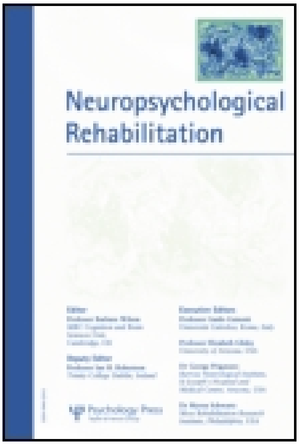

Neuropsychological Rehabilitation: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pnrh20

Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale

Robyn L Tate [a] [b] , Michael Perdices [c] [d] , Ulrike Rosenkoetter [a] , Donna Wakim [a] , Kali Godbee [a] , Leanne Togher [e] & Skye McDonald [f]

[a] Rehabilitation Studies Unit, Sydney Medical School - Northern , University of Sydney , Australia

[b] Royal Rehabilitation Centre Sydney , Australia

[c] Department of Neurology, Royal North Shore Hospital , Sydney , Australia

[d] Department of Psychological Medicine , University of Sydney , Australia

[e] Speech Pathology, Faculty of Health Sciences , University of Sydney , Australia

[f] School of Psychology , University of New South Wales , Sydney , Australia

Published online: 09 Sep 2013.

To link to this article: http://dx.doi.org/10.1080/09602011.2013.824383

Neuropsychological Rehabilitation, 2013
Vol. 23, No. 5, 619–638, http://dx.doi.org/10.1080/09602011.2013.824383

Revision of a method quality rating scale for single-case experimental designs and *n*-of-1 trials: The 15-item Risk of Bias in *N*-of-1 Trials (RoBiNT) Scale

Robyn L Tate[1,2], Michael Perdices[3,4], Ulrike Rosenkoetter[1], Donna Wakim[1], Kali Godbee[1], Leanne Togher[5], and Skye McDonald[6]

[1]Rehabilitation Studies Unit, Sydney Medical School - Northern, University of Sydney, Australia
[2]Royal Rehabilitation Centre Sydney, Australia
[3]Department of Neurology, Royal North Shore Hospital, Sydney, Australia
[4]Department of Psychological Medicine, University of Sydney, Australia
[5]Speech Pathology, Faculty of Health Sciences, University of Sydney, Australia
[6]School of Psychology, University of New South Wales, Sydney, Australia

Recent literature suggests a revival of interest in single-case methodology (e.g., the randomised *n*-of-1 trial is now considered Level 1 evidence for treatment decision purposes by the Oxford Centre for Evidence-Based Medicine). Consequently, the availability of tools to critically appraise single-case reports is of great importance. We report on a major revision of our method quality instrument, the Single-Case Experimental Design Scale. Three changes resulted in a radically revised instrument, now entitled the Risk of Bias in *N*-of-1 Trials (RoBiNT) Scale: (i) item content was revised and increased to 15 items,

Correspondence should be addressed to Professor Robyn L Tate, Rehabilitation Studies Unit, Sydney Medical School, University of Sydney, PO Box 6, RYDE 1680, Australia. E-mail: rtate@med.usyd.edu.au

This project was funded in part by the Motor Accidents Authority of New South Wales, Australia, and greatly supported by the library resources of the University of Sydney and the Royal Rehabilitation Centre Sydney. We thank Linda Sigmundsdottir and Louise Hunt for their feedback in the course of rating papers for PsycBITE.

(ii) two subscales were developed for internal validity (IV; 7 items) and external validity and interpretation (EVI; 8 items), and (iii) the scoring system was changed from a 2-point to 3-point scale to accommodate currently accepted standards. Psychometric evaluation indicated that the RoBiNT Scale showed evidence of construct (discriminative) validity. Inter-rater reliability was excellent, for pairs of both experienced and trained novice raters. Intraclass correlation coefficients of summary scores for individual (experienced) raters: $\text{ICC}_{\text{TotalScore}} = .90$, $\text{ICC}_{\text{IVSubscale}} = .88$, $\text{ICC}_{\text{EVISubscale}} = .87$; individual (novice) raters: $\text{ICC}_{\text{TotalScore}} = .88$, $\text{ICC}_{\text{IVSubscale}} = .87$, $\text{ICC}_{\text{EVISubscale}} = .93$; consensus ratings between experienced and novice raters ($\text{ICC}_{\text{TotalScore}} = .95$, $\text{ICC}_{\text{IVSubscale}} = .93$, $\text{ICC}_{\text{EVISubscale}} = .93$. The RoBiNT Scale thus shows sound psychometric properties and provides a comprehensive yet efficient examination of important features of single-case methodology.

Keywords: *n*-of-1 trial; Single-case experimental design; Internal validity; Risk of bias; Methodology; Evidence-based medicine.

INTRODUCTION

In 2011, the Oxford Centre for Evidence-Based Medicine posted “Levels of Evidence 2” on their website (http://www.cebm.net) describing a revised table of levels of evidence (Howick et al., 2011). The randomised *n*-of-1 trial, previously absent from standard evidence tables, was ranked as Level 1 evidence for treatment decision purposes in individual patients, alongside systematic reviews of multiple randomised controlled trials (RCTs). The single RCT was classified as Level 2 evidence. This development is in accordance with recommendations that Guyatt, Jaeschke, and McGinn (2002) had been advocating for many years. The inclusion of the randomised *n*-of-1 trial as Level 1 evidence is likely to have major implications for what constitutes the evidence base of health interventions.

In the medical setting, the *n*-of-1 trial arose in the mid-1980s in response to limitations that were apparent in applying the findings of RCTs to the individual patient when making treatment decisions. Guyatt and colleagues (1986; 1988; 1990; Keller, Guyatt, Roberts, Adachi, & Rosenbloom, 1988) were pioneers in adapting methodology that had been developed in clinical and educational psychology in order to experimentally investigate the effect of interventions in the individual medical patient. In clinical psychology, these methods were more commonly referred to as single-case experimental designs (e.g., Hersen & Barlow, 1976; Kazdin, 1982). Single-case methodology has two basic defining features: (i) the prospective and intensive study of the individual during multiple discrete phases—at minimum two phases, generally baseline (by convention designated with the letter, A) and treatment or intervention (designated with the letter, B), in which (ii) a specific and operationally-defined behaviour targeted for intervention is measured repeatedly and frequently during all phases.

Following Guyatt and colleagues, in medicine the *n*-of-1 trial generally refers to the double-blind, randomised, multiple crossover A-B trial in a single patient. The interventions are commonly pharmacotherapies, and more recently, other naturopathic/homoeopathic remedies (Johnston & Mills, 2004) that are generally ingested, injected, inhaled or topically applied. The *n*-of-1 design is identified by three principles: (i) blinding of patient and therapist, (ii) randomisation of paired sequences (A-B) of the control and experimental treatments in a series of cross-over periods or phases, and (iii) measurement of outcomes (i.e., the symptom/behaviour targeted for intervention). Guyatt and colleagues recommend that the *n*-of-1 trial is best suited for interventions that have dramatic "on-off" effects, short wash-out periods and that are applied to chronic conditions. Drugs and other substances are the obvious examples of interventions well suited to this design. In neurological rehabilitation, interventions using equipment, aids and technical devices are also appropriate. Blinding of patient and therapist, however, is always difficult in behavioural (and some non-pharmacological medical) interventions (Boutron et al., 2007; Boutron, Tubach, Giraudeau, & Ravaud, 2004).

The *n*-of-1 trial is one type of the large variety of single-case designs. In the behavioural sciences, Barlow, Nock, and Hersen (2009), for example, describe a broad range of 19 separate designs allowing the experiment to be tailored to the nature of the intervention and meet specific challenges in scientifically evaluating treatment effect in different circumstances of the individual patient, and new designs continue to appear in the literature (e.g., McDougall, 2013). In some types of interventions used for health conditions, the very aim of treatment is to effect long-lasting changes in the target behaviour that endure after treatment has been withdrawn (e.g., training patients in anger management techniques, improving communication and social skills, remediating gait dysfunction), rather than have the target behaviour revert to baseline, as is the principle underlying the *n*-of-1 trial. Complex experimental designs (e.g., multiple-baseline designs) have been developed to address interventions where "carry-over" effects are expected or when the intervention relies on effecting change by having the participant internalise cognitive/behavioural strategies. In addition, other sophisticated designs can be used in certain circumstances. For example alternating treatment designs are an efficient way to directly compare two treatments simultaneously, and are also expedient when time constraints are an issue; changing criterion designs, using achievement of increasingly demanding performance in sequential phases, can be used when shaping of behaviour towards a desired criterion is indicated. When specific and necessary conditions are met, all of these designs have the capacity to demonstrate cause–effect relationships between the independent variable (treatment/intervention) and the dependent variable (symptoms, signs, target behaviours) that are the object of intervention.

The published literature using a single participant (or a series of single participants), however, contains a much broader array of methods than occurs in standard textbooks on single-case methodology. Consequently, there is the potential for confusion as to what does and does not constitute single-case methodology. Figure 1 presents a taxonomy of common designs that use a single participant which we use in our in-house manuals for training in critical appraisal of single-case reports. It was derived from the results of a survey of papers published in *Neuropsychological Rehabilitation* which used a single participant (Perdices & Tate, 2009); informed by discussion at a consensus conference in 2009 to develop a reporting guideline for medical *n*-of-1 trials convened by Dr Sunita Vohra of the University of Alberta, Canada; and refined during the course of classifying more than 1000 single-participant reports in the published neurorehabilitation literature for the purpose of an archival record on our PsycBITE database (http://www.psycbite.com). PsycBITE (the Psychological database for Brain Impairment Treatment Efficacy; Tate et al., 2004) is an evidence-based database that archives all of the published non-pharmacological interventions for acquired brain impairment sourced from six electronic databases and meeting five eligibility criteria. As at April 2013, PsycBITE contained more than 4000 records, including

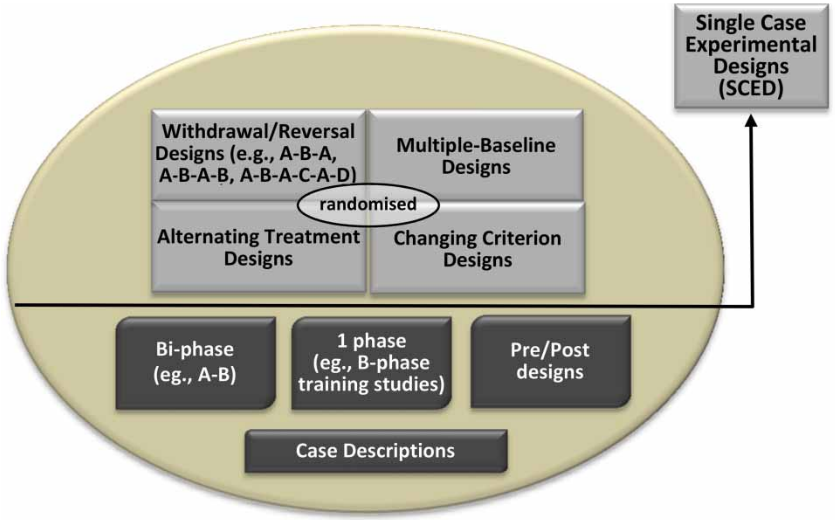
Single Case Experimental Designs (SCED)
Withdrawal/Reversal Designs (e.g., A-B-A, A-B-A-B, A-B-A-C-A-D)
Multiple-Baseline Designs
randomised
Alternating Treatment Designs
Changing Criterion Designs
Bi-phase (eg., A-B)
1 phase (eg., B-phase training studies)
Pre/Post designs
Case Descriptions

Figure 1. A taxonomy of common designs using a single participant

systematic reviews, randomised controlled trials, non-randomised controlled trials, case series, and single-participant studies.

As shown in Figure 1, common designs using a single participant include those that do *not* meet basic criteria for single-case methodology, as previously defined, such as clinical case descriptions/case reports, which are "a description of clinical practice that does not involve research methodology" (Backman & Harris, 1999; p. 171), "pre-test/post-test" type of studies of an individual (where data are not collected *during* the treatment phase), and B-phase training studies. Studies which do meet criteria for single case methodology can be further subdivided into those that do and do not have experimental control; the bi-phasic A-B design (also known as "phase change without reversal", Shadish & Sullivan, 2011) is an example of the latter group. Single-case designs that *do* have experimental control, as depicted above the horizontal line in Figure 1, include the medical *n*-of-1 trial (classified as a subset of the "withdrawal" category of designs).

Method quality of single-case reports is extremely variable (Maggin, Chafouleas, Goddard, & Johnson, 2011; Tate et al., 2010) and this has prompted development of resources to improve the conduct and report of single-case studies. Kratochwill et al. (2010; 2013) have published standards relating to the design of and evidence for single-case experiments. In addition, two sets of reporting guidelines following the CONSORT tradition are currently under development: the CONSORT Extension for *N*-of-1 Trials (CENT; personal communication, S Vohra, 26 April 2013) for the medical *n*-of-1 trial, and the Single-Case Reporting guideline In BEhavioural interventions (SCRIBE; Tate et al., in preparation) for the behavioural sciences. In the field of special education, Wolery, Dunlap, and Ledford (2011) have also published reporting guidelines for single-case designs. Finally, method quality rating scales assist in discriminating among the rigor of reports. An increasing number of such instruments is available, including the Single-Case Experimental Design (SCED) Scale developed by our team (Tate et al., 2008), which was described as "perhaps the only psychometrically validated tool for assessing the rigor of SCED methodology" (Smith, 2012, p. 512).

The 11-item SCED Scale evaluates whether a report satisfactorily addresses the following features which are pertinent to method quality of single-case designs: participant characteristics, specification of the target behaviour, research design, sampling of behaviour during each phase, presentation of raw data, reliability of measurement and independence of its assessment, analysis of results, replication, and generalisation. The scale has excellent inter-rater reliability and it discriminates well among reports of varying quality. In using the scale for the PsycBITE project rating many hundreds of single-case reports, however, we saw ways in which it could be improved; in addition we wanted to align the scale with the more recently published standards for single-case research (Kratochwill et al., 2010), thus necessitating a revision. This report is in two parts: first, description of the revised method quality scale, now entitled the Risk of Bias in *N*-of-1 Trials (RoBiNT) Scale and second, examination of its psychometric properties.

PART 1: SCALE REVISION

Three major areas of change to the SCED Scale were addressed: item content was revised and increased, two subscales were developed, and the item scoring format was changed from a binary scale (0 or 1 point) to a 3-point scale (0, 1 or 2 points—see section on scoring procedures). A detailed manual, available from the corresponding author, was developed for the RoBiNT Scale, with all items and scoring levels operationally defined, along with examples of items meeting/not meeting criteria at the various levels. The manual is used as a basis for a 2-day training workshop in which trainees learn how to apply the scale with scoring examples; proficiency is also evaluated with mastery tests. In order to ensure reliable and accurate use of the scale, we advocate that users are trained in its use and consult the manual in applying the scale to critically appraise single-case reports. We are currently developing an on-line rater training programme for the RoBiNT Scale, which will be available through PsycBITE.

Item content

Table 1 presents a comparison between the items of the original SCED and revised RoBiNT Scales. Five new items were added to the scale: three items to strengthen internal validity and two items to more comprehensively evaluate external validity. Two of the original SCED items were amalgamated into a single item (items 4 and 5 on sampling behaviour in the baseline and intervention phases, respectively).

Two of the new internal validity items are those that are widely regarded in the medical literature to be key characteristics to minimise risk of bias (randomisation and blinding of patient and therapist; Higgins, Altman, & Sterne, 2011; Moher et al., 2010). In this context, we also strengthened item 8 of the SCED Scale (independence of assessor) to require blinding of the assessor in the revised scale.

Randomisation, in the context of single-case methodology, refers to “the random assignment of treatment times to treatments” (Edgington, 1987, p. 439), rather than individuals being randomly allocated to treatment groups, as the concept is applied in research designs using groups of participants. More specifically, in single-case methodology, randomisation is used to determine (i) the sequence (or order) of phases, and/or (ii) the onset (or start-point) of treatment phase/s (see also Kratochwill & Levin, 2010).

TABLE 1
Commonality of items between the RoBiNT and SCED Scales

Items in RoBiNT Scale	Comparable SCED Scale items
Internal validity subscale	
1.Design	Yes; item 3
2.Randomisation	No; new item for RoBiNT Scale
3.Sampling behaviour (all phases)	Yes; item 4 (baseline) and item 5 (treatment)
4.Blinding patient/therapist	No; new item for RoBiNT Scale
5.Blinding assessors	Yes, Independence of assessors; item 8
6.Inter-rater reliability	Yes; item 7
7.Treatment adherence	No; new item for RoBiNT Scale
External validity and interpretation subscale	
8.Baseline characteristics	Yes, Clinical history; item 1
9.Therapeutic setting	No; new item for RoBiNT Scale
10.Dependent variable (target behaviour)	Yes; item 2
11.Independent variable (intervention)	No; new item for RoBiNT Scale
12.Raw data record	Yes; item 6
13.Data analysis	Yes; item 9
14.Replication	Yes; item 10
15.Generalisation	Yes; item 11

On the RoBiNT Scale the maximum score (i.e., 2 points) is awarded when either sequence and/or onset of all phases is randomised. Withdrawal and alternating treatment designs are amenable to randomising sequence of phases, and each of the withdrawal, multiple-baseline and changing criterion designs are amenable to randomising the time at which phase change occurs. Randomisation of onset is made feasible by, for example, employing a window of time (e.g., session 4 to 10) during which change of phase is randomly determined (e.g., onset of the next phase to occur at session 8). Wolery (2013), however, points to a number of difficulties with randomisation in the context of single-case designs and researchers need to consider potential problems when designing a study.

In incorporating the *blinding items*, to our knowledge, this issue has not been raised previously with respect to single-case methodology in the behavioural sciences. We acknowledge that blinding of participants and therapists to phase of intervention (required for a 2-point score on the RoBiNT Scale) is not easily achievable in behavioural interventions, and accordingly they were amalgamated into a single item. Including items on blinding, however, serves as a reminder of their critical role in minimising bias. Moreover, blinding of participants will become increasingly possible in some fields of neurorehabilitation with the development of new technologies in which sham treatments can be applied, such as the use of transcranial magnetic stimulation. Blinding of the assessor to phase of intervention, however, is feasible in many behavioural interventions.

The final new item introduced to strengthen the internal validity of the scale is *treatment adherence* (a component of treatment fidelity). Many years ago Peterson, Homer, and Wonderlich (1982) had observed that the focus of attention in single-case research was on "ensuring the integrity of the dependent variable", without due regard for the independent variable (the intervention). This problem is also encountered in group-based research (Glasziou, Meats, Heneghan, & Sheppherd, 2008). Certainly this was the orientation of the SCED Scale, with 8/11 items involving measurement of the target behaviour and other outcome variables. Peterson and colleagues argued that accurate and reliable description of the independent variable is equally important in terms of establishing a functional (causal) relationship between independent and dependent variables. Because the results of a study pertain solely to the treatment that was administered, it is necessary to know, at the minimum, how closely the actual administration of treatment corresponds to the planned administration of treatment.

Treatment fidelity has a number of components. Borrelli (2011) describes five domains: (i) study design (the treatment reflects its underlying theoretical principles); (ii) training, monitoring and maintaining the therapist's skill; (iii) delivery of treatment; (iv) receipt of treatment by the patient; and (v) enactment (i.e., what is used in real-life settings). Borelli's group has developed a 30-item treatment fidelity checklist (Borrelli et al., 2005). The treatment adherence item in the RoBiNT Scale focuses on part of one of Borelli's domains, delivery of treatment, specifically treatment adherence or integrity. Using the literature as a guide to select criteria, the item evaluates four components of treatment adherence and 2 points are awarded when all of the following criteria are met: (i) the person rating adherence is independent of the therapist; (ii) an explicit statement is made regarding the aspect of the intervention that is being rated and the method used to rate it, which must involve a direct, quantitative measure; (iii) a minimum of 20% of the intervention sessions are rated; and (iv) the adherence check must result in at least 80% compliance with the rating protocol.

The final two new items of the scale relate to external validity, and also address the independent variable: description of (a) the therapeutic *setting* and (b) the *intervention*. Horner et al. (2005) discuss the significance of these areas for methodological rigor. Specifically, both the setting and treatment should be described in precise and operational terms, as well as in sufficient detail to allow replication. For 2 points, the setting needs to provide information on the specific environment (i.e., configuration of the space in which the intervention is provided, e.g., description of the therapy room, classroom etc., including placement of equipment, social context, etc), and may also include the general location (eg., hospital, school). Similarly, the intervention needs to be described in detail, including the number, duration and periodicity of sessions.

Subscales

The original SCED Scale had a mix of items evaluating internal validity (e.g., items 3, 7, 8 for design, observer bias, and independence of assessor, respectively), external validity (e.g., items 10, 11 for replication and generalisation, respectively), and interpretation (e.g., item 9, statistical analysis). The PsycBITE database publishes (advisedly) a summary total score, along with the score for each of the 11 individual items. With the extra items added to the RoBiNT Scale it was possible to consider internal and external validity separately. Accordingly, the items were subdivided into two scales, the Internal Validity (IV) subscale with seven items and the External Validity and Interpretation (EVI) subscale with eight items, (see Table 1).

Scoring procedures

The third major revision to the SCED Scale involved the scoring system. The SCED Scale, following the PEDro Scale (Maher, Sherrington, Herbert, Moseley, & Elkins, 2003), used a binary scoring system (criterion met/not met), with criteria for each item being operationally defined in our in-house manual. For example, the SCED Scale item 4 (sampling of behaviour during baseline) required sufficient sampling of behaviour in the baseline phase, operationally defined in the SCED manual as a minimum of 3 data points, this being the recommended minimum number of data points suggested by authorities to demonstrate stability in the data (Barlow & Hersen, 1984; Beeson & Robey, 2006). More recently, the design standards in the field of special education have recommended at least 5 data points per phase (Kratochwill et al., 2010; 2013). We considered it important that, as with our original SCED Scale, the revised scale should be consistent with recommendations made by authorities in the field. Yet if the binary scoring system were retained, reports not meeting stringent criteria of the design standards (e.g., at least 5 data points per phase) would be discounted by being awarded a 0 score. For this reason, a 3-point rating scale was introduced, whereby 2 points were awarded for meeting the recommended stringent criteria and 1 point generally corresponded to the original SCED Scale criteria (and similarly for 0 points). The decision to change to a 3-point scale is compatible with Kratochwill and colleagues, who have an intermediate category ("meets standards with reservations") which generally corresponds to criteria that were considered adequate in the past (e.g., a minimum of 3 data points per phase).

Changing the scoring system required providing a 3-point scale for all 15 items. Ten of the RoBiNT Scale items related to the original 11 SCED items (as noted, two of the SCED items were amalgamated into a single item for the revised scale, corresponding to 10 RoBiNT Scale items) and five new items were introduced. Criteria for the award of the 2-point score for the new items were described in the foregoing section on item content; for the original SCED items used in the RoBiNT Scale, 2 or 1 points were awarded where the item met recommended criteria as follows. It should be noted that the criteria described in this report are summary notation only, and it is not advisable to score reports from this description alone. A manual is available for the purpose of applying the RoBiNT scale.

- *Design (item 1):* At least three demonstrations of the treatment effect (e.g., A-B-A-B; 6-phase multiple-baseline; Horner et al., 2005; Kratochwill et al., 2010; 2013); 1 point awarded for two demonstrations of the treatment effect
- *Sampling of behaviour (item 3):* At least 5 data points in every phase (Horner et al., 2005; Kratochwill et al., 2010; 2013); 1 point for less than 5 data points in any phase but at least 3 data points in each phase
- *Blinding of assessor (item 5):* Increasingly, single-case reports use an assessor who is independent of the therapist (score 1 point), however, to score 2 points the assessor needs to be blind to the phase of intervention
- *Inter-rater reliability (item 6):* Although Kazdin (2011) notes that there are no precise rules for the frequency with which agreement should be checked, the standards described by Kratochwill et al. (2013) specify 20% of data sampled, analysed and reported per condition, and percent agreement (or equivalent) 80% or higher; percent agreement of 70% to 79%, even if no other information on sampling is provided, score 1 point. Use of machine-generated data, free from human judgement, or "reasonably objective measures" (as defined in the manual) are awarded 2 and 1 points, respectively
- *Baseline characteristics (item 8):* The importance of understanding the baseline conditions and characteristics which serve to maintain the target behaviour has been stressed (Horner et al., 2005; Lane, Wolery, Reichow, & Rogers, 2007). Recommended practice is that these conditions and variables should be considered and evaluated prior to commencing the experiment, and the way that they inform the intervention should be articulated in the report (required for 2 points). A functional analysis is one way of achieving this end, but other less formal procedures may be used as well. The evaluation goes beyond provision of demographic, medical and functional status variables, or a clinical profile of test scores (score 1 point).
- *Target behaviour (item 10):* Many reports provide an operational definition of the target behaviour (score 1 point), but fail to describe

and/or to use a precise and repeatable measure of the target behaviour, nor specify what constitutes a correct/incorrect response. All three elements are required to score 2 points

- *Raw data record (item 12)*: A natural division occurs between those reports which provide a complete record of the raw data at a session-by-session level, versus those which provide incomplete data (e.g., multiple probe designs; aggregated data). Only the former are awarded 2 points; whereas the latter score 1 point. An ad-hoc selection of data, whether provided in graphed or tabular format, does not score any points
- *Data analysis (item 13):* Controversy remains about whether the appropriate method of analysis in single-case reports is visual or statistical. Nonetheless, 2 points are awarded if systematic visual analysis is used according to steps specified by Kratochwill et al. (2010; 2013), or visual analysis is aided by quasi-statistical techniques, or statistical methods are used where a rationale is provided for their suitability. One point is awarded if systematic/aided visual analysis is incomplete/not conducted for every phase change or no rationale is provided for statistical analysis
- *Replication (item 14):* In the context of this item, replication refers to the repetition of the entire experiment, which may be direct intersubject and/or systematic replication. This item does not refer to replication of the experimental effect, which is covered by item 1 (design). For the award of 2 points at least three replications are required (i.e., original + 3 replications; Barlow et al., 2009); one or two replications (i.e., original + 1 or 2 replications) are awarded 1 point
- *Generalisation (item 15):* Generalisation measures, whether response generalisation to other behaviours or stimulus response to other settings, need to be intentionally programmed into the design of the study, rather than a passive expectation of a “train and hope” approach (Stokes & Baer, 1977). Generalisation measures need to be evaluated throughout all phases of the experiment (Schlosser & Braun, 1994) in order to score 2 points. An accepted level of generalisation that would score 1 point is to evaluate generalisation measures prior to and at the conclusion of treatment

PART 2: PSYCHOMETRIC EVALUATION OF THE ROBINT SCALE

Method

The study used the 15-item RoBiNT Scale as described above.

The development and psychometric examination of the original SCED Scale (Tate et al., 2008) had been trialled on more than 100 single-case reports archived on PsycBITE. The revised scale was formally trialled on an additional 40 reports, which included a random sample of 5% ($n = 24/433$) of papers published between 2002 and 2011 for either stroke or traumatic brain injury archived on PsycBITE. Trialling involved independent rating of papers by the authors, meetings to discuss ratings, and, where necessary, emending the phrasing of items, clarifying rating criteria, and so forth. The revision process occurred over approximately a three-year period. Inter-rater reliability of the RoBiNT scale was then examined with an additional independent set of 20 single-case reports randomly selected from the remaining pool of 409 reports.

Two pairs of raters were used in the inter-rater reliability study: Pair 1 (authors RT and MP) were "experienced" raters and Pair 2 (authors UR and DW) were "novice" raters who were knowledgeable about evidence-based principles and single-case methodology and were trained in the use of the scale. Raters in each pair independently rated the 20 papers and subsequently met to resolve discrepancies and provide consensus ratings. The consensus ratings of the two pairs were then compared. Inter-rater reliability between the individual raters of each pair was calculated with the intraclass correlation, using the random effects model, both for the total score for the scale as a whole, as well as the two subscale scores. Intraclass correlation coefficients (ICC) were interpreted using the criteria of Cicchetti (1994, 2001), where ICC of .75 or higher is classified as excellent, between .6 and .74 as good, .4 to .59 as fair, and less than .4 as poor. Percent agreement was used to calculate inter-rater agreement for the individual items. The conventional criterion for acceptable percent agreement of 80% or higher (Anastasi & Urbina, 1997) was adopted.

Results

Using data from one of the experienced raters (Pair 1; author MP), the RoBiNT Scale took an average of 26.2 minutes ($SD = 11.28$) per paper to rate. After consensus, the mean score for Pair 1 was 10.60/30 ($SD = 4.50$; range = 2–18) for the total score, 1.90/14 ($SD = 2.27$; range = 0–6) for the IV subscale, and 8.70/16 ($SD = 2.64$; range = 2–12) for the EVI subscale.

Table 2 presents results of the inter-rater reliability analyses. For the experienced raters (Pair 1), ICC was .90 (95% CI: .76–.96) for the total score, .88 (95% CI: .70–.95) for the IV subscale, and .87 (95% CI: .67–.95) for the EVI subscale. For the trained novice raters (Pair 2), the very high ICCs were replicated (ICC = .88, 95% CI: .70–.95 for the total score; ICC = .87, 95% CI: .66–.95 for the IV subscale; and ICC = .93, 95% CI: .81–.97 for the EVI subscale). Reliability of consensus ratings between the experienced and novice raters was also excellent for the total score (ICC = .95, 95% CI: .88–.98), as well as the subscales, both ICC = .93 (95% CI: .82–.97).

TABLE 2
Inter-rater reliability of the RoBiNT scale

	Pair 1 "experienced raters"		**Pair 2 "novice raters"**	
Item	**Base rate (% scoring 1 or 2)**	**% agreement**	**Base rate (% scoring 1 or 2)**	**% agreement**
Internal validity (IV) subscale				
1: Design	40	70	40	85
2: Randomisation	10	95	0	95
3: Sampling	35	90	30	85
4: Blind participant/therapist	0	100	0	100
5: Blind assessors	0	90	15	80
6: Inter-rater reliability	35	85	30	80
7: Treatment adherence	10	95	10	80
External validity and interpretation (EVI) subscale				
8. Baseline characteristics	85	55	95	75
9. Therapeutic setting	55	95	50	70
10. Target behaviour	85	45	75	50
11. Intervention	95	70	95	80
12. Raw data record	65	95	60	80
13. Data analysis	55	80	55	80
14. Replication	65	90	65	95
15. Generalisation	50	70	65	80
Total (ICC)		.90		.88
95% confidence interval		(.76–.96)		(.70–.95)
IV subscale (ICC)		.88		.87
95% confidence interval		(.70–.95)		(.66–.95)
EVI subscale (ICC)		.87		.93
95% confidence interval		(.67–.95)		(.81–.97)

Base rate refers to the proportion both raters agreed that the criterion was met. Percent agreement refers to the proportion of times the raters agreed on a score (the criterion may or may not have been met).

At the item level, agreement between Pair 1 on the IV subscale ranged from 70% (item 1: design) to 100% (item 4: blind patients/therapists); on the EVI subscale agreement ranged from 45% (item 10: target behaviour) to 95% (items 9 and 12: setting and raw data, respectively). A similar pattern of results was demonstrated for Pair 2 (see Table 2).

Using consensus ratings from Pair 1, Figures 2a (IV subscale) and 2b (EVI subscale) show base rate data (the percentage of reports meeting criteria on items of the scale). A larger proportion of reports met criterion (either 1 or 2 points) on items from the EVI than the IV subscales. On the IV subscale, all reports scored 0 for items 4 and 5 (blind patients/therapists and blind assessors, respectively), as did 90% for items 2 and 7 (randomisation and treatment adherence, respectively), 65% for items 3 and 6 (sampling of behaviour and inter-rater reliability, respectively) and 60% for item 1 (design). By contrast, on the EVI Scale, 50% or less of all reports obtained the 0 score on any item. Conversely, very few reports obtained the stringent 2-point score on items from the IV subscale (*n* = 4 items; range 5% for items 6 and 7, inter-rater reliability and treatment adherence respectively, to 30% for item 1, design). Results were again more favourable for the EVI subscale, in which the stringent 2-point criterion was met by all items, ranging from 20% for items 9 and 15 (setting and generalisation, respectively) to 75% for item 11 (intervention).

Construct validity of the RoBiNT Scale was examined by comparing the design types using consensus ratings from Pair 1. In this randomly selected series, 7/20 reports did not use single-case methodology, as defined (i.e., a study of a single individual which contained phases, along with repeated and frequent measurement of the target behaviour during all phases). They comprised case description with data (*n* = 1), B-phase training study (*n* = 1),

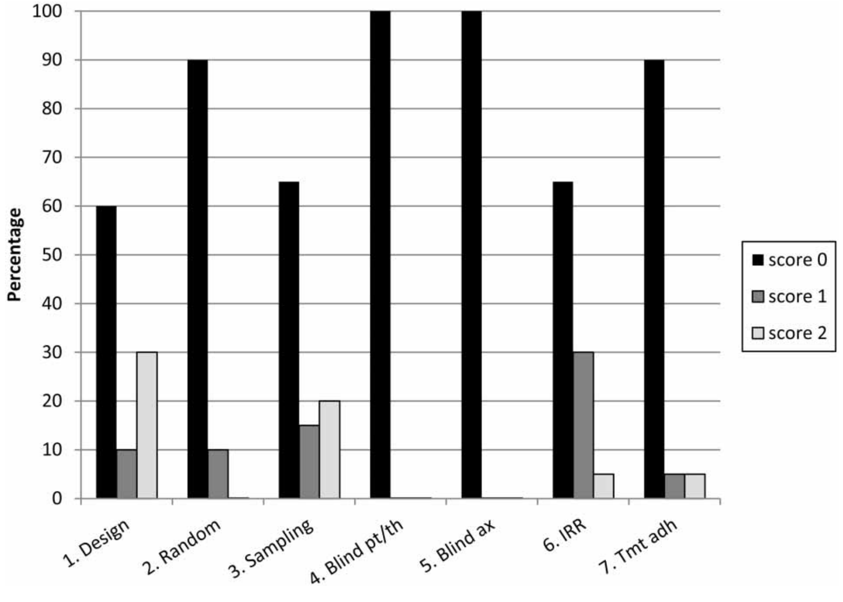

Figure 2a. Percentage of single-case reports meeting criteria on RoBiNT Internal Validity items (Pair 1). Random = randomisation; ax = assessor; pt/th = participant/therapist; IRR = inter-rater reliability; Tmt adh = treatment adherence.

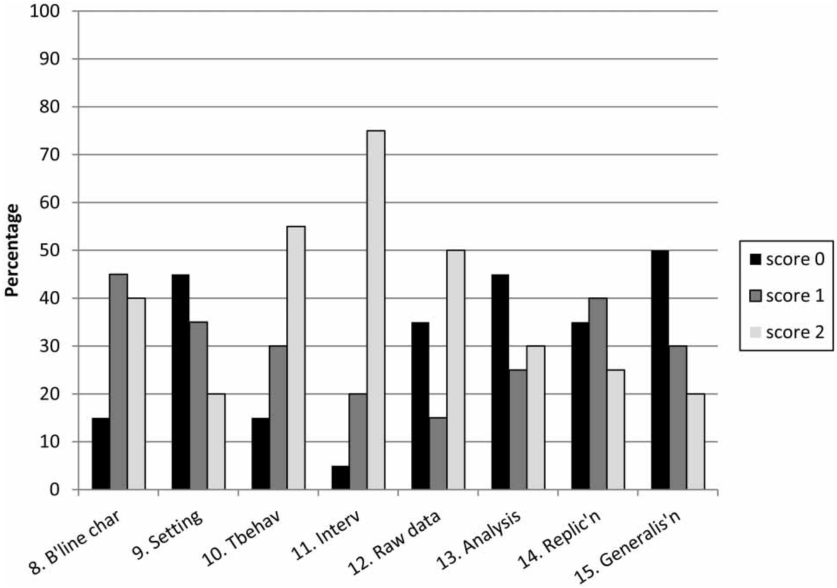

Figure 2b. Percentage of single-case reports meeting criteria on RoBiNT External Validity and Interpretation items (Pair 1). B'line char = baseline characteristics for participants; Tbehav = target behaviour; Interv = intervention; Replic'n = replication; Generalis'n = generalisation.

and "pre-test/post-test" type designs ($n = 5$). The remaining 13/20 reports used single case methodology (A-B, $n = 1$; A-B with additional phase change but without reversal or interpolated baseline, $n = 3$; A-B-A, $n = 3$; multiple-baseline, $n = 5$; alternating treatment, $n = 1$). These two subgroups were compared on a total score of 14/15 items (item 1, design, which was a de facto grouping variable, was excluded from analysis). Designs using single-case methodology obtained a significantly higher total score than reports that did not use single-case methods ($M = 12.0/28$, $SD = 2.71$ vs. $M = 6.0/28$, $SD = 2.24$, respectively; $z = -3.54$, $p < .001$).

Discussion

This study demonstrated that the revised method quality rating scale for single-case studies, the RoBiNT Scale, shows evidence of construct (discriminative) validity in that there was a significant difference in the total score between reports that did ($n = 13$) and did not ($n = 7$) use single-case methodology. The scale also has excellent inter-rater reliability for the total score for both experienced and trained novice raters (ICC = .90 and .88, respectively), as well as the subscales for each of IV (ICC = .88 and .87, respectively) and EVI (ICC = .87 and .93, respectively). The ICCs of the revised scale are better than those of the original SCED Scale for the (experienced) individual raters' total score (ICC = .90 vs .83, respectively). Moreover, the results surpass inter-rater reliability studies of other method quality rating scales, such as the PEDro scale for RCTs (individual raters' total score ICC = .56; Maher et al., 2003), and the AMSTAR scale for systematic reviews (total score ICC = .84; Shea et al., 2009).

Agreement at the item level on the RoBiNT Scale had a larger range (45% to 100% agreement for Pair 1, experienced raters) than occurred with the original SCED scale (77% to 97%; Tate et al., 2008), and five of the 15 items had agreement below the traditionally accepted criterion of 80%. There are a number of possible reasons to account for this difference in level of agreement between the original and revised scales. The most obvious factor relates to use of a 3-point scale for the revised scale as opposed to the binary score system of the original scale, thereby introducing more scope for rater disparity. Yet there are other contributing factors to some of the low rates of agreement. For example, the item with the lowest agreement between the experienced raters, at 45%, was item 10 (target behaviour) from the EVI subscale. In spite of the clear and operationalised criteria for all items in the RoBiNT Scale manual, in the course of conducting the inter-rater reliability study the raters observed that many of the reports themselves provided unclear and/or incomplete reporting of the dependent variable (target behaviour), as well as the independent variable (intervention, item 11, where rater agreement was 70%). In these cases it was necessary for raters to scour the report to find evidence to substantiate a particular score which often proved difficult (and was time consuming).

Item 1 (design) from the IV subscale, at 70% agreement, was also below the threshold for acceptable inter-rater agreement. Yet, in the course of the rating process, it was noted that only 7/20 reports (35%) correctly specified the methodological design used in the study. In 45% of papers ($n = 9$), the authors did not provide any information on the specific design used in the study and consequently raters had to infer the type of design from other information contained in the report. Even more alarming, in 20% of papers ($n = 4$) the authors incorrectly described the design (e.g., stating they used a multiple baseline design when in fact a "pre-test/post-test" type of design was used in which one or more target behaviours were evaluated at the pre-intervention assessment on a number of occasions but the intervention was not introduced to the behaviours in a time-lagged fashion, nor were measures of the target behaviours taken during the intervention phase). In these cases, the raters had to override the authors' description of the design, thereby adding to the complexity and difficulty of the inter-rater reliability process. The foregoing difficulties highlight the pressing need for the introduction of reporting guidelines such as CENT and SCRIBE that are currently under development, which will provide authors with guidance on issues that should be addressed in their reports, as well as model examples of good reporting.

Principles and procedures used in developing the original 11-item SCED Scale have been described in detail elsewhere (Tate et al., 2008). The revised RoBiNT Scale incorporates five new items, strengthens one of the original items, uses a 3-point scale to discriminate between reports meeting adequate versus stringent criteria, and scoring criteria are cast within the framework of presently accepted design and evidence standards. This makes the scale more comprehensive, rigorous and discriminating than the SCED Scale. The base rate data in Figures 2a and 2b demonstrate that, as a whole, this small random selection of reports did *not* have good method quality. This was especially the case for internal validity, with between 60% and 100% of reports obtaining a zero score on at least one of the items of the IV subscale. In this respect, we acknowledge that the reports used in this psychometric study are at a disadvantage because they are being evaluated against a scale that was not available at the time that their study was conducted. Nonetheless, the scale provides guidance on ways in which internal and external validity of single-case methods can be strengthened.

The revised scale still adheres to the principles of our original scale in that we aimed to produce an instrument that was (a) feasible to administer, (b) incorporated a minimum set of essential features of single-case designs required for their validity, (c) discriminative in terms of method quality, and (d) able to be administered reliably. In so doing, we acknowledge some limitations of the scale. In spite of the increased number of items, the universe of items pertinent to internal and external validity of single-case designs is not included (e.g., evaluation of social validity, procedural fidelity of the baseline conditions). The number of items was intentionally restricted to a minimum core set to meet our primary purpose for developing the scale; namely, to have an instrument that was feasible to use in the ongoing rating of a very large number of single-case reports archived on our PsycBITE database. Another limitation pertains to the scoring system. At this stage of its development the scale uses an ordinal level of measurement whereby the items are weighted equally, but in reality the inherent methodological importance of the items inevitably differs. Consequently, we advise that methodological robustness of a study should not be decided solely on the basis of summed scores, but rather in conjunction with scores at the individual item level. In addition, as Wilson (2011) notes, "studies can be highly rated for methodology but still report on trivial, meaningless or unimportant aspects" and we agree that this aspect of external validity always needs to be taken into account in determining the value of a study.

In conclusion, the 15-item RoBiNT Scale, which builds on our original SCED Scale, is a reliable and valid instrument to measure risk of bias in single-case reports. It was originally intended for rating studies that used single-case methodology, but in this study it has also proved useful in rating method quality of all levels of studies using a single participant. The RoBiNT Scale is an improvement on the original scale in that item content is more comprehensive, the more detailed scoring system enables greater discrimination among reports, and major components of methodology are evaluated separately. Use of method quality scales, along with reporting guidelines that are currently under development such as the CENT and SCRIBE, will provide long overdue guidance for clinicians and researchers to plan, implement, and report on single-case methodology. Hopefully, this will result in a better quality literature in this field.

REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall International Inc.

Backman, C. L., & Harris, S. R. (1999). Case studies, single-subject research, and *n* of 1 randomized trials: Comparisons and contrasts. *American Journal of Physical Medicine and Rehabilitation*, *78,* 170–176.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior for change.* (2nd ed.). Boston: Allyn and Bacon.

Barlow, D. H., Nock, M. K, & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior for change.* (3rd ed.). Boston: Pearson/Allyn and Bacon.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review*, *16,* 161–169.

Borrelli, B. (2011). The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. *Journal of Public Health Dentistry*, *71,* S52–S63.

Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A. J., Czajkowski, S., Breger, R., … Orwig, D. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology*, *73*(5), 852–860.

Boutron, I., Guittet, L., Estellat, C., Moher, D., Hróbjartsson, A., & Ravaud, P. (2007). Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. *PLoS Medicine*, *4*(2), 370–380.

Boutron, I., Tubach, F., Giraudeau, B., & Ravaud, P. (2004). Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. *Journal of Clinical Epidemiology*, *57,* 543–550.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6,* 284–290.

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology*, *23*(5), 695–700.

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology*, *34*(4), 437–442.

Gast, D. L. (2010). *Single subject research methodology in behavioural sciences*. New York, NY: Routledge.

Glasziou, P., Meats, E., Heneghan, C., & Sheppherd, S. (2008). What is missing from descriptions of treatment in trials and reviews? *British Medical Journal*, *336,* 1472–1474.

Guyatt, G., Jaeschke, R., & McGinn, T. (2002). *N* of 1 randomized controlled trials. In G. Guyatt, D. Rennie, M. O. Meade, & D. J. Cook (Eds.), User's guides to the medical literature (pp. 179–192). New York, NY: McGrawHill Medical and American Medical Association.

Guyatt, G. H., Keller, J. L., Jaeschke, R., Rosenbloom, D., Adachi, J. D., & Newhouse, M. T. (1990). The *n*-of-1 randomised controlled trial: Clinical usefulness. Our three-year experience. *Annals of Internal Medicine*, *112,* 293–299.

Guyatt, G., Sackett, D., Adachi, J., Roberts, R., Chong, J., Rosenbloom, D., & Keller, J. (1988). A clinician's guide for conducting randomized trials in individual patients. *Canadian Medical Association Journal*, *139,* 497–503.

Guyatt, G., Sackett, D., Taylor, D. W., Chong, J., Roberts, R., & Pugsley, S. (1986). Determining optimal therapy—randomised trials in individual patients. *New England Journal of Medicine*, *314*(14), 889–892.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behaviour change*. New York, NY: Pergamon.

Higgins, J. P. T., Altman, D. G., & Sterne, J. A. C. (2011). Assessing risk of bias in included studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions*, Version 5.1.0 (updated March 2011). The Cochrane Collaboration. Available from http://www.cochrane-handbook.org.

Horner, R. H., Carr, E. C., 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*(2), 165–179.

Howick, J., Chalmers, I., Glasziou, P., Greenhaigh, T., Heneghan, C., Liberati, A., . . . Thornton, H. (2011). *The 2011 Oxford CEBM Evidence Table* (Introductory Document). Oxford: Oxford Centre for Evidence-Based Medicine. Available from: http://www.cebm.net/index.aspx?o=5653

Johnston, B. C., & Mills, E. (2004). n-*of-1* randomized controlled trials: An opportunity for complementary and alternative medicine evaluation. *Journal of Alternative and Complementary Medicine*, *10*(6), 979–984.

Kazdin, A. E. (1982). *Single case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

Kazdin, A. E. (2011). *Single-case research designs. Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Keller, J. L., Guyatt, G. H., Roberts, R. S., Adachi, J. D., & Rosenbloom, D. (1988). An *N* of 1 service: Applying the scientific method in clinical practice. *Scandinavian Journal of Gastroenterology (Suppl.)*, *147,* 22–29.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. What Works Clearinghouse website:http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*(2), 124–144.

Lane, K., Wolery, M., Reichow, B., & Rogers, L. (2007). Describing baseline conditions: Suggestions for study reports. *Journal of Behavioral Education*, *16,* 224–234.

Maher, C. G., Sherrington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of RCTs. *Physical Therapy*, *83,* 713–721.

Maggin, D. M., Chafouleas, S. M., Goddard, K. M., & Johnson, A. H. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behaviour. *Journal of School Psychology*, *49,* 529–554.

McDougall, D. M. (2013). Applying single-case design innovations to research in sport and exercise psychology. *Journal of Applied Sport Psychology*, *25,* 33–45.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsch, P. C., Devereaux, P. J., … Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, *63*(8), e1–e37.

Perdices, M., & Tate, R. L. (2009). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognised and undervalued? *Neuropsychological Rehabilitation*, *19*(6), 904–927.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behaviour analysis. *Journal of Applied Behavior Analysis*, *15,* 477–492.

Schlosser, R. W., & Braun, U. (1994). Efficacy of AAC interventions: Methodologic issues in evaluating behaviour change, generalization, and effects. *AAC Augmentative and Alternative Communication*, *10,* 207–223.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–980.

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., … Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology*, *62*(10), 1013–1020.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis*, *10,* 349–367.

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Godbee, K., Cassel, A., … Rosenkoetter, U. (2010). Single-case experimental designs and *n*-of-1 trials in rehabilitation research: How good is the research in acquired brain impairment? Paper presented at the mid-year meeting of the International Neuropsychological Society, Krakow, Poland; abstract in *Journal of the International Neuropsychological Society*, *16* (Supplement 2), 23.

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and *n*-of-1 trials: Introducing the Single-case Experimental Design (SCED) Scale. *Neuropsychological Rehabilitation*, *18*(4), 385–401.

Tate, R. L., Perdices, M., McDonald, S., Togher, L., Moseley, A., Winders, K., … Smith, K. (2004). Development of a database of rehabilitation therapies for the psychological consequences of acquired brain impairment. *Neuropsychological Rehabilitation*, *15*(5), 517–534.

Tate, R. et al. (in preparation). The Single-Case Reporting guideline In BEhavioural interventions (SCRIBE): Explanation and elaboration.

Wilson, B. A. (2011). "Cutting edge" developments in neuropsychological rehabilitation and possible future directions. *Brain Impairment*, *12*(1), 33–42.

Wolery, M. (2013). A commentary: Single-case design technical document of the What Works Clearinghouse. *Remedial and Special Education*, *34*(1), 39–43.

Wolery, M., Dunlap, G., & Ledford, J. R. (2011). Single-case experimental methods. Suggestions for reporting. *Journal of Early Intervention*, *33*(2), 103–109.

Wolery, M., & Ezel, H. K. (1993). Subject descriptions and single-subject research. *Journal of Learning Disabilities*, *26*(10), 642–647.

Manuscript received June 2013
Revised manuscript received July 2013
First published online August 2013