FOCUS

TECHNICAL BRIEF NO. 27 2010

A Publication of the National Center for the Dissemination of Disability Research (NCDDR)

Low-Cost and No-Cost Steps in Research Design to Improve the Quality of Evidence

Marcel P. J. M. Dijkers, PhD

Introduction

Evidence-based practice (EBP) in medicine, psychology, rehabilitation, and other fields has put the focus on the scientific evidence underlying the activities professionals undertake for the benefit of their patients and clients. Whether selecting diagnostic tools and other assessments, identifying optimal treatments, or predicting outcomes and costs/benefits of courses of action, professionals are now expected to base their prognoses and decisions on the best scientific evidence available. Their patients' and clients' values and their own training and clinical expertise always should play a role as well (American Psychological Association, 2005; Buettner & Fitzsimmons, 2007; Davidson, Trudeau, & Smith, 2006; Dodd, 2007; Ebenbichler, Kerschan-Schindl, Brockow, & Resch, 2008; Geil, 2009; Guyatt et al., 2000; Law, 2002; Mullen, Bledsoe, & Bellamy, 2008; Pierce, 2007; Schreiber & Stern, 2005; Straus, Richardson, Glasziou, & Haynes, 2005; Welch, 2002).

Although many tools have been developed for bringing existing evidence to bear on a clinical question, the major mechanism for collecting, evaluating, and synthesizing evidence is the systematic review. In drawing conclusions and making recommendations for practitioners, systematic reviewers evaluate the quality, quantity, and diversity of all the existing evidence that addresses specific questions relevant to practice.

Systematic reviewers give preference to better evidence—giving it more weight or even relying on it exclusively. In this case, better refers to evidence from studies using stronger research designs and better outcome measures applied to larger and more representative samples. Many schemes for grading, or categorizing, the strength of evidence have been developed by professional organizations, government agencies, and individual researchers. These systems distinguish grades of evidence (typically from 3 to 10), which reflect one's level of confidence that a study's results are not affected by bias and can be generalized to others in the population from which study subjects were sampled. These schemes base an assigned grade solely or primarily on the study design. For example, in determining the effectiveness of a treatment, a randomized controlled trial (RCT) is deemed stronger than a study with historical controls.

Additionally, design and implementation characteristics may be used to assign subgrades and raise or lower a study's grade. A large RCT gives more definitive evidence than a small one and therefore is graded

Department of Rehabilitation Medicine
Mount Sinai School of Medicine, New York, NY

This paper reflects the efforts and experiences from research activities carried out under Grants H133B040033 and H133A080053 from the National Institute on Disability and Rehabilitation Research (NIDRR), Office of Special Education and Rehabilitative Services (OSERS), U.S. Department of Education to the Mount Sinai School of Medicine. Margaret Brown, PhD, deserves thanks for clearheaded editing of earlier versions; Wayne Gordon, PhD, suggested the topic; and Lisa Spielman, PhD, referred me to literature on the power implications of using measures with reduced random error.

ADVANCING RESEARCH SEDL IMPROVING EDUCATION

The National Center for the Dissemination of Disability Research (NCDDR) is a project of SEDL. It is funded by the National Institute on Disability and Rehabilitation Research (NIDRR).

higher. However, if randomization is implemented poorly, an RCT degenerates into a comparison of two nonrandomized, self-selected (or treater-selected) groups and would be graded lower.

A number of checklists have been developed to help systematic reviewers grade studies by methodically and efficiently characterizing the quality of their design, implementation, and reporting. Because a checklist does not result in a simple number that translates into a quality rating, quite a few authors have also taken the step to develop a rating scale (e.g. Downs & Black, 1998). Overviews of such checklists and quality rating scales have been provided by a number of authors (Katrak, Bialocerkowski, Massy-Westropp, Kumar, & Grimmer, 2004; Moher et al., 1995; Olivo et al., 2008; Sanderson, Tatt, & Higgins, 2007; West et al., 2002). Because the criteria for assessing the quality of a study differ, depending on the question the study aims to answer, checklists and rating scales have been developed for the three major study types: treatment/intervention (Jadad et al., 1996; Maher, Sherrington, Herbert, Moseley, & Elkins, 2003; Tate et al., 2008), diagnosis and assessment (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003), and prognosis (Bossuyt et al., 2003; Stroup et al., 2000; von Elm et al., 2007).

To optimize the potential impact of their research, grantees funded by the National Institute on Disability and Rehabilitation Research (NIDRR), as well as other researchers, should do all they can to ensure that systematic reviewers assign a high grade to their research. Researchers can do this by designing and implementing research of the highest quality (evidence grade) possible within the limitations of the resources available and the ethical standards for the treatment of research participants.

Once the research is completed, researchers should report in the professional or research literature the design, implementation, and results of that research as clearly and completely as possible to enable others, including systematic reviewers, to use the evidence. In previous articles, we addressed how to report research so it is more likely to be included in systematic reviews (Dijkers, 2009; Dijkers, Brown, & Gordon, 2008). However, a perfect report that honestly and completely presents a poorly designed and/or haphazardly implemented study is not likely to impress anyone, systematic reviewer or otherwise.

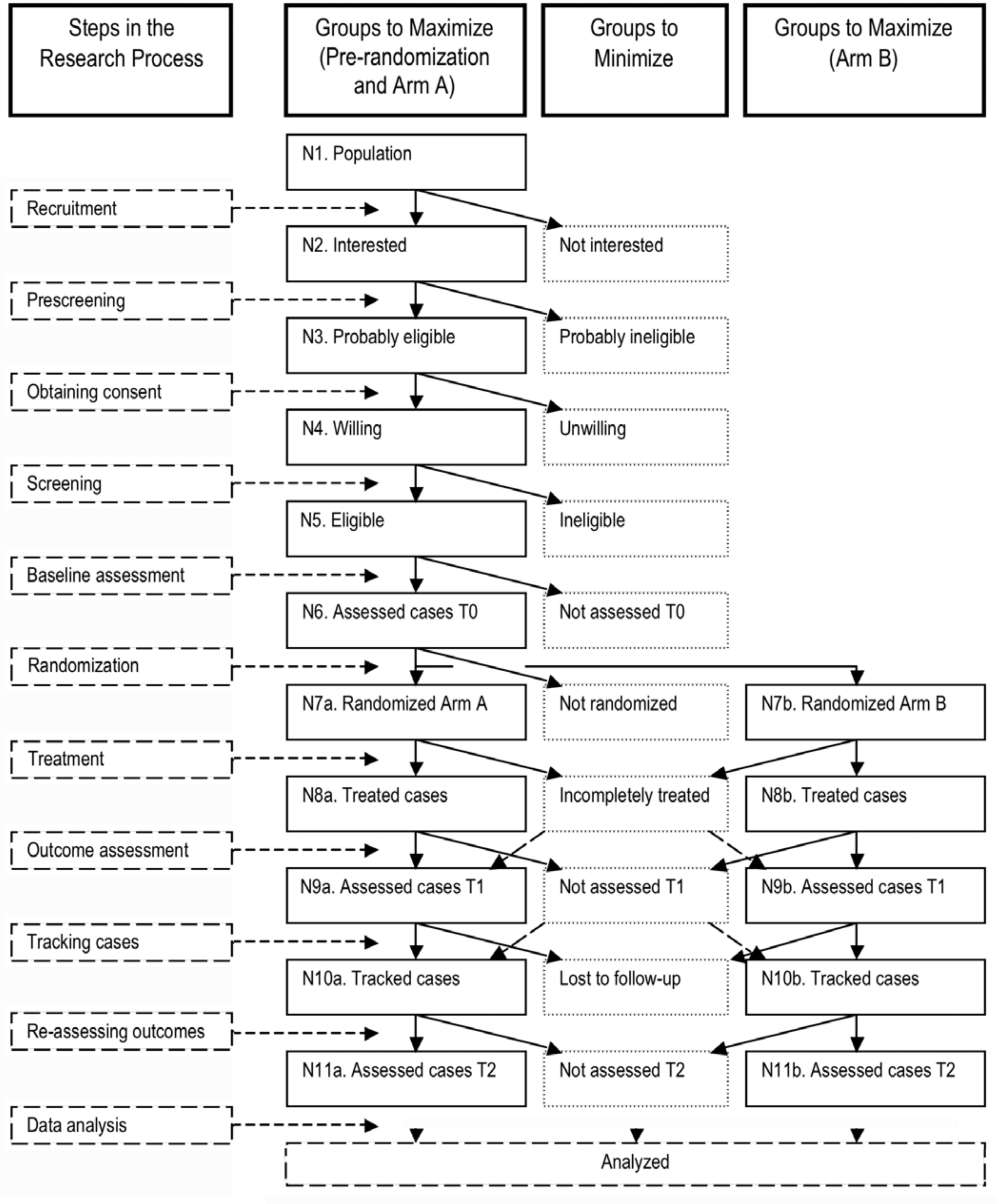
This paper offers low-cost and no-cost steps that rehabilitation researchers can take to strengthen the quality of their evidence and, thereby, the likelihood of their work receiving a high evidence grade and being included in systematic reviews. Clearly, a study's inclusion in a systematic review extends the impact of the research on the field, whether the systematic review targets practitioners, researchers, consumers, or policymakers.

Quite often, researchers know their research is not as strong as it could be and what should be done to improve it, but lack the funds to do so. This paper offers low-cost and no-cost steps that rehabilitation researchers can take to strengthen the quality of their evidence and, thereby, the likelihood of their work receiving a high evidence grade and being included in systematic reviews. Clearly, a study's inclusion in a systematic review extends the impact of the research on the field, whether the systematic review targets practitioners, researchers, consumers, or policymakers.

This paper focuses solely on intervention research. Research aiming to appraise the quality of assessment methods; to evaluate the cost-benefit ratios of individual interventions and entire treatment programs; or to describe the natural history of a particular disability, such as stroke or autism, is just as important but uses different designs that have their own sets of potential biases and other weaknesses. Future issues of *FOCUS* may address the problems in providing high-quality evidence for other types of research questions (i.e., diagnostic and prognostic) and present possible inexpensive solutions.

Figure 1

Simplified Representation of a Randomized Controlled Trial with Short-Term and Long-Term Outcome Assessment



EBP has been spearheaded in the field of medicine, where high-quality RCTs comparing a drug with a placebo are easy to realize because of the relative ease of implementing randomization, blinding, and other study design elements that minimize bias. In contrast, rehabilitation researchers typically deal with lengthy, individualized behavioral interventions for which a placebo is not possible and where, typically, participants and treaters in the innovative and comparison treatments cannot be blinded to group assignment. Thus, rehabilitation researchers often cannot design perfect research, even with unlimited resources. However, researchers in other areas, such as surgery, psychotherapy, and public health, deal with some of these same problems. Even if some of the aspects of research design that would provide the highest evidence grade are not possible, much can still be done to optimize the quality of research design and implementation.

For this paper, we examined systematic reviewers' checklists and rating scales, and identified those elements that concern design and implementation of RCTs and other intervention research. From the literature and our own experience, we then pinpointed possible solutions that rehabilitation researchers may want to consider to improve the quality of the evidence they produce.

Figure 1 presents the typical stages in a simple RCT. The steps in the research process, which are listed in the left-hand column, are used to organize this discussion. The research design depicted has two treatment arms. Subjects are assessed immediately before and after treatment, and also complete a second outcome assessment at a later point. The solid boxes in the second and fourth columns refer to groups of (potential) subjects that we want to maximize; the dotted boxes in the third column refer to groups of (potential) subjects that we want to minimize or avoid altogether. Studies with historical or contemporaneous controls may lack the randomization step, pre-post studies may also lack column 4 entirely, and other common designs in rehabilitation research may not use some steps or columns. However, some parts of the figure will apply to most intervention research, as will the suggestions presented here.

Power Considerations

A research study should have exactly as many subjects as are needed to answer the research question or test the most power-hungry hypothesis, with the investigators' specified level of confidence in the results. Having fewer participants than needed is a waste of time and resources, including subjects' goodwill, because an underpowered study cannot address the key question of the effectiveness of the treatment. And calling an underpowered study a "pilot study" is incorrect because the study is not piloting anything but is simply an inadequate version of what should have been. Having more subjects than needed also wastes time and resources, and unnecessarily exposes subjects to the risks involved in being part of the study, however minimal those risks may be.

From the perspective of systematic reviewers, underpowered studies are problematic because whatever effect the treatment of interest may have cannot be generalized to the population at large. If a meta-analysis is possible, the joint studies may have enough power, but that does not justify publishing underpowered studies.

Stand-alone statistical packages that perform power analysis for a number of research designs are readily available, and some analysis software now includes simple power analysis routines. Much harder than doing the calculations is determining the values for the assumptions underlying them: the expected effect size for the treatment, the minimal effect size that is clinically worthwhile, and the risk of Type I (false positive) and Type II (false negative) errors one is willing to accept. Additionally, the number produced by the power analysis has to be multiplied by two inflation factors. Factor 1 is an estimate of the loss of subjects between recruitment (N2 in Figure 1) and randomization (N7a/b). Factor 2 is an estimate of the loss of subjects between the number of cases randomized (N7a/b) and the number available for final assessment (N11a/b).

Often, there is a very limited basis for making these estimates, and investigators just take a stab at it—usually a very optimistic stab. Fortunately, it is legitimate to re-estimate these inflation factors on the basis of the experience in the trial. If the elimination of cases between N1 and N7a/b is less than was anticipated based on screening criteria and potential subjects' expected willingness to consent, randomization of the number needed may occur in less time than estimated and make it possible to complete the study early. If attrition after randomization is larger than anticipated, the number of cases randomized can be increased to partially make up for this, although intent-to-treat (ITT) analysis (discussed later) requires that all randomized participants who failed to make it to the finish are still included in the analysis.

The only parameter that cannot be adjusted on the basis of the data flowing from the study is the effect size—in fact, the desirability of a blinded analysis (discussed later) suggests not even looking at the actual effect size as the trial progresses. The only basis for re-estimating power is new, independent information that provides a more dependable estimate of the effect size, such as the publication of a new study in the same area.

In the "good old days," investigators would look at the outcomes of their intervention study on a regular basis, calculate the statistical significance of Arm A versus Arm B outcomes, and call it a day when the magic "p<0.05" rolled out of their computer. They certainly did not use any more resources than needed, but they also did not realize that they ran the risk of terminating the study because of a difference between Treatment A and Treatment B that was statistically significant purely by chance. We now know that repeatedly looking at the data is inappropriate. If there is a potential need to terminate a trial early (for futility, unacceptable side effects, or an unexpectedly strong effect of the experimental treatment), an interim analysis needs to be preplanned.

In the "good old days," investigators would look at the outcomes of their intervention study on a regular basis, calculate the statistical significance of Arm A versus Arm B outcomes, and call it a day when the magic "p<0.05" rolled out of their computer … We now know that repeatedly looking at the data is inappropriate.

In rehabilitation research, with its low risk of significant side effects and often moderate effect sizes, interim analyses are not common. Moreover, because they "cost degrees of freedom" (i.e., impose a need for recruiting and studying additional subjects), they may not be appropriate in rehabilitation research with its modest sample sizes. Investigators who are confident their treatment may have an effect well above that assumed in the power analysis have the option of building an interim analysis into their plan and, thus, potentially saving resources. All investigators should be aware, though, that systematic reviewers take a dim view of interim analyses that are not preplanned, because of their potential to increase bias. Nature abhors a vacuum, and systematic reviewers abhor fishing expeditions!

Narrow Versus Broad Inclusion Criteria

The proportion of potential subjects excluded from a study, as well as any differences between the treatment groups in demographic, clinical, and other characteristics, is of interest to systematic reviewers because these elements are relevant to the issue of generalization, or the external validity of a study. The major criticism of RCTs is that, however powerful they are in demonstrating the effectiveness of an intervention, they tend to involve a minor slice of the target population. Medical RCTs sometimes enroll less than 10% of the "available" group of subjects (N1), because the studies have so many and such stringent inclusion and exclusion criteria.

The resulting homogeneity of the group eventually randomized (N7a/b) may make demonstrating an effect easier, because the variance in the outcome measures is likely to be reduced, but it also makes generalizing to the entire population that may benefit from the intervention more problematic. This issue is of special concern in the exclusion of patients who have health issues in addition to the one targeted by the treatment(s) of interest. In such a case, the research report would not tell clinicians and other readers how the treatment works in patients with comorbidities or what adverse effects it may have on other disorders.

In addition, RCTs are typically done in academic health-care centers, which attract an atypical patient population. One alternative to RCTs is *Practical Clinical Trials* (PCTs), the term coined for studies relevant to clinicians and decision makers. PCTs (1) compare clinically relevant interventions (2) in a diverse population of study participants (3) in heterogeneous practice settings, and (4) collect data on a broad range of health outcomes (Glasgow, Magid, Beck, Ritzwoller, & Estabrooks, 2005; Tunis, Stryer, & Clancy, 2003).

For every new study, a number of factors should be considered to determine the necessary balance between rigor and relevance, as epitomized by the distinction between RCTs and PCTs. (Note that PCTs are not necessarily cheaper than RCTs.) What rehabilitation researchers can do inexpensively is collect data on the number and characteristics of the individuals in the successive boxes N1–N5 (or at least N2–N5) in Figure 1. This information will help researchers demonstrate that the participants actually studied (N6–N11a/b) are not a small and selective subset of the population, or at least of those expressing an interest in being a subject in the study (Group N2). Those researchers who submit to journals that call for adherence to the CONSORT requirements will also have all the information needed to complete the CONSORT flowchart (Altman et al., 2001; Moher, Schulz, Altman, & CONSORT Group, 2001).

> ***One possible step to reduce chances of attrition is not to push potential subjects too hard into entering a study. In fact, it may be a good idea to put some barriers in their way. Once subjects are randomized, the researcher is committed to them and has to report what happens to them, warts and all.***

Recruitment Versus Retention

Few intervention studies in rehabilitation have a surplus of possible subjects. More common is that investigators must scramble to find the number of subjects their power analysis told them they needed and that they promised in their funding proposal. As a consequence, study staff often exert pressure on potential subjects (see Figure 1, N3) to consent, even if someone is not that interested in being part of the study after hearing a recital of the risks and obligations. While researchers may be permitted to do everything within the parameters set by the responsible Institutional Review Board (IRB) to sway a person to consent, and such efforts may be successful in obtaining the number of subjects needed (N6), these subjects may not make it to the last assessment (N11a/b). Attrition is a serious problem in studies that last months, if not years, or require subjects to be available for a large number of treatment or follow-up sessions over a number of weeks.

Systematic reviewers take a dim view of any study that has more than 15% attrition. This number is fairly arbitrary but, even so, it is part of the quality criteria in quite a few checklists, such as the Physiotherapy Evidence Database (PEDro) scale (www.pedro.org.au/english/downloads/pedro-scale/) and the American Academy of Neurology's assessment framework (Edlund, Gronseth, So, & Franklin, 2004).

If there is attrition in both treatment arms, chances are that the group evaluated for the final effect of the two compared interventions (N11a/b) differs (in important but probably unknown ways) from the group recruited and randomized (N7a/b). In this situation, even providing a table comparing clinical and demographic characteristics of the cases that entered into the study with those that completed it will not fix the problem. Even more serious, Murphy's Law states that attrition most likely will affect the group randomized to Arm A (N7a) differently from the Arm B group (N7b). This can potentially confound the assessment of any differences between the arms that is deemed to be the result of the superiority of Treatment A over Treatment B (or the placebo or sham), even if the subjects are successfully blinded.

One possible step to reduce chances of attrition is not to push potential subjects too hard into entering a study. In fact, it may be a good idea to put some barriers in their way. Once subjects are randomized, the researcher is committed to them and has to report what happens to them, warts and all. For one drug study we are conducting, where subjects need to complete daily pain diaries over a 14-week period, we inserted a 1-week delay between the day of passing screening and the day of actual randomization, a delay that was unnecessary from any design or clinical standpoint. However, we require subjects to submit a diary each day of this week, and those who fail to do so (other than for reasons we deem acceptable) are not randomized, protecting us against the damage caused by the likelihood of their dropping out after randomization.

More Power Using Fewer Subjects

Recruitment and attrition (discussed in detail in the next section) are major problems in longitudinal rehabilitation research, and efforts to obtain a large enough sample and then retain those participants can consume a large share of available resources. The intervention(s) being trialed probably consumes another lion's share, especially if they involve multiple treatment sessions conducted by expensive clinicians. If one could decrease the number of subjects, more resources would be available to do an excellent job of recruiting, treating, and following subjects.

An entirely different research design may require fewer subjects—for example, using a single subject design (SSD) with replication across participants instead of an RCT. However, in many instances, SSDs are not suitable in rehabilitation research that focuses on interventions used during a period of natural recovery. Additionally, many evidence grading schemes place SSDs far below the top-ranked RCTs, which may discourage researchers from using SSDs.

An option for more power with fewer subjects is to select a more appropriate statistical analysis. *T* tests, or ANOVAs with repeated measures, are some of the most popular methods of analyzing group comparison trials. However, variations on these old standbys may give a somewhat higher power for the same number of subjects—for instance, a variation in ANOVAs is to use a characteristic that is strongly associated with the outcome measure as a covariate.

Yet another option is to use better outcome measures. Every measure has random error, and Spearman-Brown and similar formulas tell us that longer measures will, other factors being equal, have less random error. With less random error, the standard deviation within any group of research participants will be reduced, allowing for a larger effect size in a comparison with another group and a more precise estimate of that effect size. The size of this bonus depends on a number of factors, including the two group mean values involved, the standard deviations, the power required, the alpha value used, and the statistical test used. Although using a better measure may hardly pay off in some situations, with small samples and measures involving considerable random error, the payoff may be big.

There are several ways of getting a better outcome measure. One is to compare coefficients of variation (the ratio of the standard deviation to the mean) for alternative measures, such as the Life 3 (Andrews & Withey, 1976) and the Satisfaction With Life Scale (Diener, Emmons, Larsen, & Griffin, 1985) as alternatives for measuring quality of life. (The means and standard deviations should be derived from the same sample to ensure that the comparison is valid.) Another option is to stick with one's preferred outcomes or with those recommended by systematic reviewers, but to administer them twice (Kopriva & Shaw, 1991; Rogers & Hopkins, 1988; Sutcliffe, 1980). A double baseline doubles the number of items in one's measure, which has a predictable effect on coefficient alpha. Researchers would not necessarily need the subject to return at a 1-week interval but might administer the key outcome measure at the beginning and again at the end of an assessment session. The Life 3 is such a "double" instrument, consisting of one item that is given at both the beginning and the end of an interview, and then averaged. The double assessment could be repeated at the post-intervention follow-up sessions.

Tables from Kopriva and Shaw (1991) enable researchers to estimate power in various situations involving improved outcome measures. For example, with 200 subjects, using *t* tests with an alpha of 0.05, an effect size of 0.5, and the measure's reliability of 0.30, the power is 0.78. If reliability were improved to 0.60 (not an impossibility), 100 subjects would be sufficient to achieve a slightly higher power of 0.82 (Kopriva & Shaw, 1991). Of course, doubling or tripling the administration of the key measures will take more subject time and staff time per subject, but that increase will likely be more than offset by the time saved by having to recruit, screen, treat, assess, and reassess fewer participants.

Preventing Attrition

As mentioned, systematic reviewers may lower the evidence grade of any study that has more than 15% attrition among the subjects who have been randomized. Consequently, researchers should go to great lengths to keep subject drop-out below this level, which is not easy to do if the assessment of outcomes is removed from the treatment period by more than a few weeks.

Much of rehabilitation research is interested in how a patient performs not just at the end of treatment but often months, if not years, later. At the end of treatment we may know about a reduction in impairment and an increase in activities, but participation and quality of life should be assessed at some remove of time. This, however, creates the problem of staying in contact with subjects and keeping them interested enough in the research project to remain available for a follow-up assessment.

Loss of subjects before the final assessment of outcomes has a number of negative consequences. Researchers may worry about wasted effort and lack of resources to recruit replacements. Investigators may also worry about the need to extend the duration of the trial, with all that would mean in terms of obligations to the sponsor, staff burnout, and other penalties of subject loss.

Loss of subjects before the final assessment of outcomes has a number of negative consequences. Researchers may worry about wasted effort and lack of resources to recruit replacements. Investigators may also worry about the need to extend the duration of the trial, with all that would mean in terms of obligations to the sponsor, staff burnout, and other penalties of subject loss.

From the perspective of systematic reviewers, the major issues resulting from attrition are three: the confounding of the analysis of the treatment's effect, decreased precision of the estimate of the study's effect size, and inability to determine the size and nature of self-selection that affects external validity. If the subjects dropping out differ among the various study arms in terms of demographic, disability, or other important variables, the analysis may find that the intervention has a greater effect than actually exists, or may find an effect where none exists. Or, the opposite may occur, and the analysis may show a decreased effect or no effect, even when the intervention is truly effective. Second, because of a smaller sample (if per-protocol analysis, defined later, is used) or larger standard deviations resulting from conservative imputed outcome values (in the case of ITT analysis), the precision of the estimate (i.e., the confidence interval around the effect size) will almost always be larger than it would be in the case of minimal attrition. This makes it more difficult to get an impression of the effectiveness of the intervention. Finally, if attrition in the several study arms is quantitatively the same and the dropouts in the various treatment groups do not differ from one another in any relevant characteristics, internal validity is not affected (a fact that can be assumed but never proved). The result of the attrition could still show greater differences between the study sample and the population from which it was drawn. This difference affects the generalizability of the findings. Thus, with good reason, systematic reviewers take a dim view of studies with high attrition—and more than 15% is generally considered high.

To counteract subject loss, two steps are needed: (1) tracking participants so they can be contacted whenever an assessment is due, and (2) maintaining participants' willingness to remain subjects. One inexpensive technique of tracking subjects that we use is to harvest as many addresses and phone numbers of participants and their immediate family and friends as we can at the time of enrollment. We then use newsletters and holiday cards to get address corrections and to develop a bond between participants and the research team that may work in our favor when it comes time to conduct follow-up interviews.

To find those participants who still disappear from the radar screen, several options are available. Internet searches take limited staff time. Sending registered letters to last-known addresses may result in new addresses. Also, searching proprietary databases such as Accurint® (www.accurint.com) is expensive if purchased on a per-case basis, but flat monthly user fees may be cost-effective for larger projects. For smaller projects, these fees would fall into the moderate-cost category.

Once a subject's new phone number or address has been found, it is "merely" an issue of flexibility (making phone calls during business hours, at night, and on weekends) and persistence (making at least two calls in every category) to re-establish contact. This effort involves moderate staff time. Attempts to locate subjects need to begin well before the window for follow-up data collection opens, because the process can be drawn out and may not succeed before the window closes.

Suitable low- or moderate-cost methods of convincing subjects, once re-found, to agree to a telephone or in-person assessment are too varied and numerous to list here. Coday et al. (2005) identified 61 methods used within the Behavior Change Consortium. Robinson and Trochim identified 368 in a systematic review of the literature focused on strategies for retaining participants in health-care research (Robinson & Trochim, 2007). Fortunately, a number of good reviews of this literature have been published in recent years, from which investigators may select the steps and strategies that best fit their situations (Knobf et al., 2007; Levkoff & Sanchez, 2003; Mody et al., 2008; Robinson & Trochim, 2007; Tansey, Matté, Needham, & Herridge, 2007; Yancey, Ortega, & Kumanyika, 2006).

The literature tends to stress that researchers should not wait until the first follow-up to worry about attrition. Instead, research should be designed with issues of follow-up in mind, which may involve such steps as setting up community advisory boards, "packaging" the intervention and all materials used to communicate with participants so as to fit into their culture, and establishing systems so that clinicians who have ongoing relationships with the subjects help with tracking. Critical characteristics for data collection staff include interpersonal skills, willingness to work flexible hours, patience, and persistence.

In our experience, offering fees or stipends to participants for completing follow-up assessments plays a minimal role in retention. Although appreciated as an acknowledgment that participants' time has value, increasing the size of fees has no or minimal effect on improved retention. The major factors working against retention are a lack of interest (see previous plea about not pushing subjects into consenting), disorganization in subjects' lives, too busy lives (even among individuals with disabilities who may not work or attend school), and transportation problems. Together, these factors constitute a powerful cocktail working against the researcher.

In our experience, offering fees or stipends to participants for completing follow-up assessments plays a minimal role in retention. Although appreciated as an acknowledgment that participants' time has value, increasing the size of fees has no or minimal effect on improved retention.

The Nature of the Assessments

Systematic reviews of intervention studies focus first on whether individual studies and the full group of studies investigating a particular treatment demonstrate the intervention to be effective when compared with a placebo, "routine," or even "best" care. How much difference there is (absolute or in the context of costs, risks, and adverse effects) is the second question for which effect sizes are calculated. Systematic reviewers prefer to compare studies that use the same outcome measure, although effect sizes can be calculated and combined whatever specific outcome measures studies use. In recent years, many papers have been published that systematically review measures of impairment, activity limitation, and participation restriction (as well as other key outcomes used in rehabilitation) with the goal of making authoritative recommendations to investigators. These efforts are focused on specific outcomes either across (e.g. Dworkin et al., 2005; Turk et al., 2003) or within diagnostic groups (e.g. Anderson et al., 2008; Bryce et al., 2007; Merkel et al., 2003; Sherwin et al., 2006; Tugwell et al., 2007). Rehabilitation researchers would do well to review these recommendations and use these measures in their studies. Not only would they be using the best measures available (in a field where there may be dozens of potentially relevant instruments in the literature), but they would more likely be using the same outcome measures as other investigators, allowing for the barrier-free combination of the findings of their investigations.

Randomization

Randomization means assigning subjects to study arms in a way that is independent of any characteristics of the subjects (especially their preference for being in Arm A or Arm B), of the preferences of the treating clinicians (especially their thoughts as to what intervention is best for any specific client or patient), and of anything or anybody else. However, experience has taught systematic reviewers that the mention of randomization in a research paper may actually cover many situations where strict adherence to the scientific dictums of random assignment was sacrificed because of sloppiness or a desire to "help fate" in a way that would generate a more satisfying difference between Arm A and Arm B cases. Any time the people who prescreen, consent, screen, or assess at baseline know or can reasonably predict whether a person they are processing will end up in Arm A, they may have an opportunity to steer him or her away from that assignment by using a narrower interpretation of the inclusion criteria, overemphasizing the risks and discomforts of being in the trial, and so on. This type of influence has been demonstrated in a number of reviews that found a larger-than-average effect size in trials that had inadequate concealment of allocation compared with those that used adequate methods. Thus, biased and sloppy research in this way tends to overstate the effect of an intervention and may understate the risks or adverse events.

Systematic reviewers insist on randomization methods and timing that make it impossible for anyone to bias the results. This means concealing the assignment (e.g., by using opaque envelopes, dialing in to an automated telephone system, or going to a Web site) until the potential subject has been fully qualified.

Consequently, systematic reviewers insist on randomization methods and timing that make it impossible for anyone to bias the results. This means concealing the assignment (e.g., by using opaque envelopes, dialing in to an automated telephone system, or going to a Web site) until the potential subject has been fully qualified. Or, if randomizing is done after that point, using a method that cannot be influenced by humans and ensuring that there is no chance for the research staff to roll the dice again if they do not like the first result. These methods are all fairly simple and inexpensive to implement. Opaque envelopes are the traditional mechanism; a list maintained by a disinterested third party (the research pharmacist in many medical RCTs) often works as well, and both single-site and multisite studies may use assignments made and revealed by Web-resident programs.

Alternatives to Strict Randomization

In several of our research projects, we have needed to choose between passing up a subject (and subjects are always in short supply in our lines of research) and having potentially less than stellar randomization. Assume you are comparing Behavioral Treatment A with Behavioral Treatment B. You have one therapist, Ann, who delivers Treatment A and has slots for 5 patients at a time, and a second therapist, Bob, who delivers Comparator Treatment B and also has 5 slots. (The issue discussed here would not exist if Ann and Bob both delivered A as well as B, but that may not always be feasible given training requirements and the risk of one treatment contaminating the other.) Qualified patients who have agreed to participate are assigned randomly to Bob or Ann. Now, imagine that Ann is treating 3 patients and Bob 5, and a new potential subject shows up. You cannot randomize in this case because Bob cannot accept another patient. Adding the person's name to the waiting list until one of Bob's patients is discharged is equally unsatisfactory, because it is a disservice to the patient, who may go elsewhere for help and then be lost as a study participant.

In such instances, we have assigned the patient to Ann. Our defense is that the patient, when expressing interest in participating in the study, did not have a preference for A or B (or Ann or Bob) and did not know the number of open slots for each therapist. The combination of the typical duration of treatment, the rolls of the dice for all 8 subjects currently in treatment, and a number of other factors determined the current number of open slots for A and B. As long as the person who is screening and obtaining consent from this new participant is not aware of the number of open slots and works under the assumption that both Bob and Ann can accept a new patient in their caseload, no harm is done by bypassing randomization. Whenever both Bob and Ann do have a slot open, we use opaque envelopes.

Although randomization will create two groups that are perfectly balanced on every characteristic of interest in a study (gender, impairment level at baseline, etc.), the law of large numbers states that such equivalence will be found only if the number of cases randomized approaches infinity. Researchers will not have enough money or time to study an infinite number of cases, even with unbiased randomization procedures. Therefore, an imbalance at baseline may occur either on the key outcome or on other variables that may affect the subjects' ability to improve or to benefit from treatment. Systematic reviewers dislike such imbalances because they have the potential to confound the results. Although we know this potential exists only if the imbalanced variables have a strong association with the outcome, no one can specify how strong the baseline imbalances need to be to affect the outcome, either singly or in combination.

The impact of such imbalances can be mitigated by analyzing the difference between Arm A and Arm B on the key outcome while controlling for the unbalanced variables, but this solution reduces degrees of freedom and potentially, power. The smaller a study, the more likely one or more imbalances will arise, and the less room for trading away degrees of freedom. The most obvious solution to prevent imbalances is stratification of the sample on those variables that are thought to strongly affect the outcome. However, stratification creates its own problems, especially if it produces a large number of cells. If every cell has to have multiples of two (one subject for randomization to Arm A and one to Arm B), there is the potential for loss of subjects because an appropriate match is not found.

In various studies where we faced the problem of potential lack of balance between the treatment arms because of a small number of subjects, we have used minimization to avoid the imbalances while not getting caught in the net of stratification or post-hoc analysis (Treasure & MacRae, 1998; Treasure & MacRae, 1999). Minimization allows one to identify key factors that are thought to affect outcomes of a treatment—for example, the baseline status on the outcome of interest—or other factors that can likely influence the potential to benefit from treatment. We use a minimization statistical program to randomize subjects while taking into account their status on the key factors. The program then assigns the subjects to Arm A or Arm B to optimize the balance of the two groups on all specified factors, as long as this process does not require certainty. The chance of each case being randomized to Group A is always greater than 0.00 and smaller than 1.00, and so is the chance of randomization to Group B. The hardest part is deciding which factors are important, how many categories need to be distinguished for each of these factors, and what cut-points to use if a continuous variable is involved. The software does the rest, producing group assignments for all cases.

Treatment Integrity

The treatments and interventions that rehabilitation researchers study are highly diverse, but from a systematic reviewer's point of view they all have one thing in common: They need to be delivered completely, competently, and on the time schedule specified by the protocol in order to impact the outcomes as strongly as expected by the estimates that underlie the power analysis. Although ITT analysis (described in detail in a later section) calls for disregarding the quality and quantity of treatment each subject received, systematic reviewers prefer studies that can demonstrate intervention integrity, that is, a correspondence between the treatment called for in the protocol and what was actually delivered. This correspondence may also be referred to as program fidelity or integrity, treatment fidelity or integrity, independent variable accuracy, or procedural reliability. Demonstrating treatment integrity requires writing a detailed protocol (an operational definition of the treatment); taking steps to encourage clinicians to master it (training, feedback on performance, etc.); encouraging patients and therapists to adhere to it (reminders; bonuses; encouragement; provision of time, space, and other resources); and finally, quantifying the correspondence between the protocol and the treatment delivered.

In an analysis of 171 rehabilitation intervention articles, we found that almost half did not report intervention integrity (Dijkers et al., 2002). We presumed that this meant that intervention integrity received little or no attention in these studies, although it is not unknown for researchers to omit information from their reports that would improve the grade that systematic reviewers assign to their studies (Devereaux et al., 2004; Hill, LaValley, & Felson, 2002). Assessment of treatment integrity can sometimes be done very cheaply (e.g., by doing a pill count in a medication trial or by asking a therapist to complete a simple checklist as to the elements of the protocol she actually delivered). It can also be very expensive, sometimes costing more than the treatment delivery itself—for example, the cost of having two experts independently code every minute of a videotaped treatment session (Mullis et al., 2006). The intricacy and cost of assessing intervention integrity is determined by a number of factors, with the primary factor being the complexity and duration of the treatment being studied. A variety of approaches can be implemented and include electronic pill boxes, diaries completed by patients, quality scoring of audiotapes of psychotherapy sessions, counters or other gauges put on subjects (pedometers) or equipment, and in-person observations by the investigator. Because methods for determining intervention integrity must complement the specific treatment being investigated by a rehabilitation researcher, it is not appropriate to make specific recommendations here.

In any case, systematic reviewers consider treatment integrity an important issue in evaluating the quality of studies; thus, researchers need to provide the relevant information. Doing so starts with operationalizing one's treatment in a treatment manual (Hart, 2009), followed by deciding on feasible, valid, and reliable ways of assessing quantity and quality of treatment. Because the issue of treatment manuals for behavioral and other "complex interventions" (Campbell et al., 2007; Craig et al., 2008) and the appropriate evaluation of such studies are emergent issues in the field of research methodology, systematic reviewers are far from having reached consensus on what quantity and type of information on treatment integrity is required to judge protocol adherence. However, it is safe to assume that in the near future intervention integrity will play a larger role in assessing the quality of studies than it has in the past.

Blinded Assessment and Alternatives

Patients and clinicians have specific expectations about what particular interventions will and will not accomplish. These expectations can powerfully influence what both groups observe and report regarding the effects of the interventions delivered in Arm A and Arm B. In fact, about half of patients experience at least some benefit from a placebo pill. Consequently, expectations need to be eliminated, or at least equalized, between the two treatments compared in a trial. As every researcher knows, the solution is double blinding—where neither patients nor treaters are aware of treatment assignments. In many rehabilitation interventions, however, double blinding may be difficult, if not impossible to implement. While drugs can be disguised (although sometimes they can be identified by their side effects) and devices can be made to deliver a sham treatment, the core of rehabilitation involves therapists working with patients for an extended period. We cannot hide from patients whether they get physical therapy (PT). We probably cannot even hide whether they get PT intervention Type A or Type B—unless to the layperson both interventions look the same and the IRB allows the use of uninformative language in the consent document along the lines of "comparing two types of physical therapy." Even if we can blind the patients, we cannot blind therapists as to the work they are doing with patients.

Two possible, although less than perfect, solutions to this dilemma suggest themselves: (1) use of objective outcome measures and (2) assessment by an independent assessor who is blinded to patients' assignment to Arm A or Arm B.

An objective outcome measure is one where the patient's status is assessed by a machine that provides a report without human intervention, such as sending information directly to a database. Although in some rehabilitation research these types of outcomes are of interest, in most they are inapplicable or may, at best, make up a secondary or tertiary outcome. This situation will continue until we have machines that can observe and "measure" human functioning, such as the level of independence in dressing that a subject has achieved as a result of treatment.

Two approaches are available for the analysis of the data of RCTs and other study designs involving the comparison of groups: (1) per-protocol (PP), also called as-treated and on-treatment, and (2) intent-to-treat (ITT), also called intention-to-treat.

However, the value of measures delivered by machines as a proxy for human observation and classification should not be overlooked. If, for example, we think that patients may underreport their mobility because they were in the control group, we can provide them with pedometers, which cost only a few dollars and deliver step counts that correlate strongly with distance walked, as determined by a number of studies. Often, equipment is available that can deliver information that adequately supplements the study participant's self-report or the clinician's report. Unfortunately, such equipment is frequently expensive or requires additional expertise. However, creative thinking may suggest a pedometer or a similar technological solution that cannot be swayed by preferences for one treatment over another.

The other partial solution to the problem of being unable to blind patients and treaters is to engage an outcome assessor who is blind to patient assignment. This solution may not always be affordable, especially if an assessment requires a trained professional rather than a research assistant. It often is simpler and more straightforward for the treating clinician to do the post-test at the end of the last treatment session. Not only does engaging and training a separate assessor carry extra costs, but there is also no guarantee that he or she will remain blind. However, sometimes at least part of an evaluation can be assigned to an assessor who is blinded to the nature of the treatment participants have received. The difficulty then may be in keeping the blinding intact. Participants are told not to reveal anything about their treatment or treater while they are with the assessor. Sometimes, additional efforts are necessary. In one study in which we were involved, a patch of skin on the upper arm was harvested to prepare cells used in the treatment. As the scar remained visible, the spot on the arm was covered with a large bandage for both the subjects who received the treatment and the controls, whenever they met with the assessor.

Data Analysis

Two approaches are available for the analysis of the data of RCTs and other study designs involving the comparison of groups: (1) per-protocol (PP), also called as-treated and on-treatment, and (2) intent-to-treat (ITT), also called intention-to-treat:

- In a PP analysis, one compares only the subjects in Arm A and Arm B who received all of their assigned treatment in exactly the way and at exactly the time specified in the protocol. Cases with partial treatment, those who erroneously received all or part of the treatment of the other arm, and those who (contrary to instructions) sought additional treatment outside of the research setting are all discarded from the analysis. Also excluded is anyone with incomplete outcome data.

- In an ITT analysis, everyone assigned to a treatment group is analyzed in the treatment arm to which each participant had been assigned. If necessary, (conservative) estimates for missing outcome information are used. One method is referred to as *last observation carried forward* (LOCF), in which, for example, Time 2 outcome information is used to estimate Time 3 data if the case was lost to follow-up at Time 3.

In practice, the distinction between the two types of analyses is not as black and white as it may seem. Research indicates that some researchers use the term ITT even when they exclude some cases from the analysis, such as those that never started receiving the assigned treatment because of a clerical error (Gravel, Opatrny, & Shapiro, 2007; Herman, Botser, Tenenbaum, & Chechick, 2009; Pagoto et al., 2009; Polit & Gillespie, 2009). A standard interpretation of the term ITT may develop in future years.

The justification for ITT analysis is that it does not allow for a back door through which biases can creep in after having been excluded in the research design through randomization and allocation concealment. If analyzing cases that received less than the full treatment results in an underestimate of the effectiveness of the intervention of interest, so be it.

The justification for PP analysis is that researchers only want to evaluate the effectiveness of an entire treatment, with all its components delivered on time by competent personnel. However, PP analysis may allow subject or clinician preference to slip back into a study and bias the results. For instance, subjects who had really wanted Treatment A but were assigned to Treatment B may drop out of that arm, creating an imbalance between the two groups based on whatever makes people want Treatment A.

The justification for ITT analysis is that it does not allow for a back door through which biases can creep in after having been excluded in the research design through randomization and allocation concealment. If analyzing cases that received less than the full treatment results in an underestimate of the effectiveness of the intervention of interest, so be it. In fact, because patients in nonresearch settings often get less than the perfect treatment (e.g., they do not take all their pills, their physical therapist is sick one day, they fail to receive one treatment session because the babysitter didn't show), the ITT estimate of effect size may be a better guideline for what a new treatment can do in real life than is the PP estimate.

However, both approaches may give the researcher insufficient statistical power: the PP analysis, because of a reduced number of subjects; the ITT analysis, because of a larger standard deviation for outcomes than was estimated in the study design phase. Systematic reviewers would suggest choosing the ITT approach because the effect size estimate will be unbiased. They can handle problems of low power by means of meta-analysis of similar studies.

In fact, systematic reviewers have expressed a strong preference for studies that use ITT analysis to the point of excluding PP studies from consideration. Because the cost of doing the analysis does not differ significantly between PP and ITT, rehabilitation researchers who want their reports to be included in systematic reviews should report ITT analysis results. However, it often is possible to have one's cake and eat it too by reporting both the ITT and the PP results—providing effect size estimates for a "perfect" world, where every patient gets flawless treatment, as well as for the real world, where even the best-laid plans can go awry.

Another issue in the analysis of trial data is the blinding of the analyst. Generally, research proposals specify how the data will be analyzed, such as analysis of variance on the primary outcome with baseline status on the same characteristic used as a covariate (analysis of covariance–ANCOVA). The proposal generally does not specify rules for making the multiple minor decisions that typically need to be made to prepare the data set for analysis, such as the following:

- What should be done with extreme values on the outcome or baseline variable, which almost certainly are errors? Should they be recoded to missing? Recoded to the mean plus 3 standard deviations? Replaced by an imputed value?
- What should be done with cases that are missing information on the primary outcome variable? Should they be excluded (see the previous discussion comparing PP and ITT analyses)? Should a value be imputed? And if so, using LOCF or another algorithm?
- What should the analyst do if baseline imbalance is noted between the groups receiving Treatment A and those receiving Treatment B on characteristic *x*, which is thought to strongly affect the outcome of interest (a "failure of randomization")? Nothing? Make an ad hoc decision to use *x* as yet another covariate in the analysis?

All of these decisions have the potential to "nudge" the outcome of the main analysis toward favoring Arm A (presumably the experimental, hoped-to-be-better treatment) over Arm B. The solution that systematic reviewers prefer for minimizing this potential for bias is to blind the analyst as to which group in the dataset is which. The analyst can just prepare the file and run the analyses, blind as to whether Group 1 is Arm A and Group 2 is Arm B, or vice versa. The effect size may give the game away, but it can be arranged that the analyst not look at the Group 1 versus Group 2 difference on any outcomes until all other decisions about how to clean up and analyze the data have been made.

Conclusion

"There is nothing new under the sun" (Ecc. 1:9, New International Version), and that saying holds true for the issues of research quality that systematic reviewers evaluate. Even before the evidence-based practice movement started emphasizing such issues, we knew that blinding subjects and treaters might be a good idea, and that randomization should be taken seriously in the sense that group assignment has to be concealed until the subject and clinician are irrevocably committed to the study. What has changed is that the strictures of good research design and implementation have been built into systematic reviews, which have become the all-important means for corralling the evidence that practitioners rely on in making decisions about diagnosis, treatment, prognosis, and long-term management. Researchers who want to see their findings included in that evidence base need to start paying attention to those design elements that systematic reviewers call commendable, and to build those elements into their research design and implementation.

In addition to the specific suggestions provided in this technical brief, we have a general recommendation for rehabilitation researchers: Stay up to date on what systematic reviewers are thinking and doing. Before writing your next grant proposal, take time to study some of the checklists that systematic reviewers use to judge study quality. Then ask yourself, How can I design and implement this research in such a way that it will be coded "yes" on as many of these criteria as possible? With unlimited resources (and by disregarding the ethical standards in the IRB regulations) anyone could produce near-perfect research, but we do not live in such a world. There are ethical restrictions, and resources are always limited. Creative thinking in designing the research, combined with a fair dose of obsessive-compulsive behavior in implementation, will go a long way toward obtaining research results that are convincing not just to the practitioner-reader but also to the systematic reviewer.

References

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134*(8), 663–694.

American Psychological Association, 2005 Presidential Task Force on Evidence-Based Practice. (2005). *Draft policy statement on evidence-based practice in psychology*. Washington, DC: American Psychological Association.

Anderson, K., Aito, S., Atkins, M., Biering-Sorensen, F., Charlifue, S., Curt, A., Ditunno, J., et al. (2008). Functional recovery measures for spinal cord injury: An evidence-based review for clinical practice and research. *The Journal of Spinal Cord Medicine, 31*(2), 133–144.

Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being: Americans' perceptions of life quality*. New York: Plenum Press.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clinical Chemistry, 49*(1), 1–6.

Bryce, T. N., Budh, C. N., Cardenas, D. D., Dijkers, M. P., Felix, E. R., Finnerup, N. B., et al. (2007). Pain after spinal cord injury: An evidence-based review for clinical practice and research. Report of the National Institute on Disability and Rehabilitation Research Spinal Cord Injury Measures meeting. *The Journal of Spinal Cord Medicine, 30*(5), 421–440.

Buettner, L. L., & Fitzsimmons, S. (2007). Introduction to evidence based recreation therapy. *Annual in Therapeutic Recreation, 15*, 12–19.

Campbell, N. C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., et al. (2007). Designing and evaluating complex interventions to improve health care. *BMJ (British Medical Journal), 334*, 455–459.

Coday, M., Boutin-Foster, C., Goldman Sher, T., Tennant, J., Greaney, M. L., Saunders, S. D., et al. (2005). Strategies for retaining study participants in behavioral intervention trials: Retention experiences of the NIH Behavior Change Consortium. *Annals of Behavioral Medicine, 29*(2 Suppl.), 55–65.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., Petticrew, M., et al. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ, 337*, a1655.

Davidson, K. W., Trudeau, K. J., & Smith, T. W. (2006). Introducing the new Health Psychology series "Evidence-Based Treatment Reviews": Progress not perfection. *Health Psychology, 25*(1), 1–2.

Devereaux, P. J., Choi, P. T., El-Dika, S., Bhandari, M., Montori, V. M., Schunemann, H. J., et al. (2004). An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. *Journal of Clinical Epidemiology, 57*(12), 1232–1236.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71–75.

Dijkers, M. P. (2009). Ensuring inclusion of research reports in systematic reviews. *Archives of Physical Medicine & Rehabilitation, 90*(11 Suppl.), S60–S69.

Dijkers, M. P., Brown, M., & Gordon, W. A. (2008). Getting published and having an impact: Turning rehabilitation research results into gold. *FOCUS* Technical Brief (19). Austin, TX: SEDL, National Center for the Dissemination of Disability Research.

Dijkers, M. P., Kropp, G. C., Esper, R. M., Yavuzer, G., Cullen, N., & Bakdalieh, Y. (2002). Quality of intervention research reporting in medical rehabilitation journals. *American Journal of Physical Medicine & Rehabilitation, 81*(1), 21–33.

Dodd, B. (2007). Evidence-based practice and speech-language pathology: Strengths, weaknesses, opportunities and threats. *Folia Phoniatrica et Logopaedica, 59*(3), 118–129.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*(6), 377–384.

Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., et al. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain, 113*(1-2), 9–19.

Ebenbichler, G., Kerschan-Schindl, K., Brockow, T., & Resch, K. L. (2008). The future of physical & rehabilitation medicine as a medical specialty in the era of evidence-based medicine. *American Journal of Physical Medicine & Rehabilitation, 87*(1), 1–3.

Edlund, W., Gronseth, G., So, Y., & Franklin, G. (2004). *Clinical practice guideline process manual* (2004 ed.). St. Paul, MN: American Academy of Neurology (AAN).

Geil, M. D. (2009). Assessing the state of clinically applicable research for evidence-based practice in prosthetics and orthotics. *Journal of Rehabilitation Research and Development, 46*(3), 305–314.

Glasgow, R. E., Magid, D. J., Beck, A., Ritzwoller, D., & Estabrooks, P. A. (2005). Practical clinical trials for translating research to practice: Design and measurement recommendations. *Medical Care, 43*(6), 551–557.

Gravel, J., Opatrny, L., & Shapiro, S. (2007). The intention-to-treat approach in randomized controlled trials: Are authors saying what they do and doing what they say? *Clinical Trials, 4*(4), 350–356.

Guyatt, G. H., Haynes, R. B., Jaeschke, R. Z., Cook, D. J., Green, L., Naylor, C. D., et al. (2000). Users' Guides to the Medical Literature: XXV. Evidence-based medicine: Principles for applying the Users' Guides to patient care. *JAMA: The Journal of the American Medical Association, 284*(10), 1290–1296.

Hart, T. (2009). Treatment definition in complex rehabilitation interventions. *Neuropsychological Rehabilitation, 19*(6), 824–840.

Herman, A., Botser, I. B., Tenenbaum, S., & Chechick, A. (2009). Intention-to-treat analysis and accounting for missing data in orthopaedic randomized clinical trials. *Journal of Bone and Joint Surgery American, 91*(9), 2137–2143.

Hill, C. L., LaValley, M. P., & Felson, D. T. (2002). Discrepancy between published report and actual conduct of randomized clinical trials. *Journal of Clinical Epidemiology, 55*(8), 783–786.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavaghan, D. J., et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1–12.

Katrak, P., Bialocerkowski, A. E., Massy-Westropp, N., Saravana Kumar, V. S., & Grimmer, K. A. (2004). A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology, 4*(1), 22.

Knobf, M. T., Juarez, G., Shiu-Yu, K. L., Sun, V., Sun, Y., & Haozous, E. (2007). Challenges and strategies in recruitment of ethnically diverse populations for cancer nursing research. *Oncology Nursing Forum, 34*(6), 1187–1194.

Kopriva, R. J., & Shaw, D. G. (1991). Power estimates: The effect of dependent variable reliability on the power of one-factor ANOVAs. *Educational and Psychological Measurement, 51*(3), 585–595.

Law, M. C. (2002). *Evidence-based rehabilitation: A guide to practice*. Thorofare, NJ: Slack.

Levkoff, S., & Sanchez, H. (2003). Lessons learned about minority recruitment and retention from the Centers on Minority Aging and Health Promotion. *The Gerontologist, 43*(1), 18–26.

Maher, C. G., Sherrington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713–721.

Merkel, P. A., Clements, P. J., Reveille, J. D., Suarez-Almazor, M. E., Valentini, G., Furst, D. E., et al. (2003). Current status of outcome measure development for clinical trials in systemic sclerosis. Report from OMERACT 6. *The Journal of Rheumatology, 30*(7), 1630–1647.

Mody, L., Miller, D. K., McGloin, J. M., Freeman, M., Marcantonio, E. R., Magaziner, J., et al. (2008). Recruitment and retention of older adults in aging research. *Journal of the American Geriatrics Society, 56*(12), 2340–2348.

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16*(1), 62–73.

Moher, D., Schulz, K. F., Altman, D., & CONSORT Group (Consolidated Standards of Reporting Trials). (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA *285*(15), 1987–1991.

Mullen, E. J., Bledsoe, S. E., & Bellamy, J. L. (2008). Implementing evidence-based social work practice. *Research on Social Work Practice, 18*(4), 325–338.

Mullis, R., Dziedzic, K. S., Lewis, M., Minns Lowe, C. J., Main, C. J., Watson, P. J., et al. (2006). Validation of complex interventions in a low back pain trial: Selective video analysis cross-referenced to clinical case notes. *Contemporary Clinical Trials, 27*(5), 404–412.

Olivo, S. A., Macedo, L. G., Gadotti, I. C., Fuentes, J., Stanton, T., & Magee, D. J. (2008). Scales to assess the quality of randomized controlled trials: A systematic review. *Physical Therapy, 88*(2), 156–175.

Pagoto, S. L., Kozak, A. T., John, P., Bodenlos, J. S., Hedeker, D., Spring, B., et al. (2009). Intention-to-treat analyses in behavioral medicine randomized clinical trials. *International Journal of Behavioral Medicine, 16*(4), 316–322.

Pierce, L. L. (2007). Evidence-based practice in rehabilitation nursing. *Rehabilitation Nursing, 32*(5), 203–209.

Polit, D. F. , & Gillespie, B. M. (2009). The use of the intention-to-treat principle in nursing clinical trials. *Nursing Research, 58*(6), 391–399.

Robinson, J. M., & Trochim, W. M. (2007). An examination of community members', researchers' and health professionals' perceptions of barriers to minority participation in medical research: An application of concept mapping. *Ethnicity & Health, 12*(5), 521–539.

Rogers, W. T., & Hopkins, K. D. (1988). Power estimates in the presence of a covariate and measurement error. *Educational and Psychological Measurement, 48*(3), 647–656.

Sanderson, S., Tatt, I. D., & Higgins, J. P. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology, 36*(3), 666–676.

Schreiber, J., & Stern, P. (2005). A review of the literature on evidence-based practice in physical therapy. *Internet Journal of Allied Health Sciences & Practice, 3*(4), 17.

Sherwin, E., Whiteneck, G., Corrigan, J., Bedell, G., Brown, M., Abreu, B., et al. (2006). Domains of a TBI minimal data set: Community reintegration phase. *Brain Injury, 20*(4), 383–389.

Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). Edinburgh; New York: Elsevier/Churchill Livingstone.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. JAMA*, 283*(15), 2008–2012.

Sutcliffe, J. P. (1980). On the relationship of reliability to statistical power. *Psychological Bulletin, 88*(2), 509–515.

Tansey, C. M., Matté, A. L., Needham, D., & Herridge, M. S. (2007). Review of retention strategies in longitudinal studies and application to follow-up of ICU survivors. *Intensive Care Medicine, 33*(12), 2051–2057.

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.

Treasure, T., & MacRae, K. D. (1998). Minimisation: The platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does. *BMJ, 317*(7155), 362–363.

Treasure, T., & MacRae, K. D. (1999). Minimisation is much better than the randomised block design in certain cases. *BMJ, 318*(7195), 1420.

Tugwell, P., Boers, M., Brooks, P., Simon, L., Strand, V., & Idzerda, L. (2007). OMERACT: An international initiative to improve outcome measurement in rheumatology. *Trials, 8*, 38.

Tunis, S. R., Stryer, D. B., & Clancy, C. M. (2003). Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. JAMA*, 290*(12), 1624–1632.

Turk, D. C., Dworkin, R. H., Allen, R. R., Bellamy, N., Brandenburg, N., Carr, D. B., et al. (2003). Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. *Pain, 106*(3), 337–345.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., Vandenbroucke, J. P., et al. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *BMJ, 335*(7624), 806–808.

Welch, A. (2002). The challenge of evidence-based practice to occupational therapy: A literature review. *Journal of Clinical Governance, 10*(4), 169–176.

s. Dr. Dijkers is a former President of the Ameri

West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., et al. (2002). *Systems to rate the strength of scientific evidence*. AHRQ Evidence Report/Technology Assessment, No. 47. Rockville, MD: Agency for Healthcare Research and Quality.

Whiting, P., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2003). The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 3*, 25.

Yancey, A. K., Ortega, A. N., & Kumanyika, S. K. (2006). Effective recruitment and retention of minority research participants. *Annual Review of Public Health, 27*, 1–28.

Recommended Citation

Dijkers, M. P. J. M. (2010). Low-cost and no-cost steps in research design to improve the quality of evidence. *FOCUS* Technical Brief (27). Austin, TX: SEDL, National Center for the Dissemination of Disability Research.

Address Correspondence

Marcel Dijkers, PhD
Department of Rehabilitation Medicine, Box 1240
Mount Sinai School of Medicine
One Gustave L. Levy Place
New York, NY 10029-6574
Telephone: [00-1-] 212-659-8587
Fax: [00-1-] 212-348-5901
e-mail: marcel.dijkers@mssm.edu

Author

Marcel Dijkers, PhD, FACRM, is a research professor at Mount Sinai School of Medicine, Department of Rehabilitation Medicine and is a senior investigator on the NIDRR-funded New York TBI and SCI Model System centers at Mount Sinai. He is facilitator of the NCDDR's Task Force on Systematic Review and Guidelines. Dr. Dijkers is a former President of the American Congress of Rehabilitation Medicine. He sits on the editorial board of the *Journal of Head Trauma Rehabilitation* and the *Journal of Physical Medicine and Rehabilitation Sciences*, and is an editor of *Rehabilitation Research and Practice*.

FOCUS: A Technical Brief From the National Center for the Dissemination of Disability Research was produced by the National Center for the Dissemination of Disability Research (NCDDR) under grant H133A060028 from the National Institute on Disability and Rehabilitation Research (NIDRR) in the U.S. Department of Education's Office of Special Education and Rehabilitative Services (OSERS). NCDDR is funded 100% by NIDRR at $750,000 per project year.

NATIONAL CENTER FOR THE
DISSEMINATION OF DISABILITY RESEARCH

Advancing Research, Improving Education

SEDL

NCDDR's scope of work focuses on developing systems for applying rigorous standards of evidence in describing, assessing, and disseminating outcomes from research and development sponsored by NIDRR. The NCDDR promotes movement of disability research results into evidence-based instruments such as systematic reviews as well as consumer-oriented information systems and applications.

Available in alternate formats upon request. Available online: www.ncddr.org/kt/products/focus/focus27/