Reproducibility in small-N treatment research: a tutorial using examples from aphasiology [preprint]

Robert Cavanaugh,[1,2] Yina M. Quique,[3] Alexander M. Swiderski,[1,2,4]
Lydia Kallhoff,[5] Lauren Terhorst,[6] Julie Wambaugh,[5]
William D. Hula,[2,1] William S. Evans[1]

1. University of Pittsburgh, Department of Communication Sciences and Disorders
2. VA Pittsburgh Healthcare System, Audiology and Speech Pathology Program
3. Center for Education in Health Sciences, Northwestern University & Shirley Ryan Ability Lab
4. Carnegie Mellon University, Center for Neural Basis of Cognition
5. University of Utah, Department of Communication Sciences and Disorders
6. University of Pittsburgh, Department of Occupational Therapy

The authors have no disclosures.

Key words: aphasia treatment reproducibility replication

Corresponding Author:
Robert Cavanaugh M.S. CCC-SLP
rob.cavanaugh@pitt.edu

Purpose: Small-N studies are the dominant study design supporting evidence-based interventions in communication science and disorders, including treatments for aphasia and related disorders. However, there is little guidance for conducting reproducible analyses or selecting appropriate effect sizes in small-N studies, which has implications for scientific review, rigor, and replication. This tutorial aims to (1) demonstrate how to conduct reproducible analyses using effect sizes common to research in aphasia and related disorders and (2) provide a conceptual discussion to improve the reader's understanding of these effect sizes.

Methods: We provide a tutorial on reproducible analyses of small-N designs in the statistical programming language R using published data from Wambaugh et al. (2017). Additionally, we discuss the strengths, weaknesses, reporting requirements, and impact of experimental design decisions on effect sizes common to this body of research.

Results: Reproducible code demonstrates implementation and comparison of within-case standardized mean difference, proportion of maximal gain, Tau-U, and frequentist and Bayesian mixed-effects models. Data, code, and an interactive web-application are available as a resource for researchers, clinicians, and students.

Conclusion: Pursuing reproducible research is key to promoting transparency in small-N treatment research. Researchers and clinicians must understand the properties of common effect size measures to make informed decisions in order to select ideal effect size measures and act as informed consumers of small-N studies. Together, a commitment to reproducibility and a keen understanding of effect sizes can improve the scientific rigor and synthesis of the evidence supporting clinical services in aphasiology and in communication sciences and disorders more broadly. :::

Reproducibility in small-N treatment research: a tutorial using examples from aphasiology

Researchers make many choices in the design, conduct, and reporting of their research. These "researcher degrees of freedom" are inherent to the scientific method, but have the potential to increase false-positive findings, inflate effect sizes, and impair successful replication (Simmons et al., 2011). Study pre-registration, sharing of data and analysis scripts, and replication efforts are potential solutions for methodically integrating these researcher degrees of freedom within the scientific endeavor. However, sharing of reproducible analysis scripts and data, which promotes transparency in scientific decision-making, remains scarce in Communication Sciences and Disorders (CSD), as acknowledged by this special issue. Reasonable concerns about confidentiality, substantial training investments required, and/or lack of guidance for best practices may contribute to low rates of data and code sharing. Nonetheless, a lack of transparency negatively impacts the scientific review, replication, synthesis, and real-world impact of research in CSD. It is essential that the field increases the uptake of open science practices and fosters inclusiveness and constructive guidance while doing so.

One area where this lack of transparency may have a large impact is in small-N treatment studies. Small-N studies, including experimental and non-experimental single-case designs (also referred to as single-subject designs) and within-subject case series designs, are the "dominant" intervention design across the field of CSD (Murray et al., 2013; Togher et al., 2009). Small-N studies typically focus on treatment response at the individual level and establish experimental control within each participant rather than via a control group (Thompson, 2015). While single-case experimental studies typically include at least 2-4 participants, these designs are often extended to within-subject case-series designs with upwards of thirty participants (e.g., Gilmore et al., 2020), which are useful for testing psycholinguistic theories and exploring individual differences in treatment response (Nickels et al., 2015).

In contrast to group-level studies (e.g., randomized controlled trials), small-N designs confer advantages such as reduced cost, lower recruitment demands (Kratochwill & Levin, 2014), and the ability to evaluate patterns of treatment response at the individual level. They can offer a cost-effective means of piloting novel interventions as a precursor to large-scale trials while minimizing concerns related to statistical power. Insights into individual-level responses to treatment are also crucial for studying heterogeneous populations and for clinical providers who provide intervention at the individual level (Portney & Watkins, 2015).

However, statistical analysis of small-N studies varies widely, and there is little guidance for the selection, implementation, and reporting of reproducible analyses in small-N studies. Effect sizes, which characterize the magnitude of treatment response, are a particular source of consternation and disagreement (Howard et al., 2015). The choices researchers make in selecting and implementing effect size measures can influence outcome interpretation and negatively affect study replication. Moreover, these choices are often insufficiently reported or acknowledged, such that their impact on study reproducibility and replication is not readily apparent.

To address these challenges, this tutorial has two aims: (1) Provide a guide to conducting reproducible analyses of small-N treatment studies using the statistical programming language R (R Core Team, 2020) and (2) Discuss how the selection and implementation of effect sizes can affect the interpretation, replication, and synthesis of small-N studies, with a focus on the rich history of small-N studies in aphasia and related disorders. To achieve these aims, we reanalyze published data from a recent series of multiple-baseline single-case experimental design studies on Sound Production Treatment for post-stroke apraxia of speech and aphasia (Wambaugh et al., 2014, 2016, 2017).

This work is not intended to provide a comprehensive tutorial on small-N experimental design methodology, statistical programming, or the analytical methods within (e.g., mixed-effects models or Bayesian statistics). For each of these, we provide recommendations for further reading throughout the paper. Instead, our intent is that this paper will serve as a practical starting place for researchers engaged in small-N studies to begin incorporating reproducible analyses into their regular workflow. Moreover, we hope that this work will enable researchers to make more informed choices of effect sizes and help researchers and clinicians be more informed consumers of small-N design methodology.

Reproducibility in small-N designs

Reproducibility is defined as consistently producing the same results from the raw data gathered in a study (Nosek & Errington, 2017). In the small-N design context, reproducibility requires a well-documented processing stream that begins with individual session-level data collected at each probe, and ends with finalized figures demonstrating performance over time and statistical results reporting the certainty and magnitude of change. In small-N studies, this processing stream often includes manual entry and manipulation of individual probe data in spreadsheets and manual creation of figures. However, this approach risks failures in *process reproducibility*, where the original analysis cannot be repeated because of underspecified or missing procedural details necessary to reproduce the analysis (Nosek et al., 2022). Even in cases where the analytical process is well-documented, human-mediated procedures leave room for errors (Strand, 2021), risking failures of *outcome reproducibility* - when a reanalysis obtains a different result than originally reported (Nosek et al., 2022).

Script-based analyses using statistical programming languages are one solution for improving reproducibility in small-N designs (Hardwicke et al., 2018; Kidwell et al., 2016). Using analysis scripts allows researchers to document each step of their data pipeline and statistical approach. Scripts also allow peers and collaborators to review and validate the analytical approach as part of the research workflow. When data and scripts are shared, external researchers can reproduce study results and can more easily replicate the analysis in future studies. However, script-based analyses and statistical programming may be unfamiliar to researchers in CSD, and may pose a high barrier to entry. Therefore, the first aim of this tutorial is to demonstrate how to use the statistical programming language R to reproducibly calculate effect size measures common to aphasiology (and which are applicable to other areas of CSD research) using published data from Wambaugh and colleagues (2017).

Effect sizes in small-N designs in Aphasiology

For small-N studies, characterizing the response to treatment is central to understanding intervention efficacy, candidacy, and the theoretical mechanisms that underlie treatment success (Kratochwill & Levin, 2014). However, the optimal methodology for measuring treatment response in small-N studies in aphasia and related disorders remains an area of disagreement (Howard et al., 2015). In general, clinical researchers seek to establish 1) whether a treatment effect exists (i.e., statistical significance testing) and 2) an effect size that characterizes the magnitude of treatment response. Effect sizes are essential for validating the clinical relevance of interventions, where the magnitude of the effect (i.e., treatment response) is arguably at least as important as its statistical significance. Greater evidence for an intervention can be established by meta-analysis of multiple small-N studies, which typically focus on synthesizing effect sizes. Within-subject case-series experimental designs often rely on precise estimates of individual effect sizes to evaluate relationships between individual factors and treatment response (Rapp, 2011). Precise effect sizes are also important for estimating statistical power in subsequent trials.

There are also domain-specific considerations for estimating effect sizes in small-N designs in aphasia and related disorders. Heterogeneity in language abilities and performance variability is inherent to the nature of aphasia, and interventions often engender a wide range of treatment responses. Trends during the baseline phase are common, making it difficult for researchers to differentiate treatment response from repeated testing effects. Finally, the outcome variables of interest in most small-N studies in aphasiology are often generated from closed sets of stimuli or treatment targets, which benefit from careful modeling to appropriately characterize the data and promote the generalizability of study findings (Wiley & Rapp, 2018).

Effect sizes common to small-N studies are often sensitive to experimental design choices, which can obscure successful conceptual replication - our ability to support the same hypothesis through different experimental approaches. Effect sizes may be sensitive to experimental elements such as the difficulty, nature, or number of experimental stimuli, how stimuli are matched to participant characteristics, or the number of observations in the baseline and treatment phases. As a result, even when investigators take steps to ensure that workflows are reproducible, the choice of effect size can influence interpretation and replicability within and across small-N studies. This challenge motivates the second purpose of the present study, which is to help researchers and clinicians make informed decisions and be informed consumers of effect sizes common to small-N research.

In the following sections, we will review the conceptual definitions of effect sizes common to the small-N research in aphasia and related disorders and demonstrate their implementation in R using data from Wambaugh et al. (2017). Afterward, we will calculate and compare each effect size for all cases in Wambaugh et al. to motivate a discussion of the strengths and limitations of each measure.

Case example: reproducible reanalysis of Wambaugh et al., 2017

Wambaugh et al. (2017) reported the effects of Sound Production Treatment for apraxia of speech and aphasia for 20 individuals in a multiple-baseline design under two experimental conditions: blocked and random practice. Sound Production Treatment uses “therapeutic techniques of modeling and repetition, contrastive practice, orthographic cueing, integral stimulation, and articulatory cueing” in a “response-contingent hierarchy” to target phoneme production (p. 1744). The expectation is that repeated practice based on principles of motor learning will improve the production of target phonemes in the treated context, and (ideally) generalize to the production of those phonemes in words that are not explicitly treated.

In Wambaugh et al., (2017), participants received treatment on two lists, one in each experimental condition (random and blocked). Each list contained treated items and untreated generalization items for two target phonemes. Items consisted of a single word or occasionally a 2-3 word phrase. For 16/20 participants, treated lists contained 20 items (10 for each phoneme), and untreated lists contained 10 items (5 for each phoneme). For these participants, accuracy was determined based on the production of the target phoneme within the item. For 4/20 participants, treated lists contained 10 items (5 for each phoneme), and untreated lists contained 6 items (3 for each phoneme). For these participants, accuracy was determined based on the production of the entire item (see Wambaugh et al., 2016 for details on these four participants). All items were probed in a multiple-baseline crossover design with at least 5 baseline observations. For participants with 20-item treatment lists, accuracy was determined based on the production of the target phoneme within the item. For participants with 10-item treatment lists, the entire item was scored (see Wambaugh et al., 2016 for details on these four participants). While Wambaugh et al., (2017) aimed to compare the effects of randomized versus blocked practice on treatment outcomes, this tutorial will focus on calculating the following effect sizes: within-case standardized mean difference (Beeson & Robey, 2006), proportion of maximal gain (Lambon Ralph et al., 2010), Tau-U (Parker et al., 2011), and effect sizes based on frequentist and Bayesian mixed-effects models (e.g., Evans et al., 2021).

A reproducible analysis of data from Wambaugh et al., (2017) in R begins by loading necessary packages and setting a seed for reproducibility.[1] The data are stored in separate probe files for each participant and session, as small-N data are typically collected. By programmatically reading and combining raw probe data from each session, we avoid modifying the data by hand and minimize the chance for errors when combining the data manually.[2]

```r
library(tidyverse)       # data wrangling
library(SingleCaseES)    # calculating SMD, Tau-U
library(lme4)            # frequentist mixed-effects models
library(emmeans)         # estimating effect sizes from lme4
library(brms)            # bayesian mixed-effects models
library(tidybayes)       # estimating effect sizes from brms
library(here)            # for locating files

set.seed(42)             # set a seed for reproducibility

# create a list of files
```

[1] Users need to install packages prior to loading. We recommend the free book “Hands-On Programming with R for those completely unfamiliar with R (https://rstudio-education.github.io/hopr/) or needing assistance getting started. Installation of *brms* includes additional steps which can be found here: https://paul-buerkner.github.io/brms/.

[2] For the sake of brevity, some helper functions are omitted from the manuscript. However, all code is available in fully annotated form at https://github.com/rbcavanaugh/reproducibility-aphasia-JSLHR and https://osf.io/7fp3x/.

```
files <- list.files(
                 here("data"), # look in the study-data folder
                 full.names = TRUE,  # use the full paths of the files
                 pattern = ".csv",   # only read in .csv files
                 recursive = TRUE)   # include files within subfolders

# read in the files and combine them together
# save the resulting dataframe in an abject called "df"
df <- files %>%
  map_dfr(read_csv, show_col_types = FALSE)
```

The data is organized in a "tidy" format where each column represents a variable and each row represents an observation, or a single response to a single item (Wickham, 2014). See Table 1. for a codebook for column names and Table 2. for the first 5 rows of the data.

Table 1: Variables and descriptions for study data from Wambaugh et al., (2017)

Variable	Description
participant	de-identified participant ID
condition	probe schedule (blocked or random)
phoneme	target_phoneme
itemType	item condition (treatment or generalization)
phase	treatment phase
session	session number from Wambaugh 2017
item	item identifier
trials	number of items in the list (per phoneme)
spt2017	phase used to calcualte effect sizes in Wambaugh et al., 2017
response	accuracy of participant response
n_baselines	Number of baseline sessions

Table 2: The first 5 rows of data

participant	condition	phoneme	itemType	phase	session	item	trials	spt2017	response	n_baselines
P1	blocked	pr	tx	baseline	1	pr-1	10	pre	0	5
P1	blocked	pr	tx	baseline	1	pr-12	10	pre	0	5
P1	blocked	pr	tx	baseline	1	pr-4	10	pre	0	5
P1	blocked	pr	tx	baseline	1	pr-15	10	pre	0	5
P1	blocked	pr	tx	baseline	1	pr-5	10	pre	0	5

In the following sections, we demonstrate how to calculate each effect size for a single case participant (participant 10, blocked condition) from Wambaugh et al., (2017). Participant 10's performance on the blocked condition is shown in Figure 1. R code for calculating effect sizes for all participants, item-types, and conditions in the study is available in the supplemental material S2. We can subset and summarize performance data for participant 10 as follows.

```r
# Create a dataframe holding only data for participant 10
# The new dataframe is stored in an objected called "P10"
P10 <- df %>%
  # filter for participant 10, treated condition, blocked condition
  filter(participant == "P10",
         itemType == "tx",
         condition == "blocked") %>%
  # calculate the sum for each level of session, phase and spt2017
  group_by(session, phase, spt2017) %>%
  summarize(sum_correct = sum(response), .groups = "drop")
```

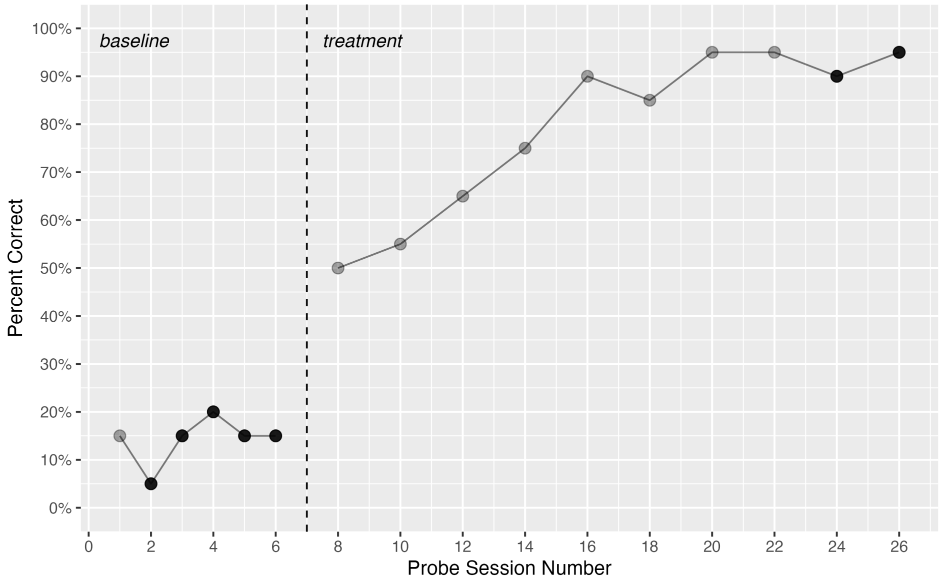

Figure 1: Participant 10 performance during baseline and treatment phase for the blocked condition. Dark circles indicate data points used to calculate dBR and PMG.

Within-case Standardized Mean Difference

Beeson and Robey (2006) advocated for using a within-case standardized mean difference in single-subject designs and meta-analyses of aphasia single subject research. It is one of, if not the most used measures in the field (Antonucci & Gilmore, 2019). The within-case standardized mean difference was initially proposed by Gingerich (1984) and later, Busk & Serlin (1992) as an individual-level effect size measurement that could be synthesized in meta-analysis. It was originally defined as the difference in means between the treatment and baseline phase divided by the standard deviation of the baseline phase. Beeson and Robey (2006) modified the measure, advocating for subtracting the mean of the baseline phase $\bar{x}_{A_1}$ from a post-treatment phase $\bar{x}_{A_2}$ in the context of an ABA design with multiple probes during the baseline and post-treatment phases (henceforth, d_{BR}). This within-case d_{BR} statistic represents the mean change between the end of the treatment phase and the baseline phase divided by the amount of variability during the baseline phase. It relies on the assumption that observations (i.e., probe sessions) are mutually independent, and that variability is present and constant within the baseline phase. d_{BR} is unbounded, with values greater than zero indicating a positive response to treatment. When there is no variability in the baseline phase, researchers must decide whether to pool the standard deviation across phases, use the baseline variability from another condition or participant, or omit the measure for that series. d_{BR} is typically interpreted based on established benchmarks of “small,” “medium,” and “large”.

$$d_{BR} = \frac{\bar{x}_{A_2} - \bar{x}_{A_1}}{S_{A_1}} \tag{1}$$

To calculate d_{BR} for a single set of data in R, we can calculate the mean of the baseline scores (x_A) , the mean of the treatment scores (x_B), and the standard deviation of the baseline scores (S_A). The d_{BR} statistic is then calculated by subtracting the baseline scores from the treatment scores and dividing by the standard deviation. In Wambaugh et al., (2017) the d_{BR} statistic was calculated using the last five baseline time points leading up to treatment and the last two time points in the treatment phase (it is not uncommon for study designs which do not collect multiple consecutive post-treatment probes to use probes from the last 2-3 treatment sessions). This information is already included in the dataset (column “spt2017”), which has values “pre” for baseline, and “post” for treatment for time points that were included in this calculation, and NA for values not included in the calculation.

While it is relatively straightforward to write a function to calculate d_{BR} in R, here we will use a function from a published package, SingleCaseES, which includes a variety of effect size methods common to single-case experimental designs (Pustejovsky et al., 2021). We can use the `SMD()` function from the SingleCaseES package to calculate d_{BR} for participant 10. We start by extracting the number of correct responses in the baseline and treatment phases, filtering for probe sessions used by Wambaugh et al., (2017) to calculate d_{BR}. Those data are then used to calculate d_{BR}.

```r
# Extract the outcomes for the pre- and post-treatment
# Save the outcomes in two objects, called "A" and "B"
A <- P10 %>% filter(spt2017 == "pre") %>% pull(sum_correct)
B <- P10 %>% filter(spt2017 == "post") %>% pull(sum_correct)
# Calculate d BR using the two vectors above
SMD(A_data = A, B_data = B)$Est
```

```
[1] 11.46566
```

Note that while the `SMD()` function also returns a confidence interval, its not clear how appropriate this confidence interval for d_{BR}, as the standard error is based on the original within-case standardized mean difference.

Proportion of Potential Maximal Gain.

Lambon Ralph and colleagues (2010) proposed the proportion of potential maximal gain (PMG) as a method for describing the relative magnitude of improvement, accounting for baseline performance. PMG was intended to be used in analyses where participants received a different number of treated items or to account for baseline severity when the same items were assigned to all participants (Lazar et al., 2010; Snell et al., 2010). This feature makes PMG particularly relevant to the present dataset, where a fifth of participants received a modified SPT with fewer treated and untreated items.

PMG is defined as the difference in the average number of correct responses between a post-treatment phase $\bar{x}_{A_2}$ and the baseline phase $\bar{x}_{A_1}$ divided by the number of items available to gain during the treatment phase (i.e., the number of items treated less the average number correct during the baseline phase; Equation 2). PMG ranges between -1 and 1, where values near zero indicate no change and positive values indicate improvement. PMG can be interpreted as the proportion of improvement relative to the amount of possible improvement after the baseline phase.

$$PMG = \frac{\bar{x}_{A_2} - \bar{x}_{A_1}}{n_{Items} - \bar{x}_{A_1}} \tag{2}$$

There is no R package that includes a function to calculate PMG to our knowledge. However, creating such a function is relatively straightforward. A function that calculates PMG similar to the `SMD()` function from the SingleCaseES package might take the following form, first calculating the mean of the baseline phase ($\bar{x}_{A_1}$), the mean of the post-treatment phase ($\bar{x}_{A_2}$), the potential improvement after the baseline phase, and then calculating the PMG statistic. Similar to d_{BR} above, PMG is calculated below using the last five observations in the baseline phase and the last two observations in the treatment phase.

```
# A function for calculating PMG that takes 3 arguments:
# Vectors of the pre-treatment data and the post-treatment data
# and the number of treated items

PMG <- function(a_data, b_data, nitems){
  mean_a <- mean(a_data) # calculate mean of the pre-treatment data
  mean_b <- mean(b_data) # calculate mean of post-treatment data
  change_score <- mean_b-mean_a # calculate the change score
  potential_gain <- nitems-mean_a # calculate the potential gain
  pmg <- change_score / potential_gain # calculate PMG
  return(pmg)
}

# Use the new function with the A and B data from above
PMG(a_data = A, b_data = B, nitems = 20)
```

```
[1] 0.9127907
```

Tau-U

Tau-U was proposed by Parker et al. (2011) as a collection of non-parametric effect size measures that use Kendall's Rank Correlation to evaluate the independence of performance between study phases. Unlike other approaches discussed in this article, Tau-U is intended to evaluate the degree of non-overlap between treatment phases rather than the total magnitude of change between phases (Parker et al., 2011). The Tau statistics are essentially a rescaling of non-overlap of all pairs (Tarlow, 2017) to the range [-1, 1], where 0 indicates no change and positive values indicate increasing independence between study phases. In aphasia and related disorders, Tau-U has generally referred to the case of Tau-U with a correction for baseline trends (Tau-U$_{\text{A VS. B} - \text{TREND A}}$). The baseline correction is typically applied if the baseline slope exceeds a cut-off ideally set a-priori, depending on the researcher's preference (Lee & Cherney, 2018). To be consistent with the aphasia literature, we will refer to Tau-U as the general statistic, specifying Tau-U$_{\text{A VS. B} - \text{TREND A}}$ or Tau-U$_{\text{A VS. B}}$ where relevant.

We can calculate Tau-U as outlined by Lee and Cherney (2018) in R using SingleCaseES First, the `lm()` function (linear regression) is used to calculate the slope of any baseline trend.

```
# start with the dataframe for participant 10
P10 %>%
    # filter for only baseline observations
    filter(phase == "baseline") %>%
    # run a linear regression to calculate the slope of performance
    lm(data = ., sum_correct~session) %>%
    # extract the coefficients of the regression
    coef()
```

```
(Intercept)     session
   2.133333    0.200000
```

Using the conservative benchmark of 0.33 recommended by Lee and Cherney (2018), we would calculate Tau-U$_{\text{A VS. B}}$ (without a baseline trend correction), as the slope of the baseline phase is only 0.2. To calculate Tau-U$_{\text{A VS. B}}$, we can use the Tau() function. Note that for Tau-U, we use all observations in the baseline and treatment phases.

```
# Extract the outcomes for the baseline and treatment phases
# Save the outcomes in two objects, called "A" and "B"
A <- P10 %>% filter(phase == "baseline") %>% pull(sum_correct)
B <- P10 %>% filter(phase == "treatment") %>% pull(sum_correct)
# Calculate Tau-U without trend correction using the two vectors above
Tau(A_data = A, B_data = B)
```

```
   ES Est         SE CI_lower CI_upper
1 Tau   1 0.02710291        1        1
```

However, if we had elected to correct for baseline trends and use Tau-U$_{\text{A VS. B} - \text{TREND A}}$, we can use the `Tau_U()` function. In this case, Tau-U$_{\text{A VS. B}}$ = 1, as there are no treatment observations equal to or less than any one baseline observation, but Tau-U$_{\text{A VS. B} - \text{TREND A}}$ = 0.95 due to a small baseline trend correction.

```
# Calculate Tau-U with trend correction using the two vectors above
Tau_U(A_data = A, B_data = B)
```

```
     ES  Est
1 Tau-U 0.95
```

Generalized Linear Mixed-effects Models

Linear mixed-effects models (also called hierarchical models, multilevel models) have grown in popularity over the past decade and confer several advantages over traditional repeated measures analyses. Such advantages include the ability to analyze trial-level responses (e.g., correct, incorrect) rather than overall session accuracy (Jaeger, 2008), accommodate unbalanced designs and missing outcome data, and account for variation in both participants and stimuli simultaneously, thereby producing more generalizable findings (Baayen et al., 2008). In small-N designs, mixed-effects models can adjust for baseline trends in performance, and allow researchers to evaluate interactions between treatment effects and other variables such as treatment condition or disorder severity, and can characterize non-linear changes in performance. Finally, the generalization of mixed-effects models beyond the linear case (generalized linear mixed-effects model; GLMM) permits researchers to appropriately characterize the dependent variable using a more appropriate probability distribution and link function, e.g., a binomial distribution and logistic link for the number of successes in a given number of trials, or a Poisson distribution and log link function to analyze count data. We direct readers to Wiley and Rapp (2018) for a primer on mixed-effects models in aphasia research and multiple recent tutorials focused on research in communication sciences and disorders (Gordon, 2019; Harel & McAllister, 2019).

There are a number of different approaches to modeling longitudinal or repeated measures data for one or more participants in small-N designs using mixed-effects models. In this paper, we will review an approach we have previously used in small-N designs in aphasiology (Evans et al., 2021; Swiderski et al., 2021) that we find to align well with our conceptual model of multiple baseline designs, noting that the general concepts likely apply to similarly structured models. This interrupted time series model was originally advocated for by Huitema & McKean (2000) in the form of a standard linear model (equation 3).

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 [T_t - (n_1 + 1)] D_t + \epsilon_t \tag{3}$$

The model describes the outcome *Y* at time *t* using fixed parameters for a baseline slope (β_1), level change (β_2) immediately following the onset of treatment, and slope change (β_3), representing the trend difference between the baseline and treatment phases. In this case, T represents the probe session number at time *t*, and n_1 represents the number of baseline sessions (Huitema, 2011). The level change and slope change parameters can be interpreted as effect sizes for their respective components. Additionally, an overall effect size and 95% confidence interval can be obtained by examining the estimated difference in performance between the end of baseline and end of treatment, accounting for a baseline trend if desired.

We can extend this model to a generalized linear mixed-effects model for participant 10, by modeling the response for each item (correct or incorrect) using a binomial distribution and logistic link function. The model's fixed effects (the primary effects of interest, Searle et al., 1992) include baseline slope, level change, and slope change. The random effects (effects for which there is interest in generalizing to the underlying population - e.g., all potential treated items) allow each item to have its own intercept and slope (see S1 for the model equation). This is considered a two-level model as repeated measures (items) are nested within observations (probe sessions) for a single participant.

To implement the two-level model, we start by filtering the original data for the same blocked condition for participant 10, but maintain the item-level data. Then, we create the level change and slope change parameters according to equation 3. The baseline slope variable is equivalent to the probe session number. The level change variable is a categorical dummy variable with values 0 for baseline, and 1 for treatment. The slope change variable multiplies the number of baseline sessions (6) plus 2 by level change;[3] this value is then subtracted from the baseline slope variable. Model coefficients are visualized in Figure 2.

```
# Create a dataframe holding item-level data for participant 10
# The new dataframe is stored in an objected called "P10"
P10 <- df %>%
  # filter for P10, blocked condition, treated items,
  # baseline or treatment phases
  filter(participant == "P10",
         condition == "blocked",
```

[3] In this case, we use the number of baseline sessions plus 2, because the probe schedule at the onset of treatment changed to every other session. In a design where probes are administered at every session, one would use the number of baseline sessions plus 1, as in Equation (3).

```
       itemType == "tx",
       phase == "baseline" | phase == "treatment") %>%
# baseline slope is equivalent to the session variable
# level change is 0 for the baseline phase and 1 for treatment phase
# slope change is calculated by subtracting the total number of baseline
# sessions + 2 from the baseline slope variable, and multiplying the
# result by the level change variable. The level change variable is
# then converted to a factor (categorical) variable.
mutate(baseline_slope = session,
       level_change = ifelse(phase == "baseline", 0, 1),
       slope_change = (baseline_slope - (6+2))*level_change,
       level_change = as.factor(level_change))
```

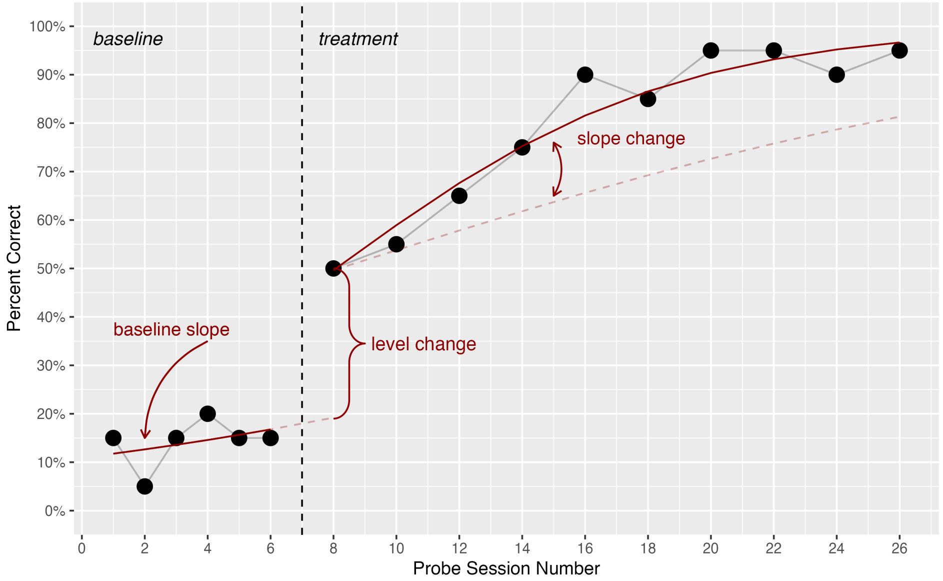

Figure 2: Participant 10 performance during baseline and treatment phase from Wambaugh et al., (2017). Plot annotations indicate Huitema & McKean (2000) model coefficients.

In the code below, the model is assigned to the object "`mod1`." The dependent variable is called "`response`", consisting of a 1 for correct and 0 for incorrect responses. The three fixed effects include baseline slope, level change, and slope change. The model also includes random intercepts for each item and random slopes for baseline slope, level change, and slope change. The family argument indicates that the dependent variable follows a binomial distribution. An additional argument specifying an optimizer is included to improve model convergence.[4]

[4] Recommendations for convergence and fit warnings: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html

```
# The resulting model is saved as an object called "mod1"
mod1 <-
    # a mixed-effects model starts with the glmer() function
    glmer(
         # response is the dependent variable (0 or 1)
         # The independent variables come after the "~" symbol
         # the fixed effects
         response ~ baseline_slope + level_change + slope_change +
         # random effects are in parentheses. item is a random intercept
         # while the three effects before the "|" symbol are random slopes
         (1 + baseline_slope + level_change + slope_change | item),
         # specify the data
          data = P10,
         # specify the distribution of the dependent variable
          family = binomial,
         # the optimizer is intended to help with convergence
          control = glmerControl(optimizer = "bobyqa"))
```

Finally, we can extract the model summary using the `summary()` command. The fixed effects portion of the model summary is reported below.

```
# remove $coefficients to display the entire summary
summary(mod1)$coefficients
```

```
                  Estimate Std. Error     z value   Pr(>|z|)
(Intercept)    -2.70899519  1.3324187 -2.03314107 0.04203827
baseline_slope  0.01742334  0.3324948  0.05240185 0.95820850
level_change1   2.50787842  1.6943323  1.48015737 0.13883125
slope_change    0.39216545  0.3965121  0.98903790 0.32264460
```

Model coefficients are returned in logits (or log-odds). Briefly, these results indicate that performance immediately prior to baseline is predicted to be about 6.2% (achieved by converting the intercept log-odds to probability). The odds of a correct response rose marginally during the baseline phase, were 12.3 times greater after the first treatment session (calculated by exponentiating the level change log-odds value of 2.5), and increased by a rate that was 1.5 times greater during treatment than it was during baseline. Neither the level change nor the slope change coefficient was statistically significant at an alpha level of .05.

An overall effect size can be calculated using the emmeans package (Lenth, 2021) by contrasting the estimated model performance at the last treatment session, with and without the level change and slope change parameters (see supplemental materials S1). This approach assumes that the baseline trend would have continued linearly in the absence of treatment and indicates that the odds of a correct response were 6.8 times greater as a result of treatment at the last treatment session (p = 0.16). Alternatively, in cases where the baseline phase is stable before the onset of treatment, we can contrast performance at the end of treatment from performance at the end of the baseline. This approach indicates that the probability of a correct response was 9.4 times greater at the end of treatment compared to the end of baseline (p = .0007).

When there are multiple participants in a study, this approach can be extended to a three-level model with random effects for both participants and items.[5] Because a three-level model can permit participants to vary in their intercept (performance at the start of the study), baseline slope, level change, and slope change, three-level models can characterize change both on average and for each participant. However, one limitation of the frequentist implementation of the three-level model is that it is difficult to estimate confidence intervals for individual participants.

Bayesian Mixed-effects Models

Bayesian implementations of mixed-effects models often resolve some challenges to implementing mixed-effects models in lme4, such as fitting models with more complex random effects structures. Bayesian models also permit estimation of individual effect sizes with an estimate of uncertainty from a group-level model. While an in-depth tutorial on Bayesian statistics is outside the scope of this work,[6] we summarize the main principles here.

Briefly, Bayesian data analysis is based on the idea that we can estimate a probability distribution for an effect (the posterior distribution) from the information contained in the data (the likelihood) and our prior knowledge of the effect (the prior, Nalborczyk et al., 2019). We can summarize an effect by calculating the mean or median of the posterior distribution and represent an effect's uncertainty by describing the tails of the posterior distribution (i.e., a Bayesian credible interval). For example, a 95% credible interval is interpreted such that there is a 95% probability that the interval contains the true effect, given the data and prior assumptions. Bayesian models can improve model estimation for small sample sizes and often permit greater complexity in the random effect structures. This is important because type-1 errors may increase in models with simpler random effect structure, a common method of addressing convergence issues that arise with complex models. In the following, we demonstrate how researchers may use the brms package (Bayesian Regression Models using Stan, Bürkner, 2017) to implement the group model. To implement the group model, we subset the data for only the treated items in the blocked condition.

```
df_glmm <- df %>%
  # select the correct phase, condition, and itemType
  filter(phase == "baseline" | phase == "treatment",
         condition == "blocked",
```

[5] Often, a minimum of 5 is recommended, though the necessary number of participants likely depends on the goals of the researcher. See https://bbolker.github.io/mixedmodels-misc/glmmFAQ

[6] For further reading on Bayesian statistics in the context of communication sciences and disorders, we refer the reader to the excellent tutorial article by Nalborczyk and colleagues (2019).

```
         itemType == "tx") %>%
  # create the Huitema model parameters
  mutate(baseline_slope = session,
         level_change = ifelse(phase == "baseline", 0, 1),
         slope_change = (baseline_slope - (n_baselines+2))*level_change,
         level_change = as.factor(level_change))
```

To fit the three-level model for multiple participants in the brms R package (Bürkner, 2017), only a few modifications are required of the lme4 code. First, the function changes from `glmer()` to `brm()`. Second, a group-level intercept for participant and slopes for baseline slope, level change, and slope change are added to permit participants to vary in their trend during baseline, initial level change, and overall slope change. Group-level intercepts are also included for items. Third, the `brm()` function family argument specifies the Bernoulli family as the special case of the binomial family, where each observation represents a single trial. Additionally, users should specify the total number of iterations, the number of iterations to remove at the beginning of sampling, the number of Markov chains, a seed for reproducibility, and importantly, prior distributions. In this case, we have included a prior distribution characterized by a mean of zero and standard deviation of 2.5 logits for the baseline slope, level change, and slope change effects, indicating that we anticipate that these effects are highly likely to fall within -5 and +5 logits (two standard deviations). At most, this prior distribution would correspond to improvements of about 85 percentage points for a participant starting around 7.5% accuracy. This prior is based on our previous use of this model structure Evans et al. (2021) and our understanding of what range of values constitute reasonable effect sizes. Similarly, we have included a prior on the intercept, with a mean of -1 and a standard deviation of 2.5. This prior expresses our knowledge of stimulus selection procedures in the study and the expectation that performance at the start of baseline will be poor (~27%) but with a wide range of plausible values (~0% accuracy to 72% accuracy). These priors help to incorporate domain knowledge into the modeling approach by establishing beforehand knowledge of the data, including the range of reasonable values for the variables in the model.

```
mod3 <-
 brm(
  # population level effects (similar to fixed effects)
  response ~ 0 + Intercept + baseline_slope + level_change + slope_change +
   # group level effects (similar to random effects)
   # by-participant group-level effects
  (1 + baseline_slope + level_change + slope_change | participant) +
   # by-item group-level effects
  (1|item),
  data = df_glmm, # the data used for the model
  family = bernoulli(), # special case of binomial with n=1 trials
  iter = 3000, # number of draws per chain
  warmup = 1000, # number of draws to toss on "warm up"
```

```
    chains = 4, # total number of chains
    seed = 42, # set a seed
    prior = c( # prior distributions
         prior(normal(-1, 2), class = b, coef = Intercept),
         prior(normal(0, 2.5), class = b)),
    # extra arguments, see rmd file
    cores = 4,
    file = here("output", "group_brm"),
    file_refit = "on_change")
```

After checking that the model demonstrates adequate fit and convergence (see supplemental materials S1), we can examine the model using `summary()`. Again, printed are the population-level effects (analogous to fixed effects), which are notably very similar to the frequentist three-level model (see Supplemental Materials S1). The model summary also returns a 95% credible interval, a convergence statistic ($\hat{R}$), and the effective sample size, an estimate of the number of independent Markov chain samples absent of autocorrelation (not pictured; the reader is referred to Nalborczyk et al. (2019) and supplemental materials S1. for interpretation guidelines).

```
# remove $fixed[,0:4] to see the entire summary
summary(mod3)$fixed[,0:4]
```

```
                  Estimate  Est.Error    l-95% CI   u-95% CI
Intercept      -3.40679200 0.23670963 -3.88543832 -2.9640354
baseline_slope  0.06295489 0.02337764  0.01649550  0.1100328
level_change1   0.90061169 0.45287149 -0.01494209  1.7760462
slope_change    0.10855856 0.03126962  0.04795764  0.1720307
```

Individual effect sizes can be obtained by contrasting the model's posterior predictions for each participant between the end of baseline and onset of the treatment phase in logits, odds-ratios, percentage gain, or items gained. This process is conceptually similar to the approaches described for the individual-level mixed-effects models, and well-commented code is available in the supplemental materials S2 (omitted here due to length). The result is a summary of an estimated distribution of effect sizes in logits for each participant, with the median of the distribution as the effect size (column ES) and a 95% credible interval (.lower and .upper). Effect sizes generated in this manner can be converted to odds ratios, percentage point gain, or an estimate of the number of items gained.

```
# Printing the top 6 rows of the effect size table reported in S2.
head(bayesian_es, n = 6)
```

```
# A tibble: 6 x 8
```

```
  participant     ES .lower .upper .width .point .interval unit
  <chr>        <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>     <chr>
1 P1            2.19 -0.368   4.99   0.95 median qi        logit
2 P10           4.01  1.52    6.75   0.95 median qi        logit
3 P11           2.97  1.52    4.61   0.95 median qi        logit
4 P12           3.06  1.99    4.13   0.95 median qi        logit
5 P13           5.86  4.31    7.51   0.95 median qi        logit
6 P14           3.96  2.66    5.34   0.95 median qi        logit
```

Researchers can interpret the magnitude of these effect sizes and examine how much of participants' credible intervals exceeds 0. Researchers can also define a range of values that are large enough to be clinically meaningful, and compare each individual's effect size distribution to this range in order to examine how many participants demonstrated clinically meaningful effects (Kruschke & Liddell, 2018).

Considerations for selecting effect sizes in small-N designs

In an ideal world, effect sizes in small-N studies are sensitive to a wide range of response patterns, provide a measure of confidence around the estimate of treatment response, and are interpretable by clinicians and researchers. Most importantly, they should be robust to experimental manipulation so that potential differences in effect sizes across cases and studies inform treatment theory and allow for conceptual replication. As Pustejovsky (2019) writes, "An effect size that is instead sensitive to such procedural features can appear to be larger (or smaller) because of how the study was conducted rather than because treatment actually produced large (or small) effects" (p. 218).

Consider recent work examining the effects of semantically-oriented anomia treatments for aphasia. Different research groups have provided converging and divergent findings for the efficacy and generalization of semantic treatments over the past three decades (e.g., Boyle, 2010; Evans et al., 2021; Gilmore et al., 2020) using a variety of related treatment approaches, but also differences in study designs and analytical methods. In order to make strong conclusions about conceptual replication or non-replication between these studies, we need to understand how methodological design decisions and analytical approaches affect conclusions about treatment response. In the following, we review the strengths and weaknesses of the different effect size measures to help clinician-researchers, and clinicians, make informed choices of effect sizes and act as informed consumers when interpreting related findings between research groups and studies.

There are substantial differences in the mathematical and conceptual approaches to the methods used to calculate effect sizes described above. Each effect size permits different conclusions about the data, may map to different formulations of the research question, is differentially sensitive to methodological decisions, and has conceptual limitations. To illustrate these differences, we calculated effect sizes for all series in Wambaugh et al., (2017).

Following Wambaugh et al., (2017), d_{BR} was calculated comparing the mean performance for the five baseline time points preceding the onset of the intervention to the mean of the last two time points of the treatment phase. We omitted series where baseline variance was zero, choosing not to pool the baseline variance or derive it from a separate series. Proportion of Potential Maximal Gain was calculated using the same observations as d_{BR}. Tau-U was calculated following the methods of Lee and Cherney (2018), including all observations in the baseline and treatment phases. Tau-$\mathrm{U}_{\text{A VS. B} - \text{TREND A}}$ was used when a linear trend line in the baseline phase exceeded 0.33, otherwise Tau-$\mathrm{U}_{\text{A VS. B}}$ was calculated. For simplicity, we collapsed performance across phonemes for d_{BR}, PMG, and Tau-U.

Because of a large number of convergence errors, even with overly simplified random-effects structures, while running individual-level models using the R package lme4 (Bates et al., 2015), we estimated mixed-effects model-based individual effect sizes using three-level Bayesian mixed-effect models (two models for each item type and two models for each condition, 4 total models). Based on our experience, this approach generally returns individual effect size estimates similar to those produced when frequentist models converge. Individual effect sizes were estimated based on the model predictions of performance differences between the end of the treatment phase and the end of the baseline phase in terms of logits and percentage point change. Priors and fitting procedures are reported in the S2.

We provide an interactive web app to allow each effect size measure to be compared along with each participant's performance: https://rb-cavanaugh.shinyapps.io/reproducible-small-N/. Effect sizes in the web app can be adjusted by modifying several researcher degrees of freedom, such as the choice to include all baseline data, set the Tau-U cutoff at 0.33 or 0.4, or extrapolate the baseline slope for the GLMM effect sizes. Throughout the remainder of the paper, we refer the reader to specific examples from Wambaugh et al., (2017), where the choice of effect size or analytical decisions can impact the interpretation of an individuals treatment response. These examples are easily viewed using the web-app. Readers can also use it to form a stronger intuitive understanding of the relationships between different measures and performance using real data. Additionally, the relationships between effect sizes are shown in Figure 3, a scatterplot matrix between the effect size measures that are the focus of this paper. This figure is available in the web-app, and will change as a result of chosen analytical decisions.

Broadly speaking the scatterplots and correlations demonstrate substantial differences between the various effect size approaches (Figure 3). We discuss the strengths and weaknesses of each effect size measure and the reasons for these differences below. A summary of the key strengths and weaknesses of each effect size is also reported in Table 3.

Table 3: Summary of major strengths and weaknesses of each effect size measure reviewed

Effect size measure	Strengths	Weaknesses
Within-case Standardized Mean (dBR)	Historically, most common method in aphasia & related disorders Easy to implement	Influenced by experimental design features (e.g., item set-size, baseline/treatment length) No clear solution to cases of low baseline variability Interpretation depends on benchmark study Lacks a measure of uncertainty
Proportion of Potential Maximal Gain (PMG)	Accommodates different item set sizes when comparing across individuals Easy to implement	Confounded by disorder severity if baseline performance is associated with disorder severity Can obscure differences in absolute change scores Lacks a measure of uncertainty
Tau-U	Non-parametric and distribution free Option to adjust for baseline trends Easy to implement	Does not fully characterize the magnitude of change Influenced by ratio of baseline/treatment observations Lack of easily interpretable scaling; not bounded between [-1, 1] Tau-UA VS. B – TREND-A lacks a measure of uncertainty
Generalized linear mixed-effects models (GLMM)	Able to adjust for baseline trends Effect size available in multiple units of measure (e.g., logits, odds-ratio, percent, items gained) each with clear interpretation Pools item-level and/or participant-level data to produce more generalizable estimates Includes confidence/credible interval Can estimate group and individual effect sizes from a single model (Bayesian)	Complex to implement Model convergence challenges are common with frequentist estimation Confidence/Credible interval width dependent on sample size

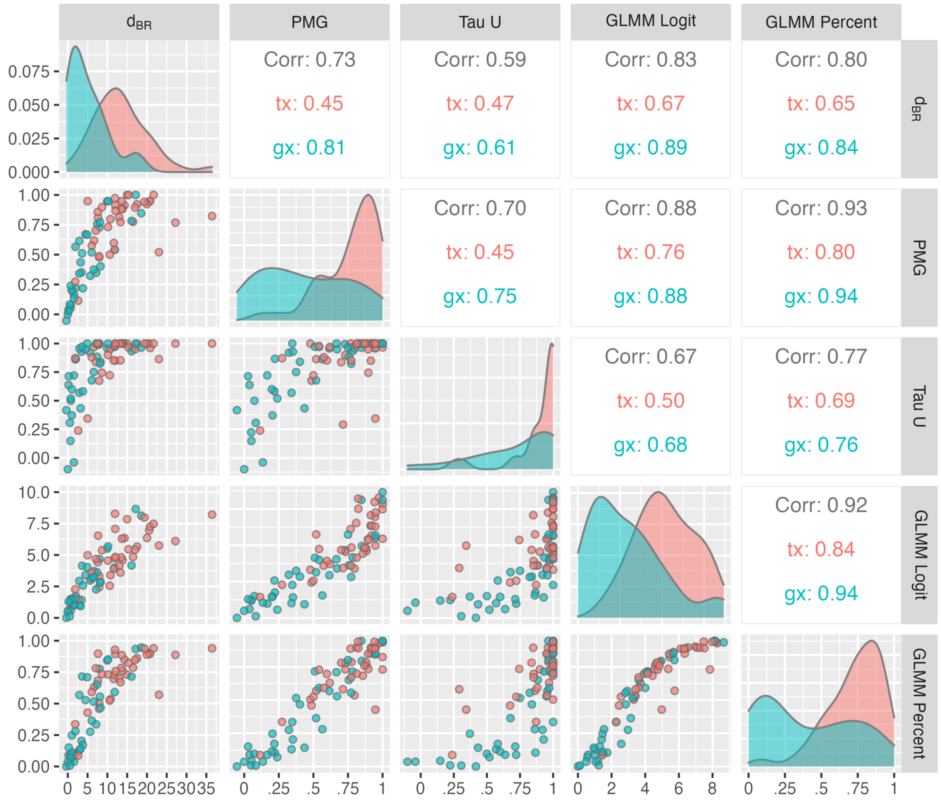

Figure 3: Relationships between individual effect size measures typically used in aphasia small-N studies.

Within-case standardized mean difference (d_{BR})

The primary feature of the scatterplots comparing d_{BR} with other effect size measures (Figure 3, first column) is marked heteroscedasticity: the variability in the relationship between d_{BR} and the other effect size measures increases as d_{BR} increases in size. This heteroscedasticity can be explained by considering that large values of d_{BR} can occur due to large changes in performance, low baseline performance variability, or both. No other effect size measure uses baseline variability to index change (though notably, the mixed-effects models use it to estimate effect size uncertainty). While the initial motivation for standardizing within-case change by the baseline variability was to create a "standardized" metric that allowed for meta-analysis and comparison across studies (Gingerich, 1984, p. 75), the within-case standardized mean difference is sensitive to study design choices that should not affect an effect size (e.g., the number of baseline probe sessions, Pustejovsky, 2019), which makes comparison between studies, or even participants within the same study, tenuous at best.[7] In aphasia and related

[7] Even measures such as Cohen's *d*, for which the original within-case standardized mean difference was based on, have been criticized for not being as comparable across as is often assumed (see Baguley, 2011).

disorders, the d_{BR} statistic is also typically compared to a meta-analytic "benchmark" study (e.g., Bailey et al., 2015; Beeson & Robey, 2006). However, because the d_{BR} statistic is sensitive to differences in study designs, the benefit of such comparisons is likely limited at best.

Further complicating comparisons across studies or to meta-analytic benchmarks is the fact that the calculation of d_{BR} often varies between studies. Authors may choose to include some or all baseline or treatment observations, accommodate near-zero baseline variability by substituting baseline variance from other conditions or participants, or average two d_{BR} scores calculated within-list versus calculating a single d_{BR} statistic for each list. Wambaugh et al., (2017) calculated d_{BR} for each phoneme within each list, consistent with the SPT benchmark study (Bailey et al., 2015). However, calculating d_{BR} collapsing across the two phonemes (as is often done with semantic category in semantically focused treatments) would have resulted in a substantially larger average d_{BR} effect sizes for the treated (13.5 vs. 8.4) and generalization items (5.1 vs. 2.9). Participant 8 (treated, random condition) provides a clear example of the effect of averaging d_{BR} across two conditions versus collapsing performance before calculating d_{BR}. On the other hand, Wambaugh et al., (2017) used the last five baseline observations preceding the onset of the intervention rather than all baseline time points regardless of phase length (Bailey et al., 2015). Including all versus the last five baseline time points can influence the mean of baseline performance and baseline variability, and thus impact d_{BR}. For example, including only the last five baseline observations nearly doubles d_{BR} for participant 9 (treated, random condition), even though the average baseline performance during the last 5 observations is higher than the mean of the entire baseline phase. These differences limit direct comparison of d_{BR} across studies, and their effects are rarely discussed even when they are reported. To make d_{BR} comparable across studies and reproducible, researchers must minimally report the absolute change, baseline variability, and any deviations from established benchmark studies.

There are additional outstanding criticisms of d_{BR} that may negatively impact its utility, which have been described previously: (1) assumptions of independent observations and constant variance required by d_{BR} are rarely met in single subject designs (Howard et al., 2015); (2) benchmarks for interpreting d_{BR} must be established before d_{BR} can be interpreted; (3) d_{BR} is often applied to binomially distributed data, where the mean and standard deviation are related, thus introducing bias dependent on the level of baseline performance. (4) d_{BR} is influenced by autocorrelation (Archer et al., 2019), where performance at one probe session is correlated with the previous session. (5) d_{BR} does not account for baseline trends. We have included d_{BR} in this tutorial given its current widespread use. However, given these considerable limitations and methodological complications, we do not recommend its continued use in future studies.

Proportion of Potential Maximal Gain

A notable characteristic of PMG in Figure 3 is the similarity between PMG and the mixed-effect model effect sizes, particularly in terms of percent gain. This relationship is expected for studies such as Wambaugh et al., (2017), which use strict stimulus-selection methods that are matched to participant ability. When baseline performance is similar across individuals, PMG will be highly correlated with absolute change (e.g., percentage point gain estimated by the mixed-effects models). The downside of this approach is that where absolute change is equivalent, differences in PMG are attributable to differences in baseline performance. If a group of participants has the same absolute change (an increase of 10 out of 20 items), PMG can vary drastically: from 0.5 if a participant averages 0% correct at baseline to 1.0 if a participant averages 50% correct at baseline. For example, examine participants 11 and 12 (treated items, blocked condition, using the last 5 baseline observations) in the web-app. Both participants improve by roughly 9 items, but PMG = 0.95 for participant 11 and PMG = 0.62 for participant 12. The difference is driven only because participant 11 averaged about 50% (10 items) correct at baseline while participant 12 averaged about 25% (5 items) correct. The consequence of this feature is that for studies that use stringent stimulus selection procedures, PMG will largely index absolute change. Alternatively, for studies with more variability in baseline performance across participants (e.g., for studies providing the same items to all participants), PMG will be correlated with baseline severity even if severity does not moderate response to treatment. For such studies, finding that treatment was more beneficial for milder participants could simply be an artifact of the choice of PMG as an effect size. Ultimately, for appropriate comparison of performance between participants and across studies, researchers must consider these features of PMG.

PMG was intended to be used in analyses where participants received a different number of treated items or to account for baseline severity when the same items were assigned to all participants. PMG serves a similar purpose in this data, providing a measure of change on the same scale for all participants, even though 4 participants received fewer treated and untreated items in a modified protocol (see Wambaugh et al., 2016). As with d_{BR}, PMG does not provide a level of certainty; thus, estimating PMG cannot distinguish whether or not change is unlikely to occur by chance alone, or whether two values of PMG might be different. Ultimately, alternative methods of estimating individual effect sizes can better account for differences in item set size (e.g., mixed-effects models), have less potential for dependence on baseline severity, and include a measure of uncertainty. For these reasons, we recommend researchers pursue other effect size measures if applicable to their study design, particularly if they are interested in the relationship between aphasia severity and treatment response. If used, best practice would include reproducible analyses that report unstandardized change scores.

Tau-U

Tau-U is a non-parametric effect size measure designed to demonstrate the degree of overlap between phases but is not intended to distinguish between the magnitude of treatment response when there is no overlap. The large degree to which Tau-U diverges from other effect sizes is readily apparent in Figure 3, where there are clear ceiling effects. In other words, where Tau-U is equal to 1, there is a wide variation in the other effect size measures. These cases are characterized by no overlap between the baseline and treatment phases but widely varying degrees of change from baseline to the end of treatment. For example, compare participant 1 and participant 2 (treated items, blocked condition). Both participants have Tau-U scores = 1, but participant 1's absolute change is less than half of participant 2. This limitation of Tau-U has been discussed previously (Wolery et al., 2010) but is often overlooked. The consequence is that interpretations of Tau-U (e.g., large: 0.60 to 0.80, very large: 0.80 to 1, Vannest & Ninci, 2015) are not comparable to other effect sizes discussed here. For example, Tau-U fell in the "large" or "very large" range for nearly all treated conditions, but ranged from no effect to a large effect size by d_{BR} standards. Researchers, reviewers, and consumers should be aware of this conceptual difference when interpreting Tau-U and comparing effect sizes across the literature.

There have been a number of additional criticisms of Tau-U recently, summarized by Tarlow (2017): (1) Tau-U has inconsistent terminology and multiple mathematical definitions, which generate different values and thus require researchers to be aware of and transparent about which Tau-U they have employed, (2) While Tau-U$_{\text{A VS. B}}$ is bounded between -1 and 1, the baseline-corrected Tau-U$_{\text{A VS. B} - \text{TREND A}}$ is not, and can return inflated values ranging from -2 to 2 (3) Tau-U data cannot be visualized graphically (4) the degree of baseline correction is influenced by the ratio of the number of observations in the baseline phase to the treatment phase. Like d_{BR}, this final point is pertinent to studies such as Wambaugh et al., 2017, where the number of baseline points varies across participants to demonstrate experimental control and also varies based on whether an intervention was provided first or second within an individual participant. Moreover, the choice of cutoff for using a baseline correction can affect Tau-U estimates between participants with varying degrees of baseline trend within and across studies. Finally, there is no clear confidence interval for Tau-U$_{\text{A VS. B} - \text{TREND A}}$ (Pustejovsky et al., 2021), which is evident when using the `Tau()` versus `Tau_U()` functions in SingleCaseES.

Generalized Linear Mixed-effects models

While frequentist mixed-effects models address some of the limitations of d_{BR}, PMG, and Tau-U, difficulty obtaining convergence with item-level logistic mixed-effects models is a common occurrence in our experience. Wiley and Rapp (2018) suggest that individual models can be run to statistically examine change and estimate effect sizes for each participant. However, a substantial number of individual-level models in the Wambaugh et al., (2017) data set failed to converge, even with overly-simplified random effects structures, which are likely to return anti-conservative standard errors for repeated measures data. The reason for non-convergence and singular fit warnings in this data likely stem from two issues: relatively few items per list (especially in the case of generalization items) and some occasions of near-complete separation (i.e., performance at floor) during the baseline phase. The best practices for dealing with convergence and fit warnings are still a matter of debate (Meteyard & Davies, 2020). How researchers accommodate non-convergence and singular fit introduces additional “researcher degrees of freedom,” which can impact replication. These decisions add to the complexity of mixed-effect models and underscore the need for reproducible analysis when they are used.

One benefit of the interrupted time series model is the ability to adjust for a baseline trend, though determining the cases in which to do so can be challenging (Manolov et al., 2019). Calculating effect sizes using the more conservative method of projecting out performance to the end of treatment based on baseline trend may underestimate change if a baseline trend levels off and stabilizes before the start of an intervention. Alternatively, failing to account for a trend may attribute too much change to the onset of treatment (see participant 20, treated items, random condition). Visual or statistical analysis of the baseline trend may be useful, nothing that if there are less than 5 baseline measurement occasions, the trend may be unreliable (Huitema, 2011). Overall baseline trends may be poor predictors of ongoing performance if stability is reached in the final 3-5 baseline sessions; it may not be appropriate to extrapolate the baseline slope in these cases (see participant 16, treated items, random condition). The use of an unrelated "untreated" condition that is simply exposed to repeated probing through the intervention may further clarify whether or not it is necessary to extrapolate the baseline phase. If there is minimal change due to repeated probing in a balanced, untreated condition, a reasonable assumption may be that baseline trends reflect noise rather than a trend.

Another benefit specific to the generalized linear mixed-effects approach is the ability to report effect sizes in different units: logits, odds-ratios, percentage point gain, or the number of items gained, which can inform different research questions. Effect sizes in units of logits and odds-ratios are insensitive to item set size, unbounded, and, relative to percent correct or number correct units, place more value on change for individuals who perform closer to floor or ceiling at baseline. For example, logits and odds ratios will indicate that a 5-item change from 0/20 to 5/20 items or 15/20 to 20/20 items is greater than a 5-item change from 8/20 to 13/20 items. Percentage point gain is similarly agnostic to item set size, but will return the same effect size for the three cases noted above (a 25 percentage point improvement). Interpreting change in terms of the number of items gained will reflect differences in item set size, and similarly considers these three cases to be equivalent (a 5 item improvement). The non-linear relationships between these scales make them helpful for answering different research questions. For example, if a researcher is interested in comparing two studies that treat substantially different numbers of items, it may be desirable to use the number of items gained since the other units may overstate relative improvement in the study with fewer items.

Additionally, when baseline performance is highly variable across participants, using logits may be advantageous because they are less subject to floor and ceiling effects whereas percentage point gain or the number of items gained may unfairly penalize participants who perform well during the baseline phase, as they have less room for improvement. Logits may also be preferred when studying moderators of treatment effects because using bounded percent or number correct metrics can lead to out-of-bounds predictions or distortions in coefficient estimates. Finally, logits have a clearer basis for interpretation (i.e., the log-odds of a correct response) than standardized mean difference, proportion of maximal gain, or Tau-U.

Bayesian models, for both individual participants and groups, are often well-suited to smaller sample sizes and recommended for convergence and fit challenges in frequentist mixed-effects models (Bates et al., 2015). Bayesian models for groups of participants readily provide estimates of effect sizes and associated credible intervals for individuals, which allows researchers not only to establish whether a treatment worked on average, but also to identify the number of participants for whom there is reliable evidence of a treatment effect - a critical need for future aphasia research (Breitenstein et al., 2022). Estimating individual-level effects from a group model can improve reliability and reduce overfitting through partial pooling, where extreme observations are pulled towards the group average (Nalborczyk et al., 2019).

One criticism of Bayesian statistics concerns the use of priors, which some researchers argue can have an outsized influence on model results. In this tutorial, and papers used in our lab, we generally advocate for using "weakly regularizing" prior distributions, which improve model sampling and estimation, only assuming a range of plausible treatment effects centered around zero. When used appropriately, this feature of the Bayesian approach helps to constrain findings to what may be considered an a priori reasonable range of effects, reduce regression to the mean, and formally incorporate researchers' existing knowledge to increase the precision of model parameter estimates (Nalborczyk et al., 2019).

Neither the frequentist nor Bayesian implementations of the interrupted time series model (Huitema & Mckean, 2000) discussed in this tutorial explicitly account for temporal autocorrelation (e.g., the correlation between performance on adjacent sessions), though it may be advantageous to do so in future work. Additionally, the ability to detect reliable changes is dependent on sample size, in terms of the number of participants, items, and measurement occasions. Simulation-based power analysis may be used to anticipate the needed sample sizes and the growing number of studies using similar models (Evans et al., 2021; Swiderski et al., 2021), and the present data, should make power analysis for these models more feasible.

General Discussion

In this tutorial, we demonstrated how to conduct reproducible analysis of small-N treatment research using the statistical programming language R and how effect size selection and implementation can affect the interpretation and replication of findings within and across studies. This tutorial aims to serve as a starting place for researchers engaged in small-N studies to begin incorporating reproducible analyses into their regular workflow and better understand how their choice of effect size measure can threaten successful replication. Discussions of effect size strengths and weaknesses are intended to help researchers and clinicians be more informed consumers of this important body of research.

While we have described a number of differences across analytical approaches, all methods described here require researchers to make many small decisions in the process of analyzing small-N data. Which data points should researchers include in their analysis? Should we pool variance across phases, and how do we adjust d_{BR} for cases of no baseline variability? When should we correct for a baseline trend? How do we address mixed-effects model non-convergence or select appropriate random effects? Many of these decisions are difficult to anticipate a-priori during study conceptualization (and ideally, preregistration), yet they are critical for promoting successful study replication. The extent and influence of these degrees of freedom underlie the importance of reporting fully reproducible analyses, promoting transparency in small-N design research. Such dissemination also reduces barriers to successful meta-analysis of small-N designs, which is necessary for reaching scientific consensus.

This tutorial has focused on effect sizes common to the small-N literature in aphasia and related disorders, but there are alternative effect sizes with desirable qualities that are likely to be of interest to the field. For example, the gradual effects model (Swan & Pustejovsky, 2018) is able to capture nonlinear change under normal, count, and binomial data generating processes, and describe change in an interpretable, unstandardized effect size. The Log-response and Log-odds ratios also have desirable properties for estimating change across the binomial and count data typical of small-N studies in aphasia (Pustejovsky, 2018).

We offer modest guidance for researchers wondering about the "best" analytical approach and effect sizes for small-N designs in aphasia and related disorders, based on our discussion and comparison of effect sizes in Wambaugh et al., (2017). First and foremost, researchers should select effect sizes that align with their research questions and are well-suited to their study design. Any discontinuity between the research question and statistical method limits the conclusion drawn from the study. Second, while all effect sizes reviewed in this tutorial have limitations, researchers should be particularly cautious in their use of d_{BR} and PMG, given the lack of established confidence intervals and sensitivity to experimental manipulations. The Tau-U statistics are well supported in the single-case experimental design literature, but only describe the degree of non-overlap between phases rather than the magnitude of treatment response.

Of the effect sizes common to small-N studies in aphasia and related disorders, we recommend using mixed-effects models, which can generate effect sizes that are accompanied by uncertainty and are more robust to experimental manipulations. While this tutorial was intended to make mixed-effects models more approachable, we recognize that they are complex and require additional statistical expertise. The alternative approaches noted (i.e., the gradual effects model, log-response and log-odds ratios) above may strike a better compromise between complexity and rigor but are outside of the scope of this tutorial. Ultimately, the reality is that choosing the "best" effect size is highly context-dependent. Researchers must be knowledgeable about the strengths and weaknesses of their chosen method and transparent about how their methodological decisions and the choice of statistical method might influence their conclusions.

Pursuing reproducible research using script-based approaches is a critical first step in addressing the challenges common to analyzing small-N studies in aphasia and related disorders. Given the impact of small-N studies on clinical rehabilitation services, we argue that sharing data and script-based analyses is research best practice and should be a minimum standard for our field. Pairing reproducible analyses with informed selection of effect sizes can improve scientific rigor and transparency and the mapping between research questions and analytical techniques and facilitate more robust tests of conceptual replications across studies.

Acknowledgements

Research reported in this manuscript was supported by the National Institute of Deafness and Communication Disorders of the National Institutes of Health under award number F31DC019853-01 awarded to Robert Cavanaugh. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The contents of this manuscript were also developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR) awarded to Yina Quique. NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this manuscript do not necessarily represent the policy of NIDILRR, ACL, or HHS, and you should not assume endorsement by the Federal Government.

References

Antonucci, S., & Gilmore, N. (2019). *Do aphasia core outcome sets require core analysis sets: Where do we go from here in single subject design research?* Roundtable presented at the 49th Clinical Aphasiology Conference, Whitefish, MT.

Archer, B., Azios, J. H., Müller, N., & Macatangay, L. (2019). Effect Sizes in Single-Case Aphasia Studies: A Comparative, Autocorrelation-Oriented Analysis. *Journal of Speech, Language, and Hearing Research*, 1–10. https://doi.org/10.1044/2019_JSLHR-L-18-0186

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bailey, D. J., Eatchel, K., & Wambaugh, J. (2015). Sound Production Treatment: Synthesis and quantification of outcomes. *American Journal of Speech-Language Pathology*, *24*(4), S798–S814.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review*, *16*(4), 161–169. https://doi.org/10.1007/s11065-006-9013-7

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: What's in a name? *Topics in Stroke Rehabilitation*, *17*(6), 411–422. https://doi.org/10.1310/tsr1706-411

Breitenstein, C., Hilari, K., Menahemi-Falkov, M., L. Rose, M., Wallace, S. J., Brady, M. C., Hillis, A. E., Kiran, S., Szaflarski, J. P., Tippett, D. C., et al. (2022). Operationalising treatment success in aphasia rehabilitation. *Aphasiology*, 1–40.

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In *Single-case research designs and analysis: New directions for psychology and education* (pp. 187–212).

Evans, W. S., Cavanaugh, R., Quique, Y., Boss, E., Starns, J. J., & Hula, W. D. (2021). Playing With BEARS: Balancing Effort, Accuracy, and Response Speed in a Semantic Feature Verification Anomia Treatment Game. *Journal of Speech, Language, and Hearing Research*, *64*(8), 3100–3126. https://doi.org/10.1044/2021_JSLHR-20-00543

Gilmore, N., Meier, E. L., Johnson, J. P., & Kiran, S. (2020). Typicality-based semantic treatment for anomia results in multiple levels of generalisation. *Neuropsychological Rehabilitation*, *30*(5), 802–828. https://doi.org/10.1080/09602011.2018.1499533

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *Journal of Applied Behavioral Science*, *20*(1), 71–79. https://doi.org/10.1177/002188638402000113

Gordon, K. R. (2019). How mixed-effects modeling can advance our understanding of learning and memory and improve clinical and educational practice. *Journal of Speech, Language, and Hearing Research*, *62*(3), 507–524.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., et al. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448.

Harel, D., & McAllister, T. (2019). Multilevel models for communication sciences and disorders. *Journal of Speech, Language, and Hearing Research*, *62*(4), 783–801.

Howard, D., Best, W., & Nickels, L. (2015). Optimising the design of intervention studies: Critiques and ways forward. *Aphasiology*, *29*(5), 526–562. https://doi.org/10.1080/02687038.2014.985884

Huitema, B. E. (2011). *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies* (2nd ed.). Wiley-Blackwell.

Huitema, B. E., & Mckean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, *60*(1), 38–58.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, *14*(5), e1002456.

Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). *Single-case intervention research: Methodological and statistical advances.* American Psychological Association. https://doi.org/10.1037/14376-000

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206.

Lambon Ralph, M. A., Snell, C., Fillingham, J. K., Conroy, P., & Sage, K. (2010). Predicting the outcome of anomia therapy for people with aphasia post CVA: Both language and cognitive status are key predictors. *Neuropsychological Rehabilitation*, *20*(2), 289–305. https://doi.org/10.1080/09602010903237875

Lazar, R. M., Minzer, B., Antoniello, D., Festa, J. R., Krakauer, J. W., & Marshall, R. S. (2010). Improvement in aphasia scores after stroke is well predicted by initial severity. *Stroke*, *41*(7), 1485–1488.

Lee, J. B., & Cherney, L. R. (2018). Tau-U: A Quantitative Approach for Analysis of Single-Case Experimental Data in Aphasia. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, *27*, 495–503. https://doi.org/10.1044/2017_AJSLP-16-0197

Lenth, R. V. (2021). *Emmeans: Estimated marginal means, aka least-squares means* [Manual]. https://CRAN.R-project.org/package=emmeans

Manolov, R., Solanas, A., & Sierra, V. (2019). Extrapolating baseline trend in single-case data: Problems and tentative solutions. *Behavior Research Methods*, *51*(6), 2847–2869.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

Murray, E., Power, E., Togher, L., McCabe, P., Munro, N., & Smith, K. (2013). The reliability of methodological ratings for speechBITE using the PEDro-P scale. *International Journal of Language & Communication Disorders*, *48*(3), 297–306.

Nalborczyk, L., Batailler, C., Loevenbruck, H., Vilain, A., & Bürkner, P. C. (2019). An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Nickels, L., Rapp, B., & Kohnen, S. (2015). Challenges in the use of treatment to investigate cognition. *Cognitive Neuropsychology*, *32*(3-4), 91–103. https://doi.org/10.1080/02643294.2015.1056652

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *Elife*, *6*, e23383.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*(2), 284–299. https://doi.org/10.1016/j.beth.2010.08.006

Portney, L. G., & Watkins, M. P. (2015). *Foundations Of Clinical Research: Applications To Practice* (2nd ed.). Prentice Hall.

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99–112.

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, *24*(2), 217.

Pustejovsky, J. E., Chen, M., & Swan, D. M. (2021). *SingleCaseES: A calculator for single-case effect sizes* (R package version 0.5.0) [Computer software]. https://CRAN.R-project.org/package=SingleCaseES

R Core Team. (2020). *R: A language and environment for statistical computing* (4.0.3 ed.). R Foundation for Statistical Computing. https://www.r-project.org/

Rapp, B. (2011). Case series in cognitive neuropsychology: Promise, perils, and proper perspective. *Cognitive Neuropsychology*, *28*(7), 435–444.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. Wiley.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Snell, C., Sage, K., & Lambon Ralph, M. A. (2010). How many words should we provide in anomia therapy? A meta-analysis and a case series study. *Aphasiology*, *24*(9), 1064–1094. https://doi.org/10.1080/02687030903372632

Strand, J. (2021). *Error tight: Exercises for lab groups to prevent research mistakes*. https://doi.org/10.31234/osf.io/rsn5y

Swan, D. M., & Pustejovsky, J. E. (2018). A gradual effects model for single-case designs. *Multivariate Behavioral Research*, *53*(4), 574–593.

Swiderski, A. M., Quique, Y. M., Dickey, M. W., & Hula, W. D. (2021). Treatment of underlying forms: A bayesian meta-analysis of the effects of treatment and person-related variables on treatment response. *Journal of Speech, Language, and Hearing Research*, *64*(11), 4308–4328.

Tarlow, K. R. (2017). An Improved Rank Correlation Effect Size Statistic for Single-Case Designs: Baseline Corrected Tau. *Behavior Modification*, *41*(4), 427–467. https://doi.org/10.1177/0145445516676750

Thompson, C. K. (2015). Establishing the effects of treatment for aphasia using single-subject-controlled experimental designs. *Aphasiology*, *29*(5), 588–597.

Togher, L., Schultz, R., Tate, R., McDonald, S., Perdices, M., Smith, K., Winders, K., & Savage, S. (2009). The methodological quality of aphasia therapy research: An investigation of group studies using the PsycBITETM evidence-based practice database. *Aphasiology*, *23*(6), 694–706.

Vannest, K. J., & Ninci, J. (2015). Evaluating Intervention Effects in Single-Case Research Designs. *Journal of Counseling & Development*, *93*(4), 403–411. https://doi.org/10.1002/jcad.12038

Wambaugh, J. L., Nessler, C., Wright, S., & Mauszycki, S. C. (2014). Sound Production Treatment: Effects of blocked and random practice. *American Journal of Speech-Language Pathology*, *23*(2), S225–S245.

Wambaugh, J. L., Nessler, C., Wright, S., Mauszycki, S. C., DeLong, C., Berggren, K., & Bailey, D. J. (2017). Effects of blocked and random practice schedule on outcomes of Sound Production Treatment for acquired apraxia of speech: Results of a group investigation. *Journal of Speech, Language, and Hearing Research*, *60*, 1739–1751.

Wambaugh, J. L., Nessler, C., Wright, S., Mauszycki, S., & DeLong, C. (2016). Sound Production Treatment for acquired apraxia of speech: Effects of blocked and random practice on multisyllabic word production. *International Journal of Speech-Language Pathology*, *18*(5), 450–464.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Wiley, R. W., & Rapp, B. (2018). Statistical analysis in Small-N Designs: Using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology*, *33*(1), 1–30. https://doi.org/10.1080/02687038.2018.1454884

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of Overlap Methods for Quantitatively Synthesizing Single-Subject Data. *The Journal of Special Education*, *44*(1), 18–28. https://doi.org/10.1177/0022466908328009

Table and Figure Captions

Figure 1. Participant 10 performance during baseline and treatment phase for the blocked condition. Dark circles indicate data points used to calculate dBR and PMG.

Figure 2. Participant 10 performance during baseline and treatment phase from Wambaugh et al., (2017). Plot annotations indicate Huitema & McKean (2000) model coefficients.

Figure 3. Relationships between individual effect size measures typically used in aphasia small-N studies.

Table 1. Variables and descriptions for study data from Wambaugh et al., (2017)

Table 2. The first 5 rows of data

Table 3. Summary of major strengths and weaknesses of each effect size measure reviewed