Review

Single-subject research design: recommendations for levels of evidence and quality rating

Lynne Romeiser Logan PT MA PCS, Tone Management and Mobility Program, Upstate Medical University, Syracuse, NY;
Robbin R Hickman PT DSc PCS, Physical Therapy, University of Nevada, Las Vegas, NV, USA.
Susan R Harris* PhD PT FAPTA FCAHS, Department of Physical Therapy, Faculty of Medicine, University of British Columbia, Vancouver, Canada.
Carolyn B Heriza EdD PT FAPTA, Pediatrics, Rocky Mountain University of Health Professions, Provo, UT, USA.

**Correspondence to third author at* Department of Physical Therapy, Faculty of Medicine, T325-2211 Wesbrook Mall, University of British Columbia, Vancouver, BC V6T 2B5, Canada.
E-mail: shar@interchange.ubc.ca

DOI: 10.1111/j.1469-8749.2007.02005.x

The aim of this article is to present a set of evidence levels, accompanied by 14 quality or rigor questions, to foster a critical review of published single-subject research articles. In developing these guidelines, we reviewed levels of evidence and quality/rigor criteria that are in wide use for group research designs, e.g. randomized controlled trials, such as those developed by the Treatment Outcomes Committee of the American Academy for Cerebral Palsy and Developmental Medicine. We also reviewed methodological articles on how to conduct and critically evaluate single-subject research designs (SSRDs). We then subjected the quality questions to interrater agreement testing and refined them until acceptable agreement was reached. We recommend that these guidelines be implemented by clinical researchers who plan to conduct single-subject research or who incorporate SSRD studies into systematic reviews, and by clinicians who aim to practise evidence-based medicine and who wish to critically review pediatric single-subject research.

Societal accountability and professional mandates across all healthcare disciplines demand that professionals engage in evidence-based medicine or evidence-based practice, namely the integration of best research evidence with clinical expertise and patient values.[1] Evaluative scales and guidelines exist to help clinicians and scientists critically review published results of group-design research as part of the decision-making process. However, researchers in rehabilitation and social sciences often utilize single-subject research designs (SSRDs)[2–4] for which no criteria exist to guide the critical review process. Therefore, the purpose of this paper is to present a set of guidelines to foster the critical review of studies using SSRD. All authors are members of a sub-committee of the American Academy for Cerebral Palsy and Developmental Medicine (AACPDM) Treatment Outcomes Committee who were charged with developing these guidelines for single-subject research designs.

Evaluation of SSRD must address many of the same criteria used to evaluate group designs, e.g. level of evidence and quality/rigor of methods used, analogous to those from the Oxford Centre for Evidence-Based Medicine.[1] Criteria for scientific quality and rigor of group designs have been developed also by Jadad et al.,[5] van Tulder et al.,[6] and the Treatment Outcomes Committee of the AACPDM.[7] These criteria facilitate assessment of a study's reliability, internal and external validity, and application to diverse populations, but categorize all types of SSRD into a single evidence level that fails to reflect the potential usefulness of this type of research. Likewise, many of the quality questions used for group

design are not appropriate for SSRD. The guidelines for level and quality presented in this article address issues common to both SSRD and group research, e.g. reliability and validity; they also deal with issues specific to SSRD, such as length and stability of baseline and intervention phases, critical design issues not usually present in group-design evaluation schemes. Tables I and II illustrate these issues.

Background to the single-subject research design

SSRDs differ dramatically from case reports, although the two are often confused. Case reports can illuminate theory, describe novel interventions, or develop hypotheses for research. Case reports or case series carefully describe the patient(s), the clinician's decision-making processes, the intervention provided and associated outcomes, but do not expose the patient to controlled experimental conditions, e.g. collecting baseline data for a prescribed period of time before initiating treatment or separating out specific elements of intervention.[8] Consequently, the case report does not provide any assurance that the change was due to the intervention rather than to history, maturation, regression, or testing, so no causal inferences can be made.[8,9]

Alternative terminology for SSRD includes within-subject methods, repeated measures designs, and intrasubject replication designs.[10] All such designs expose the subject to both treatment and control (or comparison) conditions, thus allowing subjects to act as their own controls. SSRDs may be conducted with one subject or replicated across several subjects, seek to discover whether the initial behavior being studied changed after introduction of the intervention, and what evidence exists to suggest that the intervention actually caused the observed changes.[11] If randomization is present and subject(s) and examiners are blinded, SSRD is a very powerful design. In fact, Guyatt and colleagues argued that *N*-of-1 randomized controlled trials, a type of SSRD, might represent the highest level of evidence in clinical practice.[12]

In SSRD, the outcome of interest (target behavior or dependent variable) is measured repeatedly in each condition or phase of the research process. One or more intervention periods are combined with one or more baselines (non-intervention periods) to develop conclusions about changes in the target problem and, possibly, effects of the intervention on that problem.[11] In all designs, the letter 'A' refers to the non-intervention or baseline phase and the letter 'B' to the first identified treatment or intervention phase. Different letters are used to represent subsequent intervention phases. If an intervention or baseline is repeated, the same letter is used to represent the repeated phase.

It is important to establish stability of the data within the baseline phase before introducing the intervention. Data are stable, whether in baseline or intervention phases, first, if there is consistency of the data with no wide fluctuations and second, the data predict a pattern of data into the next phase. Stable data can be flat, increasing, or decreasing.[11,13] A stable data pattern allows clear comparisons across the various phases.[14]

Table I: Levels of evidence for single-subject research designs (SSRDs)

Level	*Evidence*
I	Randomized controlled *N*-of-1 (RCT), alternating treatment (ATD), and concurrent or non-concurrent multiple baseline designs (MBDs)[a] with clear-cut results; generalizability if the ATD is replicated across three or more subjects and the MBD design consists of a minimum of three subjects, behaviors, or settings. These designs can provide causal inferences.
II	Non-randomized, controlled, concurrent MBD[a] with clear-cut results; generalizability if design consists of a minimum of three subjects, behaviors, or settings; limited causal inferences.
III	Non-randomized, non-concurrent, controlled MBD[a] with clear-cut results; generalizability if design consists of a minimum of three subjects, behaviors, or settings; limited causal inferences.
IV	Non-randomized, controlled SSRDs with at least three phases (ABA, ABAB, BAB, etc.) with clear-cut results; generalizability if replicated across five or more different subjects; only hints at causal inferences.
V	Non-randomized controlled AB single-subject research design with clear-cut results; generalizability if replicated across three or more different subjects; suggests causal inferences allowing for testing of ideas.

[a]If the intervention(s) is known to be successful, a baseline or control phase is not required.

Evaluating rigor and quality of single-subject research designs

The methods for evaluating rigor and quality of SSRD presented in this paper provide a systematic means for assessing whether the baseline and intervention have been applied under standardized conditions that guard against threats to internal validity. As in group designs, reliable assessment of outcome measures must be established. Reliability refers to the accuracy or reproducibility of the measurements taken, whereas validity refers to the confidence with which the research findings are 'believable' and meaningful.[13] Internal validity is the degree to which a causal relationship between the independent and dependent variables has been established, whereas external validity is the extent to which results can be generalized beyond the subjects included in the study.[15]

Types of single-subject research design

There are many types of SSRD that are used to collect and analyze data to judge changes in the target behavior and to decide whether the intervention can be inferred to be causally related to these changes.[11] The research question guides the choice of the SSRD, with each design having strengths and limitations. Descriptions of the various designs are available in several texts[10,13,14] and review articles.[1,3,9,16]

Studies that have used SSRD are relatively common in the literature for developmental medicine and rehabilitation sciences. The following are a few examples of relevant studies using different types of designs: A–B (simple baseline design),[17] A–B–A (withdrawal),[18,19] multiple baseline design across subjects,[20] and alternating treatment design.[21]

Visual and statistical analysis of single-subject research design

SSRD data are analyzed visually but also by using various statistical techniques,[22] each of which has strengths and limitations as well as conditions under which they should be used.[11] Use of several different statistical analyses is likely to

enhance acceptance or rejection of the outcomes.[11,23] Visual analysis of graphed data, following standard conventions, is used to evaluate differences between phases, for example, baseline and intervention. Trend, slope, and level analyses are often used.

Descriptive statistics summarize patterns of data and aid visual analysis; these include measures of central tendency, variability, trend lines, and slope of the trend lines.[11] Common inferential statistical tests for SSRD include χ^2 (Bloom et al.[11]) and *t*-tests,[11,14] the celeration line approach (also referred to as split-middle method),[11,14,16] the two- and three-SD band methods,[11,14] and the C-statistic.[11,15,24]

Methods for evaluating specific types of single-subject research designs

Within-subjects methods, such as SSRD, are advantageous in rehabilitation settings in which participants being studied are frequently heterogeneous or when few subjects are available, as in low-incidence conditions. Likewise, these methods are preferred when the researcher suspects that subjects may demonstrate variability from day to day. Because each subject is studied intensely, influences other than the target intervention can be identified.

Experimental features that contribute to establishing causality serve to distinguish the various levels of evidence and influence quality ratings in SSRD. Study design is among the most prominent feature manipulated by investigators. Numerous types of study design are possible in SSRD including: *N*-of-1 randomized controlled trials;[25] alternating treatment designs; randomized multiple-baseline designs (concurrent or nonconcurrent); replicated basic designs with at least three phases, e.g. A–B–A or A–B–C, in which C is a second intervention; and A–B or simple baseline designs. Table I summarizes the hierarchical levels of evidence yielded by various SSRD options.

In addition to varying the study design, investigators may also manipulate other study attributes to strengthen the study's ability to establish causality and to ensure its rigor or quality. Key experimental elements and the SSRD methods used to establish them are summarized in Table II. When study design and methodology are considered together in critical appraisal of SSRD, evidence-based consumers may have differential confidence in the findings of that research. Table I provides guidance in evaluation of evidence for each level of the SSRD scale. In general, clinicians should seek evidence from as high in the hierarchy as possible.[13]

Table I shows that each design 'may' yield a particular type of evidence. These statements are provisional because researchers and research consumers must also be able to evaluate the quality of the research conducted in addition to weighing the strength of the design itself. Similar to quality questions for group designs, the following questions are presented to guide researchers and research consumers of SSRD in assessing the quality of the research as an important step in evaluating the overall evidence.

Quality questions in single-subject research designs

The authors of this article (all who have conducted, published, and/or taught SSRD) developed the following 14 questions based on review of similar questions or criteria used for evaluating group designs,[5–7] as well as on the article by Horner et al.[3] Scoring of this quality test is simply done by counting 'yes' answers and ascribing 1 point to each. Because questions 5 and 8 are two-part questions, 0.5 points are assigned to each part. Based on review of evaluation scoring cut-offs for group designs, the following categories were used: strong,

Table II: Study design elements and single-subject research design (SSRD) methods

Design elements	*SSRD methods*
Subjects	Observing one or more clients before, during, and after interventions
Repeated outcome measures	Set of procedures used to observe changes in identified target behavior (a specific concern or objective of the client) measured repeatedly over time
Phases	Time periods consisting of baseline (control) phases, intervention phases, and follow-up phases during which repeated outcomes are measured
Comparison of phases to determine outcome	Baseline (control) phases, intervention phases, and follow-up phases arranged to support a decision of causality
Consistency of outcome measures across phases	Repeated measurement conditions to which clients are subjected during baseline, intervention, and follow-up phases are consistent
Random allocation	Random allocation of subjects, settings, or behaviors in multiple baseline design; random allocation of intervention in *N*-of-1 and alternating treatment designs
Concurrency	For multiple baseline designs, intervention and baseline (control) phases are investigated concurrently
Manipulation of exposure	Clients exposed to both intervention phases and baseline (control) phases
Ascertainment of exposure (compliance with control vs intervention condition)	Each assigned intervention or baseline (control) condition – and only that condition – was experienced by a client during the specified phases
Loss to follow-up	Subject attrition occurred before final collection phase
Loss of data points	Data points lost during phases or client(s) lost prior to final collection phase
Outcome evaluation	Only data from adjacent phases are compared
Statistical evaluation of the presence of a change or difference	Data are analyzed using visual/graphic analysis such as level and trend, descriptive statistics, and/or inferential statistics

11–14; moderate, 7–10; and weak, less than 7.

Interagreement analyses were conducted by the authors on the first version of these questions ($n=19$) for six SSRD articles; based on our results, those questions with low agreement (no more than 50%) were subsequently excluded. Three SSRD articles were then evaluated using the final 14 questions. Agreement among the four raters on overall methodological strength (weak, moderate, or strong) across the three studies was 75%. In the authors' experience, this level of agreement is in line with those of the group-design rating scales.

DESCRIPTION OF PARTICIPANTS AND SETTINGS

1. Was/were the participant(s) sufficiently well described to allow comparison with other studies or with the reader's own patient population?

INDEPENDENT VARIABLE

2. Were the independent variables operationally defined to allow replication?
3. Were intervention conditions operationally defined to allow replication?

DEPENDENT VARIABLE

4. Were the dependent variables operationally defined as dependent measures?
5. Was interrater or intrarater reliability of the dependent measures assessed before and during each phase of the study?
6. Was the outcome assessor unaware of the phase of the study (intervention vs control) in which the participant was involved?
7. Was stability of the data demonstrated in baseline, namely lack of variability or a trend opposite to the direction one would expect after application of the intervention?

DESIGN

8. Was the type of SSRD clearly and correctly stated, for example A–B, multiple baseline across subjects?
9. Were there an adequate number of data points in each phase (minimum of five) for each participant?
10. Were the effects of the intervention replicated across three or more subjects?

ANALYSIS

11. Did the authors conduct and report appropriate visual analysis, for example, level, trend, and variability?
12. Did the graphs used for visual analysis follow standard conventions, for example *x*- and *y*-axes labeled clearly and logically, phases clearly labeled (A, B, etc.) and delineated with vertical lines, data paths separated between phases, consistency of scales?
13. Did the authors report tests of statistical analysis, for example celeration line approach, two-standard deviation band method, C-statistic, or other?
14. Were all criteria met for the statistical analyses used?

Conclusions

Both design features and methodological quality/rigor figured prominently in developing these guidelines that parallel those for group designs. It is important to remember that evidence-based clinical decision making includes elements beyond the critical evaluation of the research design, namely integrating that evidence with clinical judgment and the unique values of each patient and family.[3,26] Therefore, the third and final step in applying these SSRD evidence guidelines to clinical situations is to place them into a context that includes clinical judgment, child preferences, and family values.

Accepted for publication 2nd August 2007.

Acknowledgments

We thank the members of the AACPDM Treatment Outcomes Committee for their input on an earlier draft of this paper.

References

1. Centre for Evidence-Based Medicine, Oxford University. www.cebm.net (accessed 11 November 2006).
2. Backman CL, Harris SR, Chisholm JA, Monette AD. (1997) Single-subject research in rehabilitation: a review of studies using AB, withdrawal, multiple baseline, and alternating treatments designs. *Arch Phys Med Rehabil* **78:** 1145–1153.
3. Horner RH, Carr EG, Halle J, McGee G, Odom S, Wolery M. (2005) The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children* **71:** 165–179.
4. Jones P. Single-case research and statistical analysis in school psychology and counseling, 2005–2006. www.unlv.edu/faculty/pjones/pj.htm (accessed 7 February 2007).
5. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. (1996) Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* **17:** 1–12.
6. van Tulder M, Furlan A, Bombardier C, Bouter L, Editorial Board of the Cochrane Collaboration Back Review Group. (2003) Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine* **28:** 1290–1299.
7. O'Donnell M, Darrah J, Adams R, Butler C, Roxborough L, Damiano DL. (2004) AACPDM methodology to develop systematic reviews of treatment interventions. www.aacpdm.org/resources/systematicReviewsMethodology.pdf (accessed 5 February 2007).
8. McEwen I, editor. (2001) *Writing Case Reports: A How-to Manual for Clinicians.* Alexandria, VA: American Physical Therapy Association.
9. Magill J, Barton L. (1985) Single subject research designs in occupational therapy literature. *Can J Occup Ther* **52:** 53–58.
10. Kazdin AE. (1982) Single-case experimental designs in clinical research and practice. *New Directions for Methodology of Social & Behavioral Science* **13:** 33–47.
11. Bloom M, Fischer J, Orme JG. (2006) *Evaluating Practice: Guidelines for the Accountable Professional.* 5th edn. Boston, MA: Allyn and Bacon, Pearson Educational, Inc.
12. Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, Wilson MC, Richardson WS. (2000) Users' guides to the medical literature. XXV. Evidence-based medicine: principles for applying users' guides to patient care. *J Am Med Assoc* **284:** 1290–1296.
13. Portney LG, Watkins MP. (2000) *Foundations of Clinical Research: Applications to Practice.* 2nd edn. Upper Saddle River, NJ: Prentice Hall Health.
14. Ottenbacher KJ. (1986) *Evaluating Clinical Change. Strategies for Occupational and Physical Therapists.* Baltimore, MD: Williams and Wilkins.
15. Domholdt E. (2000) *Physical Therapy Research: Principles and Applications.* 2nd edn. Philadelphia, PA: WB Saunders.
16. Patrick PD, Mozzoni M, Patrick ST. (2000) Evidenced-based care and the single-subject. *Infants & Young Children* **13:** 60–73.
17. Fragala MA, O 'Neil ME, Russo KJ, Dumas HM. (2002) Impairment, disability, and satisfaction outcomes after lower-extremity botulinum toxin A injections for children with cerebral palsy. *Pediatr Phys Ther* **14:** 132–144.
18. Mudge S, Rochester L, Recordon A. (2003) The effect of treadmill training on gait, balance and trunk control in a hemiplegic subject: a single system design. *Dis Rehabil Res* **17:** 1000–1007.
19. Fertel-Daly D, Bedell G, Hinojosa J. (2001) Effects of a weighted vest on attention to task and self-stimulatory behaviors in

preschoolers with pervasive developmental disorders. *Am J Occup Ther* **55:** 629–640.
20. Shumway-Cook A, Hutchison S, Kartin D, Price R, Woollacott M. (2003) Effect of balance training on recovery of stability in children with cerebral palsy. *Dev Med Child Neurol* **45:** 591–602.
21. Washington KA, Deitz JC, White OR, Schwartz IS. (2002) The effects of a contoured foam seat on postural alignment and upper-extremity function in infants with neuromotor impairments. *Phys Ther* **82:** 1064–1076.
22. Wolery M, Harris SR. (1982) Interpreting results of single-subject research designs. *Phys Ther* **62:** 445–452.
23. Institute of Medicine. (2001) *Committee on Strategies for Small-Number Participant Clinical Research Trials.* Washington, DC: Institute of Medicine.
24. Nourbaksh RN, Ottenbacher KJ. (1994) The statistical analysis of single-subject data: a comparative examination. *Phys Ther* **74:** 768–776.
25. Backman CL, Harris SR. (1999) Case studies, single-subject research, and N of 1 randomized trials: comparisons and contrasts. *Am J Phys Med Rehabil* **78:** 170–176.
26. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. (2000) *Evidence-based Medicine: How to Practice and Teach EBM.* 2nd edn. New York: Churchill Livingstone.