Calculating Effect Sizes for Single-Case Research

An Introduction to the SingleCaseES and scdhlm Web Applications and R Packages

James E. Pustejovsky
University of Wisconsin - Madison

April 26, 2023
Acknowledgement

- The work reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grants R324U190002 and R305D190023. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education.
Outline

1. Organizing and curating data from single-case designs.

2. Within-study effect sizes.
   - Background
   - The SingleCaseES app

3. Between-case standardized mean differences
   - Background
   - The scdhlm app
Organizing and curating data from single-case designs
Why organize and curate your data?

1. So that you can do statistical analysis and effect size calculations.

2. So that you can share your data.
   - Make it easily accessible for inclusion in systematic reviews!

3. Because graphing data usually involves loss of information.

4. To fully document your research study.
Tidy SCD data

- One column per variable
- One row per observation session
- Descriptive labels for
  - Case (participant)
  - Phase of design or treatment condition

<table>
<thead>
<tr>
<th>Case</th>
<th>Phase</th>
<th>Session</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deborah's</td>
<td>Baseline</td>
<td>1</td>
<td>62.63</td>
</tr>
<tr>
<td>Group</td>
<td>Baseline</td>
<td>2</td>
<td>40.22</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>3</td>
<td>54.26</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>4</td>
<td>40.26</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>5</td>
<td>46.82</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>6</td>
<td>52.45</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>7</td>
<td>25.37</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>8</td>
<td>26.32</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>9</td>
<td>7.65</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>10</td>
<td>11.41</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>11</td>
<td>13.30</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>12</td>
<td>22.66</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>13</td>
<td>13.34</td>
</tr>
<tr>
<td>Amy's Group</td>
<td>Baseline</td>
<td>1</td>
<td>16.67</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>2</td>
<td>28.43</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>3</td>
<td>29.41</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>4</td>
<td>30.39</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>5</td>
<td>45.10</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>6</td>
<td>37.25</td>
</tr>
</tbody>
</table>
Multiple dependent variables

- **Wide format**: Use separate columns for multiple outcome variables

<table>
<thead>
<tr>
<th>Case</th>
<th>Phase</th>
<th>Session</th>
<th>Problem_Behavior</th>
<th>On_Task_Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>1</td>
<td>16.7</td>
<td>56.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>2</td>
<td>20.0</td>
<td>70.0</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>3</td>
<td>26.7</td>
<td>66.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>4</td>
<td>20.0</td>
<td>86.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>5</td>
<td>16.7</td>
<td>56.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>6</td>
<td>13.3</td>
<td>70.0</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Intervention</td>
<td>7</td>
<td>16.7</td>
<td>46.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Intervention</td>
<td>8</td>
<td>20.0</td>
<td>73.3</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Intervention</td>
<td>9</td>
<td>20.0</td>
<td>56.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Intervention</td>
<td>10</td>
<td>30.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Intervention</td>
<td>11</td>
<td>30.0</td>
<td>63.3</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Intervention</td>
<td>12</td>
<td>13.3</td>
<td>63.3</td>
</tr>
</tbody>
</table>
Multiple dependent variables

- **Long format**: One row per outcome measure per session

<table>
<thead>
<tr>
<th>Case</th>
<th>Phase</th>
<th>Session</th>
<th>DV</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>1</td>
<td>On Task Behavior</td>
<td>56.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>1</td>
<td>Problem Behavior</td>
<td>16.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>2</td>
<td>On Task Behavior</td>
<td>70.0</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>2</td>
<td>Problem Behavior</td>
<td>20.0</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>3</td>
<td>On Task Behavior</td>
<td>66.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>3</td>
<td>Problem Behavior</td>
<td>26.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>4</td>
<td>On Task Behavior</td>
<td>86.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>4</td>
<td>Problem Behavior</td>
<td>20.0</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>5</td>
<td>On Task Behavior</td>
<td>56.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>5</td>
<td>Problem Behavior</td>
<td>16.7</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>6</td>
<td>On Task Behavior</td>
<td>70.0</td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>6</td>
<td>Problem Behavior</td>
<td>13.3</td>
</tr>
</tbody>
</table>
Adding more detail

- Add further details about what happened in the study.

- Some ideas:
  - Actual session date + times (YYYY-MM-DD-HH:MM)
  - Observation session lengths
  - Clinician/therapist IDs
  - Notes about events

<table>
<thead>
<tr>
<th>Case</th>
<th>Phase</th>
<th>Session</th>
<th>Problem Behavior</th>
<th>On-Task Behavior</th>
<th>Date</th>
<th>Session length</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>1</td>
<td>16.7</td>
<td>56.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>2</td>
<td>20.0</td>
<td>70.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deborah's Group</td>
<td>Baseline</td>
<td>3</td>
<td>26.7</td>
<td>66.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Share Your Data!
Within-case effect size indices

PND  PAND  IRD  PEM  LRR  POGO  SMD(within)  Tau-U  Tau(AB)  NAP
Within-case effect size indices

- Single-number summary of the **direction** and **magnitude** of intervention effect (functional relation) **for each case** within a study.

- Use these if you want to:
  - Describe results separately for each participant
  - Examine heterogeneity of effects or associations with individual-level characteristics
  - Compare results across participants and SCED studies that use various outcome measures

- Lots of proposed effect size indices. Today we'll focus on
  - Non-overlap of all pairs
  - Within-case standardized mean difference
  - Log-response ratio
SingleCaseES: Single-series calculator


- Or by opening RStudio and typing

```r
library(SingleCaseES)
SCD_effect_sizes()
```

- Two parts to the app:
  - Single-series calculator (direct data entry)
  - Multiple-series calculator (using a data file)
Simplest possible model

- Stable baseline and treatment phases (no time trends)
  - Immediate shift in level due to intervention
- Independence of outcome measurements
**Notation**

- \( n_A \) observations in phase A: \( y_1^A, \ldots, y_{n_A}^A \)

- \( n_B \) observations in phase B: \( y_1^B, \ldots, y_{n_B}^B \)

- Mean level of the outcome in each phase: \( \mu_A, \mu_B \)
  - Estimated by sample means \( \bar{y}_A, \bar{y}_B \)

- Standard deviation of the outcome in each phase: \( \sigma_A, \sigma_B \)
  - Estimated by sample standard deviations \( S_A, S_B \)
Non-overlap of all pairs

- Non-overlap measures are defined in terms of *ordinal comparisons* of outcomes.

- Non-overlap of all pairs (Parker and Vannest, 2009) is defined in terms of all pairs of one observation from phase A and one observation from phase B.

- For every pair $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$, take
  
  $$q_{ij} = \begin{cases} 
  1 & \text{if } y_j^B \text{ better than } y_i^A \\
  \frac{1}{2} & \text{if } y_j^B = y_i^A \\
  0 & \text{if } y_j^B \text{ worse than } y_i^A 
  \end{cases}$$

- NAP estimator:

  $$NAP = \frac{1}{mn} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} q_{ij}$$

- Standard error based on unbiased estimator (Sen, 1967; Mee, 1990)
  - Methods assume that observations are independent and identically distributed within each phase.
Limited range of sensitivity

- Limited range where NAP (and other non-overlap measures) sensitive to change.
Within-case standardized mean difference

- Proposed by Gingerich (1984) and Busk and Serlin (1992)
- Parameter definition:

\[ \delta = \frac{\mu_B - \mu_A}{\sigma_A} \]

- Difference in means, "standardized" by baseline variation
- NOT equivalent to between-case SMD because \( \sigma_A \) includes only within-case variation.

- Appropriate for interval-scale outcomes
  - Is variability of outcomes approximately constant for different mean levels?
  - Standardizing by within-case variation means this measure will be strongly affected by reliability of measurements
  - Problematic for outcomes with restricted range in baseline
Within-case standardized mean difference: estimation

- Originally proposed estimator:
  \[ d = \frac{\bar{y}_B - \bar{y}_A}{S_A} \]

- Estimator with small-sample bias correction:
  \[ g = \left( 1 - \frac{3}{4n_A - 5} \right) \times \frac{\bar{y}_B - \bar{y}_A}{S_A} \]

- Approximate standard error, assuming independent observations:
  \[ SE_g = \left( 1 - \frac{3}{4n_A - 5} \right) \sqrt{\frac{1}{n_A} + \frac{S_B^2}{n_B S_A^2} + \frac{d^2}{2(n_A - 1)}} \]
Proportional change in levels

- Percentage (proportional) change from baseline to intervention is an easily interpretable "informal" effect size (Campbell and Herzinger, 2010).

- The log response ratio is a formal measure of effect size that describes change in proportional terms (Pustejovsky, 2015; Pustejovsky, 2018).

- Parameter definition:

\[ \psi = \log \left( \frac{\mu_B}{\mu_A} \right) \]

  - Appropriate for ratio-scale outcomes (frequency counts, percentage duration)

  - Natural logarithm is used to make the range unrestricted.

- Transformation to percentage change:

\[ \% \text{ change} = 100\% \times (e^{\psi} - 1) \]
Log response ratio: estimation

- Basic estimator (biased if $m$ or $n$ is small):

$$ R_1 = \log \left( \frac{\tilde{y}_B}{\tilde{y}_A} \right) $$

- Bias-corrected estimator:

$$ R_2 = \log \left( \frac{\tilde{y}_B}{\tilde{y}_A} \right) + \frac{\tilde{S}_B^2}{2n_B\tilde{y}_B^2} - \frac{\tilde{S}_A^2}{2n_A\tilde{y}_A^2} $$

- Approximate standard error for $R_2$, assuming independent observations:

$$ SE_R = \sqrt{\frac{\tilde{S}_A^2}{n_A\tilde{y}_A^2} + \frac{\tilde{S}_B^2}{n_B\tilde{y}_B^2}} $$
Direction of improvement

- Two versions of LRR:
  - **LRRi**: Positive numbers represent *increases* in *desirable outcomes*
  - **LRRd**: Negative numbers represent *decreases* in *undesirable outcomes*
- Use the version that corresponds to *predominant valence* of outcomes in your data.
- For count outcomes, LRRi = -LRRd
- For proportion / percentage outcomes, the outcome valence is harmonized before calculation.
  - For proportion / percentage outcomes, LRRi ≠ -LRRd
Truncation constants

- If $\bar{y}_A = 0$ or $\bar{y}_B = 0$ then LRR is undefined.
  - If $S^2_A = 0$ or $S^2_B = 0$ then $SE_R$ is undefined
- To handle such situations, the app uses truncated mean and truncated SD estimators:

  $$\bar{y}_A = \max \left \{ \bar{y}_A, \frac{1}{2n_AD} \right \}, \quad \bar{y}_B = \max \left \{ \bar{y}_B, \frac{1}{2n_BD} \right \}$$

  and

  $$\tilde{S}^2_A = \max \left \{ S^2_A, \frac{1}{n_AD^2} \right \}, \quad \tilde{S}^2_B = \max \left \{ S^2_B, \frac{1}{n_BD^2} \right \}$$

- $D$ is a constant that depends on the outcome scale and measurement procedures
  - Number of intervals / items
  - Session length for direct observation
  - Can also define your own $D$
SingleCaseES: Multiple-series calculator

- Basic walk-through with data from Rodriguez and Anderson (2014)
- Calculating phase-pairs in ABAB designs
- Aggregating effect sizes
Rodriguez and Anderson (2014)

Integrating a social behavior intervention during small group academic instruction using a total group criterion intervention
Lambert, Cartledge, Heward et al. (2006)

Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students
Calculating phase pairs

- Might want to calculate effect sizes for adjacent pairs of baseline and intervention phases.
- SingleCaseES provides an option to determine phase pairs automatically.
Aggregating effect sizes

- After calculating effect sizes for adjacent pairs of phases, we might want to **average them together** to simplify reporting or further analysis.
  - Average across phase pairs in an ABAB design
  - Average across cases to generate an overall summary effect size estimate
- Several options for taking weighted averages
  - Equal weighting
  - Inverse-variance weighting: $\frac{1}{V}$ (use for LRR)
    - $n_A$
    - $n_B$
  - $n_A n_B$ (use for NAP)
    - $\frac{1}{n_A} + \frac{1}{n_B}$ (use for SMD)
# Load packages
library(SingleCaseES)

# Load data
library(readxl)
library(janitor)

dat <- read_excel(path = "Small-is-Beautiful-effect-size-workshop.xlsx", sheet = "Lambert") %>%
clean_names(case = "parsed")

# Clean data
library(dplyr)

dat <-
  dat %>%
    group_by(case) %>%
    mutate(phase_pair_calculated = calc_phase_pairs(treatment, session = day)) %>%
    ungroup()

# Batch calculation
res <- batch_calc_ES(dat = dat,
  grouping = c(case),
  condition = treatment,
  outcome = outcome,
  aggregate = c(phase_pair_calculated),
  weighting = "1/nA + 1/nB",
  session_number = day,
  baseline_phase = "SSR",
  interaction = "EE"
)
Between-case standardized mean differences
Premises

- **Goal:** Estimate an effect size using data from a single-case design that is *in the same metric* as the standardized mean difference effect size from a between-group experimental design.

- **Why?** *(Shadish, Hedges, Horner et al., 2015)*
  - **Translation** of single-case research for researchers who work primarily with between-groups designs
  - **Comparison** of results from single-case studies and between-groups studies, for purposes of understanding the utility and limitations of each type of design
  - **Synthesis** involving both single-case and between-groups designs
SMD in between-group experiment

• What is the SMD from a between-group experiment?

\[ \delta_{BC} = \frac{\left( \text{Average outcome if everybody gets intervention} \right) - \left( \text{Average outcome if nobody gets intervention} \right)}{\left( \text{SD of outcome if nobody gets intervention} \right)} \]

\[ \delta_{BC} = \frac{\left( \text{Average outcome if everybody gets intervention} \right) - \left( \text{Average outcome if nobody gets intervention} \right)}{\sqrt{\left( \text{Between-participant variance} \right) + \left( \text{Within-participant variance} \right)}} \]

• We aim to estimate these component quantities using data from a single-case experimental design.
The broad strategy

(Pustejovsky, Hedges, and Shadish, 2014) described a general strategy for estimating BC-SMD:

1. Develop a hierarchical linear model that describes:
   - The form of time trends and intervention effects
   - How the trends and intervention effects vary across participants

2. Imagine a hypothetical between-group experiment with the same population of participants, same intervention, same dependent variable.
   - When is treatment initiated?
   - When are outcomes assessed?

3. Use the hierarchal model to estimate the components of $\delta_{BC}$ for the hypothetical experiment.

4. Make a small-sample correction (similar to Hedges' $g$)
Design translation

A multiple baseline across participants:

<table>
<thead>
<tr>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
<th>T12</th>
<th>T13</th>
<th>T14</th>
<th>T15</th>
<th>T16</th>
<th>T17</th>
<th>T18</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

A hypothetical between-group design (with pre-test):

<table>
<thead>
<tr>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
<th>T12</th>
<th>T13</th>
<th>T14</th>
<th>T15</th>
<th>T16</th>
<th>T17</th>
<th>T18</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overview of methods literature

- BC-SMD estimators for a basic hierarchical linear model with no time trends:
  - Hedges, Pustejovsky, and Shadish (2012): Treatment reversal (ABAB) design replicated across 3+ participants
  - Hedges, Pustejovsky, and Shadish (2013): Multiple baseline / multiple probe design with 3+ participants
- Pustejovsky, Hedges, and Shadish (2014) described a more general strategy for multiple baseline / multiple probe designs across participants
  - Valentine, Tanner-Smith, Pustejovsky et al. (2016): Tutorial and practical guidance
- Swaminathan, Rogers, and Horner (2014) proposed Bayesian estimation methods
- Chen, Pustejovsky, Klingbeil et al. (2023) proposed BC-SMD methods for more complex designs:
  - Multiple baseline across behaviors, replicated across 3+ participants
  - Clustered multiple baseline design across participants (3+ clusters)
  - Multivariate multiple baseline design across 3+ participants
scdhlm web app

- Or by opening RStudio and typing

```r
library(scdhlm)
shine_scd()
```
Rodriguez and Anderson (2014)

Integrating a social behavior intervention during small group academic instruction using a total group criterion intervention.
The most basic HLM

- Level-1 model for each participant:

\[ Y_{ij} = \beta_{0j} + \beta_{1j}(Tx)_{ij} + e_{ij} \]

where \( \text{Var}(e_{ij}) = \sigma_e^2 \) and \( e_{1j}, \ldots, e_{Tj} \sim AR_1(\phi) \)

- Level-2 model:

\[
\begin{align*}
\beta_{0j} &= \theta_{00} + u_{0j}, \quad u_{0j} \sim N(0, \sigma_{u0}^2) \\
\beta_{1j} &= \theta_{10}
\end{align*}
\]

- Under this model:

  - Average outcome if nobody gets intervention: \( \theta_{00} \)
  - Average outcome if everybody gets intervention: \( \theta_{00} + \theta_{10} \)
  - SD of outcome if nobody gets intervention: \( \sqrt{\sigma_{u0}^2 + \sigma_e^2} \)
  - BC-SMD effect size: \( \delta_{BC} = \frac{\theta_{10}}{\sqrt{\sigma_{u0}^2 + \sigma_e^2}} \)
\[ \hat{\theta}_{00} = 37.5 \]
\[ \hat{\theta}_{10} = -24.7 \]
\[ \hat{\sigma}^2_e = 112.1 \]
\[ \hat{\sigma}^2_{u0} = 36.7 \]

\[ \hat{\delta}_{BC} = \frac{-24.7}{\sqrt{112.1 + 36.7}} = -2.03 \]

\[ g_{BC} (SE) = -1.99 (0.31) \]
A more flexible HLM

- Level-1 model for each participant:
  \[ Y_{ij} = \beta_{0j} + \beta_{1j}(\text{Time})_{ij} + \beta_{2j}(Tx)_{ij} + \beta_{3j}(Tx)_{ij} \times ((\text{Time})_{ij} - k_j) + e_{ij} \]
  where \( \text{Var}(e_{ij}) = \sigma_e^2 \) and \( e_{1j}, \ldots, e_{Tj} \sim AR_1(\phi) \) and \( k_j \) is last baseline session.

- Level-2 model:
  \[
  \begin{align*}
  \beta_{0j} &= \theta_{00} + u_{0j}, \quad u_{0j} \sim N(0, \sigma_{u0}^2) \\
  \beta_{1j} &= \theta_{10} + u_{1j} \\
  \beta_{2j} &= \theta_{20} + u_{2j} \\
  \beta_{3j} &= \theta_{30} + u_{3j}
  \end{align*}
  \]
  - Adding a random effect \( \rightarrow \) allowing slope / Tx effect to vary across cases
  - Omitting a random effect \( \rightarrow \) assuming slope / Tx effect is constant
  - Models with more random effects require more cases
A more flexible HLM

- Can also modify assumptions about level-1 errors
- Different variance by treatment phase:

\[
\text{Var}(e_{ij}) = \begin{cases} 
\sigma^2_{eC} & \text{if } (Tx)_{ij} = 0 \\
\sigma^2_{eT} & \text{if } (Tx)_{ij} = 1 
\end{cases}
\]

- Correlation structure of level-1 errors:
  - First order auto-regression \((AR_1(\phi))\)
  - First order moving average \((MA_1(\phi))\)
  - Independent

Model should be informed by **theoretical expectations** and **visual inspection**
Barton-Arwood, Wehby, and Falk (2005) Reading instruction for elementary-age students with emotional and behavioral disorders: Academic and behavioral outcomes
Models with time trends

- For models with time trends, we need to specify *timing* of pre-test and post-test for the hypothetical between-group design.

- **Initial treatment time**: Last session of baseline phase before being assigned to intervention or comparison condition.
  - Default: Length of shortest baseline phase

- **Focal follow-up time**: Session during which outcomes would be assessed in hypothetical experiment.
  - Default: Last measurement occasion for first case to enter intervention
  - This is not a particularly good default
  - Ideally, pick a focal follow-up time based on a meaningful or typical treatment duration
Barton-Arwood, Wehby, and Falk (2005) effect size calculations

- Model specification
  - Baseline level (random)
  - Baseline time trends (constant)
  - Treatment level change (constant)
  - Treatment trend change (random)
  - Level-1 variance differs by phase
- Initial treatment time: After 6 sessions
- Focal follow-up time of session 16 (10 sessions of treatment).
- BC-SMD estimate:

\[ g_{BC} (SE) = 0.82 (0.75) \]
Illustrative application of BC-SMDs


- Calder, Claessen, Ebbels et al. (2020): multiple baseline across nine participants
  - Data available in the Excel workbook
  - Try calculating a BC-SMD estimate after 10 weeks of intervention

- Calder, Claessen, Ebbels et al. (2021): crossover randomized trial with $N = 21$ participants
  - 10 weekly intervention sessions
  - $g = 1.97, SE = 0.11$ for expressive morphosyntax
  - $g = 0.06, SE = 0.06$ for grammaticality judgements
Limitations of between-case SMD

- Tool for translating from single-case logic to group-design logic.
  - Premised on the idea that a hypothetical group design is theoretically plausible
- Describes a **summary, average effect** across a set of cases
  - Potentially concealing individual-level heterogeneity
- For some models, magnitude depends on the features (timing) of hypothetical between-group design
- Technical limitations
  - Only available for some designs
  - Requires at least 3 participants (preferably more!)
  - Models assume normal (Gaussian) errors
  - Care needed for model selection
References


References

