Summer 7-15-2017

ESTIMATING THE IMPACT OF ASSESSMENT AND TREATMENT FIDELITY ON APHASIA TREATMENT OUTCOMES

Trisha L. Tanaka
University of New Mexico - Main Campus

Trisha Tanaka

Candidate

Speech and Hearing Sciences Department

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Jessica Richardson , Chairperson

Richard Arenas

Janet Patterson

ESTIMATING THE IMPACT OF ASSESSMENT AND TREATMENT FIDELITY ON APHASIA TREATMENT OUTCOMES

by

TRISHA TANAKA

B.S., Communication Disorders and Deaf Education
B.A., Global and International Studies

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE
SPEECH-LANGUAGE PATHOLOGY

The University of New Mexico
Albuquerque, New Mexico

JULY, 2017

ACKNOWLEDGEMENTS

I am extremely grateful for the endless time, perspective, and words of encouragement Dr. Jessica Richardson has offered me over these past two years. Her guidance and experience gave me the confidence to continue with my thesis in the midst of fruitless literature reviews, mathematical uncertainty, and mind numbing writer’s block. She instilled in me a passion for research that I cannot shake. I would like to wholeheartedly thank Dr. Janet Patterson and Dr. Richard Arenas for providing me with their invaluable expertise in assessment and treatment research to best plan, execute, and refine this thesis. I am also indebted to Sarah Grace Hudspeth Dalton for her Excel mastery during times of great need, late night feedback, and serving as a general bouncing board for ideas. She and Dr. Richardson should both be recognized for their remarkable commitment to treatment and assessment integrity. To my fiancé, Kevin, this journey would have never even started without your love and constant reminders to try and try again.

Estimating the Impact of Assessment and Treatment Fidelity on Aphasia Treatment Outcomes

Trisha Tanaka

B.A., Global and International Studies, University of California, Santa Barbara, 2011

B.S., Communication Disorders and Deaf Education, Utah State University, 2015

M.S., Speech-Language Pathology, University of New Mexico, 2017

ABSTRACT

Purpose: Calls for treatment and assessment fidelity strongly suggest the need to reduce treatment provider and assessor variance surrounding intervention research. The extent to which these sources of variance influence treatment outcomes in aphasia treatment research has yet to be examined. This simulation study sought to explore the relationships between quality of fidelity methods, sample size, power to detect treatment effects, and aphasia treatment effect sizes.

Methods: Individual participant outcomes collected from previous aphasia treatment research studies were used to simulate 200,000 participant outcomes, from which 8,000 sample treatment trials were simulated. Effect sizes were calculated for treatment outcomes related to four total assessment and treatment fidelity methods - treatment provider training, treatment provider monitoring, assessor blinding, and assessor training. Results from calculations were applied to 80,000 simulated participant trials of varying sample sizes, fidelity levels, and outcome assessments to determine effect size and power to detect effects.

Results: Simulated results found: positive effect sizes and increased power to detect effects for high fidelity treatment provider training and monitoring, with reduced effect sizes and ability to detect effects from high fidelity assessor blinding, and no effects for assessor training. Increased power was observed as sample size increased. Multidimensional assessment outcomes resulted in higher treatment effect sizes and power to detect effects than unidimensional outcomes.

Conclusions: Simulations generally support findings from previous research. With the exception of treatment provider training, few studies reported calculable outcomes related to fidelity, validating the need for this simulation and future research. High fidelity treatment provider training and monitoring are simple methods to increase ability to detect treatment effects and effect size overall, and blinding assessors helps to reduce biased reporting. Recommendations for researchers with limited resources are provided to reduce variance from assessors and treatment providers and increase confidence in results.

KEY WORDS: aphasia, assessment, treatment, intervention, fidelity, integrity, simulation, blind, training, monitor, adherence

TABLE OF CONTENTS

ACKNOWLEDGEMENTS..iii

ABSTRACT..iv

LIST OF FIGURES..viii

LIST OF TABLES..ix

Introduction..1

Method..8

Study Design..8

Extraction of Individual Treatment Outcomes from Aphasia Treatment Studies...... 9

Power Analysis Simulation for Baseline Study Trials..10

Treatment and Assessment Fidelity Article Searches..12

Research Synthesis of Fidelity Outcomes Calculated Into Effect Sizes..15

Power Simulation Including Fidelity Effect Sizes..15

Results..16

Effect Sizes Calculated From Health and Behavioral Science Literature..16

Simulated WAB and BNT Change Scores..20

Simulated Trials with WAB and BNT Participant Change Scores..21

Effect Size and Power Simulation with Treatment Provider Training..23

Effect Size and Power Simulation with Treatment Fidelity Monitoring..24

Effect Size and Power Simulation with Assessor Blinding..25

Effect Size and Power Simulation with Assessor Training..27

Effect Size and Power Simulation with Combined Effect Sizes..28

Discussion..28

Treatment Provider Training..30

Treatment Provider Monitoring..32

Assessor Blinding 33

Assessor Training..35

Assessor Errors 37

Recommendations for Aphasia Treatment Researchers with Limited Resources...38

Limitations and Future Directions 38

Conclusions 40

List of Appendices...42

References...56

LIST OF FIGURES

Figure 1. 9

Figure 2. 14

Figure 3. 18

Figure 4. 21

Figure 5. 21

Figure 6. 23

Figure 7. 25

Figure 8. 26

Figure 9. 27

LIST OF TABLES

Table 1 20

Introduction

In treatment research, variance between participants, providers, and assessors can obstruct interpretations of treatment outcomes. Fidelity measures in treatment research (i.e., methods ensuring adherence to prescribed treatment and assessment procedures) may reduce variance, also described as noise or error. Direct comparisons of studies with high and low fidelity in the health and behavioral science literature have indicated that high treatment fidelity generally increases the power to detect effects (Borrelli, 2011) and is associated with increased effect sizes overall (Claridge, 2014; Hansen, et al., 1991; Koehler, et al., 2013; Maxfield & Hyer, 2002).

Treatment research is designed to infer relationships between treatment variables and patient outcomes. Ideally, it is a vehicle for dissemination of information in which practitioners can be confident, as these inferences may ultimately lead to beneficial outcomes for clients in non-laboratory settings. The level of confidence one can have in study results relates directly to study validity, or how closely a study's inference approximates the truth, and measures what it states that it measures (Shadish, Cook, & Campbell, 2002). One component of validity, statistical conclusion validity, is key to distinguishing whether there is an association between treatment and outcome and related magnitude (Shadish, Cook, & Campbell, 2002). Inaccurate conclusions about presence of an association include Type I errors, which assume relationships exist where there are none, and Type II errors, which assume that relationships do not exist when they do. Threats to statistical conclusion validity may include low statistical power, unreliable measures of variables obtained, and unreliable implementation (Shadish, Cook, & Campbell, 2002). Another threat, sometimes labeled a Type III error, occurs when inconsistent or nonexistent implementation discredits conclusions of either significance or nonsignificance (Nigg, Allegrante, & Ory, 2002). When these aforementioned threats are not removed or evaluated to determine their influence, the accuracy of claims, or inferences, about a treatment is at increased risk.

A threat to statistical conclusion validity is variance, which arises in part from inherent differences between participants, providers, and assessors. Deviations from one person to another are natural, even expected to a degree, with some factors being more or less controllable. Factors resistant to control might include patient temperament, motivation, family support, and fatigue. Experiments often attempt to use stringent inclusion and exclusion criteria to control for or reduce the impact of these sources of variance, though at a cost to generalization meaningful to clinicians.

While enrollment criteria are more consistently used to account for patient-related noise, other more preventable provider- and interventionist- related sources of variance receive inconsistent attention. Such sources include, for example, therapist drift, or deviation from prescribed therapy protocol over time. If unchecked over the course of an intervention, therapeutic providers may drift in methodology and inadvertently include non-prescribed elements of therapy or exclude core components, making it difficult to determine whether the core treatment components are the cause of outcome change. Another source could be errors found in scoring procedures, such as counting errors, addition and subtraction of scores, and transfer of raw scores to standardized scores, which could potentially contribute to inaccurate estimates of change following intervention. This variance may impact interpretations of the significance of a treatment effect and perhaps more importantly of the magnitude of difference, or the effect size, between conditions and/or groups.

Fidelity measures in treatment research (i.e. methods ensuring adherence to prescribed treatment and assessment procedures) can remove or reduce variance from sources previously mentioned. Treatment fidelity, the most commonly discussed type of fidelity, is defined as the extent to which an intervention is delivered as intended and is distinguishable from comparative treatment condition (Borrelli, 2011). Establishing treatment fidelity may involve control of provider qualifications and training as well as monitoring of the following: therapist drift from prescribed treatment protocol, contamination of therapeutic components, removal of therapeutic components, and inclusion of non-prescribed components.

Studies that take steps to ensure high fidelity (a term often interchangeable with “integrity”) have demonstrated benefits of revealing a stronger signal, in the form of larger effect sizes, across the behavioral and health science literature (e.g., Claridge, 2014; Hansen, Graham, Wolkenstein, & Rohrbach, 1991; Koehler, Lösel, Akoensi, & Humphreys, 2013; Solomon, Battistich, Watson, Schaps, & Lewis, 2000). Meta-analyses of treatment fidelity include reports from youth programs where effect sizes increased 2 to 3 times more with programs that monitored treatment implementation compared to those that did not (Dubois et al., 2002; Smith, Schneider, Smith, & Ananiadou, 2004). Sufficient training with the use of a treatment manual, most relevant for complex, step-by-step programs such as eye movement desensitization and reprocessing (EMDR) therapy, results in larger effect sizes compared to studies that do not incorporate such training and resources (Lee & Cuijpers, 2013). Also in the field of psychotherapy, studies of treatments addressing perinatal depression that included fidelity checks for treatment adherence produced higher effect sizes than those that did not (Claridge, 2014).

Investigations of studies with high and low fidelity in the health and behavioral science literature have indicated that high treatment fidelity increases the power to detect effects that may have otherwise been obscured by variance (Borrelli, 2011) and is associated with increased effect sizes overall (Claridge, 2014; Hansen, et al., 1991; Koehler, et al., 2013; Maxfield & Hyer, 2002). For example, despite the large amount of variability inherent in programs that span many research sites in several countries, a review of correctional programs for young offenders throughout Europe revealed a 12% reduction in re-offenders participating in programs with high fidelity versus a 5% rate reduction for programs with low fidelity (Koehler, et al., 2013). Improved fidelity over the course of an intervention can be beneficial as well - a longitudinal psychoeducational study reported greater student outcomes in schools that significantly improved implementation fidelity over time (Solomon et al., 2000).

While the inclusion of treatment fidelity measures has gained traction in research intervention guidelines, with increasing efforts to monitor and provide consistent guidelines for treatment fidelity standards in particular (Bellg et al., 2004; Borrelli, 2011; Gearing et al., 2011), the same cannot be said for assessment fidelity, or guidelines to monitor adherence to assessment protocol (Richardson et al., 2016). Just as variance in the provision of core treatment components may impact outcome interpretations, measurement of outcomes is also susceptible to variance - for example, assessor errors in scoring, assessor drift from protocol, contamination of assessment criterion and methods, and lack of assessor blinding (Richardson et al., 2016). Recent recommendations for increased assessment fidelity include: predetermined assessor and rater training and qualifications, use of training manuals, video-observation of administration and scoring methods, role-play and monitoring of practice assessments and scoring with immediate feedback, booster training sessions for scoring and administration, adherence monitoring, and more (Richardson et al., 2016). Compared to treatment fidelity, assessment fidelity has received little attention, and less is known about the influence of assessment fidelity on power and effect sizes.

Perhaps the most well-known and commonly recommended practice of assessment fidelity is blinding outcome assessors for treatment condition to control for observer bias. Subjective outcome assessments are especially at high risk for inflated results, as indicated by a review of observer bias in subjective rating systems and its influence upon outcomes (Hrobjartsson et al., 2013). Aggregate analysis concluded that subjective ratings by non-blinded assessors compared to blinded assessors on the exact same measure and participant pool led to an exaggeration in effect size by 68 percent. Implications of observer bias through non-blinding suggest strong impacts on research outcomes, and mixed results when blinding is included in replication studies. Yet the practice of non-blinded assessment still occurs, as indicated in recent reports of scarce blinding in speech-language pathology and related fields (Leong, 2014; Simpson, 2014). Even if studies have self-labeled as ‘double-blind’, there is a need to critically evaluate or consider results with an air of skepticism, as further appraisal of 200 clinical trials has revealed that at least one in five studies with this label did not include participants, providers, or data collectors who were blind to conditions (Haahr & Hróbjartsson, 2006). This is not trivial - exaggeration of effect sizes due to non-blinding alone could mean the difference between a study being published, possible misinterpretation of the true nature and impact of an intervention, and adoption by health and behavioral science professionals. The influence of a single assessment fidelity dimension thus raises red flags as to the impact that other assessment fidelity dimensions may have. Nevertheless, the dearth of information about assessment fidelity in the health and behavioral science literature makes it difficult to draw any definitive conclusions about what impacts inclusion of this component has more specifically for the field of speech-language pathology.

There is a growing body of evidence supporting inclusion of fidelity monitoring in research. Without ensuring fidelity, it is difficult to determine whether a specific therapeutic component is beneficial, harmful, or insignificant. Further, some aspects of fidelity may be more detrimental to obscuring true treatment effects if not monitored compared to others, but which fidelity components should be prioritized, for example in the case of limited resources, is unknown. Studies including mixed fidelity dimensions of both assessment and treatment domains have shown that certain measures were more influential to outcomes than others, but not in a consistent manner (Maxfield & Hyer, 2002). In the field of psychology, several different dimensions of fidelity have been significant moderators of effect size, depending on the nature of the intervention (Claridge, 2014; Maxfield & Hyer, 2002).

In speech-language pathology, specifically in the aphasia literature, neither treatment nor assessment fidelity receive the attention needed (Hinckley & Douglas, 2013; Richardson et al., 2016). There are not enough studies reporting upon fidelity components to conduct a meta-analyses on the influence of fidelity on treatment effect sizes as has been performed in other related literature (e.g., Hrobjartsson et al., 2013;

Maxfield & Hyer, 2002). While methodologically sound, it would be unethical to prospectively compare outcomes of studies with varying degrees of treatment and assessment fidelity, given what we currently know about the impact of fidelity on detection of effects in various fields (psychology, education, etc.). An alternative to directly influencing and observing outcomes related to low and high fidelity would be the use of a simulation study.

Simulations can inform program decisions by demonstrating the influence of a variety of factors on possible outcomes. They can evaluate the quality, effectiveness, and efficiency of a program, leading to program decisions, and on a larger scale, recommendations for policy planners (Mielczcarek & Uziałko-Mydlikowska, 2012). Simulation studies can also ask questions that may be important to further examine, but do not compromise a participant's well-being, such as a retrospective study that simulated the accuracy of various screeners to predict survival rates of individuals with cardiovascular disease (Bailey, Berson, Handelsman, & Hodges, 2001). Further, simulation studies that include measures from real participants of previous research interventions may better recognize individual gains made that are indicative of meaningful change for those populations, which can often be washed out in a large sample of statistical analysis.

The versatile nature of simulation studies affords the opportunity to consider several different scenarios and their effects across a variety of situations. For instance, with synthetic projections applied to unpublished data that initially lacked an effect size or statistical significance, investigators found that high-risk substance abuse prevention programs could be 12 times more effective if implementation fidelity components were included (Derzon, Sale, Springer, & Brounstein, 2005). Calls for organizations to more closely examine treatment integrity also come from education research, best represented by a simulation study from Stockard (2010). This study calculated differences in effect size related to hypothetical low and high treatment integrity conditions, asserting that low treatment integrity may mask true findings of both ineffective and effective interventions (Stockard, 2010).

The purpose of this study is to:

(1) compile effect sizes from the health and behavioral science literature documenting the influence of single dimensions of fidelity on treatment outcomes,

(2) use simulation to investigate the impact of monitoring select treatment and assessment fidelity components, both individually and in combination, on power and effect sizes,

(3) use simulation to investigate the interaction of sample size and fidelity on treatment outcomes, and

(4) provide recommendations to future researchers about assessment and treatment fidelity components to include as well as strategies to compensate for variance when inclusion of certain fidelity components are impossible or not within their resources to implement.

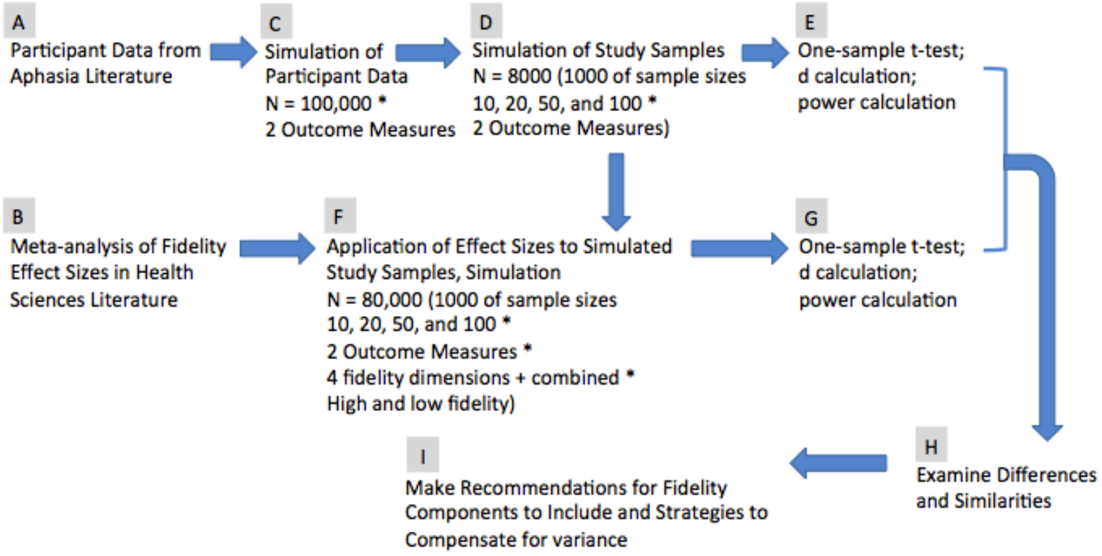
Method

Study Design

The research workflow for this study is depicted in Figure 1 and included literature review, meta-analysis, numerous simulations, and interpretation. For simulations, descriptive statistics, statistical analyses, and figure creation, SPSS 24 and Microsoft Excel were utilized. R (v. 3.3.2) and RStudio (v. 1.0.136) were used for power calculations.

Figure 1.

Workflow of Study Design


A
Participant Data from Aphasia Literature
C
Simulation of Participant Data
N = 100,000 *
2 Outcome Measures
D
Simulation of Study Samples
N = 8000 (1000 of sample sizes
10, 20, 50, and 100 *
2 Outcome Measures)
E
One-sample t-test;
d calculation;
power calculation
B
Meta-analysis of Fidelity Effect Sizes in Health Sciences Literature
F
Application of Effect Sizes to Simulated Study Samples, Simulation
N = 80,000 (1000 of sample sizes
10, 20, 50, and 100 *
2 Outcome Measures *
4 fidelity dimensions + combined *
High and low fidelity)
G
One-sample t-test;
d calculation;
power calculation
H
Examine Differences and Similarities
I
Make Recommendations for Fidelity Components to Include and Strategies to Compensate for variance

Extraction of Individual Treatment Outcomes from Aphasia Treatment Studies

Individual participant outcomes from aphasia treatment studies were obtained so that ecologically valid change scores of persons with varying types and severities of aphasia could be entered into the simulation to exemplify the non-normal distribution of treatment outcomes for this population (Figure 1A). Outcomes were obtained from recent treatment studies (between 2000-2015) listed on the Aphasia Treatment Evidence Tables at the ANCDS Aphasia Treatment Website (http://aphasiatx.arizona.edu). Studies spanning a variety of aphasia treatment categories were examined (e.g., speech production/fluency and lexical retrieval). Of the 122 studies reviewed, 33 reported individual Western Aphasia Battery, or Western Aphasia Battery-Revised, Aphasia Quotient (WAB-AQ and WAB-R-AQ, hereafter labeled as WAB) and/or Boston Naming Test (BNT and BNT-2, hereafter labeled as BNT) change scores (see Appendix A for references). For consistency across measures, pre- and immediately post- treatment scores were extracted (i.e., not long-term follow-up). For within-group crossover designs, data from the first treatment phase were extracted. A total of 108 WAB and 94 BNT individual participant change scores were extracted, with 75 WAB and BNT scores from the same participants.

Previous reviews have revealed that information about both assessment and treatment fidelity in current aphasia treatment research is limited (Hinckley & Douglas, 2013; Richardson et al., 2016), highlighting the need for a simulation to understand impacts on treatment study effect sizes and power to detect effects. Fidelity may not be reported for a number of reasons (e.g., lack of guidelines, oversight in journal requirements, low awareness by investigators, or inclusion of fidelity but no description). It is probable that some of the change scores included were from studies performed with high fidelity, while some were from studies performed with moderate and low fidelity. We considered that a simulation including participant change scores from aphasia treatment studies that likely spanned the spectrum of fidelity would serve well as a middle ground baseline of fidelity from which we extracted our information.

Power Analysis Simulation for Baseline Study Trials

A Monte Carlo simulation of participant change scores on the WAB was conducted to generate 100,000 each pseudo-random participant treatment-induced change scores into a distribution determined by the original samples (Figure 1C). This process was repeated for BNT change scores. Descriptive statistics (means, medians, standard deviations, variance, skewness and kurtosis) were examined to ensure similarities between the original individual participant change scores and the generated simulated values. To facilitate investigation of the influence of fidelity on effect size and power as a function of sample size, we randomly extracted 10 participant change scores 1,000 times to represent 1,000 simulated study trials of *n*=10 each. We repeated this for *n*=20, 50, and 100 (Figure 1D). This process was performed for participant change scores measured by both the WAB and BNT. We then conducted statistical analyses, including *t*-tests (Equation 1), effect size (Equations 2 and 3), and power (Equation 4) (Figure 1E).

Using a one-sample *t*-test, we generated *t*-test values to determine whether the simulated change scores, reflecting treatment-induced change, differed from a hypothesized population where treatment did not result in change (i.e., where the population mean [μ] = 0).

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \; where \; s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{1}$$

where $\bar{x}$ = sample mean, s = sample standard deviation, and n = sample size.

Effect sizes were calculated using *t*-test statistics divided by the square root of the sample size, as in the following formula:

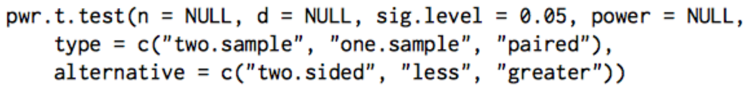
$$d = \frac{t}{\sqrt{n}} \tag{2}$$

where *d* = Cohen's measure of sample effect size for comparing two sample means.

This was validated by using the following formula in excel (again where the population mean [μ] = 0):

$$d = \frac{\bar{x}-\mu}{s}. \tag{3}$$

Post-hoc power to detect effects was calculated using effect size, sample size, one-sample t-test, and alpha (less than 0.05) with the following R code structure:



```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL,
    type = c("two.sample", "one.sample", "paired"),
    alternative = c("two.sided", "less", "greater"))
```

(4)

Treatment and Assessment Fidelity Article Searches

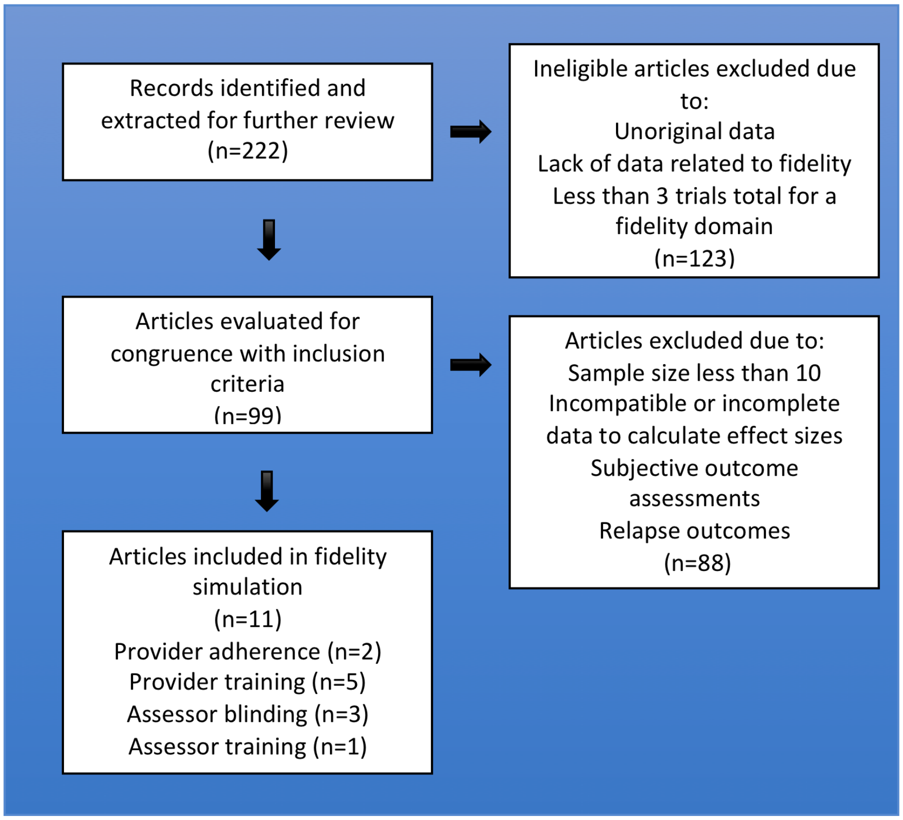
Peer-reviewed articles (January 2000 – February 2017) were appraised for discrete treatment and assessment fidelity dimensions and the influence of their relative presence or absence on treatment outcomes (Figure 1B). Searches using Google Scholar, Linguistics and Language Behavior Abstracts (LLBA), and PubMed were conducted including a combination of terms: *fidelity, validity, reliability, adherence, integrity, treatment, implementation, intervention, assessment, assessor drift, variance*, and *noise*. Initial searches yielding treatment outcomes with measurable levels of fidelity (e.g., high versus low-to-no fidelity, or a continuum of adherence to core components) were examined for utility. Potential treatment and assessment fidelity dimensions considered were assessor and provider qualifications, training, skills and knowledge, contamination, and delivery monitoring, as well as inter-rater reliability and external vs. internal evaluators (Gearing, 2011). An inclusion criterion of at least 3 journal articles for each dimension was required to further pursue additional searches of a dimension. The investigators discussed inclusion of the two treatment fidelity and two assessment fidelity dimensions with the largest source of data most applicable to aphasia outcomes until agreed upon. The resulting dimensions included assessor blinding, assessor/rater training, provider training, and provider adherence.

Following the identification of candidate fidelity dimensions, further searches within behavioral and health science fields included the following search terms and derivations: *fidelity, validity, adherence, integrity, treatment, implementation, intervention, provider, outcome, assessor, rater, training, blind/unblind, mask/unmask, psychology, education, applied behavioral analysis, occupational therapy, speech-language pathology,* and *physical therapy*. References from reviews and articles were examined for pertinent information related to assessment and/or treatment fidelity dimensions (e.g., Durlak & Dupre, 2008; Hrobjartsson et al., 2013; Reed & Sturges, 2012).

A total of 222 outcome studies and reviews were identified and extracted for further review from behavioral and health sciences, none of which included aphasia treatment studies. With an inclusion criteria of original data related to fidelity and more than 3 study trials for a fidelity domain, 99 outcome studies were then considered for inclusion (Figure 2). Due to the limited outcome data related to assessment and treatment fidelity, the following exclusion criteria were considered post-hoc: a small sample size (*n* =10 or less), complex or incompatible data for analysis (e.g., use of confirmatory or growth models or post-assessment data only), and outcomes not comparable to WAB and BNT measurements (e.g., subjective global depression outcome scales, relapse rates, aggressive behavior, or a Likert rating scale of attitudes towards drug abuse). Grounds for inclusion/exclusion of a study were discussed amongst the authors until consensus was reached. Due to incompatible data, articles with medical treatments for multiple sclerosis, Parkinson's disease, and cerebral palsy as well as interventions for early childhood school readiness, social skills, and reading were excluded. Articles with outcome assessments considered too subjective for inclusion were from the fields of school psychology and youth services. Fidelity data from a total of 11 articles were included in this study, with 2 articles reporting upon treatment provider monitoring, 5 articles addressing treatment provider training, 3 addressing assessor blinding, and 1 article reporting upon assessor training.

Figure 2.

Flow Chart of Fidelity Studies Meeting Inclusion Criteria



Records identified and extracted for further review (n=222)

Ineligible articles excluded due to:
Unoriginal data
Lack of data related to fidelity
Less than 3 trials total for a fidelity domain
(n=123)

Articles evaluated for congruence with inclusion criteria
(n=99)

Articles excluded due to:
Sample size less than 10
Incompatible or incomplete data to calculate effect sizes
Subjective outcome assessments
Relapse outcomes
(n=88)

Articles included in fidelity simulation
(n=11)
Provider adherence (n=2)
Provider training (n=5)
Assessor blinding (n=3)
Assessor training (n=1)

Research Synthesis of Fidelity Outcomes Calculated Into Effect Sizes

Data extracted for the four selected fidelity dimensions were translated into effect sizes in the form of Cohen's *d* (Equations 5) using an online effect size calculator (https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD24.php) (Figure 1B). Results were validated with Excel calculators. Whenever possible, we used means and standard deviations (19/23 outcomes from 9/11 studies), and controlled for direction of effects (e.g., when reduced participant scores reflected positive outcomes) (Borenstein, Hedges, Higgins, & Rothstein, 2009):

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_{pooled}} \quad \text{where} \quad S_{pooled} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}. \tag{5}$$

If not available, we used correlation coefficients (Equation 7) and translated to Cohen's *d* (Cortina & Nouri, 2000) as performed with one study (Benner, Nelson, Stage, & Ralston, 2011):

$$d = \sqrt{\frac{2r}{(1-r^2)}} \tag{6}$$

where *r* = estimate of the Pearson product-moment correlation coefficient. For one study (Hamre et al., 2010), the authors did not provide any of the above, but reported Cohen's *d* for two outcomes, which were included in the simulations.

Power Simulation Including Fidelity Effect Sizes

Effect sizes for all dimensions were translated into forest plots to aid visual representation of the potential moderating factors when inclusion/exclusion of fidelity dimensions occurred (Figure 1B) (Borenstein et al., 2009; Cooper, Hedges, & Valentine, 2009). Effect sizes derived from the meta-analysis above were used to solve for the difference in average change scores between our baseline simulated samples and those with high and low fidelity (Equation 8). The following formula was used to solve for $\bar{X}_1$:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s}. \quad (8)$$

Across the 1,000 study trials for each sample size, effect sizes for each fidelity domain were sequenced as listed in Figure 3 and iteratively applied, using the positive or negative sign available in Figure 3 for high fidelity simulations, and reversing the sign for low fidelity simulations (e.g., *d* = 0.43 for high fidelity treatment provider training and *d* = -0.43 for low fidelity treatment provider training) (Figure 1F). Effect sizes were matched appropriately to WAB and BNT by subjective and objective qualities of the outcome assessments used (Figure 3). Scores for the BNT are more objective in nature, relying upon whether the individual names an item pictured, with limited room for interpretation. Aside from a complement of relatively objectively scored scales, two scores for the WAB include rating scales with criteria on verbal fluency that require rater judgment and are more subjective in nature. Effect sizes using objective outcomes were applied to BNT simulations; effect sizes using subjective and objective outcomes were applied to WAB simulations. One-sample *t*-tests were again conducted with the new sample mean, allowing for computation of effect size and power (Figure 1G). Effect sizes and power calculations of all base simulation, high fidelity simulations, and low fidelity simulations were compared (Figure 1H).

Results

Effect Sizes Calculated From Health and Behavioral Science Literature

Forest plots of effect sizes for each fidelity domain were created to visually represent overall negative or positive effect sizes related to a fidelity domain (Figure 3).

Five studies (with 15 relevant outcomes) reported outcomes related to provider training; positive effect sizes as a result of increased provider training were observed for 11/15 outcomes (Figure 3). Two studies (with 4 relevant outcomes) reported outcomes related to treatment provider monitoring, both with positive effect sizes as a result of treatment provider monitoring. Three studies (with 3 relevant outcomes) reported outcomes related to assessor blinding; negative effect sizes as a result of assessor blinding were observed for 2/3 outcomes. Negative effect sizes included outcomes in favor of the unblinded assessor. One study reported outcomes related to assessor training, with an effect size of 0 when comparing results to waitlist controls.

Types of outcome measures were frequently language- and literacy-based consisting mostly of children (e.g., 20/23 study outcomes). The three outcome measures unrelated to speech and language were more subjective and also consisted of adult participants, including: medical examiner performance (Cook, et al., 2009), psychological well-being (Westbrook, Sedgwick-Taylor, Bennett-Levy, Butler, & McManus, 2008), and movement, cognition, and activities of daily living for individuals with Parkinson's disease (Ulm & Schüler, 1999) (Figure 3). Due to their more subjective nature, these outcomes were only applied to the Monte Carlo simulations of participant outcomes as measured by the WAB. The remaining 20 outcome measures were applied to both simulated outcomes as measured by the WAB and BNT. Results of simulations with applied effect sizes are further described between pages 23-28.

Figure 3.

Effect Sizes for Change Scores Related to Fidelity Dimensions

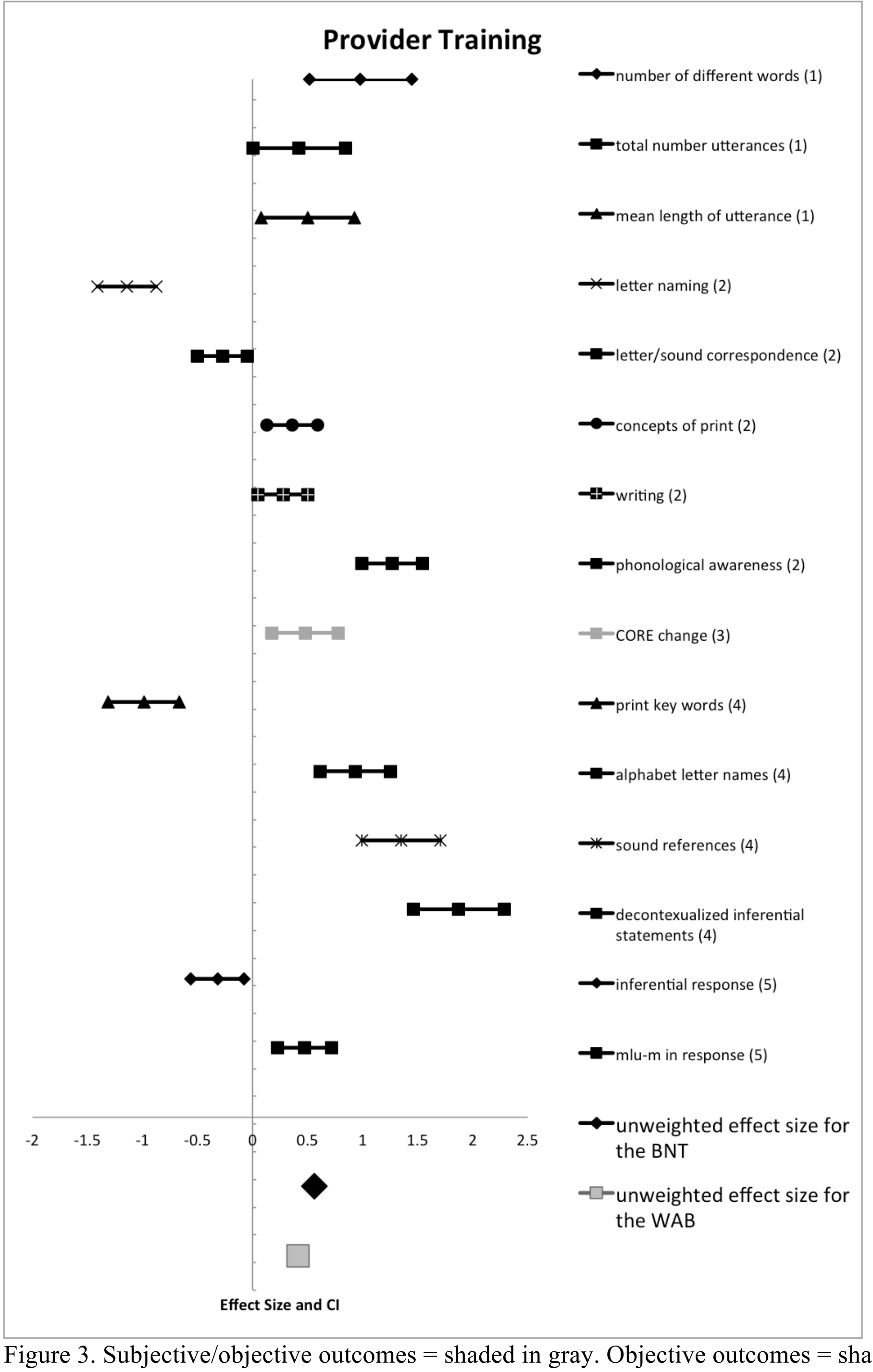


Figure 3. Subjective/objective outcomes = shaded in gray. Objective outcomes = shaded in black. CI = confidence interval. CORE = Clinical Outcome in Routine Evaluation. mlu-m = mean length of utterance in morphemes. (1) Piasta, et al. (2012); (2) Milburn, et al. (2015); (3) Westbrook, et al. (2008) (4) Girolametto, et al. (2012); (5) Rezzonico, et al. 2015.

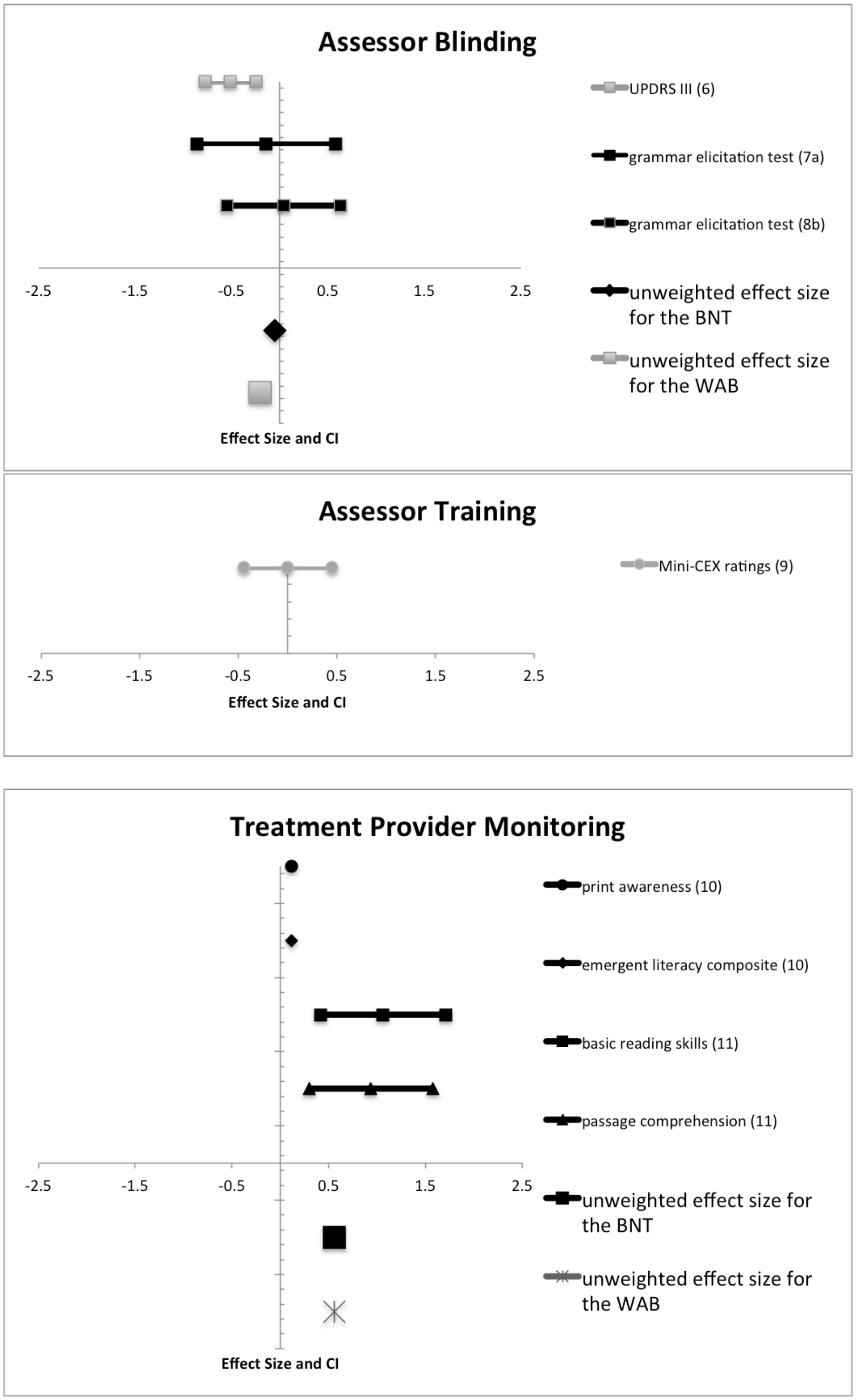


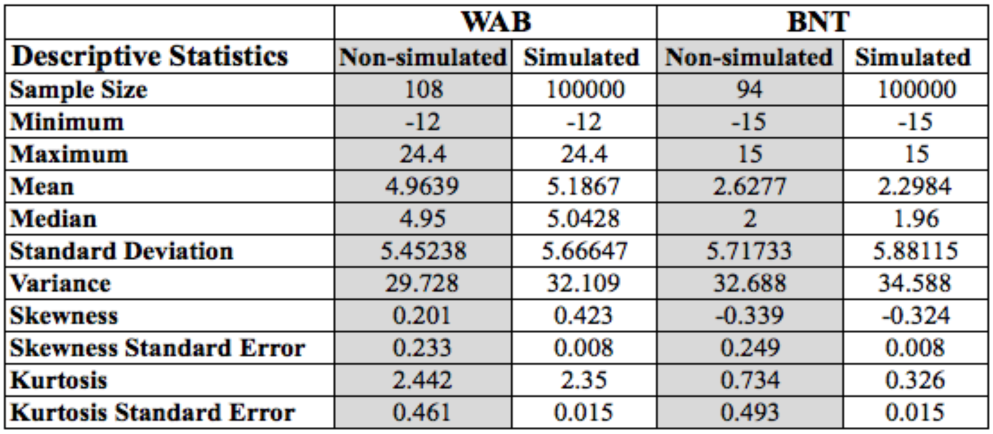
Figure 3. Subjective/objective outcomes = shaded in gray. Objective outcomes = shaded in black. CI = confidence interval. UPDRS III = Unified Parkinson's Disease Rating Scale III. Mini-CEX = Mini-Clinical Evaluation Exercise for Trainees. (6) Ulm & Schüler (1999); (7) Smith-Lock, et al. (2013a); (8) Smith-Lock, et al. (2013b); (9) Cook, et al. (2009); (10) Hamre, et al. (2010); (11) Benner, et al. (2011).

Simulated WAB and BNT Change Scores

Individual participant WAB change scores extracted from aphasia treatment studies (*M* = 4.96, *SD* = 5.45) were comparable to simulated data (*M* = 5.19, *SD* = 5.67) (Table 1; Figure 4). Individual participant BNT change scores extracted from aphasia treatment studies (*M* = 2.63, *SD* = 5.71) were comparable to simulated data (*M* = 2.3, *SD* = 5.88), with a slight decrease in gain scores and increase in standard deviations for the simulated data sets (Table 1; Figure 5). Skewness and kurtosis were similar for both real and simulated conditions as confirmed by visual inspection and statistical analysis.

Table 1

Descriptive Statistics of Non-simulated and Simulated WAB and BNT Change Scores



	WAB		BNT	
Descriptive Statistics	**Non-simulated**	**Simulated**	**Non-simulated**	**Simulated**
Sample Size	108	100000	94	100000
Minimum	-12	-12	-15	-15
Maximum	24.4	24.4	15	15
Mean	4.9639	5.1867	2.6277	2.2984
Median	4.95	5.0428	2	1.96
Standard Deviation	5.45238	5.66647	5.71733	5.88115
Variance	29.728	32.109	32.688	34.588
Skewness	0.201	0.423	-0.339	-0.324
Skewness Standard Error	0.233	0.008	0.249	0.008
Kurtosis	2.442	2.35	0.734	0.326
Kurtosis Standard Error	0.461	0.015	0.493	0.015

Figure 4.

Real Participant and Simulated WAB-AQ Change Scores

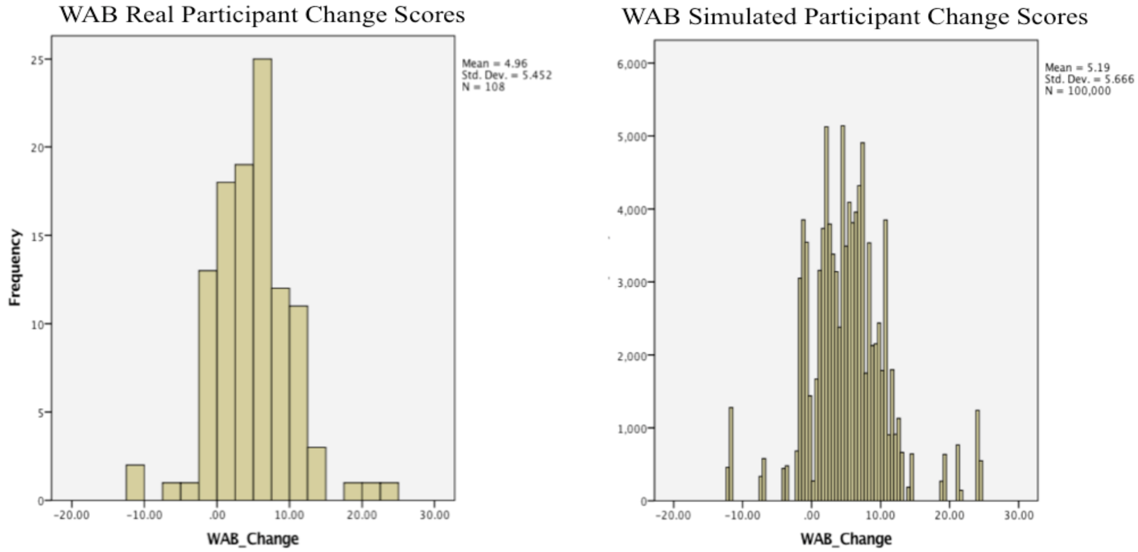
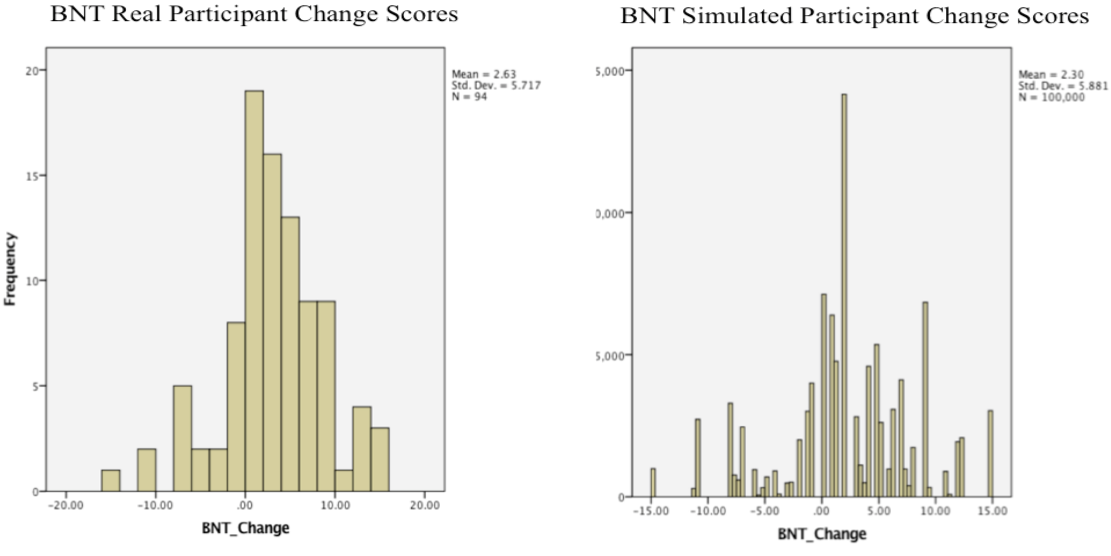


Figure 5.

Real Participant and Simulated BNT Change Scores



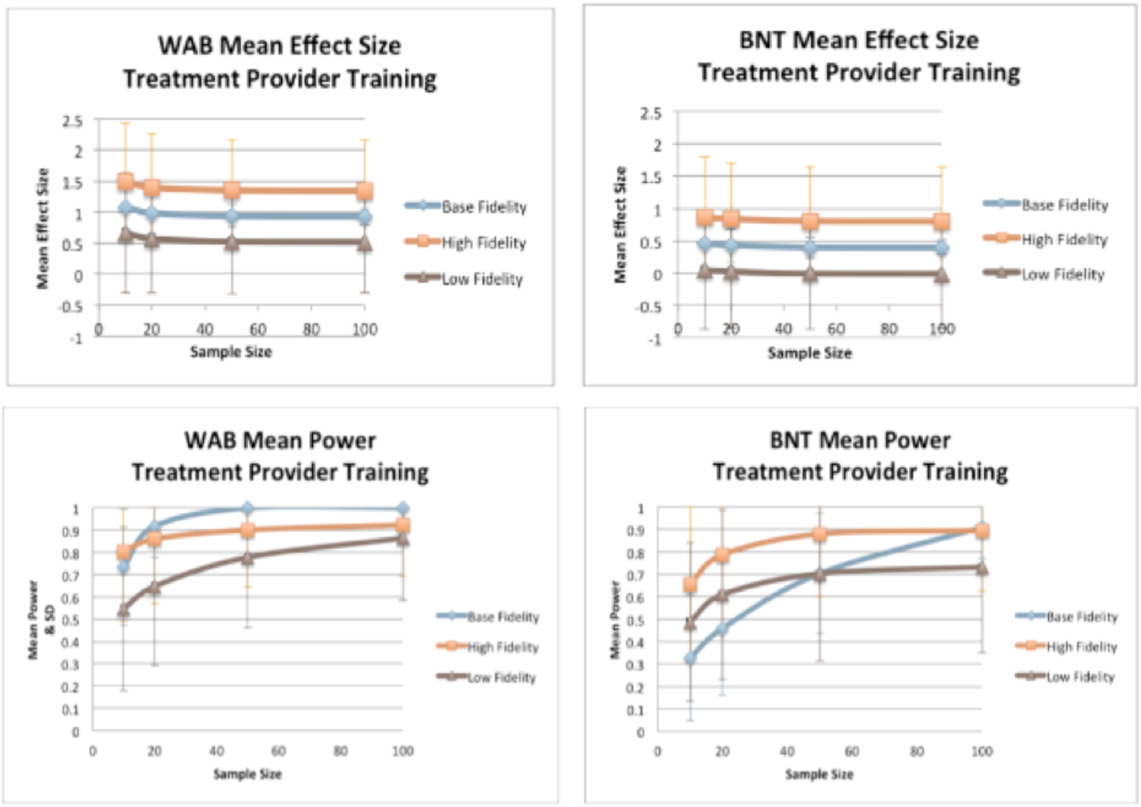
Simulated Trials with WAB and BNT Participant Change Scores

Baseline participant WAB and BNT change scores across simulated study trials of increasing sample size revealed decreases in the following: 1) average effect sizes, 2) range of results for simulated study averages, and 3) standard deviations for simulated study averages (Appendix B; Appendix D; Appendix E). From a sample size of 10 to 100, mean effect sizes (with standard deviations in parentheses) for the WAB changed between 1.08 (0.50) and 0.93 (1.3), while mean effect sizes for BNT change scores were lower, changing from 0.46 (0.40) to 0.40 (0.11), respectively. The lowest sample size, n=10, resulted in the largest range of possible effect sizes for both WAB (between -0.04 to 3.879) and BNT (between -0.54 and 2.44) outcomes.

For both simulated trials using WAB and BNT change scores, an increase in sample size directly correlated to an increase in power to detect effects (Appendix C, Appendix F), with an inverse relationship to range and standard deviations (i.e., higher sample sizes experienced reduced variance around mean power and higher power overall). From a sample size of 10 to 100, trials including WAB change scores resulted in mean power and standard deviation between 0.73 (0.26) and 0.99 (.0000014); trials with BNT change scores resulted in a comparably lower and wider range of mean power between 0.33(0.28) and 0.91(0.14). For the base simulation, all sample sizes except for *n*=10 for the WAB met and exceeded common standards for power (where adequate power = .80). BNT outcomes did not meet standards until a sample size of 100 was reached.

Figure 6.

Comparison of Treatment Provider Training Effect Size and Power at Base, High, and Low Fidelity Conditions for WAB and BNT Outcomes



Effect Size and Power Simulation with Treatment Provider Training

Simulated trials with high fidelity treatment provider training conditions resulted in an overall higher mean effect size compared to base simulation (reported above) and low fidelity conditions. At high fidelity levels of treatment provider training, large to very large mean effect sizes were observed compared to small and large mean effect sizes at base simulation (using descriptors by Sawilowsky, 2009) (Figure 6; Appendix B). Very small and medium mean effect sizes were found with low fidelity.

With high fidelity effect sizes applied, WAB outcomes met and exceeded standards for power for all sample sizes, including *n*=10 which did not have satisfactory power in base simulation. A sample size of 50 was necessary to meet standards for BNT outcomes, compared to *n*=100 for base simulation. While power increased with sample size at low fidelity conditions also, good power standards were met with a sample size of 100 for WAB outcomes only.

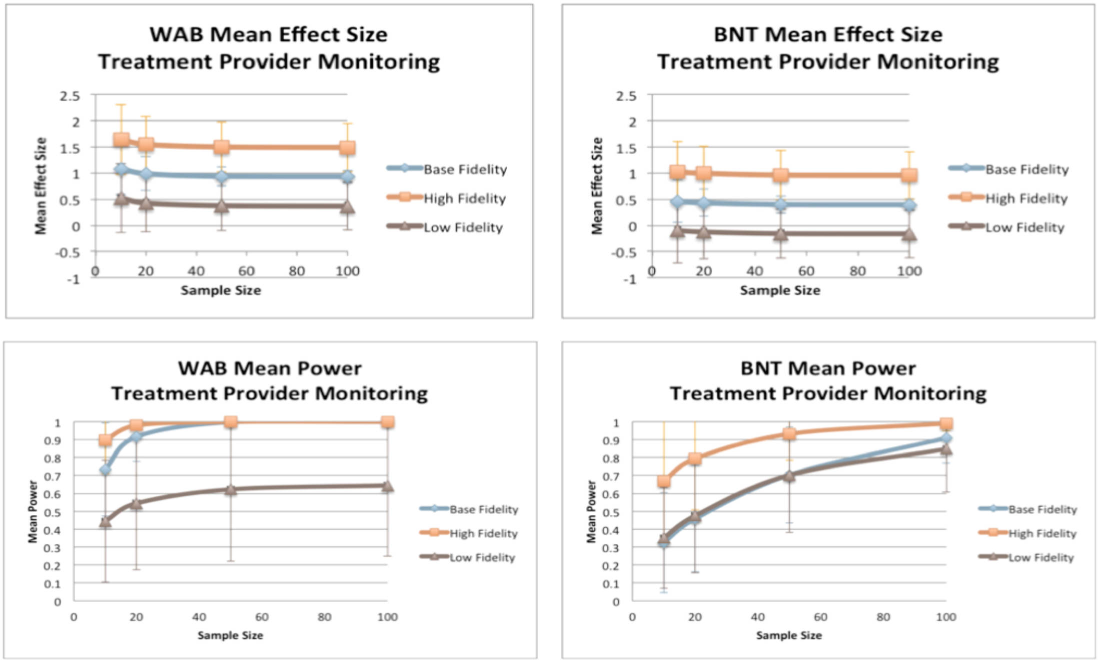
Effect Size and Power Simulation with Treatment Fidelity Monitoring

With high fidelity treatment provider monitoring, large to very large effects were observed compared to small and large effects at base simulation (Figure 7; Appendix B). Low fidelity simulations were characterized by small negative mean effect sizes for BNT outcomes and medium mean effect sizes for WAB outcomes.

Mean power in high treatment fidelity monitoring conditions met and exceeded standards (power = .80) for WAB outcomes, reaching a mean power of 1 with a sample size of 100; with BNT outcomes, a sample size of 20 and above was necessary to closely approximate and exceed standards (Figure 7; Appendix C). At low fidelity, power standards were not met at any sample size for WAB outcomes; a sample size of 100 with BNT outcomes yielded acceptable power standards.

Figure 7.

Comparison of Treatment Provider Monitoring Effect Size and Power at Base, High, and Low Fidelity Conditions for WAB and BNT Outcomes



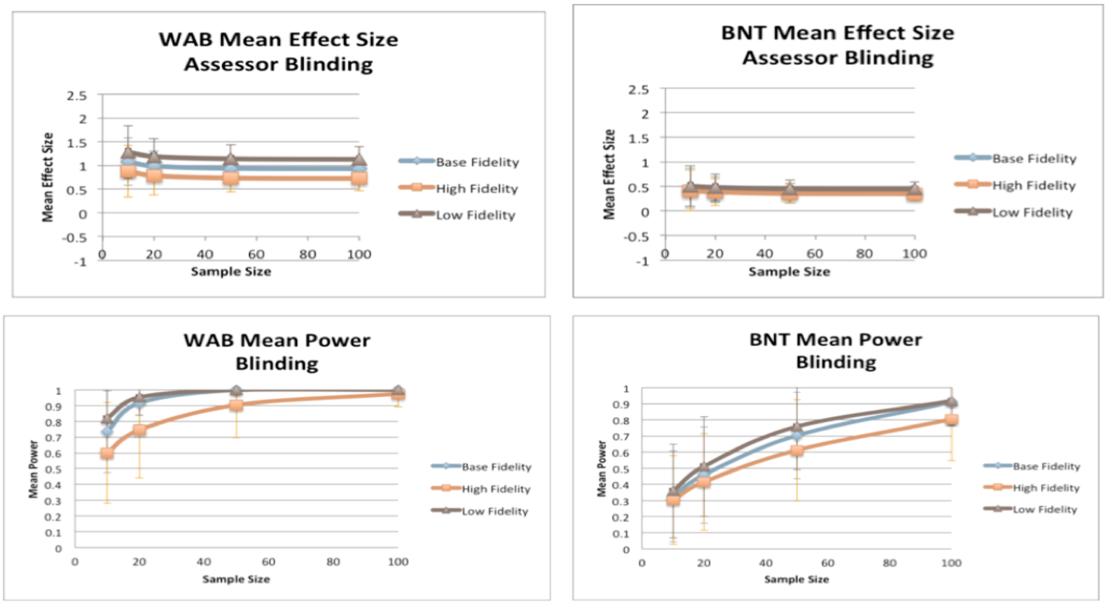
Effect Size and Power Simulation with Assessor Blinding

With assessor blinding, direction of results differed from treatment provider training and monitoring, in that high fidelity conditions experienced decreased outcomes compared to base and low fidelity conditions (Figure 8; Appendix B). At high fidelity assessor blinding for WAB outcomes, medium-to-large mean effect sizes were observed, compared to large effects at base fidelity, and large-to-very large mean effect sizes at low fidelity. All BNT mean effect size outcomes at high, base, and low fidelity conditions were small, except one medium mean effect size observed at low fidelity conditions with a sample size of 10. Differences may be due to the characteristics of outcomes extracted from the health and behavioral science literature that were matched to the WAB and BNT outcome simulations for their subjective and objective qualities (Figure 3). Fidelity outcomes applied to WAB change scores included 1 subjective outcome, where blinding seemed to be more influential, and 2 objective outcomes. In contrast, only the 2 objective fidelity outcomes, which seemed to be less influenced by blinding, were applied to the simulated BNT change scores. For all assessor blinding fidelity conditions, variance around the mean effect size reduced as sample size increased.

With high fidelity assessor blinding effect sizes applied, WAB outcomes did not meet standards of power for sample sizes below 20, compared to low fidelity simulation where standards were met at all sample sizes (Figure 8; Appendix C). Regardless of high, base, or low fidelity, BNT outcomes did not meet standards until a sample size of 100 was observed.

Figure 8.

Comparison of Assessor Blinding Effect Size and Power at Base, High, and Low Fidelity Conditions for WAB and BNT Outcomes

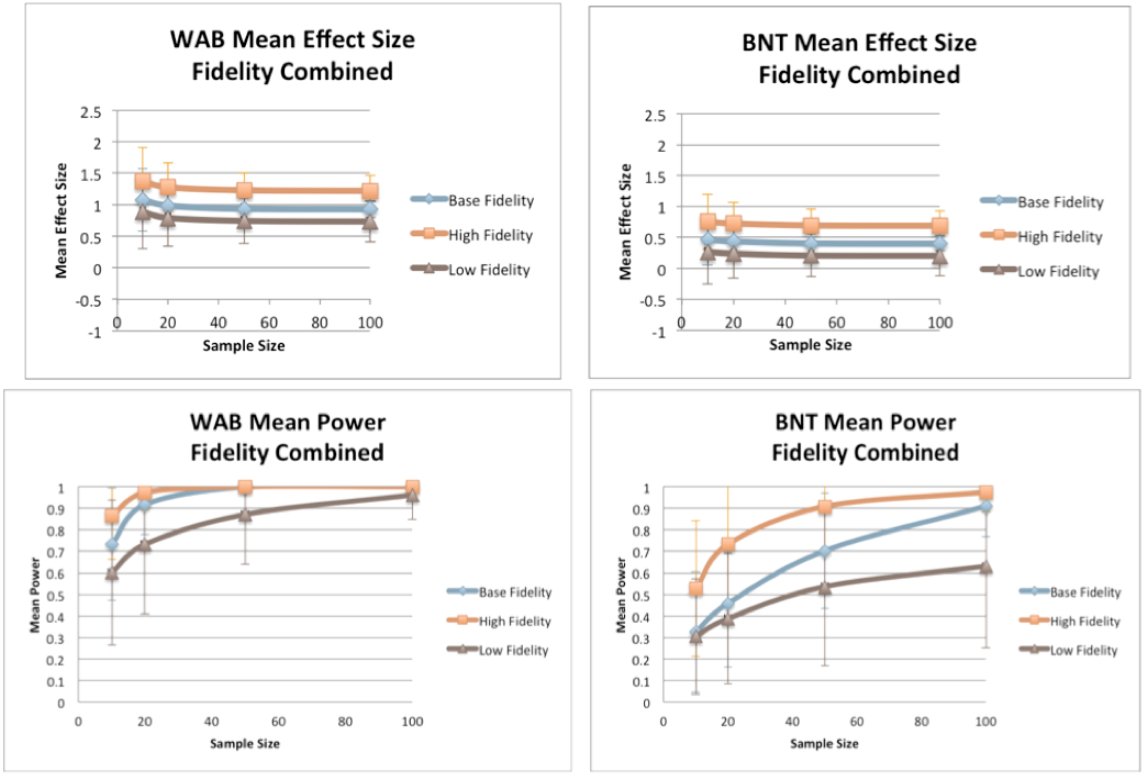


Effect Size and Power Simulation with Assessor Training

Effect sizes and power simulations were not performed with assessor training data, as the one trial able to meet inclusion criteria for assessor training reported no difference in ratings between trained and untrained conditions. With an effect size of zero, results are equal to that of the base simulation (Appendix B).

Figure 9.

Comparison of Combined Fidelity Effect Size and Power at Base, High, and Low Fidelity Conditions for WAB and BNT Outcomes



Effect Size and Power Simulation with Combined Effect Sizes

To evaluate the combined influence of fidelity, the unweighted effect sizes within each fidelity domain were averaged so that the aggregate effect sizes could be combined. For WAB outcomes, these were 0.413, 0.5596, -0.20173, and 0; for BNT outcomes, these were 0.4084, 0.5596, -0.0473, and 0 (Figure 3).

With combined high fidelity effect sizes, medium and very large effects were observed compared to small and large effects at base simulation (Figure 9; Appendix B). Low fidelity simulations were characterized by small effects for BNT outcomes and medium to large effects for WAB outcomes.

With high fidelity combined effect sizes applied, WAB outcomes met and exceeded standards for power for all sample sizes, including *n* = 10 which did not reach standards at base fidelity levels. A sample size of 50 and above was necessary to reach standard power for BNT outcomes at high fidelity conditions, compared to base fidelity conditions requiring *n* = 100 (Figure 9; Appendix C). At low fidelity, power increased with sample size, but BNT outcomes did not meet power standards.

Discussion

The importance of fidelity is often overlooked, and the impact of assessor- and treatment provider- related noise can conceal the true connection between treatment and outcomes. This study sought to examine the relationships between fidelity measures, sample size, treatment effect sizes, and power to detect treatment effects for individuals with aphasia. This simulation was the first of its kind to synthesize measurable data related to fidelity from the health and behavioral science literature into effect sizes, apply effect sizes to simulated aphasia treatment data, and simulate various levels of fidelity and sample size to determine impact on treatment outcomes. Simulation results for 3 out of 4 fidelity dimensions - treatment provider training, treatment provider monitoring, and assessor blinding - suggest that low and high fidelity levels influence treatment outcomes in the form of effect sizes and power to detect effects. Across all conditions, mean effect sizes and related variance decreased and power increased as sample size increased as expected.

Some of the biggest differences between simulated levels of fidelity that seemed to impact change in treatment outcomes for the studies using the WAB included treatment provider training and monitoring. As a result of low fidelity, low power to detect effects was found for both dimensions, while high fidelity conditions resulted in meeting and exceeding power standards at all sample sizes. Compared to large effects found at base fidelity, both treatment provider training and monitoring resulted in very large effects at high fidelity levels, and small-to-medium effects at low fidelity levels.

Some of the biggest differences between simulated levels of fidelity that seemed to impact change in treatment outcomes for the studies using the BNT included combined fidelity, treatment provider training, and treatment provider monitoring. As a result of low fidelity, low power to detect effects was found for all three dimensions except at the highest sample size for treatment provider monitoring. High fidelity conditions resulted in meeting and exceeding power standards for sample sizes of 50 and 100. Compared to small effects found at base fidelity, both treatment provider training and monitoring resulted in medium-to-large effects at high fidelity levels, and negative-to-small effects at low fidelity levels.

Meta-analysis and visual inspection of forest plots for each fidelity domain's effect sizes contributed to and thus predicted the outcomes of simulations. The results of the simulation also revealed changes related to sample size that may have been otherwise concealed in a meta-analysis. Study trials with smaller sample sizes were less able to detect change in outcomes depending upon the type of assessment used. This informs our interpretation of results in two other ways. First, the nature of the assessment instrument, and the behavior under scrutiny, matters. Many of the real treatment outcomes measured by the multidimensional WAB and the unidimensional BNT change scores were extracted from the same participants, yet the descriptive statistics of the original change scores and simulated outcomes related to each assessment were quantitatively different. Second, small but perhaps meaningful change scores, as exemplified in BNT outcomes, run the risk of poor detection when a singular behavior (e.g., naming) is assessed and when sample sizes are small. This is particularly relevant since treatment studies in speech-language pathology often rely upon small sample sizes for many valid reasons (e.g., funding, participant pool, length of treatment, transportation barriers).

The direction and amount of influence differed across fidelity dimensions. For example, high fidelity efforts achieved through blinded assessors were more likely to result in reduced effect sizes and power to detect effects, and objective outcome measures were less impacted by lack of blinding than combination subjective-objective measures. Conversely, higher fidelity of treatment provider training and treatment provider monitoring resulted in increased effect sizes and power to detect effects than poor fidelity conditions.

Treatment Provider Training

Of all fidelity dimensions reviewed, studies most frequently reported the benefits of treatment provider training. Possibly as a result of multiple outcome measures in relation to training, effect sizes reported ranged from large to small and negative, which highlighted responses to specific core components of treatment as a result of training (Girolametto, Weitzman, & Greenberg, 2012; Milburn, et al., 2015; Rezzonico, et al., 2015), but did not detract from the positive effects observed overall. As exemplified in our simulations, treatment provider training for intervention studies made the difference between no-to-small effects and large effects. While increasing sample size may have improved the likelihood of detecting effects regardless of treatment provider training, the required increase is likely not attainable for most researchers. In the face of poor treatment provider training, a sample size of 50 may be necessary to approximate power of 0.8 for studies including the WAB and BNT, while high quality training may achieve the same power with 10 to 20 participants. Depending on the outcome assessment, effect sizes with low fidelity levels can be nonexistent or small, regardless of sample size. Treatments with good provider training and large sample sizes but low effect sizes may benefit from post-hoc analyses to determine the most and least active treatment specific components.

Recommendation. Training for treatment providers varied by type (e.g., coaching sessions, workshops, and case study discussion) and amount (e.g., 5 surplus coaching sessions, 20 hour workshop, and a 10 week training course). Regardless of variety, training seemed to be effective for increasing effect sizes and improving ability to detect effects. Because of this probable effectiveness, and because training is a relatively simple aspect of treatment fidelity to implement, it is recommended that researchers systematically provide provider training and report operational details. Future trials should identify core training elements and core outcomes to allow precise measurement of relationship between training and patient outcomes.

Treatment Provider Monitoring

Studies meeting inclusion criteria for effect sizes used in treatment provider monitoring simulations consisted of 2 literacy-based interventions for at-risk preschoolers (Hamre et al., 2010) and students identified with reading difficulties (Benner, et al., 2011). Effect sizes extracted from both studies indicated that treatment provider monitoring was positively associated with student reading outcomes. As a result, high fidelity levels resulted in large to very large effects, and the highest power to detect effects of all dimensions at low sample sizes. These simulation results should be interpreted with caution, as further exploration of studies across the health and behavioral sciences that did not meet inclusion criteria suggest that the relationship between outcomes and treatment adherence, a measure of treatment provider monitoring, may not be straightforward and may vary according to field of study and nature of intervention.

Associations between high treatment fidelity and positive participant outcomes were common in the applied behavioral analysis literature (Arkoosh et al., 2007; Carroll, Kodak & Fisher, 2013; DiGenarro Reed, Reed, Baez, & Maguire, 2011; Groskreutz, Groskreutz, & Higbee, 2011; Jenkins, Hirst, & Reed, 2015; Pence & Peter, 2015), and related behavioral interventions (Villodas, McBurnett, Kaiser, Rooney, & Pfiffner, 2014), where increased provider adherence to treatment components was associated with an increase in the target behavior(s). In other fields such as psychotherapy, more complex relationships between treatment outcomes, therapist adherence, experience, alliance, and client severity are thought to exist (Tschuschke et al., 2015). Studies reporting variable or minimal-to-no change related to provider adherence included two psychotherapy trials, whose treatment outcomes were related to therapist experience and patient level of severity, known moderators of provider adherence (Tschuschke et al., 2015; Webb et al., 2012).

Recommendation. Treatment provider monitoring has been accomplished via several methods (e.g., self and observer report, fidelity checklists, performance feedback, and video observation) and certain aspects of monitoring may reduce bias, maintain adherence over time, or reveal changes to treatment protocols that enhanced or reduced effectiveness (Benner et al., 2011; Hamre et al., 2010; Lillehoj, Griffin, & Spoth, 2004; Perepletchikova & Kazdin, 2005; Reinke, Lewis-Palmer, & Merrell, 2008). While results are mixed across the health and behavioral sciences literature, the studies used to guide our simulations and our subsequent simulation results support the claim that monitoring treatment adherence in language interventions is related to improved outcomes. It is recommended that accurate descriptions and measures of the methods used for fidelity monitoring be included. Descriptions of potentially related factors (i.e., patient, provider, and program characteristics) are recommended to better understand barriers to treatment provider adherence (Perepletchikova & Kazdin, 2005).

Assessor Blinding

Lack of blinding assessors can influence treatment results, often in the form of inflated outcomes for participants. This study included trials with subjective rating measures (Ulm & Schüler, 1999) related to cognitive, behavioral, and movement-related presentations of Parkinson's disease as well as more objective criterion-based measures of grammatical structures for children with specific language impairments (Smith-Lock, et al., 2013a; Smith-Lock, Leitao, Lambert, & Nickels, 2013b). The current simulation findings support results from other studies that subjective outcome assessments, may be most exposed to bias compared to objective outcome assessments (Wood et al., 2008), trials with assessments that are more objective in nature may still contain bias (Liu, LaValley, & Latham, 2011).

Most surprising to the investigators was the lack of minable data from studies reporting blinding outcomes in order to fit this study's parameters. When reviewing trials in the Hrobjartsson et al. (2013) meta-analysis for information most similar to aphasia treatment outcomes, only one study (Ulm & Schüler, 1999) included accessible pre-treatment assessment information related to blinding. Some studies reported that there was no difference between blinded and unblinded assessors, but did not include data to support claims (e.g., Tewuerbati et al., 2015).

Blinding can impact other design components beyond pre- and post-treatment outcomes. Pressure for unblinded assessors to ensure high numbers of participants fit inclusion criteria for a higher severity may not only bias results in favor of the experimental group, but the control group as well, effectively washing the results and incorrectly determining a responder status for both groups (Kobak et al., 2010). Even waitlist outcomes are at risk. As discussed in a meta-analysis (Steinert, Stadter, Stark, & Leichsenring, 2016), participants evaluated by unblinded assessors demonstrated less change during waitlist period compared to blinded assessors, though this should be interpreted with caution due to a small sample size (blinded = 5; nonblinded = 3).

Recommendation. Consistent with recommendations in speech language pathology research, risk of bias should be limited by blinding assessors, especially with subjective measures (Ebbels, 2017). Studies should explicitly report who is blinded and unblinded due to open interpretation of labels such as ‘double-blind’ (Haahr & Hróbjartsson, 2006) and whether blinding was maintained before enrollment and during the waitlist period (Kobak, Kane, Thase, & Nierenberg, 2007; Steinert et al., 2016) as well as throughout the study (Bennett, Hughes, & Johnson, 2011). In studies where a blinded assessor is no longer available or an assessor has become unblinded, results of the difference in outcomes between the two conditions should be included (Smith-Lock, et al., 2013a; Smith-Lock, et al., 2013b), or a second rater blinded to condition may be necessary to code a majority of the assessment information (Pennington, Goldbart, & Marshall, 2004). All of the above solutions are likely to help reduce participant variance and inflated or washed results.

Assessor Training

Studies reviewed consistently identified assessor training as important for high assessment fidelity, but the effects of assessor training are unclear at this point. One trial fit inclusion criteria for its effect size to be used in this simulation study showing no difference in overall accuracy and inter-rater reliability between training and no training conditions (Cook, Dupras, Beckman, Thomas, & Pankratz, 2009). Several studies were not included in this simulation due to the nature of the assessment (e.g., depression scale) or incompatible data (e.g., lack of information to determine direction of change in ratings impacted by the training).

Assessor experience and/or qualifications may moderate the influence of study-specific assessor training. Experience may influence the need for reliability training, as not all assessors may require training nor may some meet prerequisite standards despite training due to lack of experience (Cook, et al., 2009; Kobak, Lipsitz, Williams, Engelhardt, & Bellew, 2005; Stitt, Simonds, & Hunt, 2003). Assessor training may level out the playing field for novices, but leveling may not occur quickly or at all for every assessor (Hansen, Elholm Madsen, & Sørensen, 2016). Assessor training may improve the precision and reliability of administration and scoring. For example, with training, ratings of students' communicative performance were more stringent overall and resulted in improved inter-rater reliability (Stitt et al., 2003). Demonstrating the positive effects of training sessions on inter-rater reliability, Müller & Szegedi (2002), applied results from previous studies to calculate both power as a function of reliability and sample size necessary to compensate for low reliability and reach standard power. Müller & Szegedi's (2002) study suggested that 3 to 5 training sessions adequately met study needs for inter-rater reliability to demonstrate group difference and that false negatives in studies may be due to low reliability in psychopharmacology trials examined.

Recommendation. Differences in training program qualities may include: method of delivery (e.g., live or online), components of training (e.g., review of criteria, behavioral observations of video performance, common assessor errors) and intensity or amount of time devoted to training (e.g., number of training sessions, half or full day workshops) (Cook, et al., 2009; Stitt, Simonds, & Hunt, 2003). It is recommended that details of assessor training as well as assessor characteristics (e.g., values and experience) are reported, and the efficacy of assessor training further explored. As with treatment provider training, assessor training is likely a relatively simple fidelity domain to implement that may have a positive and substantial trade-off for effect sizes and power. To further simplify such processes, web-based rater training has been studied as an alternative to in-person training, though training may not be sufficient to generalize knowledge into applied performance, and results vary by field (Elder, Barkhuizen, Knoch, & Von Randow, 2007; Kobak et al., 2005; Rosen, et al., 2008). Supplemental "live" and applied trainings are recommended until this is further researched.

Assessor Errors

Due to difficulty analyzing direction of impact, the influence of rater errors was not included in this simulation. However, we decided to discuss it here because the impact of assessor errors is of high concern, particularly in high stakes situations concerning incorrect diagnosis or treatment/placement decisions for an individual. For example, 91% of test packets from an oral reading fluency trial (Reed & Sturges, 2013) had at least one correctable error and 8% of test packets were administered in such a way that they were rendered insufficient for inclusion. A trial by Loe, Kadlubek, & Marks (2007) was the only study found to report direction of scoring errors on an intelligence scale by school psychology graduate students, with an average of 5 points higher and 8 points lower than expected of the true score. Moreover, rater errors may be reduced with training but not fully resolved (Platt, Zachar, Ray, Underhill, & LoBello, 2007; Reed & Sturges, 2013), and the complexity of an assessment may incrementally increase likelihood of errors (Charter, Walden, & Padilla, 2000).

Recommendation. Assessor errors differ in scoring (e.g., addition, transfer of scores, conversion, and plotting) and administration (e.g., unnecessary cues and prompts, or missing instructions), and these errors can often be avoided (Richardson, Dalton, Shafer, & Patterson, 2016). It is recommended that planning stages of a treatment study include predetermined rater qualifications, amount of expected training sessions to calibrate raters, and rater testing criterion to include assessors in a treatment study. Assessors should be familiar with common errors, and score assessments twice or investigators should ensure an additional assessor rescores items. The impact of assessor errors on treatment outcomes is less known and should be reported to better understand impacts and methods of remediation.

Recommendations for Aphasia Treatment Researchers with Limited Resources

Sample size increased power to detect effects across all fidelity dimensions, at high and low fidelity as well as base simulations. Trials with sample sizes below 20 were at most risk for low effect sizes and power. Increased sample size is not feasible for most researchers, particularly those in aphasia treatment research with limited resources and patient databases. Outcomes studies with sample sizes below 20 will have the highest chances of success to detect and report high effects if a multidimensional assessment is supported with high quality treatment provider training and/or monitoring. Bias in more subjective outcomes should be reduced with blinding to increase the chance of reporting true effects.

Limitations and Future Directions

As with all meta-analyses, the treatment data used to generate the base simulations as well as the treatment-related effect sizes used to manipulate fidelity is likely influenced by reporting bias and/or the "file drawer problem" (Borenstein et al., 2009). We only have access to published findings, which are likely to have larger effects, not those findings that were deliberately suppressed or those that were banished to a file drawer because of little to no effects. Therefore, our estimates may indeed be overestimates of reality.

Further, the scope of this study and lack of easily extractable data and consistently labeled information regarding fidelity measures made it difficult to obtain high quantities and quality of information closely related to aphasia treatments. This has been described as the "apples and oranges" criticism (Borenstein et al., 2009), and until information is reported upon more consistently in speech-language and aphasia literature, we should utilize caution when interpreting these findings. Attempts to include studies with adult participants who received speech and language interventions were restricted due to limited data for this population. Treatment outcomes including fidelity appear to be biased towards younger participants who may react to type and amount of fidelity differently than their older counterparts. Attempts were made to include interventions and assessments most representative of treatment and diagnostic options for individuals with aphasia as possible, but it is difficult to determine how different the reported gains may be with individuals post-stroke vs. populations without an acquired speech or language disorder. Future simulations including more fidelity information specific to older populations and individuals commonly served by speech-language pathologists are recommended.

Studies analyzed included an inevitable variance in the type and amount of provider training, as well as intervention adherence amongst providers, and types of assessments used for blinded vs. nonblinded assessors (e.g., subjective Likert scales vs. discrete rate of behaviors). The studies included were analyzed for relative levels of high and low fidelity, as opposed to pre-determined quantities and qualities of fidelity, and as such varied in the amount and type of fidelity included. With increased reports of fidelity and related measures, the influence of a fidelity dimension's distinct characteristics should be examined more closely.

Conclusions

In an important rehabilitation field that cannot afford the costs of research waste and in which it is difficult to recruit a high volume of participants, variability in the form of providers, assessors, and patients must be prevented and/or measured in order to draw stronger treatment conclusions that researchers and consumers have confidence in. This study and previous research suggest that fidelity guidelines and measures are a useful tool for more accurate effect sizes and power. Fidelity should be considered at all stages of a treatment study, including planning the design, and troubleshooting the potential sources of variance, or ineffective qualities, in a post-hoc manner.

Before fidelity factors were introduced, power to detect true effects was heavily influenced by sample size and difference in overall change scores by assessment type. Lower sample sizes increased variability of effect sizes calculated for each trial and reduced power to detect effects was observed. When high fidelity treatment provider monitoring and training were applied to simulated trials with small sample sizes, power and effect sizes increased. High fidelity assessor blinding resulted in deflated outcomes compared to base and low fidelity conditions. No observed difference was found for assessor training. Combined fidelity outcomes were observed to have less variance but also were slightly less influential on power and effect size than treatment provider training and monitoring alone, possibly due to inclusion of blinded assessors and rater training. Type of outcome assessment was also a strong moderator of treatment results. Results should be considered preliminary as type, amount, and field-specific reports of both assessment and treatment fidelity are not comprehensively reported in research studies. As more detailed information about the fidelity elements included in this study begin to emerge, we anticipate increased ability to interpret fidelity-specific components that are most resourceful to researchers for a particular treatment.

List of Appendices

Appendix A **43**

Appendix B **48**

Appendix C.. **50**

Appendix D **52**

Appendix E 53

Appendix F 54

Appendix G 55

Appendix A

References for Aphasia Treatment Studies Including Participant Outcomes Extracted

Archibald, L. M., Orange, J. B., & Jamieson, D. J. (2009). Implementation of computer-based language therapy in aphasia. *Therapeutic Advances in Neurological Disorders*, *2*(5), 299-311.

Ball, A. L., de Riesthal, M., Breeding, V. E., & Mendoza, D. E. (2011). Modified ACT and CART in severe aphasia. *Aphasiology*, *25*(6-7), 836-848.

Beeson, P. M., Higginson, K., & Rising, K. (2013). Writing treatment for aphasia: A texting approach. *Journal of Speech, Language, and Hearing Research*, *56*(3), 945-955.

Beeson, P. M., King, R. M., Bonakdarpour, B., Henry, M. L., Cho, H., & Rapcsak, S. Z. (2011). Positive effects of language treatment for the logopenic variant of primary progressive aphasia. *Journal of Molecular Neuroscience*, *45*(3), 724-736.

Boo, M., & Rose, M. L. (2011). The efficacy of repetition, semantic, and gesture treatments for verb retrieval and use in Broca's aphasia. *Aphasiology*, *25*(2), 154-175.

Breier, J. I., Juranek, J., & Papanicolaou, A. C. (2011). Changes in maps of language function and the integrity of the arcuate fasciculus after therapy for chronic aphasia. *Neurocase*, *17*(6), 506-517.

Breier, J. I., Maher, L. M., Novak, B., & Papanicolaou, A. C. (2006). Functional imaging before and after constraint-induced language therapy for aphasia using magnetoencephalography. *Neurocase*, *12*(6), 322-331.

Breier, J. I., Maher, L. M., Schmadeke, S., Hasan, K. M., & Papanicolaou, A. C. (2007).

Changes in language-specific brain activation after therapy for aphasia using magnetoencephalography: a case study. *Neurocase*, *13*(3), 169-177.

Cherney, L. R., Halper, A. S., Holland, A. L., & Cole, R. (2008). Computerized script training for aphasia: Preliminary results. *American Journal of Speech-Language Pathology*, *17*(1), 19-34.

Edmonds, L. A., & Babb, M. (2011). Effect of verb network strengthening treatment in moderate-to-severe aphasia. *American Journal of Speech-Language Pathology*, *20*(2), 131-145.

Edmonds, L. A., Mammino, K., & Ojeda, J. (2014). Effect of Verb Network Strengthening Treatment (VNeST) in persons with aphasia: Extension and replication of previous findings. *American Journal of Speech-Language Pathology*, *23*(2), S312-S329.

Estes, C., & Bloom, R. L. (2011). Using voice recognition software to treat dysgraphia in a patient with conduction aphasia. *Aphasiology*, *25*(3), 366-385.

Falconer, C., & Antonucci, S. M. (2012). Use of semantic feature analysis in group discourse treatment for aphasia: Extension and expansion. *Aphasiology*, *26*(1), 64-82.

Faroqi-Shah, Y., & Virion, C. R. (2009). Constraint-induced language therapy for agrammatism: Role of grammaticality constraints. *Aphasiology*, *23*(7-8), 977-988.

Ferguson, N. F., Evans, K., & Raymer, A. M. (2012). A comparison of intention and pantomime gesture treatment for noun retrieval in people with aphasia. *American Journal of Speech-Language Pathology*, *21*(2), S126-S139.

Furnas, D. W., & Edmonds, L. A. (2014). The effect of computerised Verb Network

Strengthening Treatment on lexical retrieval in aphasia. *Aphasiology*, *28*(4), 401-420.

Goff, R. A. (2013*). Examining the effectiveness of intensive language action therapy in individuals with nonfluent aphasia* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (Accession Order No. 3602296).

Hashimoto, N. (2012). The use of semantic-and phonological-based feature approaches to treat naming deficits in aphasia. *Clinical Linguistics & Phonetics*, *26*(6), 518-553.

Hashimoto, N., & Frome, A. (2011). The use of a modified semantic features analysis approach in aphasia. *Journal of Communication Disorders*, *44*(4), 459-469.

Hough, M. S. (2010). Melodic intonation therapy and aphasia: Another variation on a theme. *Aphasiology*, *24*(6-8), 775-786.

Kendall, D. L., Pompon, R. H., Brookshire, C. E., Minkina, I., & Bislick, L. (2013). An analysis of aphasic naming errors as an indicator of improved linguistic processing following phonomotor treatment. *American Journal of Speech-Language Pathology*, *22*(2), S240-S249.

Kendall, D. L., Rosenbek, J. C., Heilman, K. M., Conway, T., Klenberg, K., Rothi, L. J. G., & Nadeau, S. E. (2008). Phoneme-based rehabilitation of anomia in aphasia. *Brain and Language*, *105*(1), 1-17.

Kurland, J., Baldwin, K., & Tauer, C. (2010). Treatment-induced neuroplasticity following intensive naming therapy in a case of chronic Wernicke's aphasia. *Aphasiology*, *24*(6-8), 737-751.

Kurland, J., Pulvermüller, F., Silva, N., Burke, K., & Andrianopoulos, M. (2012). Constrained versus unconstrained intensive language therapy in two individuals with chronic, moderate-to-severe aphasia and apraxia of speech: behavioral and fMRI outcomes. *American Journal of Speech-Language Pathology*, *21*(2), S65-S87.

Kurland, J., Wilkins, A. R., & Stokes, P. (2014, February). iPractice: Piloting the effectiveness of a tablet-based home practice program in aphasia treatment. In *Seminars in Speech and Language, 35* (1), 51-64.

Lacey, E. H., Lott, S. N., Snider, S. F., Sperling, A., & Friedman, R. B. (2010). Multiple oral re-reading treatment for alexia: the parts may be greater than the whole. *Neuropsychological Rehabilitation*, *20*(4), 601-623.

Lott, S. N., Sperling, A. J., Watson, N. L., & Friedman, R. B. (2009). Repetition priming in oral text reading: a therapeutic strategy for phonologic text alexia. *Aphasiology*, *23*(6), 659-675.

Maher, L. M., Kendall, D., Swearengin, J. A., Rodriguez, A., Leon, S. A., Pingel, K., Holland, A., & Rothi, L. J. G. (2006). A pilot study of use-dependent learning in the context of constraint induced language therapy. *Journal of the International Neuropsychological Society*, *12*(6), 843-852.

Murray, L. L., & Karcher, L. (2000). A treatment for written verb retrieval and sentence construction skills. *Aphasiology*, *14*(5-6), 585-602.

Peach, R. K., & Reuter, K. A. (2010). A discourse-based approach to semantic feature analysis for the treatment of aphasic word retrieval failures. *Aphasiology*, *24*(9), 971-990.

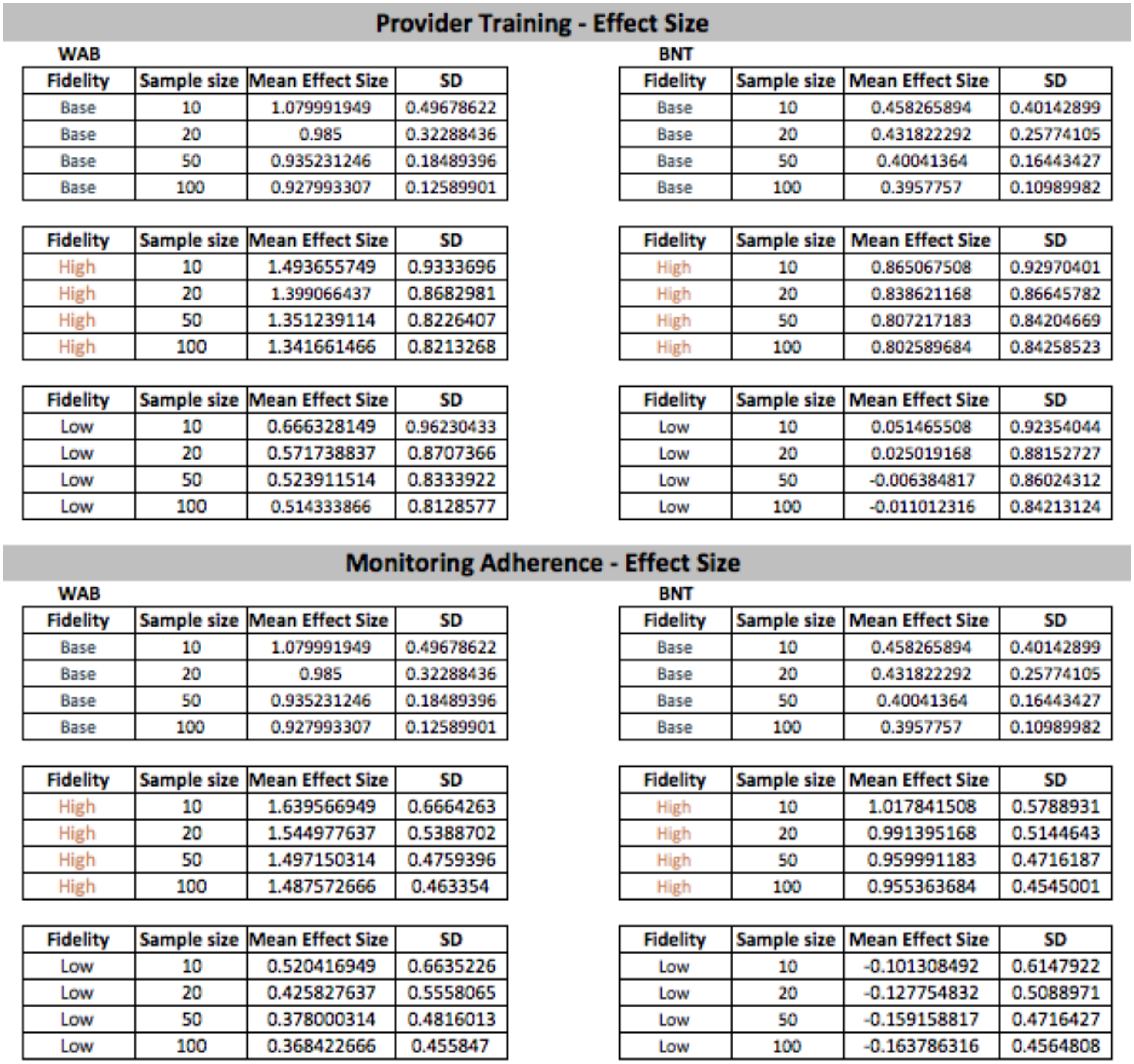
Raymer, A. M., McHose, B., Smith, K. G., Iman, L., Ambrose, A., & Casselton, C. (2012). Contrasting effects of errorless naming treatment and gestural facilitation for word retrieval in aphasia. *Neuropsychological Rehabilitation*, *22*(2), 235-266.

Rogalski, Y., Edmonds, L. A., Daly, V. R., & Gardner, M. J. (2013). Attentive reading and constrained summarisation (ARCS) discourse treatment for chronic Wernicke's aphasia. *Aphasiology*, *27*(10), 1232-1251.

Rose, M. L., Attard, M. C., Mok, Z., Lanyon, L. E., & Foster, A. M. (2013). Multi-modality aphasia therapy is as efficacious as a constraint-induced aphasia therapy for chronic aphasia: A phase 1 study. *Aphasiology*, *27*(8), 938-971.

Appendix B

Figure 1. Comparison of Mean Effect Size With Base, High, and Low Fidelity Conditions



Provider Training - Effect Size

WAB

Fidelity	Sample size	Mean Effect Size	SD
Base	10	1.079991949	0.49678622
Base	20	0.985	0.32288436
Base	50	0.935231246	0.18489396
Base	100	0.927993307	0.12589901

Fidelity	Sample size	Mean Effect Size	SD
High	10	1.493655749	0.9333696
High	20	1.399066437	0.8682981
High	50	1.351239114	0.8226407
High	100	1.341661466	0.8213268

Fidelity	Sample size	Mean Effect Size	SD
Low	10	0.666328149	0.96230433
Low	20	0.571738837	0.8707366
Low	50	0.523911514	0.8333922
Low	100	0.514333866	0.8128577

BNT

Fidelity	Sample size	Mean Effect Size	SD
Base	10	0.458265894	0.40142899
Base	20	0.431822292	0.25774105
Base	50	0.40041364	0.16443427
Base	100	0.3957757	0.10989982

Fidelity	Sample size	Mean Effect Size	SD
High	10	0.865067508	0.92970401
High	20	0.838621168	0.86645782
High	50	0.807217183	0.84204669
High	100	0.802589684	0.84258523

Fidelity	Sample size	Mean Effect Size	SD
Low	10	0.051465508	0.92354044
Low	20	0.025019168	0.88152727
Low	50	-0.006384817	0.86024312
Low	100	-0.011012316	0.84213124

Monitoring Adherence - Effect Size

WAB

Fidelity	Sample size	Mean Effect Size	SD
Base	10	1.079991949	0.49678622
Base	20	0.985	0.32288436
Base	50	0.935231246	0.18489396
Base	100	0.927993307	0.12589901

Fidelity	Sample size	Mean Effect Size	SD
High	10	1.639566949	0.6664263
High	20	1.544977637	0.5388702
High	50	1.497150314	0.4759396
High	100	1.487572666	0.463354

Fidelity	Sample size	Mean Effect Size	SD
Low	10	0.520416949	0.6635226
Low	20	0.425827637	0.5558065
Low	50	0.378000314	0.4816013
Low	100	0.368422666	0.455847

BNT

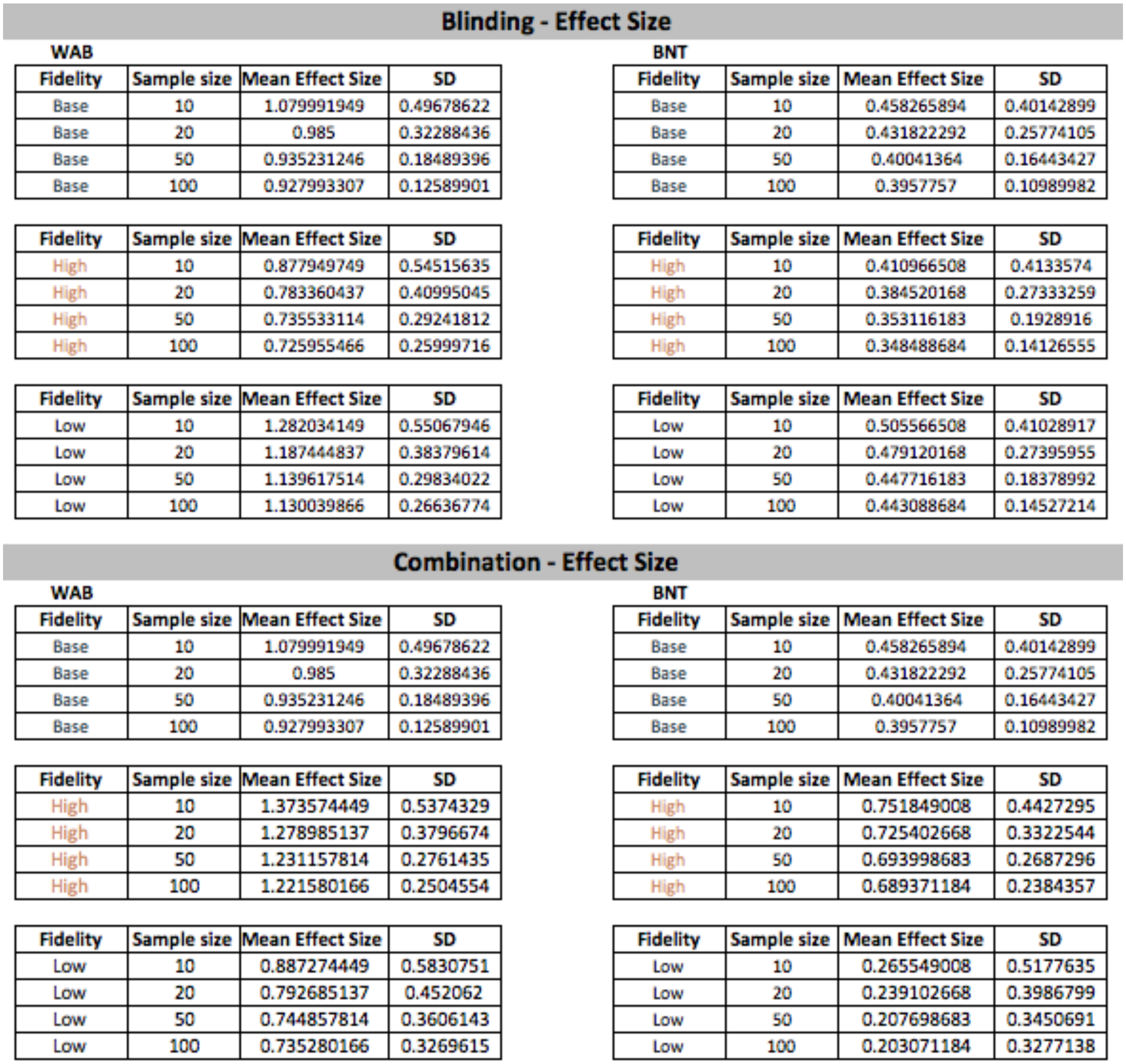
Fidelity	Sample size	Mean Effect Size	SD
Base	10	0.458265894	0.40142899
Base	20	0.431822292	0.25774105
Base	50	0.40041364	0.16443427
Base	100	0.3957757	0.10989982

Fidelity	Sample size	Mean Effect Size	SD
High	10	1.017841508	0.5788931
High	20	0.991395168	0.5144643
High	50	0.959991183	0.4716187
High	100	0.955363684	0.4545001

Fidelity	Sample size	Mean Effect Size	SD
Low	10	-0.101308492	0.6147922
Low	20	-0.127754832	0.5088971
Low	50	-0.159158817	0.4716427
Low	100	-0.163786316	0.4564808

Appendix B Continued

Figure 1. Comparison of Mean Effect Size With Base, High, and Low Fidelity Conditions



Blinding - Effect Size

WAB

Fidelity	Sample size	Mean Effect Size	SD
Base	10	1.079991949	0.49678622
Base	20	0.985	0.32288436
Base	50	0.935231246	0.18489396
Base	100	0.927993307	0.12589901

Fidelity	Sample size	Mean Effect Size	SD
High	10	0.877949749	0.54515635
High	20	0.783360437	0.40995045
High	50	0.735533114	0.29241812
High	100	0.725955466	0.25999716

Fidelity	Sample size	Mean Effect Size	SD
Low	10	1.282034149	0.55067946
Low	20	1.187444837	0.38379614
Low	50	1.139617514	0.29834022
Low	100	1.130039866	0.26636774

BNT

Fidelity	Sample size	Mean Effect Size	SD
Base	10	0.458265894	0.40142899
Base	20	0.431822292	0.25774105
Base	50	0.40041364	0.16443427
Base	100	0.3957757	0.10989982

Fidelity	Sample size	Mean Effect Size	SD
High	10	0.410966508	0.4133574
High	20	0.384520168	0.27333259
High	50	0.353116183	0.1928916
High	100	0.348488684	0.14126555

Fidelity	Sample size	Mean Effect Size	SD
Low	10	0.505566508	0.41028917
Low	20	0.479120168	0.27395955
Low	50	0.447716183	0.18378992
Low	100	0.443088684	0.14527214

Combination - Effect Size

WAB

Fidelity	Sample size	Mean Effect Size	SD
Base	10	1.079991949	0.49678622
Base	20	0.985	0.32288436
Base	50	0.935231246	0.18489396
Base	100	0.927993307	0.12589901

Fidelity	Sample size	Mean Effect Size	SD
High	10	1.373574449	0.5374329
High	20	1.278985137	0.3796674
High	50	1.231157814	0.2761435
High	100	1.221580166	0.2504554

Fidelity	Sample size	Mean Effect Size	SD
Low	10	0.887274449	0.5830751
Low	20	0.792685137	0.452062
Low	50	0.744857814	0.3606143
Low	100	0.735280166	0.3269615

BNT

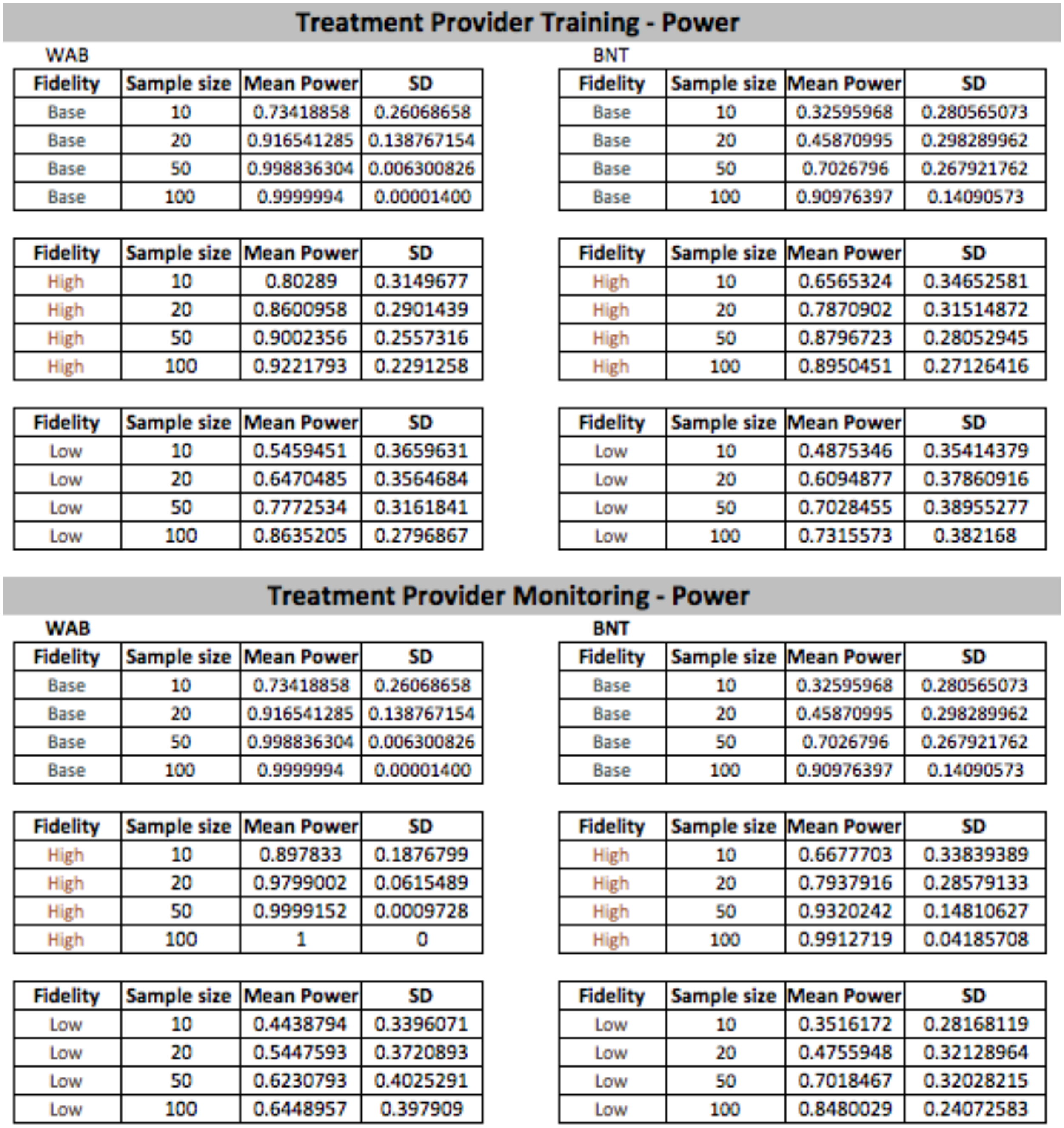
Fidelity	Sample size	Mean Effect Size	SD
Base	10	0.458265894	0.40142899
Base	20	0.431822292	0.25774105
Base	50	0.40041364	0.16443427
Base	100	0.3957757	0.10989982

Fidelity	Sample size	Mean Effect Size	SD
High	10	0.751849008	0.4427295
High	20	0.725402668	0.3322544
High	50	0.693998683	0.2687296
High	100	0.689371184	0.2384357

Fidelity	Sample size	Mean Effect Size	SD
Low	10	0.265549008	0.5177635
Low	20	0.239102668	0.3986799
Low	50	0.207698683	0.3450691
Low	100	0.203071184	0.3277138

Appendix C

Figure 1. Comparison of Mean Power With Base, High, and Low Fidelity Conditions



Treatment Provider Training - Power

WAB

Fidelity	Sample size	Mean Power	SD
Base	10	0.73418858	0.26068658
Base	20	0.916541285	0.138767154
Base	50	0.998836304	0.006300826
Base	100	0.9999994	0.00001400

Fidelity	Sample size	Mean Power	SD
High	10	0.80289	0.3149677
High	20	0.8600958	0.2901439
High	50	0.9002356	0.2557316
High	100	0.9221793	0.2291258

Fidelity	Sample size	Mean Power	SD
Low	10	0.5459451	0.3659631
Low	20	0.6470485	0.3564684
Low	50	0.7772534	0.3161841
Low	100	0.8635205	0.2796867

BNT

Fidelity	Sample size	Mean Power	SD
Base	10	0.32595968	0.280565073
Base	20	0.45870995	0.298289962
Base	50	0.7026796	0.267921762
Base	100	0.90976397	0.14090573

Fidelity	Sample size	Mean Power	SD
High	10	0.6565324	0.34652581
High	20	0.7870902	0.31514872
High	50	0.8796723	0.28052945
High	100	0.8950451	0.27126416

Fidelity	Sample size	Mean Power	SD
Low	10	0.4875346	0.35414379
Low	20	0.6094877	0.37860916
Low	50	0.7028455	0.38955277
Low	100	0.7315573	0.382168

Treatment Provider Monitoring - Power

WAB

Fidelity	Sample size	Mean Power	SD
Base	10	0.73418858	0.26068658
Base	20	0.916541285	0.138767154
Base	50	0.998836304	0.006300826
Base	100	0.9999994	0.00001400

Fidelity	Sample size	Mean Power	SD
High	10	0.897833	0.1876799
High	20	0.9799002	0.0615489
High	50	0.9999152	0.0009728
High	100	1	0

Fidelity	Sample size	Mean Power	SD
Low	10	0.4438794	0.3396071
Low	20	0.5447593	0.3720893
Low	50	0.6230793	0.4025291
Low	100	0.6448957	0.397909

BNT

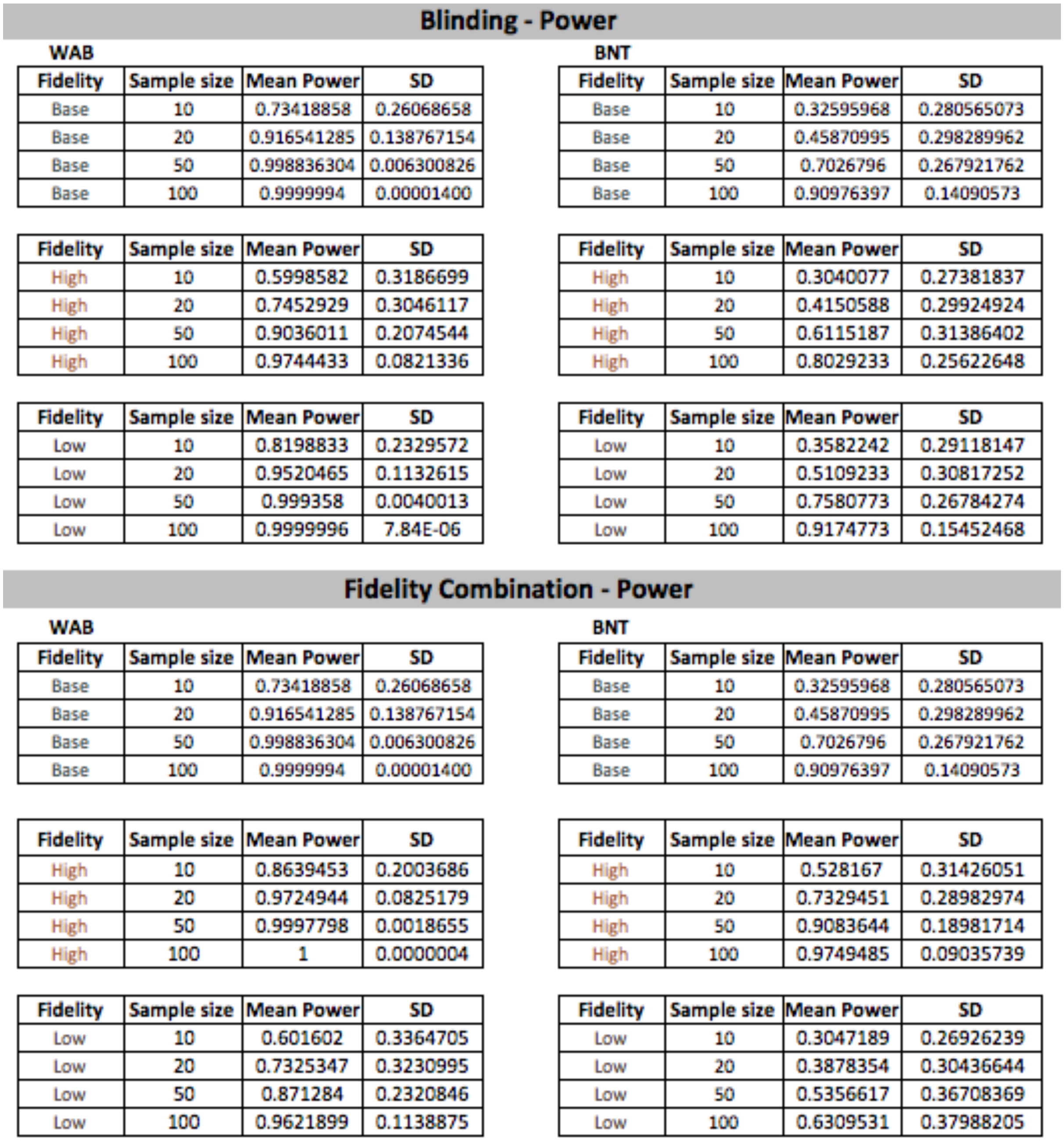
Fidelity	Sample size	Mean Power	SD
Base	10	0.32595968	0.280565073
Base	20	0.45870995	0.298289962
Base	50	0.7026796	0.267921762
Base	100	0.90976397	0.14090573

Fidelity	Sample size	Mean Power	SD
High	10	0.6677703	0.33839389
High	20	0.7937916	0.28579133
High	50	0.9320242	0.14810627
High	100	0.9912719	0.04185708

Fidelity	Sample size	Mean Power	SD
Low	10	0.3516172	0.28168119
Low	20	0.4755948	0.32128964
Low	50	0.7018467	0.32028215
Low	100	0.8480029	0.24072583

Appendix C Continued

Figure 1. Comparison of Mean Power With Base, High, and Low Fidelity Conditions



Blinding - Power

WAB

Fidelity	Sample size	Mean Power	SD
Base	10	0.73418858	0.26068658
Base	20	0.916541285	0.138767154
Base	50	0.998836304	0.006300826
Base	100	0.9999994	0.00001400

Fidelity	Sample size	Mean Power	SD
High	10	0.5998582	0.3186699
High	20	0.7452929	0.3046117
High	50	0.9036011	0.2074544
High	100	0.9744433	0.0821336

Fidelity	Sample size	Mean Power	SD
Low	10	0.8198833	0.2329572
Low	20	0.9520465	0.1132615
Low	50	0.999358	0.0040013
Low	100	0.9999996	7.84E-06

BNT

Fidelity	Sample size	Mean Power	SD
Base	10	0.32595968	0.280565073
Base	20	0.45870995	0.298289962
Base	50	0.7026796	0.267921762
Base	100	0.90976397	0.14090573

Fidelity	Sample size	Mean Power	SD
High	10	0.3040077	0.27381837
High	20	0.4150588	0.29924924
High	50	0.6115187	0.31386402
High	100	0.8029233	0.25622648

Fidelity	Sample size	Mean Power	SD
Low	10	0.3582242	0.29118147
Low	20	0.5109233	0.30817252
Low	50	0.7580773	0.26784274
Low	100	0.9174773	0.15452468

Fidelity Combination - Power

WAB

Fidelity	Sample size	Mean Power	SD
Base	10	0.73418858	0.26068658
Base	20	0.916541285	0.138767154
Base	50	0.998836304	0.006300826
Base	100	0.9999994	0.00001400

Fidelity	Sample size	Mean Power	SD
High	10	0.8639453	0.2003686
High	20	0.9724944	0.0825179
High	50	0.9997798	0.0018655
High	100	1	0.0000004

Fidelity	Sample size	Mean Power	SD
Low	10	0.601602	0.3364705
Low	20	0.7325347	0.3230995
Low	50	0.871284	0.2320846
Low	100	0.9621899	0.1138875

BNT

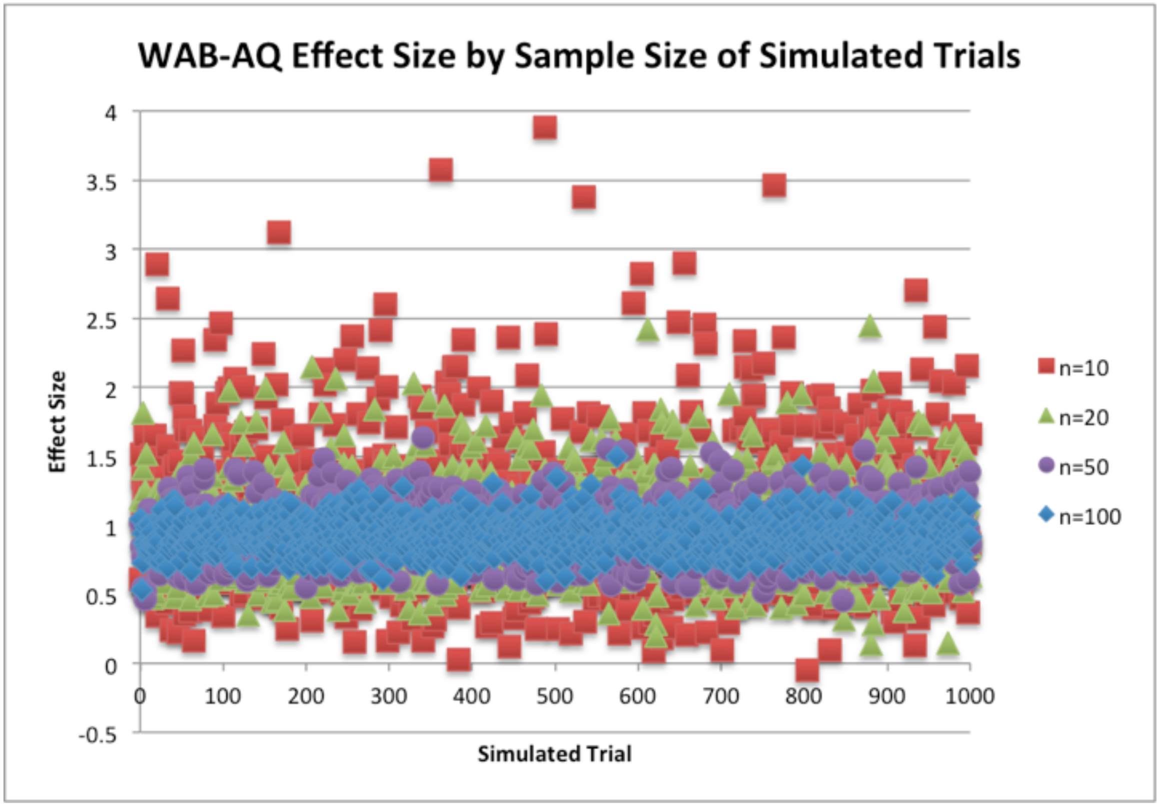
Fidelity	Sample size	Mean Power	SD
Base	10	0.32595968	0.280565073
Base	20	0.45870995	0.298289962
Base	50	0.7026796	0.267921762
Base	100	0.90976397	0.14090573

Fidelity	Sample size	Mean Power	SD
High	10	0.528167	0.31426051
High	20	0.7329451	0.28982974
High	50	0.9083644	0.18981714
High	100	0.9749485	0.09035739

Fidelity	Sample size	Mean Power	SD
Low	10	0.3047189	0.26926239
Low	20	0.3878354	0.30436644
Low	50	0.5356617	0.36708369
Low	100	0.6309531	0.37988205

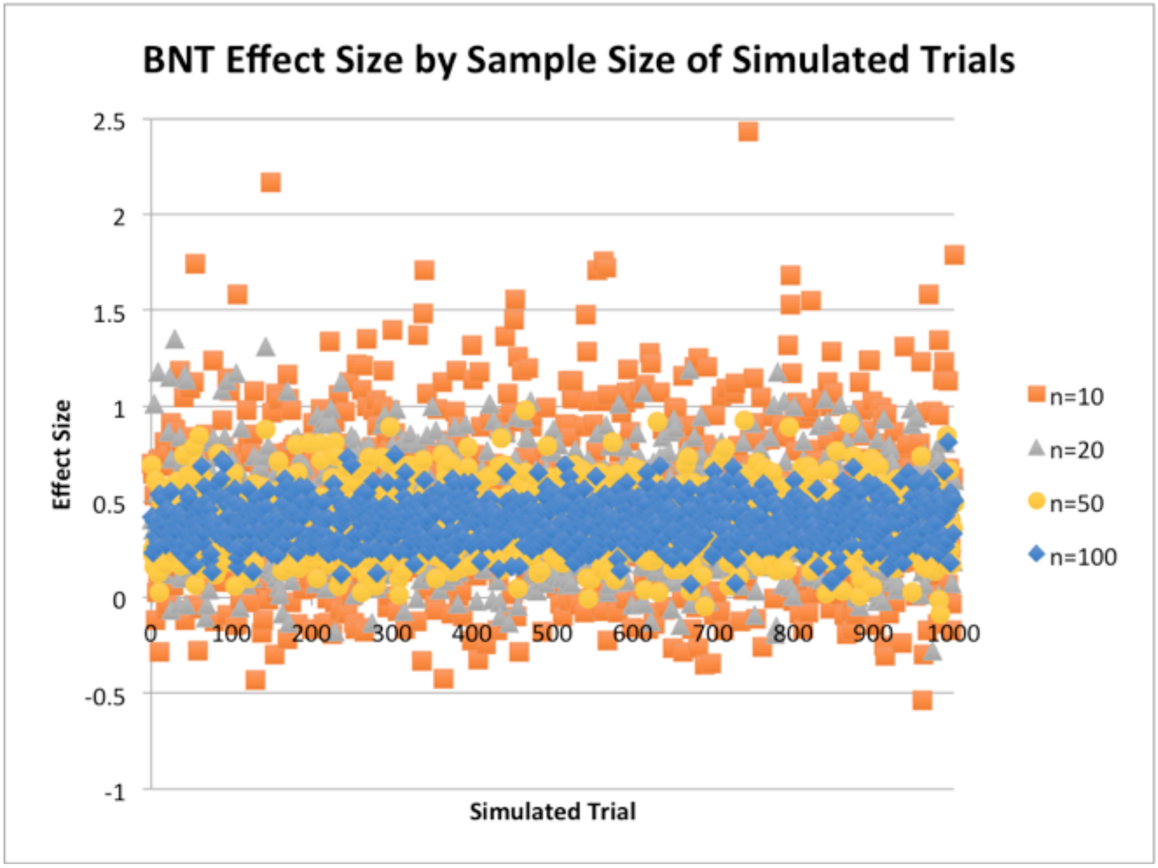
Appendix D

Figure 1. Effect Size as a Function of Sample Size for Base Fidelity WAB-AQ Change Scores



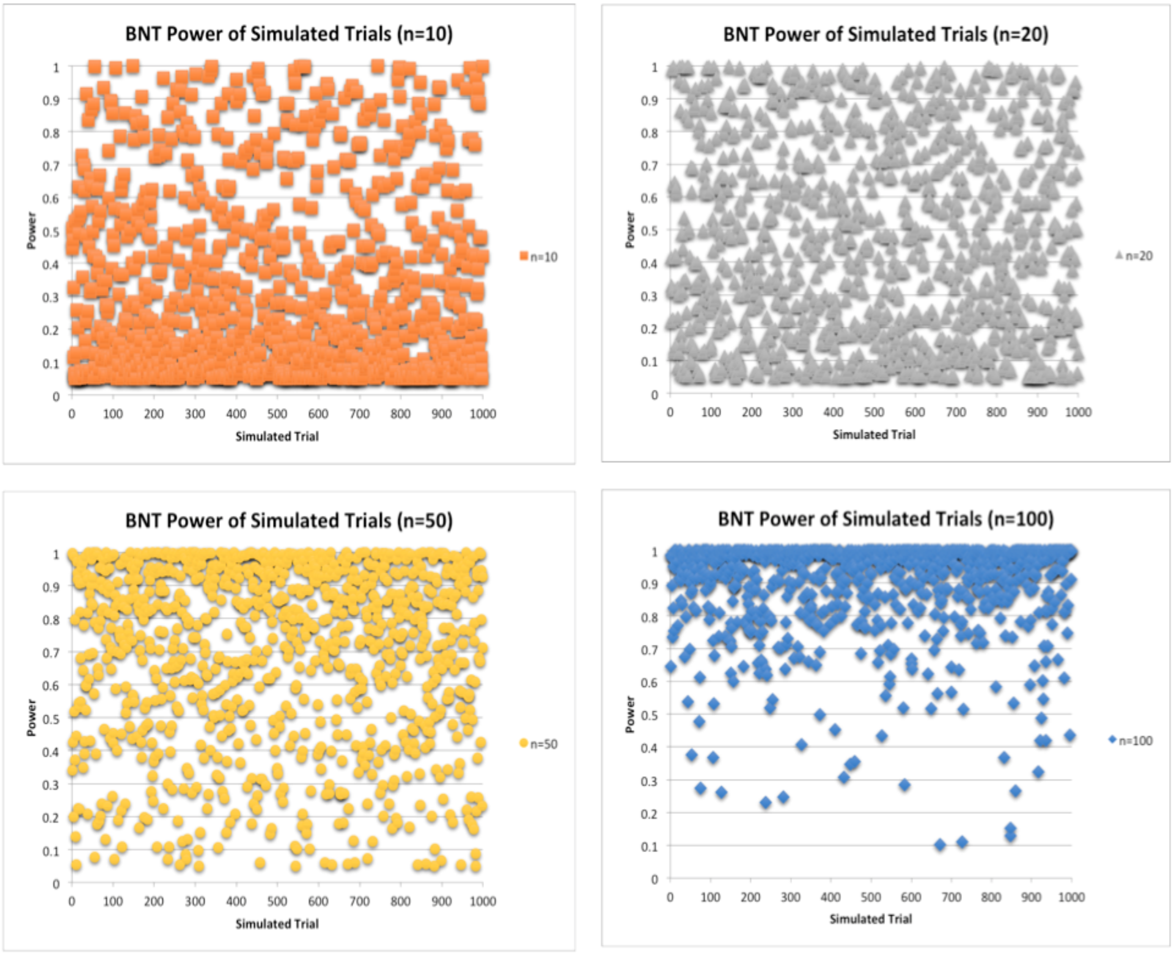
Appendix E

Figure 1. Effect Size as a Function of Sample Size for Base Fidelity BNT Change Scores



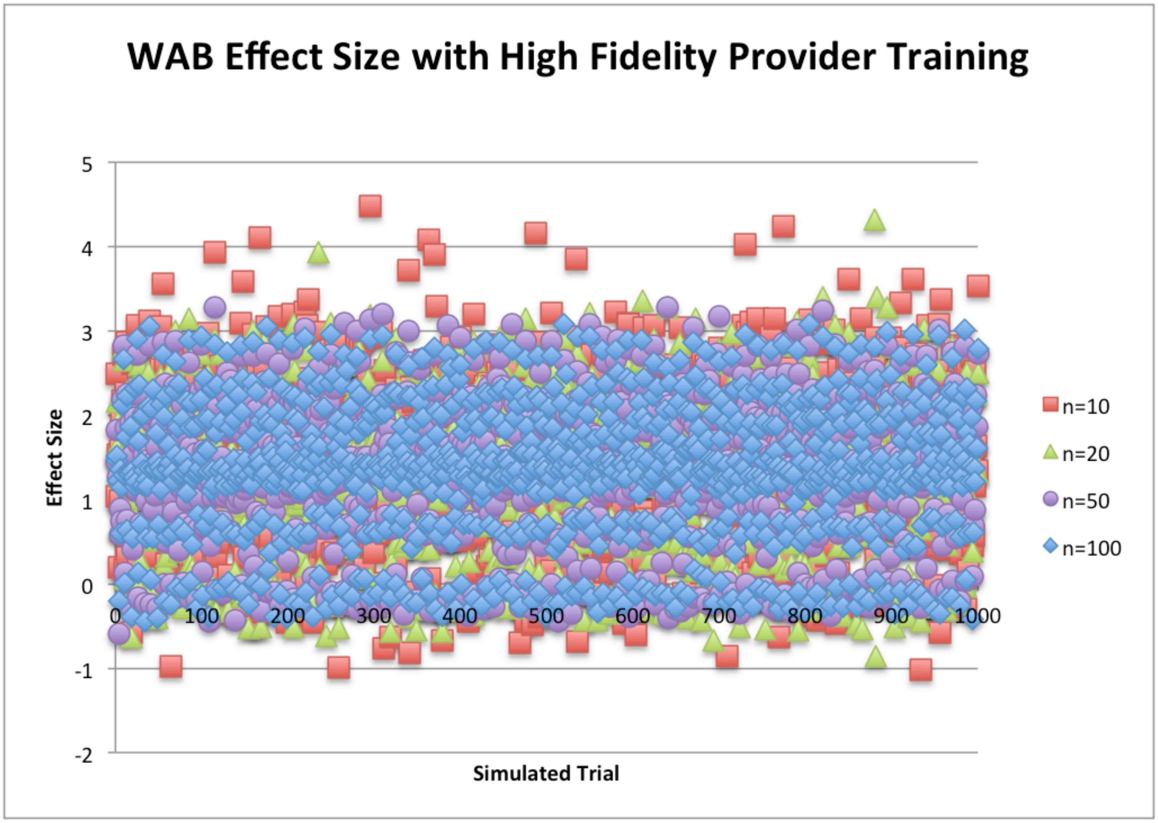
Appendix F

Figure 1. Power as a Function of Sample Size for BNT Change Scores



Appendix G

Figure 1. Scatter Plot of Effect Size as a Function of Sample Size for WAB Scores with High Fidelity Provider Training



References

Arkoosh, M. K., Derby, K. M., Wacker, D. P., Berg, W., McLaughlin, T. F., & Barretto, A. (2007). A descriptive evaluation of long-term treatment integrity. *Behavior Modification*, *31*(6), 880-895.

Bailey, J., Berson, A., Handelsman, H., & Hodges, M. (2001). Utility of current risk stratification tests for predicting major arrhythmic events after myocardial infarction. *Journal of the American College of Cardiology*, *38*(7), 1902-1911.

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 practice and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*(5), 443-451.

Benner, G. J., Nelson, J. R., Stage, S. A., & Ralston, N. C. (2011). The influence of fidelity of implementation on the reading outcomes of middle school students experiencing reading difficulties. *Remedial and Special Education*, *32*(1), 79-88.

Bennett, M. I., Hughes, N., & Johnson, M. I. (2011). Methodological quality in randomised controlled trials of transcutaneous electric nerve stimulation for pain: Low fidelity may explain negative findings. *Pain*, *152*(6), 1226-1232.

Borenstein, M. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111.

Borrelli, B. (2011). The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. *Journal of Public Health Dentistry, 71*(S1), S52-S63.

Carroll, R. A., Kodak, T., & Fisher, W. W. (2013). An evaluation of programmed treatment-integrity errors during discrete-trial instruction. *Journal of Applied*

Behavior Analysis, *46*(2), 379-394.

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*(4), 571-574.

Claridge, A. (2014). Efficacy of systemically oriented psychotherapies in the treatment of perinatal depression: a meta-analysis. *Archives Of Women's Mental Health*, *17*(1), 3-15.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *Journal of General Internal Medicine*, *24*(1), 74.

Cortina, J. (2000). *Effect size for ANOVA designs* (Sage University Papers Series. Quantitative Applications in the Social Sciences, series no. 07-129). Thousand Oaks, CA: Sage.

Derzon, J. H., Sale, E., Springer, J. F., & Brounstein, P. (2005). Estimating intervention effectiveness: Synthetic projection of field evaluation results. *Journal Of Primary Prevention*, *26*(4), 321-343.

DuBois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002). Effectiveness of mentoring programs for youth: A metaanalytic review. *American Journal of Community Psychology*, 30, 157–198.

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*(3-4), 327-350.

Ebbels, S. H. (2017). Intervention research: Appraising study designs, interpreting findings and creating research in clinical practice. *International Journal of Speech-Language Pathology*, 1-14.

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37-64.

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*(1), 79-88.

Girolametto, L., Weitzman, E., & Greenberg, J. (2012). Facilitating emergent literacy: Efficacy of a model that partners speech-language pathologists and educators. *American Journal of Speech-Language Pathology*, *21*(1), 47-63.

Groskreutz, N. C., Groskreutz, M. P., & Higbee, T. S. (2011). Effects of varied levels of treatment integrity on appropriate toy manipulation in children with autism. *Research in Autism Spectrum Disorders*, *5*(4), 1358-1369.

Haahr, M. T., & Hróbjartsson, A. (2006). Who is blinded in randomized clinical trials? A study of 200 trials and a survey of authors. *Clinical Trials, 3*(4), 360-365.

Hamre, B. K., Justice, L. M., Pianta, R. C., Kilday, C., Sweeney, B., Downer, J. T., & Leach, A. (2010). Implementation fidelity of MyTeachingPartner literacy and language activities: Association with preschoolers' language and literacy growth. *Early Childhood Research Quarterly*, *25*(3), 329-347.

Hansen, T., Elholm Madsen, E., & Sørensen, A. (2016). The effect of rater training on scoring performance and scale-specific expertise amongst occupational therapists participating in a multicentre study: A single-group pre–post-test study. *Disability and Rehabilitation*, *38*(12), 1216-1226.

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*(6), 568-79.

Hróbjartsson, A., Thomsen, A.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. *CMAJ : Canadian Medical Association Journal = Journal De L'Association Medicale Canadienne*, *185*(4), E201-11.

Jenkins, S. R., Hirst, J. M., & Reed, F. D. D. (2015). The effects of discrete-trial training commission errors on learner outcomes: An extension. *Journal of Behavioral Education*, *24*(2), 196-209.

Kobak, K. A., Kane, J. M., Thase, M. E., & Nierenberg, A. A. (2007). Why do clinical trials fail? The problem of measurement error in clinical trials: Time to test new paradigms. *Journal of clinical psychopharmacology*, *27*(1), 1-5.

Kobak, K. A., Leuchter, A., DeBrota, D., Engelhardt, N., Williams, J. B., Cook, I. A., & Alpert, J. (2010). Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response. *Journal of Clinical Psychopharmacology*, *30*(2), 193-197.

Kobak, K. A., Lipsitz, J. D., Williams, J. B., Engelhardt, N., & Bellew, K. M. (2005). A new approach to rater training and certification in a multicenter clinical trial.

Journal of Clinical Psychopharmacology, *25*(5), 407-412.

Koehler, J. A., Lösel, F., Akoensi, T. D., & Humphreys, D. K. (2013). A systematic review and meta-analysis on the effects of young offender treatment programs in Europe. *Journal Of Experimental Criminology*, *9*(1), 19-43.

Lee, C. W., & Cuijpers, P. (2013). A meta-analysis of the contribution of eye movements in processing emotional memories. *Journal Of Behavior Therapy And Experimental Psychiatry*, *44*(2), 231-239.

Lillehoj, C. J., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education & Behavior*, *31*(2), 242-257.

Liu, C. J., LaValley, M., & Latham, N. K. (2011). Do unblinded assessors bias muscle strength outcomes in randomized controlled trials of progressive resistance strength training in older adults?. *American Journal of Physical Medicine & Rehabilitation*, *90*(3), 190-196.

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*(3), 237-247.

Maxfield L., & Hyer L. (2002). The relationship between efficacy and methodology in studies investigating EMDR treatment of PTSD. *Journal Of Clinical Psychology*, *58*(1), 23-41.

McCurtin A, & Roddam H. (2012). Evidence-based practice: SLTs under siege or opportunity for growth? The use and nature of research evidence in the profession. *International Journal Of Language & Communication Disorders,* *47*(1), 11-26.

Metcalfe, C., Lewin, R., Wisher, S., Perry, S., Bannigan, K., & Moffett, J. K. (2001). Barriers to implementing the evidence base in four NHS therapies: Dietitians, occupational therapists, physiotherapists, speech and language therapists. *Physiotherapy, 87*(8), 433-441.

Mielczarek, B., & Uziałko-Mydlikowska, J. (2012). Application of computer simulation modeling in the health care sector: a survey. *Simulation*, 88(2), 197-216.

Milburn, T. F., Hipfner-Boucher, K., Weitzman, E., Greenberg, J., Pelletier, J., & Girolametto, L. (2015). Effects of coaching on educators' and preschoolers' use of references to print and phonological awareness during a small-group craft/writing activity. *Language, Speech, and Hearing Services in Schools*, *46*(2), 94-111.

Müller, M. J., & Szegedi, A. (2002). Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *Journal of Clinical Psychopharmacology*, *22*(3), 318-325.

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*(5), 670-679.

Pence, S. T., & St. Peter, C. C. (2015). Evaluation of treatment integrity errors on mand acquisition. *Journal of Applied Behavior Analysis*, *48*(3), 575-589.

Pennington, L., Goldbart, J., & Marshall, J. (2004). Interaction training for conversational partners of children with cerebral palsy: A systematic review. *International Journal of Language & Communication Disorders*, *39*(2), 151-170.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change:

Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*(4), 365-383.

Piasta, S. B., Justice, L. M., Cabell, S. Q., Wiggins, A. K., Turnbull, K. P., & Curenton, S. M. (2012). Impact of professional development on preschool teachers' conversational responsivity and children's linguistic productivity and complexity. *Early Childhood Research Quarterly*, *27*(3), 387-400.

Platt, T. L., Zachar, P., Ray, G. E., Underhill, A. T., & LoBello, S. G. (2007). Does Wechsler Intelligence Scale administration and scoring proficiency improve during assessment training?. *Psychological reports*, *100*(2), 547-555.

Reed, F. D. D., Reed, D. D., Baez, C. N., & Maguire, H. (2011). A Parametric Analysis of Errors of Commission During Discrete-Trial Training. *Journal of Applied Behavior Analysis*, *44*(3), 611-615.

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*(5), 259-268.

Reinke, W. M., Lewis-Palmer, T., & Merrell, K. (2008). The Classroom Check-up: A classwide teacher consultation model for increasing praise and decreasing disruptive behavior. *School Psychology Review*, *37*(3), 315.

Rezzonico, S., Hipfner-Boucher, K., Milburn, T., Weitzman, E., Greenberg, J., Pelletier, J., & Girolametto, L. (2015). Improving preschool educators' interactive Shared book reading: Effects of coaching in professional development. *American Journal of Speech-Language Pathology*, *24*(4), 717-732.

Richardson, J. D., Dalton, S. G. H., Shafer, J., & Patterson, J. (2016). Assessment fidelity in aphasia research. *American Journal of Speech-Language Pathology*, *25*(4S), S788-S797.

Rosen, J., Mulsant, B. H., Marino, P., Groening, C., Young, R. C., & Fox, D. (2008). Web-based training and interrater reliability testing for scoring the Hamilton Depression Rating Scale. *Psychiatry Research*, *161*(1), 126-130.

Sawilowsky, S (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods. 8* (2): 467–474.

Shadish, W., Cook, T., & Campbell, D. (2002). Statistical Conclusion Validity and Internal Validity. In Shadish, W., Cook, T., & Campbell, D. (Eds). *Experimental and Quasi Experimental Designs for Generalized Causal Inference* (pp. 33-63). Boston, MA: Houghton Mifflin.

Solomon, D., Battistich, V., Watson, M., Schaps, E., & Lewis, C. (2000). A six-district study of educational change: direct and mediated effects of the child development project. *Social Psychology Of Education : An International Journal*, *4*(1), 3-51.

Smith, J. D., Schneider, B. H., Smith, P. K., & Ananiadou, K. (2004). The effectiveness of whole-school antibullying programs: A synthesis of evaluation research. *School Psychology Review, 33,* 547–560.

Smith-Lock, K., Leitão, S., Lambert, L., Prior, P., Dunn, A., Cronje, J., Newhouse, S., & Nickels, L. (2013). Daily or weekly? The role of treatment frequency in the effectiveness of grammar treatment for children with specific language impairment. *International Journal of Speech-Language Pathology*, *15*(3), 255-267.

Smith-Lock, K. M., Leitao, S., Lambert, L., & Nickels, L. (2013). Effective intervention for expressive grammar in children with specific language impairment. *International Journal of Language & Communication Disorders*, *48*(3), 265-282.

Steinert, C., Stadter, K., Stark, R., & Leichsenring, F. (2016). The effects of waiting for treatment: A meta-analysis of waitlist control groups in randomized controlled trials for Social Anxiety Disorder. *Clinical Psychology & Psychotherapy*.

Stitt, J. K., Simonds, C. J., & Hunt, S. K. (2003). Evaluation fidelity: An examination of criterion-based assessment and rater training in the speech communication classroom. *Communication Studies, 54* (3) 341-353.

Stockard J. (2010). An analysis of the fidelity implementation policies of the what works clearinghouse. *Current Issues In Education, 13*(4), 1-24.

Tewuerbati, S., Maimaitili, M., Zhu, G., Du, G., Liu, B., Sailike, D., Fan, Y., & Dangmurenjiafu, G. (2015). Timing of endoscopic third ventriculostomy in pediatric patients with congenital obstructive hydrocephalus: Assessment of neurodevelopmental outcome and short-term operative success rate. *Journal of Clinical Neuroscience*, *22*(8), 1292-1297.

Tschuschke, V., Crameri, A., Koehler, M., Berglar, J., Muth, K., Staczan, P., Von Wyl, A., Schulthess, P., & Koemeda-Lutz, M. (2015). The role of therapists' treatment adherence, professional experience, therapeutic alliance, and clients' severity of psychological problems: Prediction of treatment outcome in eight different psychotherapy approaches. Preliminary results of a naturalistic study. *Psychotherapy Research*, *25*(4), 420-434.

Ulm, G., & Schüler, P. (1999). Cabergolin versus Pergolid. *Aktuelle Neurologie*, *26*(08), 360-365.

Villodas, M. T., McBurnett, K., Kaiser, N., Rooney, M., & Pfiffner, L. J. (2014). Additive effects of parent adherence on social and behavioral outcomes of a collaborative school–home behavioral intervention for ADHD. *Child Psychiatry & Human Development*, *45*(3), 348-360.

Webb, C. A., DeRubeis, R. J., Dimidjian, S., Hollon, S. D., Amsterdam, J. D., & Shelton, R. C. (2012). Predictors of patient cognitive therapy skills and symptom change in two randomized clinical trials: the role of therapist adherence and the therapeutic alliance. *Journal of consulting and clinical psychology*, *80*(3), 373.

Westbrook, D., Sedgwick-Taylor, A., Bennett-Levy, J., Butler, G., & McManus, F. (2008). A pilot evaluation of a brief CBT training course: Impact on trainees' satisfaction, clinical skills and patient outcomes. *Behavioural and Cognitive Psychotherapy*, *36*(5), 569-579.

Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Jüni, P., Altman, D. G., Gluud, C., Martin, M., Wood, A.J.G., & Sterne, J. A. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *British Medical Journal*, *336*(7644), 601-605.