Global Evidence and Implementation Summit 2018
Melbourne, Australia

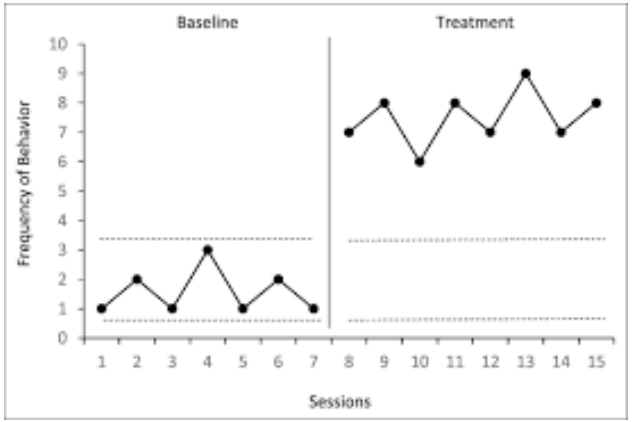
Quality Appraisal of Single-Case Experimental Designs: Current Issues and Future Directions

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 1
Current Quality Appraisal Tools for Single-Subject Experimental Designs (SSEDs)

Characteristics/ Properties*	Tool (maximum quality score)						
	Certainty Framework (no max. score)	Evaluative Method (max. = 12)	EVIDAAC Scales (max. = 10 or 19)	Logan et al. Scale (max. = 14)	SCED Scale (max. = 10)	Smith et al. Scale (max. = 15)	WWC Standards (no max. score)
Composition of tool	Ranks certainty of evidence as "conclusive" (highest), "preponderant", "suggestive", or "inconclusive" (lowest), based on research design, interobserver agreement of dependent variable, and treatment integrity	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	One treatment scale: 10 items; two or more treatments scale: 19 items; higher score = higher quality	14 questions containing 16 items; studies are rated "strong" (11-14 points), "moderate" (7-10 points), or "weak" (less than 7 points)	11-item rating scale; item 1 assesses clinical history information; items 2-11 allow calculation of quality score; higher score = higher quality	15-item rating scale; higher score = higher quality	*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 "
Content validity established	No	Yes	No	No	Yes	No	No
Inter-rater reliability provided	No	Yes, including expert and novice raters	No	Yes, including the four authors of the scale	Yes, including expert and novice raters	No	No

Note. EVIDAAC = Evidence in Augmentative and Alternative Communication; SCED = Single-Case Experimental Design; WWC = What Works Clearinghouse. *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

Article (Authors; Year)	Single-Subject Research Design	Quality Appraisal Scores and Rankings based on SSED Appraisal Tools (maximum quality score)*						
		Certainty Framework	Evaluative Method (max.=12)	EVIDAAC Scales (max.=10)	Logan et al. Scale (max.=14)	SCED Scale (max.=10)	Smith et al. Scale (max.=15)	WWC Standards
Crozier & Tincani, 2005	Withdrawal (A-B-A-C)	"Inconclusive"	58% "weak"	80%	71% "moderate"	70%	87%	"Does not meet evidence standards"
Ganz & Sigafoos, 2005	Changing Criterion	"Suggestive"	50% "weak"	90%	54% "moderate"	60%	53%	"Meets standards with reservations": "Moderate evidence"
Ozdemir, 2008	Multiple Baseline Across Participants	"Preponderant"	67% "adequate"	80%	39% "weak"	60%	60%	"Meets standards with reservations": "Strong evidence"
Tincani, 2004	Alternating Treatment	"Preponderant"	67% "adequate"	79% CSSEDARS	61% "moderate"	70%	80%	"Meets standards": "Moderate evidence"

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)

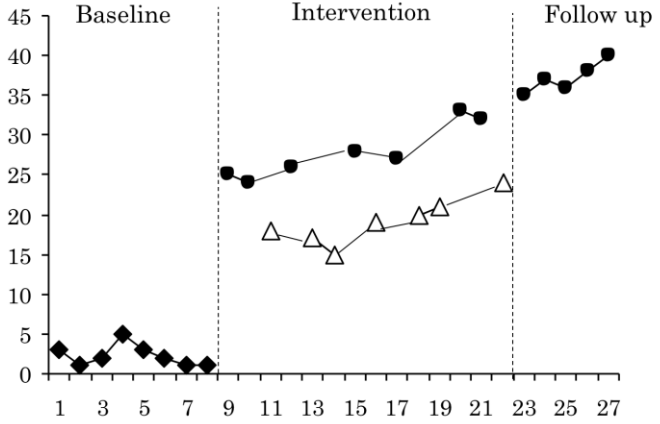
(Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015)

GEIS 2018: Quality Appraisal of Single-Case Experimental Designs

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 2
Agreements, Disagreements, and Consensus Ratings of Sample Studies.

Studies	Flores et al. (2012)				Bock et al. (2005)				Gregory et al. (2009)				Trembath et al. (2009)			
Items	PB	RS	A/D	F	PB	RS	A/D	F	PB	RS	A/D	F	PB	RS	A/D	F
# of participants	5	5	X	5	6	6	X	6	6	6	X	6	3	3	X	3
# of analyses	5	5	X	5	6	6	X	6	6	6	X	6	1	1	X	1
1. Participants	Y	Y	X	Y	Y	N		N	N	Y		Y	Y	Y	X	Y
2. Settings	Y	Y	X	Y	N	Y		Y	Y	Y	X	Y	Y	N		N
3. DV operationalized	Y	Y	X	Y	N	N	X	N	Y	Y	X	Y	N	Y		N
4. DV repeatedly	N	N	X	N	Y	Y	X	Y	N	N	X	N	Y	Y	X	Y
5. IOA	Y	Y	X	Y	Y	Y	X	Y	Y	Y	X	Y	Y	Y	X	Y
6. IV defined	Y	N		N	Y	N		N	Y	Y	X	Y	N	Y		N
7. Treatment integrity	Y	Y	X	Y	N	Y		N	N	N	X	N	N	N	X	N
8. Baseline/Probes same	N	N	X	N	Y	N		N	N	N	X	N	Y	N		N
9. Participants not biased	N	N	X	N	Y	Y	X	Y	N	N	X	N	Y	Y	X	Y
10. DV same across	Y	Y	X	Y	Y	Y	X	Y	Y	Y	X	Y	Y	Y	X	Y
11. IOA comparable	N	N	X	N	Y	Y	X	Y	N	Y		Y	N	N	X	N
12. TI comparable	Y	Y	X	Y	Y	Y	X	Y	N	N	X	N	N	N	X	N
13. Baselines/condition	N	N	X	N	Y	Y	X	Y	N	N	X	N	Y	N		N
14. Baseline consistent	N	N	X	N	Y	Y	X	Y	N	N	X	N	Y	Y	X	Y
15. Design carryover	N	N	X	N	Y	Y	X	Y	N	N	X	N	Y	Y	X	Y
16. Safeguard carryover	Y	Y	X	Y	Y	Y	X	Y	N	N	X	N	Y	N		Y
17. Design order	N	N	X	N	Y	Y	X	Y	N	Y		Y	Y	Y	X	Y
18. Safeguards order	N	N	X	N	Y	Y	X	Y	N	Y		N	N	N	X	N
19. Sets equivalence	Y	N		N	N	Y		Y 0.33	N	N	X	N	N	N	X	N
20. Stimuli/sets random	N	N	X	N	Y	Y	X	Y	N	N	X	N	N	N	X	N
21. Stimuli/sets allocate	N	N	X	N	N	N	X	N	N	N	X	N	N	N	X	N
22. Experimental control	Y	N		N	Y	N		N	N	N	X	N	Y	N		Y
23. Clear separation	N	N	X	N	Y	Y	X	Y	N	N	X	N	N	N	X	N
TOTALS			22/3	8Y 15N			18/7	16Y 7N			21/4	8Y 15N			18/7	10Y 13N
Inter-rater Agreement			88%				72%				84%				72%	

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

Results & Future Directions

- 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

References

- Wendt, O., & Miller, B. (2012). Quality appraisal of single subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*, 235-268.
- Heyvaert, M., Wendt, O., Van den Noortgate, & Onghena, P. (2015). Randomization and data-analysis items in tools for reporting and evaluating single-case experimental studies. *Journal of Special Education, 49*, 146-156.
- Schlosser, R.W., Belfiore, P., Sigafoos, J., Briesch, A., & Wendt, O. (2018). Appraisal of comparative single-case experimental designs for instructional interventions with non-reversible target behaviors: Introducing the CSCEDARS ("Cedars"). *Research in Developmental Disabilities, 79*, 33-52.
- Gregori, E., Wendt, O., Gerow, S., Peltier, C., Genc, D., Lory, C, & Gold, Z. S. (2018, Manuscript under review). *Functional communication training for adults with autism spectrum disorder: A systematic review and quality appraisal.*