Quality Appraisal of Single-Case Experimental Designs: Current Issues and Future Directions

Oliver Wendt, PhD

Campbell Disability Coordinating Group
Department of Communication Sciences and Disorders
University of Central Florida
Single-case Experimental Designs (SCEDs)

- Examines pre- versus post-treatment performance within a small sample (Kennedy, 2005)
- Employs repeated and reliable measurement, within- and between-subject comparisons
- Participants serve as their own control
- Compares performance prior to intervention to performance during/after intervention
- Basis for determining treatment efficacy, used to establish evidence-based practice (Horner et al., 2005)
Importance of Quality Appraisal

- Process of examining reliability, internal and external validity of research report
- Consumers of research need to evaluate the methodological rigor of any single-case experimental design (SCED)
- Applied researchers aim to synthesize SCEDs in a systematic review to assess study quality and assign more weight to sound studies (Petticrew & Roberts, 2006)
- Critical appraisal of SCED has been largely overlooked until recently
Three Major Projects

1. **Comparison of Quality Appraisal Tools**
   - Evaluating quality of SCDs crucial for research synthesis and documenting evidence-based practice

2. **Review of Randomization and Data-Analysis Items**
   - More emphasis on statistical analysis and synthesis of SCEDs

3. **Appraisal of Comparative SCEDs**
   - Comparative designs demand specific criteria
GEIS 2018: Quality Appraisal of Single-Case Experimental Designs

COMPARISON OF QUALITY APPRAISAL TOOLS
<table>
<thead>
<tr>
<th>Characteristics/Properties*</th>
<th>Tool (maximum quality score)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Composition of tool</strong></td>
<td></td>
</tr>
<tr>
<td>Ranks certainty of evidence as “conclusive” (highest), “preponderant”, “suggestive”, or “inconclusive” (lowest), based on research design, interobserver agreement of dependent variable, and treatment integrity</td>
<td>EVIDAAC Scales (max. = 10 or 19) One treatment scale: 10 items; two or more treatments scale: 19 items; higher score = higher quality. Logan et al. Scale (max. = 14) 14 questions containing 16 items; studies are rated “strong” (11-14 points), “moderate” (7-10 points), or “weak” (less than 7 points). SCED Scale (max. = 10) 11-item rating scale; item 1 assesses clinical history information; items 2-11 allow calculation of quality score; higher score = higher quality. Smith et al. Scale (max. = 15) 15-item rating scale; higher score = higher quality. WWC Standards (no max. score) Design Standards, rank internal validity as “Meets Standards”, “Meets Standards with Reservations”, and “Does not Meet Standards”; Evidence of Effect Standards rate effects strength as (1) “Strong Evidence”, (2) “Moderate Evidence,” or (3) “No Evidence”.</td>
</tr>
<tr>
<td><strong>Certainty Framework (no max. score)</strong></td>
<td>Evaluative Method (max. = 12) 12-item rating scale divided into primary and secondary indicators; strength of research ranked “strong”, “adequate”, or “weak” based on number and level of indicators achieved.</td>
</tr>
<tr>
<td>Content validity established</td>
<td>No</td>
</tr>
<tr>
<td>Inter-rater reliability provided</td>
<td>No</td>
</tr>
</tbody>
</table>

*An extended version of this table containing further details on the various tools is available from the first author upon request.*
Application of Appraisal Tools

Small field test was conducted to compare seven appraisal tools:

- Four SCED treatment articles
- Each one representing one of the major design types:
  - Withdrawal design (Crozier & Tincani, 2005),
  - Changing criterion design (Ganz & Sigafoos, 2005)
  - Multiple baseline design (Ozdemir, 2008)
  - Alternating treatment design (Tincani, 2004).
- All of the articles from the field of treatment efficacy in autism
- The first and second author independently applied each appraisal tool to each article
- Calculated inter-rater agreement using percentage agreement yielded an agreement rate of 85%.
# Table 3
Comparison of Quality Appraisal Tools When Applied to Four Different Types of Single-Subject Experimental Designs

<table>
<thead>
<tr>
<th>Article (Authors; Year)</th>
<th>Single-Subject Research Design</th>
<th>Quality Appraisal Scores and Rankings based on SSED Appraisal Tools (maximum quality score)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Certainty Framework</td>
</tr>
<tr>
<td>Crozier &amp; Tincani, 2005</td>
<td>Withdrawal (A-B-A-C)</td>
<td>“Inconclusive”</td>
</tr>
<tr>
<td>Ganz &amp; Sigafoos, 2005</td>
<td>Changing Criterion</td>
<td>“Suggestive”</td>
</tr>
<tr>
<td>Ozdemir, 2008</td>
<td>Multiple Baseline Across Participants</td>
<td>“Preponderant”</td>
</tr>
<tr>
<td>Tincani, 2004</td>
<td>Alternating Treatment</td>
<td>“Preponderant”</td>
</tr>
</tbody>
</table>

*Note. CSSEDARS = Comparative Single-Subject Experimental Design Rating Scale; EVIDAAC = Evidence in Augmentative and Alternative Communication; IOA = Interobserver agreement; IV = Independent variable; SCED = Single-Case Experimental Design; SSED = Single-Subject Experimental Design; WWC = What Works Clearinghouse. *An extended version of this table containing further appraisal details is available from the first author upon request.*
Results

- The different appraisal instruments vary remarkably in evaluation results of SCED studies.
- Some tools appear to be stronger and more rigorous in quality assessment than others.
- The four soundest tools are listed first, in hierarchical order starting with the more rigorous ones. The last three tools are not listed in any hierarchical sequence.

- Evaluative Method
- Certainty Framework
- WWC Standards
- EVIDAAC Scales
- Logan et al. Scale
- SCED Scale
- Smith et al. Scale

(Wendt & Miller, 2012)
Systematic Review – Different Outcomes

- Review: Efficacy of Functional Communication Training for Adults with Autism (Gregori, Wendt, & Gerow, under review)
- 13 SCED studies
- Two appraisal tools:
  - What Works Clearinghouse (WWC) Design Standards
  - Council for Exceptional Children (CEC) Quality Indicators
- WWC: 6 studies rated as strong/moderate evidence
- CEC: no studies meet criteria

→ Differences in the outcomes based on the appraisal rubric applied
Conclusions / Future Directions

- Different tools yield variable quality appraisals
- No agreement on a “gold standard” against which to compare a newly developed tool
- Keep context, focus, and limitations of the tool in mind
- We recommend that applied researchers and practitioners carefully select among these four and distinguish different purposes:
  - **The Evaluative Method**: For comprehensive systematic reviews that aim to inform both clinical/educational practice and policy.
  - **The Certainty Framework**: For time-efficient literature reviews such as rapid evidence reviews (United Kingdom Civil Service, 2011) or critically appraised topics (Wendt, 2006).
  - **The WWC Standards**: When reviews are particularly aiming for a thorough assessment of internal validity.
  - **The EVIDAAC Scales**: The user-friendliness of the scale—that is, an easily accessible format and clear instructions how to use the instrument—also make it an option for the less experienced reviewer. (Wendt & Miller, 2012)
GEIS 2018: Quality Appraisal of Single-Case Experimental Designs

REVIEW OF RANDOMIZATION AND DATA ANALYSIS ITEMS
Randomization and Data Analysis

Items in SCED Reporting Tools

Goal:

- Focus on randomization and data-analysis items included in SCED quality appraisal tools.
  - Increased acknowledgement of the importance of randomization in SCED research (cf. Kratochwill & Levin, 2010; Onghena & Edgington, 2005)
  - Growing development of more adequate data-analysis procedures for SCE designs
  - Insight that visual analysis of SCEs alone offers unreliable conclusions (cf. Kazdin, 2011)
Randomization in SCED

11 tools were retrieved for this extension to Wendt & Miller (2012).

- 7 tools previously found (Wendt & Miller, 2012)
- 4 additional standards/guideline papers not operationalized into check-lists

- Major discrepancies between the tools and the state-of-the-art data analysis procedures

- Only 2 out of the 11 retrieved tools include an item on randomization and/or data analysis
Conclusion

Inclusion of the criteria:

- Express the size of the effect
- Use an appropriate statistical analysis
- Random assignment of measurement occasions to the levels of the independent variable(s)

(Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015)
APPRAISAL OF COMPARATIVE SCEDS
Background

Shortcomings of previously developed SCED scales:

- No effort to provide separate criteria relative to design type
- Some designs have specific requirements and design tactics
- Exclusion of comparative designs or treating them like single intervention
- BUT: comparative design have own unique internal validity constraints and considerations
Aims

- Develop new checklist: Comparative Single-Case Experimental Design Appraisal Rating System (CSCEDARS)

- Iterative process of tool development:
  - Draft items and operational definitions for version 1;
  - Pilot coding of sample studies using the revised version 1
  - Development of a draft of version 2 based on prior feedback and Horner et al. (2005)
  - Development of version 3 as a result of peer-review
  - Initial reliability assessments; and development of version 4 as a result of reliability assessments. (Schlosser et al., 2018)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PB  RS  A/D  F</td>
<td>PB  RS  A/D  F</td>
<td>PB  RS  A/D  F</td>
<td>PB  RS  A/D  F</td>
</tr>
<tr>
<td># of participants</td>
<td>5   5   X  5</td>
<td>6   6   X  6</td>
<td>6   6   X  6</td>
<td>3   3   X  3</td>
</tr>
<tr>
<td># of analyses</td>
<td>5   5   X  5</td>
<td>6   6   X  6</td>
<td>6   6   X  6</td>
<td>1   1   X  1</td>
</tr>
<tr>
<td>1. Participants</td>
<td>Y   Y   X  Y</td>
<td>N   Y   Y  Y</td>
<td>N   Y   Y  Y</td>
<td>Y   Y   X  Y</td>
</tr>
<tr>
<td>2. Settings</td>
<td>Y   Y   X  Y</td>
<td>N   Y   Y  Y</td>
<td>Y   Y   X  Y</td>
<td>Y   N   N  N</td>
</tr>
<tr>
<td>3. DV operationalized</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   Y   N  N</td>
</tr>
<tr>
<td>4. DV repeatedly</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
</tr>
<tr>
<td>5. IOA</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
</tr>
<tr>
<td>6. IV defined</td>
<td>Y   N   N  N</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
<td>N   Y   N  N</td>
</tr>
<tr>
<td>7. Treatment integrity</td>
<td>Y   Y   X  Y</td>
<td>N   Y   N  N</td>
<td>N   N   X  N</td>
<td>N   N   X  N</td>
</tr>
<tr>
<td>8. Baseline/Probes same</td>
<td>N   N   X  N</td>
<td>Y   N   N  N</td>
<td>N   N   X  N</td>
<td>Y   N   N  N</td>
</tr>
<tr>
<td>9. Participants not biased</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
</tr>
<tr>
<td>10. DV same across</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
</tr>
<tr>
<td>11. IOA comparable</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   Y   Y  Y</td>
<td>N   N   X  N</td>
</tr>
<tr>
<td>12. TI comparable</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>N   N   X  N</td>
</tr>
<tr>
<td>13. Baselines/condition</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>Y   N   N  N</td>
</tr>
<tr>
<td>14. Baseline consistent</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
</tr>
<tr>
<td>15. Design carryover</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
</tr>
<tr>
<td>16. Safeguard carryover</td>
<td>Y   Y   X  Y</td>
<td>Y   Y   X  Y</td>
<td>N   Y   N  N</td>
<td>Y   N   Y  Y</td>
</tr>
<tr>
<td>17. Design order</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   Y   Y  Y</td>
<td>Y   Y   X  Y</td>
</tr>
<tr>
<td>18. Safeguards order</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   Y   N  N</td>
<td>N   N   X  N</td>
</tr>
<tr>
<td>19. Sets equivalence</td>
<td>Y   N   N  N</td>
<td>N   Y   N  Y</td>
<td>N   N   X  N</td>
<td>N   N   X  N</td>
</tr>
<tr>
<td>20. Stimuli/sets random</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>N   N   X  N</td>
</tr>
<tr>
<td>21. Stimuli/sets allocate</td>
<td>N   N   X  N</td>
<td>N   N   X  N</td>
<td>N   N   X  N</td>
<td>N   N   X  N</td>
</tr>
<tr>
<td>22. Experimental control</td>
<td>Y   N   N  N</td>
<td>N   Y   N  N</td>
<td>N   N   X  N</td>
<td>Y   N   N  N</td>
</tr>
<tr>
<td>23. Clear separation</td>
<td>N   N   X  N</td>
<td>Y   Y   X  Y</td>
<td>N   N   X  N</td>
<td>N   N   X  N</td>
</tr>
<tr>
<td><strong>TOTALS</strong></td>
<td>22/3  8Y</td>
<td>18/7   16Y</td>
<td>21/4  8Y</td>
<td>18/7  13N</td>
</tr>
<tr>
<td></td>
<td>15N</td>
<td>7N</td>
<td>15N</td>
<td>13N</td>
</tr>
<tr>
<td><strong>Inter-rater Agreement</strong></td>
<td>88%</td>
<td>72%</td>
<td>84%</td>
<td>72%</td>
</tr>
</tbody>
</table>
Results & Future Directions

- Interrater Agreement (using percentage agreement) mean 79% (range 72-88%)

- Three applications:
  1. Appraise internal validity of comparative SCEDs
  2. Assess risk of bias of included SCEDs in systematic reviews
  3. Use prospectively for better design of comparative SCEDs
Currently, no empirical guidelines to interpret “strong”, “moderate”, or “weak”; higher score is better

Psychometric properties yet to establish:
- Content validity
- Further types of reliability

Best used by raters with considerable expertise

Future effort could also be linked to establishing reporting characteristics
Questions
Contact Information

Oliver Wendt, Ph.D.
School of Communication Sciences and Disorders
College of Health Professions and Sciences
University of Central Florida
4364 Scorpius Street, HPA-2
Orlando, FL 32816-2215, USA
Phone: (+1) 407-823-4772
Fax: (+1) 765-823-4816
E-mail: oliver.wendt@ucf.edu

