Small Sample Research Designs for Evidence-Based Rehabilitation: Issues and Methods

James E. Graham, PhD, DC, Amol M. Karmarkar, PhD, OTR, Kenneth J. Ottenbacher, PhD, OTR


Conventional research methods, including randomized controlled trials, are powerful techniques for determining the efficacy of interventions. These designs, however, have practical limitations when applied to many rehabilitation settings and research questions. Alternative methods are available that can supplement findings from traditional research designs and improve our ability to evaluate the effectiveness of treatments for individual patients. The focus on individual patients is an important element of evidenced-based rehabilitation. This article examines one such alternate approach: small-N research designs. Small-N designs usually focus on 10 or fewer participants whose behavior (outcomes) are measured repeatedly and compared over time. The advantages and limitations of various small-N designs are described and illustrated using 3 examples from the rehabilitation literature. The challenges and opportunities of applying small-N designs to enhance evidence-based rehabilitation are discussed.

Key Words: Evaluation studies as topic; Evidence-based practice; Rehabilitation; Research design. © 2012 by the American Congress of Rehabilitation Medicine

Randomized controlled trials (RCTs) are the most widely recommended research approach to evaluate treatment efficacy in biomedical research.1-2 Efficacy studies are designed to demonstrate causal relationships between treatment and outcomes under strictly controlled conditions. RCTs are also the most highly valued source of research evidence, selecting treatments for individual patients based on evidence-based practice guidelines such as those established by the Center for Evidence Based Medicine (http://www.cebm.net/) and Cochrane Collaboration (http://www.cochrane.org/).

Some of the same qualities that make RCTs the criterion standard for efficacy research may limit their application in assessing the effectiveness of a given intervention for an individual patient. Effectiveness studies are designed to examine the effects of an intervention with typical patients in everyday situations wherein an investigator cannot control all the extraneous factors. RCTs tend to have strict inclusion and exclusion criteria and typically report average treatment effects obtained from statistical comparisons of group-level (aggregate) data from experimental and control groups. The article by Horn et al3 in this issue, as well as earlier reviews by Grimmer,4 Kravitz,2 and colleagues provide more details on why it is sometimes difficult to extrapolate findings from RCTs to everyday clinical practice.

An alternative research approach, broadly classified as small-N designs, involves serial observations of single persons or small groups before, during, and after an intervention period. This patient-level focus can facilitate evidence-based practice in 2 broad ways. First, it enables researchers to provide clinicians with practical information for making decisions to improve the care of individual patients. Second, it provides a potential avenue for including evaluation and research design in clinical practice and building the foundation for evidence-based rehabilitation at the level of the individual patient in actual treatment settings.5 Guyatt et al6 have suggested a revision in the hierarchy of best evidence to include N-of-1 randomized trials as the highest or best level of evidence for individual decision making. The purposes of this article are to: (1) provide an overview and select examples of small-N designs, (2) describe analysis and interpretation options for these designs, and (3) discuss the opportunities and challenges for small-N designs in rehabilitation research.

Small-N designs are not new. This approach is well-established in education and the behavioral sciences and is increasingly present in the clinical literature.7-9 The textbook by Bloom et al10 is an excellent resource for using small-N designs to evaluate and inform clinical practice. This text addresses many of the issues (eg, statistical vs visual analysis) that are not adequately covered in this article due to space limitations.

Several different terms are used to describe small-N methodology, including single-subject, N-of-1 trials, single-case, and single-system designs, among others. Regardless of the terminology, the design framework is essentially the same: (1) studying a single person or small group of persons over time, (2) repeated measurement of the outcome, and (3) the sequential application and withdrawal of (or variation in) the intervention.

Unlike traditional parallel group RCTs where the independent variable (intervention) is standardized, small-N designs may involve a dynamic intervention that is manipulated across phases of the design.10 The manner in which the intervention is applied distinguishes different categories of small-N designs. While the science of small-N research continues to evolve,11 there are 3 fundamental design frameworks frequently described in the small-N literature: basic and experimental

List of Abbreviations

| AFO | ankle-foot orthosis |
| RCT | randomized controlled trial |
drawal designs, multiple baseline designs, and changing intensity or alternating treatments designs.\textsuperscript{5,12} It is important to understand that the small-N designs described in the next 3 sections vary in their ability to draw causal inferences linking the treatment (independent variable) and outcome (dependent variable), in the same manner that traditional designs based on group comparisons also vary in their ability to control threats to internal validity.

**BASIC (A-B) AND EXPERIMENTAL WITHDRAWAL (A-B-A . . .) DESIGNS**

The basic design in this classification involves an A-B structure, where A represents the baseline (nontreatment) phase and B refers to the intervention phase. The outcome measure is recorded repeatedly over both phases. This model is appealing and easy for clinicians to understand because it is closely related to what happens in day-to-day clinical practice.\textsuperscript{8} However, it is important to differentiate the A-B design from a case study, which is principally a detailed after-the-fact descriptive summary without systematic manipulation of the independent variable; that is, in a case study there is no established baseline and transition to an intervention phase.

The key strengths of the A-B design are its applicability to almost any clinical setting or problem, as well as its simplicity in evaluating whether changes occurred in the outcome after the transition from baseline to intervention. The primary weakness of the A-B design is its inability to differentiate coincidence from causality with only a single transition between the baseline (nontreatment) and treatment phase.\textsuperscript{5} Because the information regarding patient performance is recorded repeatedly over time, A-B designs can be used to document patterns of clinical change, but they cannot unequivocally demonstrate a causal connection between the change in the outcome variable and a specific treatment.

There are several experimental withdrawal designs within this framework (eg, A-B-A, A-B-A-B, B-A-B) that increase the ability to study effects associated with establishing a causal inference. Horner et al\textsuperscript{10} suggest that at least 3 transitions are necessary to begin to establish causality and/or internal validity. For example, in the A-B-A-B design, the targeted outcome measure is recorded repeatedly over all phases, the intervention (B) applied for a period, then withdrawn (second A phase), and then reintroduced (second B phase). The premise is to observe changes in the patterns of the outcome(s) with both the initiation and withdrawal of the intervention. Substantial improvements during the intervention phase(s) and subsequent reversal to baseline levels after withdrawal of the intervention increase confidence in the validity of the inference that a causal relationship exists.\textsuperscript{13}

While more powerful than the basic A-B design, these experimental withdrawal designs may have ethical limitations (eg, removing a seemingly successful intervention) or practical (design) issues (eg, interventions that produce carryover effects, or outcomes that are stable behaviors that cannot be expected to return to pretreatment levels).\textsuperscript{2} An experimental withdrawal design may be an efficient approach to examining the effects of an adaptive device or assistive technology, but these designs are less useful in studies of rehabilitation interventions leading to enduring gains in performance,\textsuperscript{3} such as increased strength.

Figure 1 shows an A-B design with follow-up (for 1 of 5 participants) in a study on the effects of mass practice using constraint-induced movement therapy on lower-extremity function in patients with stroke.\textsuperscript{14} Performance on the Timed Up & Go test was observed repeatedly during 6 sessions of the 2-week baseline phase (A), 6 sessions of the 2-week intervention phase (B), and 2 follow-up sessions at 3 and 6 months. The baseline phase demonstrates some variability in performances across the 6 sessions, but no discernable change (or slope) in the data points. The dashed horizontal line indicates the value that is 2 SDs below the mean of the baseline phase. Assumptions of normality notwithstanding, for data that show little to no trend over the baseline phase, the 2 SD bands are considered thresholds for statistical significance for \( \alpha = .05.\textsuperscript{15} \) A gradual, but consistent slope (decrease in time needed to perform the test) can be seen across the 2-week intervention phase. In addition, the faster performances in the Timed Up & Go test appear to be retained at both 3- and 6-month follow-up. Note, the third phase was intentionally designed to be a follow-up rather than a baseline (A) withdrawal phase, because the strength and functional gains achieved through the intervention are lasting effects that would not be expected to return to baseline values immediately after terminating the training program.

**MULTIPLE BASELINE DESIGNS**

Multiple baseline designs are viable alternatives to the experimental withdrawal designs in situations where either removing the intervention is inappropriate ( unethical) or the treatment effects are long-lasting. Multiple baseline designs can target changes across 3 areas: (1) multiple baseline across subjects (the most common form of this design), (2) multiple baseline across behaviors (outcomes), and (3) multiple baseline across settings. The structure for these designs is a direct extension of the basic A-B design, where 2 or more target A-B pairs are created with the timing of the B phases systematically staggered across the different targets. In other words, baseline observations are made across all targets (persons, outcomes, or settings). A strength of this design is that the intervention is introduced at different times across each of the (multiple) baselines. The assumption is that the pattern of performance during the baseline will remain similar across phases (A, B, etc) if the intervention has no impact on the target outcome. If the pattern of performance changes each time the intervention is introduced across persons, settings, or related outcomes,\textsuperscript{5,16} this is viewed as evidence of a potential causal connection between the intervention and the outcome.

---

**Fig. 1.** A-B design with follow-up for 1 person. The y axis presents time to complete the Timed Up & Go test; the x axis presents the sequence of observations across 6 baseline (A), 6 intervention (B), and 2 follow-up (FU) sessions. The dashed horizontal line indicates the value that is 2 SDs below the mean of baseline phase scores. Marklund I, Klassbo M. Clinical Rehabilitation. Volume 20, Issue Number 7, pp. 568-76, copyright © 2006 by SAGE Publications. Reprinted by Permission of SAGE.\textsuperscript{14}
A strength of the multiple baseline approach is that it includes a built-in series of replications involving the treatment of interest. Replications are important in establishing generalizability in both small-N research and large group-comparisons studies (without random selection of participants). The role of replication in establishing generalizability is described in more detail in the Opportunities and Challenges section below. The sequential introduction of the treatment also helps to control some threats to internal validity that are present in the basic A-B design, such as a coincidental change in the outcome that occurs at the same time the intervention is introduced.\(^1\)\(^6\)

There are 2 potential challenges in conducting and interpreting multiple baseline design studies.\(^1\)\(^7\) The first, which applies primarily to the multiple baseline across subjects version, is related to the time and resources required to concurrently identify, recruit, and measure several participants. A potential solution to this limitation is to use a modified version of the design referred to as the nonconcurrent multiple baseline design. This adaptation requires the investigator to determine possible baseline length and randomly assign these different baseline periods to persons as they are identified and become available to be studied. The second challenge, which applies to the multiple baseline design, arises when the various outcome measures are positively correlated. If changes in one outcome cause changes in another, then it is not possible to assess the cause-effect relationship between the intervention and the second outcome. This situation requires a revision of the outcome measures or the selection of another research design, for example A-B-A-B.\(^1\)\(^7\)

Figure 2 shows a multiple baseline design for 1 person. The figure is adapted from Carey and Matyas' study\(^1\)\(^8\) on direct and transfer effects of stimulus-specific training on joint proprioception in 5 patients with stroke. Two outcome measures (patients’ perception of wrist position in the flexion-extension and ulnar-radial deviation planes) were recorded across 3 consecutive phases of 10 sessions each. In phase 1 (baseline), both outcomes were monitored with no intervention. Phase 2 (intervention 1) included flexion-extension stimuli provided every 2 to 3 days. In phase 3 (intervention 2), ulnar-radial deviation stimuli were added along with a maintenance program for the flexion-extension stimuli. Lastly, follow-up assessments were conducted 12 to 14 weeks after phase 3. The figure shows considerable variability overall and substantial error in proprioceptive awareness for both planes of movement during the baseline phase. Stimulus-specific training led to immediate reductions in both the variability and magnitude of errors, yet there appeared to be no transfer of flexion-extension training to accuracy of ulnar-radial perception. The results of the follow-up assessments suggest that the improvements in proprioceptive awareness were long-lasting.

### CHANGING INTENSITY AND ALTERNATING TREATMENTS DESIGNS

The broad, flexible framework of the changing intensity and alternating treatment designs is perhaps most reflective of everyday rehabilitation clinical practice. RCTs aside, rarely does a patient receive a single intervention for a fixed duration regardless of their interim progress. Rather, rehabilitation typically entails the use of several concurrent interventions (procedures, devices, activities, etc), and the parameters of these interventions (type or intensity or duration, etc) are adjusted as a patient’s needs and abilities change.

As with the other small-N designs described above, the names essentially reflect the design parameters. Changing intensity designs can be illustrated with the following formula: \(AB^1B^2B^3\), where the Bs with superscripts indicate successive intervention phases in which stepwise changes in intervention parameters are implemented.\(^1\)\(^6\) This design is not effective for establishing causality but can provide useful information in demonstrating change in patient performance over time. Additional baseline phases (withdrawal designs) can be incorporated into this design to increase internal validity, but this is neither necessary nor realistic in everyday clinical practice. Another caveat with respect to this design is that it is best suited for outcomes that can be gradually improved over time.\(^5\)

Alternating treatments designs can be illustrated with the following formulas: \(A-B-C\), \(A-B-A-C\), \(A-B-A-C-A\), or any other feasible combination of baseline/standard care (A) and intervention phases (B = intervention 1, C = intervention 2, etc). Interpreting the results of alternating treatment designs must be done with caution. For example, order effects may influence the magnitude of the responses and confound interpretation. Also, it is important to understand that only adjacent phases can be compared, because there may be unmeasured factors or interactions among treatments that occur during the intervening phase. Patterns or relationships observed for non-adjacent phases can, however, facilitate the development of hypotheses that can be tested in future studies.\(^5\)

Figure 3 shows an alternating treatment design\(^1\)\(^9\) comparing 2 treatments to improve gait in a patient with hemiparetic stroke. The treatments included 2 different types of ankle-foot orthoses (AFOs) designed to reduce foot-drop. The baseline phase consisted of barefoot walking recorded during 5 sessions over 7 days. The intervention phase spanned 1 month during which 12 sessions each of 3 randomly ordered conditions were observed: barefoot, traditional AFO, and tone-inhibiting dynamic AFO. Four additional barefoot-condition assessments
Performances during successive phases (eg, baseline and intervention, or intervention and follow-up) are compared. The use of a control phase rather than a control group shifts the emphasis from between-group variance to within-person variance. The fact that repeated measurements are gathered from the same person increases the likelihood of serial dependency or autocorrelation within the data. This potential for dependency, along with other aspects of small-N research (eg, small number of observations, nonnormal distributions, unequal variances), violate basic assumptions for common statistical techniques such as $t$ tests and analysis of variance.

Statistical analysis options in small-N designs range from simple (eg, celeration line, C statistic, running medians, 2 SD bands) to complex (eg, hierarchical linear modeling). Detailed descriptions of these techniques are beyond the scope of this article, but comprehensive analytic approaches for small-N designs are available in the literature. It is important to note that effect size has gained popularity because it overcomes many of the limitations of $P$ values in small-N research. Manolov et al\textsuperscript{10,11} provide examples and describe the strengths and limitations of several effect size calculations, including the common standardized mean difference approach, regression-based approaches, and visual-based approaches.

**OPPORTUNITIES AND CHALLENGES FOR SMALL-N DESIGNS IN REHABILITATION**

Health care reform, a growing, aging, and diversifying population, and technologic and scientific advances are all transforming the health care system and the way rehabilitation services are (and will be) delivered. One important characteristic of evidence-based practice is the focus on research information relevant to the individual patient. Sackett et al\textsuperscript{32} originally described evidence-based medicine as the “conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients.” The emerging field of personalized medicine promotes this viewpoint as well. Small-N research designs provide information directly relevant to the individual participants being studied. Lillie et al\textsuperscript{3} stated recently that “the ultimate goal of an n-of-1 trial is to determine the optimal or best intervention for an individual patient using objective data-driven criteria.” Of course, it is important to acknowledge that the use of small-N research methods can only directly determine the best treatments for persons that are involved in a specific study. Fields such as clinical and counseling psychology have proposed the scientist-practitioner model in which systematic evaluation and accountability methods involving aspects of small-N research are incorporated into daily practice. The scientist-practitioner model attempts to integrate components of small-N designs directly into clinical practice. This involves a number of logistic and ethical issues (eg, human subject protection) that are beyond the scope of this article, but discussed elsewhere.

An advantage of small-N design is that they allow investigators (and clinicians) to potentially identify characteristics relevant to individual patient performance. If an experimental group of 50 patients does statistically better than a control group of 50 patients, the difference could be due to a small number of persons in the treatment group showing large changes, while the majority of individuals show little or no change. Individual variation may be masked by the group average. In small-N designs, each participant is assessed repeatedly and comparisons within the person are made over time allowing patterns of performance to be linked to individuals with specific characteristics. Once a clinically significant difference appears within a small-N design, the practitioner can...
identify the patient variables and other relevant factors present when the result was obtained. In the small-N approach, not only are factors such as sex, age, diagnosis, level of disability, and education kept constant in the same participant over time, but so are all significant life experiences that occur before the intervention begins. This degree of individual control is only possible in large-N group comparison designs when the participants are measured repeatedly and followed over long periods. Doing this is very difficult within the context of an RCT. The efficacy of a given treatment in a traditional RCT is usually assessed only after the treatment is completed (post-test). The typical pretest, posttest RCT design precludes the continuous assessment and analysis of the patient’s performance during treatment. Small-N designs usually allow researchers to observe within-person variability and relate environmental or physical characteristics to patient performance. Repeated observations permit a systematic analysis of the course of treatment and may suggest useful modifications as the study progresses. The researcher using a small-N design can take into account and analyze the impact of day-to-day contingencies and related events that affect patient behaviors and performances.

There are several factors that have limited the growth of small-N clinical research. Not the least of these factors is the dominance of the RCT research design hierarchy, including the publication bias of journal editors and predisposition of funding agencies. Another factor involves the general lack of research skills to conduct small-N studies and appropriately interpret and apply the findings among rehabilitation researchers and practitioners. The small-N research approach includes a wide variety of designs, similar to the diversity in larger-N group comparison designs. We have introduced 3 categories of small-N designs and provided examples. There are other forms of small-N designs and many variations within each category of design. Identifying which design is the best fit for a particular research question or context depends on many factors. In general, small-N designs are practical complements to larger N trials. They can be useful in the early developmental phase of research as well as in refining the application of research findings to individual patients. Specific situations in which small-N designs may be particularly useful include (1) low-prevalence conditions or combinations of (comorbid) conditions where it is not feasible to recruit patients for a sufficiently powered between-groups analysis; (2) conditions and/or interventions with the potential for substantial variability in either the responsiveness or magnitudes of responses, for example in the early phase of a research program when there is little information about which outcomes will be sensitive to a new treatment; (3) when the sequence or duration of treatments is dependent on the achievement of certain milestones or outcome thresholds; or (4) in unique settings or situations that may limit the feasibility of group comparison clinical trials, for instance in the early stages of establishing protocols for expensive or invasive interventions.

Small-N research can be challenging and these designs are not universally applicable to all clinical questions or patients. All designs—both small-N and large-N (group comparison)—have limitations. Some of the design-specific caveats relevant to small-N methods were described in the sections above. Perhaps the most commonly identified limitation associated with small-N research is the perceived inability to generalize the study findings—the issue of external validity.

Because randomly selecting a representative sample from a large target population is not a component of small-N designs, replication is the alternative strategy used to establish the generalizability of small-N research findings. Barlow and Hersen describe 3 strategies for establishing generalizability in small-N research. The first form of generalizability involves the accumulation of a number of direct replications of the specific treatment effect on 1 well-defined outcome measure within a defined clinical setting. In this form of replication, participants are matched as closely as possible on subject characteristics. The aim is to establish, as clearly as possible, that a given intervention can have an effect on a certain kind of patient within a specific setting. If a series of direct replication small-N studies produces consistently positive results, then the replication process moves to the next level.

The second level of replication involves the systematic replication of the treatment across various participants, settings, clinicians, or a combination of these. Systematic replication helps to establish the generality of the findings over a wider range of situations than direct replication.

The final strategy of replication identified by Barlow and Hersen is clinical replication. Clinical replication involves establishing the generality of related components of the intervention. These might include issues such as intensity or duration or combining multiple components of the intervention and testing them across various patients and settings.

Using replication to establish generalizability aims to develop a scientific consensus regarding the effectiveness (or lack of effectiveness) of an intervention across different persons, clinicians, and settings. It cannot be achieved in the context of a single study, but must be developed over time and involve multiple studies, often conducted by different investigators.

Meta-analysis of small-N studies has become increasingly popular as a means of integrating the findings of multiple studies and contributing to the generalizability of small-N studies. The reader is referred to Barlow and Hersen and Bloom et al for more information on the issues of replication and generalizability in small-N research. Also, see the article by Johnston and Dijkers in this issue for background on synthesizing information from several small-N trials and integrating this information with data from RCTs.

CONCLUSIONS

Not all rehabilitation clinicians will participate in traditional large-N group comparison experimental research designed to test hypotheses or refine theory. However, every clinician does have a responsibility to document the services they provide and changes in patient performance related to the person’s individual rehabilitation goals.

Grimmer et al contend that clinicians should be a driving force in developing more appropriate and relevant evidence for rehabilitation practice. Clinically derived evidence can serve as the catalyst for investigators to design high-quality research that is relevant to clinical practice (see article by Whyte et al in this issue). Small-N designs represent one approach that is available to practitioners and that may allow them to contribute to the advancement of rehabilitation science and practice. While RCTs remain the criterion standard for establishing treatment efficacy, it is important to recognize that small-N research can supplement and/or refine the findings from large parallel group trials when making treatment decisions for an individual patient. We are optimistic that the relationship between these approaches will continue to evolve into a partnership that provides the evidence necessary to validate clinical practice and improve services provided to persons with disabilities and their families.
References