true for investigations that employ multiple assessors and/or raters, repeated measures, and/or lengthy assessment batteries that include tests with different administration procedures (e.g., item time limits, assessor cueing) and complex scoring systems. Because assessment instruments vary widely in their theoretical foundation, intended audience, and administration and scoring procedures, attention to assessment fidelity is essential. For example, if considering only timing of naming assessments, participants are allowed 5 s to respond on the Northwestern Assessment of Verbs and Sentences (Cho-Reyes & Thompson, 2012), 20 s on the Western Aphasia Battery (Kertesz, 2007) and Boston Naming Test (Kaplan, Goodglass, & Weintraub, 2001), and 30 s on the Philadelphia Naming Test (Roach, Schwartz, Martin, Grewal, & Brecher, 1996). Differences can also be observed in the specificity of administration instructions, clinician responses, attempts, and scoring (see Table 1), an important consideration because it is common for studies (and standard protocols like the AphasiaBank protocol) to incorporate several different assessment

¹Two additional types of scientific validity (construct validity and external validity) are not discussed here because these components are tied less to issues of implementation or fidelity and more to inferences and generalizations that can be made given the study design, research questions, and selected assessments and treatments.

Table 1. Sample instructions to clinicians regarding assessment administration and scoring.

Clinician response following item presentation (including cueing, feedback, repetition)	Assessment tool
If no response or incorrect response, provide cue (tactile, phonemic, semantic). Task-related reminders (e.g., say only one sentence, use all words), but no other feedback.	WAB-R, Naming NAVS, ASPT
Give feedback after response; provide correct answer after incorrect response.	PNT
If incorrect response is given, then provide cue (semantic, phonemic). If no answer is given, show multiple choice options. Document types of errors with error codes.	BNT
Timing 5 s to respond, can prompt for second attempt (additional 5 s). 20 s, cues provided (no additional time after cue). 30 s 20 s, if a cue provided give an additional 5 s.	NAVS, VCT WAB-R, Naming PNT BNT
Attempts Do not count items where regional pronunciation creates disagreement in stimulus/target rhyme. (Test control individuals to determine regional rhymes.) Score first complete attempt (at minimum CV or VC response) that is not self-interrupted and certain prosodic qualities (specified in instructions). If motor speech disorder present, allow omission, addition, or substitution. Score final response within 10 s. Score only the final attempt.	PALPA, Rhyme PNT PNT NAVS, VNT NAVS, ASPT
Scoring 3 points if correct (even if minor articulatory error) and no cue needed. 2 points if response is recognizable as target but with phonemic paraphasia and no cue needed. 1 point if cue is needed. 0 points if incorrect or no response after cueing. 6 points if correct, 5 points if the correct response is delayed, 4 points if a semantic cue is needed, 3 points if the response provided with a semantic cue is delayed, 2 points if a phonemic cue is needed, 1 point if the response provided with a phonemic cue is delayed, 0 points if no response is given. 1 point if correct, 0 points if incorrect.	WAB-R, Naming BEST BNT

Note. WAB-R = Western Aphasia Battery–Revised; NAVS = Northwestern Assessment of Verbs and Sentences; ASPT = Argument Structure Production Test (subtest of the NAVS); VCT = Verb Comprehension Test (subtest of the NAVS); VNT = Verb Naming Test (subtest of the NAVS); PNT = Philadelphia Naming Test; BNT = Boston Naming Test; PALPA = Psycholinguistic Assessments of Language Processing in Aphasia (Kay, Lesser, & Coltheart, 1992); BEST = Bedside Evaluation and Screening Test (Fitch-West, Sands, & Ross-Swain, 1988).

measures targeting the same discrete language impairment, such as confrontation naming (Benjamin et al., 2014; Carragher, Sage, & Conroy, 2013; MacWhinney, Fromm, Forbes & Holland, 2011) or semantics and phonology (de Jong-Hagelstein et al., 2011). To further illustrate, assessment fidelity is likely to be higher if utilizing an assessment battery with fewer tests, or which holds instructions mostly constant across tests (i.e., the same prompts, timing, and scoring), compared with an assessment battery for which these things vary widely from test to test. Assessment fidelity may be lower if the instructions for the assessor are not detailed or specific, making it difficult to determine what “good” adherence would look like.

The Standards for Educational and Psychological Testing (SEPT; developed jointly by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) outlines general guidelines relating to administration, scoring, reporting, and interpretation of assessments (2014; pp. 114–121). The standards specify that “assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence

to the established procedure should be monitored, and any material errors should be documented and, if possible, corrected” (p. 114). Despite these assessment fidelity recommendations by SEPT, guidelines to ensure adherence to an assessment protocol are not well developed and have received considerably less attention compared with treatment fidelity guidelines. Yet, it makes little sense to focus primarily on ensuring treatment fidelity without taking similar steps to ensure assessment fidelity, because changes on assessment performance are generally the data entered into statistical analyses; if these changes have not been reliably measured, then treatment effects cannot be accurately determined.

Findings from related fields support the importance of assessment fidelity. In reading intervention research, a recent study demonstrated that 16% of the variance in student scores on a curriculum-based oral reading test could be attributed to different examiners (Cummings, Biancarosa, Schaper, & Reed, 2014). Another study reported that 8% of reading assessments had to be discarded due to significant abnormalities, and that an alarming percentage of scored assessments (> 90%) had scoring errors that were found only after double-scoring the assessments (Reed & Sturges, 2013). Several psychoeducational studies have measured examiner errors on the Wechsler Intelligence Scale for Children (WISC, WISC–Revised, WISC–Third Edition, WISC–Fourth

Edition) and Woodcock Johnson III Test of Cognitive Abilities (WJ III COG; Alfonso, Johnson Patinella, & Rader, 1998; Loe, Kadlubek, & Marks, 2007; Ramos, Alfonso, & Schermerhorn, 2009; Sherrets, Gard, & Langner, 1979; Slate & Chick, 1989; Slate & Jones, 1990; Slate, Jones, Coulter, & Covert, 1992). These studies reported on 156 examiners (including graduate students, psychologists, and psychometricians) and found that 81% to 100% of protocols had errors, with the average number of errors ranging from 7.8 to 38.4 per protocol. These errors resulted in incorrect scores (e.g., full scale IQ) in all of the studies. A similar study measured examiner errors in scoring the Rey Figure Test and found that experienced assessors made clerical scoring errors on 18% to 24% of administrations (Charter, Walden, & Padilla, 2000). In addition to scoring errors, studies consistently and convincingly indicate that using unblinded assessors significantly overestimates treatment effects (e.g., Hrobjartsson et al., 2013; Poolman et al., 2007; Wykes, Steel, Everitt, & Tarrier, 2008). It seems clear that assessment fidelity does influence treatment research effect sizes and may cause either an over- or underestimation of the true treatment effect.

Aims

The aims of this study are to (a) document the frequency with which assessment fidelity is reported in aphasia treatment literature, (b) describe assessment fidelity reporting, and (c) provide recommendations for establishing and monitoring assessment fidelity. Applying treatment fidelity guidelines (Bellg et al., 2004; Borrelli, 2011; Chan et al., 2013; Gagnier et al., 2013; Gearing et al., 2011; Hart & Bagiella, 2012; Hildebrand et al., 2012; Hoffmann et al., 2014; Schulz et al., 2010) as well as the SEPT recommendations to the assessment process could reduce threats to assessment fidelity that include variability in assessor qualifications and training, drift, contamination, assessor turnover, and assessor bias (Gearing, et al., 2011; Shadish et al., 2002). This would in turn increase the power to detect treatment effects and maximize the effectiveness of establishing and monitoring treatment fidelity.

Method

Studies listed on the Aphasia Treatment Evidence Tables at the ANCDs Aphasia Treatment Website (<http://aphasiatx.arizona.edu/>) and published between 2010 and 2015 were reviewed. A total of 88 studies were included, ranging from single case studies to randomized controlled trials. The ANCDs evidence tables include a rating of the class (or strength of the evidence) of each study on the basis of study design, according to guidelines published by the American Academy of Neurology (Silberstein, 2000). Class 3 studies provide the weakest evidence and include expert opinions, case studies, case reports, and single-subject multiple baseline studies across behaviors. Class 2 studies provide intermediate evidence derived from observational studies with concurrent controls and single-subject

multiple baseline studies across subjects. The strongest evidence is derived from Class 1 studies (one or more randomized control trials, or meta-analysis of these). The studies reviewed here consisted of five Class 1 studies, 40 Class 2 studies, and 42 Class 3 studies.

Six components were chosen to mirror applicable aspects of treatment fidelity recommended by Bellg et al. (2004) and to expand upon SEPT recommendations: (a) assessment instruments, (b) assessor qualifications, (c) assessor/rater training, (d) adherence to assessment protocols (assessment delivery), (e) rating reliability, and (f) assessor blinding. Binary coding (*reported* = 1, *not reported* = 0) was used to determine whether or not the study reported upon the components of interest. When a study reported upon one of the components, the specific details were recorded to allow for description of the information reported (e.g., If the study reported assessor qualifications, describe the specific qualifications).

Results

Minimal information regarding assessment fidelity was reported (see Table 2). Of the 88 studies reviewed here, only 57% ($N = 50$) provided information regarding to assessment fidelity; 4.5% reported information regarding assessment instruments (e.g., psychometric properties of chosen tests), 35.2% reported assessor qualifications, 6.8% reported assessor training, 37.5% reported assessor (and/or rater) reliability, and 27.3% reported assessor blinding. No studies reported on assessment delivery. Of the 57% of studies that reported on assessment fidelity components, 26 reported on one component of assessment fidelity, 16 reported on two components, five studies reported on three components, and three studies reported on four components of assessment fidelity. The class of research did not seem to influence investigator reporting of assessment fidelity, because little variation was observed—40% of the Class 1 studies (2/5), 40% of the Class 2 studies (16/40), and 48% of the Class 3 studies (20/42) reported on assessment fidelity.

The descriptions of assessment fidelity varied widely. Psychometric properties of assessment instruments included reports of Cronbach's alpha, test-retest reliability, criterion-related validity, construct validity, interrater reliability, and goodness of fit. Reported assessor qualifications included undergraduate bilingual students (for a bilingual study), graduate students familiar with the selected treatment, speech-language pathologists, behavioral neurologists, neuropsychologists, members of the research unit, and first or second authors. Assessor training included training scoring on previously collected data, training and briefing of task requirements and participant cueing, providing oral instructions and an administration manual, and specific training for a less frequently used assessment. Methods for determining reliability of scoring or coding varied widely across studies. Differences included the percentage of data rescored (e.g., 14%, 25%, 50%, 100%), the individuals completing the reliability (e.g., trained students, first author, speech-language pathologist), the types of reliability reported

Table 2. Frequency of reporting proposed assessment fidelity components in aphasia treatment literature.

	Assessment instrument	Assessor qualifications	Assessor and rater training	Assessment delivery	Scoring reliability	Assessor blinding
Reported/total (%)	4/88 (4.5)	31/88 (35.2)	6/88 (6.8)	0/88 (0)	33/88 (37.5)	24/88 (27.3)

(interrater, intrarater), and how reliability was incorporated into the scoring procedures (e.g., forced agreement, mean of both raters' scores, second score incorporated if a disagreement occurred). Assessor blinding was also inconsistent; of the 24 studies that reported on blinding, 12 stated outright that assessors were not blinded, or that they were unblinded during the course of the assessment. For others, assessors were not blinded, but scoring was conducted by a blinded graduate student, research assistant, or speech-language pathologist.

Recommendations for Establishing and Reporting Assessment Fidelity

Our review of recent aphasia treatment peer-reviewed articles revealed that relatively little attention is paid to reporting upon the training and performance of assessment personnel, other than perhaps ensuring a minimum training standard (e.g., certified speech-language pathologist, second-year graduate clinician, etc.) or reporting intra- and interrater reliability. Although it may be the case that monitoring of assessment fidelity occurred but was not reported, consumers of research have no way of knowing or of using that knowledge to help determine their confidence level in the results and to guide inferences. Given the importance of assessment to treatment studies, procedures to ensure assessment fidelity, and reporting the steps taken to do so, are critical to advancing our knowledge of treatment efficacy. Borrowing from best practices in treatment fidelity (Bellg et al., 2004; Borrelli, 2011; Gearing et al., 2011; Hart & Bagiella, 2012; Hildebrand et al., 2012) and translating them to the assessment process, we provide the following recommendations for establishing and monitoring assessment fidelity with an accompanying checklist for researchers and readers (Table 3).

Assessment Instrument

The reliability and stability of an instrument over time, settings, and administrators are important for drawing correct conclusions about relationships between variables, for facilitating detection of false treatment effects, and for generalizing across studies using the same assessment instruments (Shadish et al., 2002). Attention to instrument selection during study design and reporting of psychometric properties of selected instruments should take place to (a) ensure investigators are selecting the instruments most appropriate and capable of testing study hypotheses and (b) to ensure readers have knowledge of and can have confidence in these decisions. There are several published guidelines and frameworks that describe necessary features of psychometrically valid instruments and can be used to aid in the selection of assessment instruments (Cicchetti, 1994; Kagan et al., 2008;

Turkstra et al., 2005). If an assessment will be used as a repeated measure to detect changes due to treatment, the specific validity of using the assessment for that purpose should be reported, because not all assessments are intended to sensitively measure change (Bothe & Richardson, 2011), nor are all able to reveal reliable change without very large, and likely not feasible, pre-post score differences (Jacobson & Truax, 1991). Whenever possible, a psychometrically sound instrument should be preferred to study specific or unvalidated instruments to allow for better comparability across studies. If an assessment is developed for the express purpose of the study, psychometric properties of its validity and reliability (such as Cronbach's alpha or split-half reliability) should be reported, and the assessment should be included as an appendix or supplemental material.

Assessor and Rater Qualifications

In treatment fidelity research, the need to monitor provider characteristics, treatment agents, and/or agent competence is emphasized (e.g., Bellg et al., 2004; Durlak & DuPre, 2008; Gearing et al., 2011), and other downstream fidelity components (e.g., training, delivery, etc.) are thought to interact with prior qualifications and experience (Bellg et al., 2004). It is important to clearly describe, and standardize when possible, assessor and rater qualifications. As with any skilled position, minimum requirements for training and experience should be decided upon for the study and reported. This could include a combination of minimum educational (e.g., master's in communication sciences and disorders or similar degree) and experience requirements (e.g., 4 years of assessment experience). Regarding the latter, investigators may also specify whether or not they require population-specific (e.g., stroke, aphasia, adult neurogenic, etc.) and assessment instrument-specific experience. It is crucial to recognize that the administrator is a vital part of the assessment instrumentation—setting the pace; monitoring time limits; giving instructions, cues, or feedback; and making rating judgments. Instruments are thus only as reliable and resistant to change as the person (or persons, or in some cases the computer program) administering them. In other words, a reliable instrument coupled with a certified speech-language pathologist or experienced research assistant may be necessary but is not sufficient to ensure high assessment fidelity.

Assessor Training

A suggestion for improving the quality of assessment measures, and thereby improving study validity and increasing power, is to conduct “better training of raters” (Shadish et al., 2002, p. 49). Study-specific training should be utilized and standardized across all assessors, beginning with

Table 3. Comprehensive assessment fidelity guide for use by readers and authors.

Assessment fidelity components and strategies	Absent/minimal consideration or reporting	Moderate consideration or reporting	Extensive consideration or reporting
Assessment instrument			
1. Report validity of instruments.	0	1	2
2. Report reliability of normative data.	0	1	2
3. Report psychometrics for study-specific instruments.	0	1	2
4. Report appropriateness of selected repeated measures.	0	1	2
5. Include study-specific assessments in publication.	0	1	2
Assessor/rater qualifications			
1. Report professional experience.	0	1	2
2. Report educational/vocational training.	0	1	2
3. Report previous experience with instruments.	0	1	2
Assessor/rater training			
1. Standardize training.	0	1	2
2. Provide assessment manuals for all instruments.	0	1	2
3. Report training requirements.	0	1	2
Assessment delivery and monitoring			
1. Video record all assessment sessions.	0	1	2
2. Examine recordings for protocol adherence.	0	1	2
3. Examine scoring/rating for protocol adherence.	0	1	2
4. Report adherence to assessment/scoring protocols.	0	1	2
Reliability			
1. Establish proportion of data used to measure reliability.	0	1	2
2. Establish minimum acceptable reliability (.80–.95).	0	1	2
3. Measure all appropriate forms of reliability.	0	1	2
4. Measure reliability of scoring and assessment adherence.	0	1	2
5. Report statistics used to measure reliability.	0	1	2
6. Report reliability coefficients for selected statistics.	0	1	2
Blinding			
1. Blind outcomes assessors/raters whenever possible.	0	1	2
2. Monitor for unblinding of assessors/raters.	0	1	2
3. Report who was blinded.	0	1	2
4. Report if unblinding occurred, and why.	0	1	2

Note. Based on Figure 2 of Gearing et al. (2011).

development of a training manual that includes administration procedures for each individual test and integrated information across the tests highlighting danger zones for contamination. This could also include assessment-specific checklists that assessors could review immediately prior to conducting the testing. Highlighting the opportunities most susceptible to drift or contamination brings the information to the forefront so it can be actively avoided.

Following development of the manual, additional training procedures for assessors should include one or more of the following: (a) independent reading of the manual followed by review with an expert (whose qualifications are described); (b) video observation of expert administration of each test, administered to persons with varying types and severity of aphasia; (c) small-group training sessions including administration manual review, highlighting similarities and differences between administration procedures via discussion and video observation, and supervised role-play with feedback; (d) at project initiation, supervised assessment sessions with expert feedback; and (e) yearly booster small-group training and occasional supervised assessment sessions with expert feedback.

Whenever possible, treatment studies should have blinded assessors who administer the assessment measure

and separate blinded raters who score the assessment measure. Additional training procedures for raters should include one or more of the following: (a) independent reading of scoring manuals followed by review with an expert; (b) co-scoring of training videos, with an expert rater narrating the decision-making process and discussing with the trainee any pertinent topics; (c) independent scoring of training videos with scores checked by an expert rater or compared with acceptable scores provided by a panel of raters; (d) feedback session reviewing independent scoring; (e) repeat of Steps b and c if high point-to-point agreement (> 95%) was not achieved; and (f) yearly booster small-group training and scoring review. Some assessment measures may require more specialized training (e.g., discourse transcription, coding of naming errors, motor speech ratings, etc.) that would need to be accounted for in training procedures.

Use of the varied teaching methods above during training will help to facilitate learning despite different learning styles (Bellg et al., 2004; Borrelli, 2011). Furthermore, the extensive training should serve as a potential equalizer in the face of different qualifications and should have considerable effect on reducing variability in performance across assessors/raters (Bellg et al., 2004; Borrelli, 2011). Recent research into the effectiveness of web-based training

(Kobak, Engelhardt, & Lipsitz, 2006; Kobak, Lipsitz, Williams, Engelhardt, & Bellew, 2005) may provide a potential avenue for assessor/rater training and monitoring for those investigators with funding to support use of such a service.

Before study initiation, investigators should have a plan for remediation of unacceptable performance and scoring during training as well as during monitoring of assessment delivery. There should also be a plan in place for assessor absence and/or turnover. This may involve training of backup assessors or some other solution that ensures that the assessment instrument (i.e., the test plus the assessor) remains essentially unchanged. The same rater should administer assessments to the same subject throughout the study. However, when this is not possible, having a backup rater trained via the same stringent training program at study onset can be key to minimizing data contamination.

Assessment Delivery and Monitoring

Adherence to the assessment procedures throughout the study should be monitored for the following: adherence to assessment scripts and administration procedures; frequency of added nonprescribed elements; frequency of omitted prescribed elements; frequency of cross-contamination; and assessor engagement (difficult to track but vital personal factors, such as empathy, carriage, and comportment; Hildebrand et al., 2012). Monitoring, preferably with checklists, can be conducted via direct supervision of assessors, video- or audio-taping, and/or assessor self-report (see Borelli, 2011, for pros and cons of these approaches). Study assessments can also be reviewed manually after administration as part of monitoring for potential adherence and scoring errors, again utilizing a standard checklist for the scale. Each checklist should begin with a general check of items related to good clinical practice: ensuring that the assessor in question has been trained; ensuring that the order of assessments as specified in the study protocol (if applicable) has been maintained; checking that any study assessment requiring caregiver input has utilized the same caregiver if possible both within and across visits; and conducting a check to ensure that patient and caregiver responses have been appropriately documented (if applicable) on the source document. After the general check has been conducted, study reviewers can then employ a more specific review of each study assessment measure. When reviewing the Western Aphasia Battery–Revised (Kertesz, 2007), for example, reviewers should check to ensure that patient responses have been documented where appropriate for spontaneous speech and the auditory/verbal comprehension items. Further, the reviewer should check to confirm that all subtest scores have been recorded and calculated as well as accurately transferred to the score sheet.

Study investigators should develop rules a priori for the percentage and frequency of assessment session time monitoring (e.g., 20% of total assessment session time monitored quarterly) and for acceptable delivery rating scores (e.g., maximum number of additions and/or omissions allowed, etc.). Further, qualifications of the person(s)

monitoring adherence should be established at study initiation and clearly described (such as level of education, years of experience with the assessments). This individual will also complete reviewer training with an experienced reviewer (e.g., study investigator) at study onset, and use a standard set of guidelines to monitor assessment fidelity.

Reliability

Reliability is an important, but commonly misunderstood, component of assessment fidelity. Many researchers and clinicians understand reliability to be a function of the assessment instrument; however, reliability is a function of the data collected (Thompson, 1994), and therefore varies by study. An instrument cannot be reliable or unreliable, but it can be used to obtain reliable or unreliable scores. For this reason, scoring or rating reliability (e.g., inter- and intrarater, intraclass correlation coefficients) should be consistently reported for a predetermined percentage of assessment items or sessions (typically between 20% and 50%), and all applicable types of reliability should be reported. Because there are many different statistics that can be used to measure reliability, specific information about the measure selected should be reported. Reliability coefficients should be equal to .80 at a minimum; however, it is important to note that .80 may not be sufficient, or even particularly good, given certain circumstances (Kazdin, 1982). Depending on the types of decisions that will be made on the basis of the data, levels of .90 or .95 may be more appropriate (Kottner, et al., 2011; Nunnally & Bernstein, 1994; Polit & Beck, 2008). This is consistent with minimum standards for high stakes placement testing in psychoeducational research (Salvia, Ysseldyke, & Bolt, 2007), and appropriate for studies seeking to establish efficacy. It should be clearly described who the raters or scorers are (if different from those administering the assessment), and their qualifications should be described.

Blinding

Blinding is critical for avoiding observer, or detection, bias, which can lead to significant exaggeration of treatment effects (Hrobjartsson et al., 2013; Pildal et al., 2007; Poolman et al., 2007; Wykes et al., 2008). Without incorporating assessor and/or rater blinding, there is a danger of reporting false positive conclusions regarding treatment efficacy, and use of interobserver reliability as a proxy is discouraged, because high and acceptable interobserver agreement has been shown to coincide with very high observer bias or exaggeration of treatment effects (Hrobjartsson et al., 2013). Investigators should consult reporting guidelines (e.g., CONSORT) and/or The Cochrane Collaboration tool for assessing risk of bias (Higgins & Green, 2011) during the study design phase to determine which measures and personnel should be blinded for unbiased detection of treatment effects. Explicit information about who was blinded, avoiding use of vague descriptions such as *double-blind* (Higgins & Green, 2011; Schulz, Chalmers, & Altman, 2002), should be provided, as should whether or not study personnel remained blinded or became unblinded.

Limitations and Future Directions

We reviewed aphasia treatment studies published between 2010 and 2015 for reporting of assessment fidelity components and presented guidelines for establishing assessment fidelity. These guidelines are consistent with various reporting requirements (e.g., CONSORT, TIDieR), which will allow researchers to clearly explicate their study design choices to the reader, and will allow readers to evaluate research effectively (Hoffmann et al., 2014; Schulz et al., 2010). It may not be possible or feasible to incorporate all recommendations into studies involving assessment, but we encourage investigators to incorporate any assessment fidelity recommendations provided above and any treatment fidelity recommendations provided by Bellg et al. (2004) and others at every opportunity. Generally speaking, if one component of assessment fidelity cannot be utilized (e.g., not blinding assessors due to lack of appropriate personnel), it may be important for other components of assessment fidelity to be more rigorously enacted to protect against the introduction of additional bias in other areas. However, it is yet to be known which fidelity components, when left unchecked, contribute the most variance and reduce power to detect true effects. Using the tools provided here, we encourage researchers to “put their best methodological foot forward” within the constraints of their resources.

Although we argue that the effects of poor assessment fidelity could negatively have an impact on interpretation and validity of research on the basis of reports of scoring errors and blinding effects in other fields (Alfonso et al., 1998; Charter et al., 2000; Cummings et al., 2014; Hrobjartsson et al., 2013; Loe et al., 2007; Poolman et al., 2007; Ramos et al., 2009; Reed & Sturges, 2013; Sherrets et al., 1979; Slate & Chick, 1989; Slate & Jones, 1990; Slate et al., 1992; Wykes et al., 2008), this has not been directly manipulated. An investigation of the relationship between effect size and assessment fidelity reporting was not undertaken due in part to the heterogeneity of the studies reviewed, uncertainty concerning whether or not assessment fidelity was monitored but not reported, and inconsistency of effect size reporting or reporting of individual data so that effect sizes could be determined. Although such a study would certainly add value to this discussion, perhaps the most feasible next step would be a simulation study to determine how large the effects of poor assessment fidelity are on treatment outcomes.

Some treatment fidelity components (e.g., treatment receipt) were not translated to the assessment fidelity guidelines here, including one treatment fidelity strategy for the component training that warrants discussion. Illustrating the relationship between the treatment and the specific project goals emphasizes the importance of adhering to prescribed procedures, and helps create clinician *meta-competence* or *buy-in* that is thought to be important in treatment fidelity (Borrelli, 2011; Gearing et al., 2011). This may seem intuitive for assessment as well—understanding the *why* may facilitate investment in the *how*, or adherence to procedures. In addition, clinician attrition (which would essentially alter the instrumentation, because the clinician is

a key component of instrumentation) may be less frequent when clinicians are invested in the project goals and feel their contribution is important. However, the need to blind assessors is in conflict with providing complete and specific information about project goals, because provision of that information can create an observer bias. Perhaps a broader explanation of the importance of improving outcomes in aphasia treatment coupled with a thorough explanation, with examples, of how assessment fidelity is crucial for such an endeavor would suffice for assessor meta-competence.

Conclusion

The importance of establishing and monitoring treatment fidelity has been emphasized in recent years by both the broader health research and the aphasia research communities. Little attention or discourse has been devoted to assessment fidelity, and current reporting of assessment fidelity in aphasia research is inadequate. This is an unfortunate circumstance because the vast majority of aphasia research involves some form of assessment, whether or not a treatment component is included. Selecting psychometrically sound assessment instruments, incorporating minimum assessor and rater qualifications, providing extensive assessor and rater training, monitoring implementation of assessment procedures and rating reliability, and blinding assessors and raters will guard against threats to statistical conclusion and internal validity that include, among others, clinician-to-clinician variability, drift, and contamination. These approaches (which should be coupled with treatment fidelity monitoring if a treatment study) will in turn increase the power to detect effects, reduce the occurrence of Type I, II, or III errors, and increase investigator and consumer confidence in the results. By doing so, aphasia researchers can avoid using valuable research resources (e.g., federal funding, investigator time, patient time and resources, etc.) to produce underpowered results with small effect sizes; results which are, at best, confusing to consumers and, at worst, misleading (Ioannidis et al., 2014). The impact of such “research waste” can only begin to be fathomed if one recognizes that the impact of it would be felt not only in a single study, but throughout the entire research community (Bellg et al., 2004; Borrelli, 2011; Ioannidis et al., 2014; Salman et al., 2014).

We encourage consumers to expect information about fidelity to be included in future research and to be critical of studies that do not address these important issues. As we continue to explore this topic, we expect there will be issues unique to assessment fidelity that have not been mentioned in guidelines for treatment fidelity. Recommendations may need to be added or removed as our understanding of assessment fidelity, and its effect on treatment outcomes, broadens.

Acknowledgment

This working group of authors was made possible by the Clinical Aphasiology Conference roundtable sessions.

References

- Alfonso, V. C., Johnson, A., Patinella, L., & Rader, D. E. (1998). Common WISC III examiner errors: Evidence from graduate students in training. *Psychology in the Schools, 35*, 119–125.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., . . . Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*, 443–451.
- Benjamin, M. L., Towler, S., Garcia, A., Park, H., Sudhyadhom, A., Harnish, S., . . . Rothi, L. J. G. (2014). A behavioral manipulation engages right frontal cortex during aphasia therapy. *Neuro-rehabilitation and Neural Repair, 28*, 545–553.
- Borrelli, B. (2011). The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. *Journal of Public Health Dentistry, 71*, S52–S63.
- Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., . . . Kressel, H. Y. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology, 277*, 826–832.
- Bothe, A. K., & Richardson, J. D. (2011). Statistical, practical, clinical, and personal significance: Definitions and applications in speech-language pathology. *American Journal of Speech-Language Pathology, 20*, 233–242.
- Brady, M. C., Kelly, H., Godwin, J., & Enderby, P. (2012). Speech and language therapy for aphasia following stroke. *The Cochrane Database of Systematic Reviews, 16*(5), CD000425.
- Carragher, M., Sage, K., & Conroy, P. (2013). The effects of verb retrieval therapy for people with non-fluent aphasia: Evidence from assessment tasks and conversation. *Neuropsychological Rehabilitation, 23*, 846–887.
- Chan, A. W., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., . . . Doré, C. J. (2013). SPIRIT 2013 statement: Defining standard protocol items for clinical trials. *Annals of Internal Medicine, 158*, 200–207.
- Charter, R. A., Walden, D. K., & Padilla, S. P. (2000). Too many simple clerical scoring errors: The Rey Figure as an example. *Journal of Clinical Psychology, 56*, 571–574.
- Cho-Reyes, S., & Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). *Aphasiology, 26*, 1250–1277.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284.
- Cummings, K. D., Biancarosa, G., Schaper, A., & Reed, D. K. (2014). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology, 52*, 361–375.
- De Jong-Hagelstein, M., Van de Sandt-Koenderman, W. M. E., Prins, N. D., Dippel, D. W. J., Koudstaal, P. J., & Visch-Brink, E. G. (2011). Efficacy of early cognitive-linguistic treatment and communicative treatment in aphasia after stroke: A randomised controlled trial (RATS-2). *Journal of Neurology, Neurosurgery, & Psychiatry, 82*, 399–404.
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327–350.
- Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*, 237–256.
- Fitch-West, J., Sands, E. S., & Ross-Swain, D. (1988). *Bedside Evaluation and Screening Test of Aphasia*. Austin, TX: Pro-Ed.
- Gagnier, J. J., Kienle, G., Altman, D. G., Moher, D., Sox, H., Riley, D., & CARE Group. (2013). The CARE guidelines: Consensus-based clinical case report guideline development. *Global Advances in Health and Medicine, 2*(5), 38–43.
- Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*, 79–88.
- Gresham, F. M., Gansle, K. A., Noell, G. H., & Cohen, S. R. (1993). Treatment integrity of school-based behavioral intervention studies. *School Psychology Review, 22*, 254–272.
- Hansen, W. B., Graham, J. W., Wolkenstein, B. H., & Rohrbach, L. A. (1991). Program integrity as a moderator of prevention program effectiveness: Results for fifth-grade students in the adolescent alcohol prevention trial. *Journal of Studies on Alcohol, 52*, 568–579.
- Hart, T., & Bagiella, E. (2012). Design and implementation of clinical trials in rehabilitation research. *Archives of Physical Medicine and Rehabilitation, 93*, S117–S126.
- Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). The Cochrane Collaboration. Retrieved from www.cochrane-handbook.org
- Hildebrand, M. W., Host, H. H., Binder, E. F., Carpenter, B., Freedland, K. E., Morrow-Howell, N., . . . Lenze, E. J. (2012). Measuring treatment fidelity in a rehabilitation intervention study. *American Journal of Physical Medicine & Rehabilitation, 91*, 715.
- Hinckley, J. J., & Douglas, N. F. (2013). Treatment fidelity: Its importance and reported frequency in aphasia treatment studies. *American Journal of Speech-Language Pathology, 22*, S279–S284.
- Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., . . . Lamb, S. E. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ, 348*, g1687.
- Holland, A. L., Fromm, D. S., DeRuyter, F., & Stein, M. (1996). Treatment efficacy: Aphasia. *Journal of Speech and Hearing Research, 39*, 27–36.
- Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., . . . Brorson, S. (2013). Observer bias in randomized clinical trials with measurement scale outcomes: A systematic review of trials with both blinded and nonblinded assessors. *Canadian Medical Association Journal, 185*, E201–E211.
- Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., . . . Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet, 383*, 166–175.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19.
- Kagan, A., Simmons-Mackie, N., Rowland, A., Huijbregts, M., Shumway, E., McEwen, S., . . . Sharp, S. (2008). Counting what counts: A framework for capturing real-life outcomes of aphasia intervention. *Aphasiology, 22*, 258–280.
- Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test*. Austin, TX: Pro-Ed.

- Kay, J., Lesser, R., & Coltheart, M. (1992). *Psycholinguistic Assessment of Language Processing in Aphasia (PALPA)*. London, United Kingdom: Erlbaum.
- Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.
- Kertesz, A. (2007). *The Western Aphasia Battery-Revised*. San Antonio, TX: Pearson.
- Kobak, K. A., Engelhardt, N., & Lipsitz, J. D. (2006). Enriched rater training using Internet-based technologies: A comparison to traditional rater training in a multi-site depression trial. *Journal of Psychiatric Research*, 40, 192–199.
- Kobak, K. A., Lipsitz, J. D., Williams, J. B. W., Engelhardt, N., & Bellew, K. M. (2005). A new approach to rater training and certification in a multicenter clinical trial. *Journal of Clinical Psychopharmacology*, 25, 407–412.
- Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., . . . Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies*, 48, 661–671.
- Loe, S. A., Kadlubek, R. M., & Marks, W. J. (2007). Administration and scoring errors on the WISC-IV among graduate student examiners. *Journal of Psychoeducational Assessment*, 25, 237–247.
- MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, 25, 1286–1307.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, 151, 264–269.
- Nigg, C. R., Allegrante, J. P., & Ory, M. (2002). Theory-comparison and multiple-behavior research: Common themes advancing health behavior research. *Health Education Research: Theory and Practice*, 17, 670–679.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.
- Pildal, J., Hróbjartsson, A., Jørgensen, K. J., Hilden, J., Altman, D. G., & Gøtzsche, P. C. (2007). Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *International Journal of Epidemiology*, 36, 847–857.
- Polit, D. F., & Beck, C. T. (2008). *Nursing research: Generating and assessing evidence for nursing practice*, (8th ed.). Philadelphia, PA: Lippincott, Williams & Wilkins.
- Poolman, R. W., Struijs, P. A., Krips, R., Sierevelt, I. N., Marti, R. K., Farrokhyar, F., & Bhandari, M. (2007). Reporting of outcomes in orthopaedic randomized trials: Does blinding of outcome assessors matter? *Journal of Bone and Joint Surgery*, 89, 550–558.
- Ramos, E., Alfonso, V. C., & Schermerhorn, S. M. (2009). Graduate students' administration and scoring errors on the Woodcock-Johnson III Tests of Cognitive Abilities. *Psychology in the Schools*, 46, 650–657.
- Reed, D. K., & Sturges, K. M. (2013). An examination of assessment fidelity in the administration and interpretation of reading tests. *Remedial and Special Education*, 34, 259–268.
- Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). Philadelphia Naming Test. *Clinical Aphasiology*, 24, 121–133.
- 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.
- Salman, R. A. S., Beller, E., Kagan, J., Hemminki, E., Phillips, R. S., Savulescu, J., . . . Chalmers, I. (2014). Increasing value and reducing waste in biomedical research regulation and management. *The Lancet*, 383, 176–185.
- Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment in special and inclusive education* (10th ed.). Boston, MA: Houghton Mifflin Company.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine*, 152, 726–732.
- Schulz, K. F., Chalmers, I., & Altman, D. G. (2002). The landscape and lexicon of blinding in randomized trials. *Annals of Internal Medicine*, 136, 254–259.
- Shadish, W., Cook, T., & Campbell, D. (2002). Statistical conclusion validity and internal validity. In W. Shadish, T. Cook, & D. Campbell (Eds.), *Experimental and quasi experimental designs for generalized causal inference* (pp. 33–63). Boston, MA: Houghton Mifflin.
- Sherrets, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. *Psychology in the Schools*, 16, 495–496.
- Silberstein, S. D. (2000). Practice parameter: Evidence-based guidelines for migraine headache (an evidence-based review): Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, 55, 754–762.
- Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. *Psychology in the Schools*, 26, 78–84.
- Slate, J. R., & Jones, C. H. (1990). Student error in administering the WISC-R: Identifying problem areas. *Measurement and Evaluation in Counseling and Development*, 23, 137–140.
- Slate, J. R., Jones, C. H., Coulter, C., & Covert, T. L. (1992). Practitioners' administration and scoring of the WISC-R: Evidence that we do err. *Journal of School Psychology*, 30, 77–82.
- Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, 54, 837–847.
- Turkstra, L., Ylvisaker, M., Coelho, C., Kennedy, M., Sohlberg, M. M., Avery, J., & Yorkston, K. (2005). Practice guidelines for standardized assessment for persons with traumatic brain injury. *Journal of Medical Speech-Language Pathology*, 13, ix–ix.
- Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., & STROBE Initiative. (2007). The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Preventive Medicine*, 45, 247–251.
- Wykes, T., Steel, C., Everitt, B., & Tarrier, N. (2008). Cognitive behavior therapy for schizophrenia: Effect sizes, clinical models, and methodological rigor. *Schizophrenia Bulletin*, 34, 523–537.