Making the Best Match: Selecting Outcome Measures for Clinical Trials and Outcome Studies

Wendy J. Coster

Selecting an appropriate outcome measure is a critical step in designing valid and useful clinical trials and outcome studies. This selection process needs to extend beyond examining basic psychometric properties to consider additional features of instruments that may affect their validity and utility for the study’s purpose. This article discusses these additional factors and their potential impact on outcome measurement. Guidelines are proposed to help clinical researchers and consumers of clinical research literature evaluate the match between the study purpose, population, and instrument.


The recent emphasis on evidence-based practice in health care has stimulated a growing literature on the design of clinical trials, especially randomized controlled trials (RCTs), and outcome studies. To meet criteria for best evidence, these guides have focused on the features of research design and procedures that maximize internal and statistical validity. Critiques and systematic reviews of the clinical literature also have concentrated their evaluation on how well particular studies adhered to these design and procedure guidelines (e.g., Stolee, Lim, Wilson, & Glenny, 2012). Much attention has been paid to how well a particular study controlled for potential sources of influence other than the intervention being tested or whether a suitably representative sample was obtained that would support generalization of results. However, another important factor affects the validity of the inferences drawn from clinical research that has received surprisingly little attention: the choice of outcome measure.

The choice of outcome measure represents how the researcher has operationalized a “successful outcome”; thus, the usefulness of the study as a contribution to clinical knowledge hinges on the adequacy of the measure for this purpose. The best design and most rigorously executed procedures cannot make up for a poorly chosen measure. Important knowledge about the impact of the intervention may be lost because the selected measure was unable to capture it or, even worse, distorted the true results. This article suggests guidelines for researchers and users of the research literature to help evaluate the suitability of particular outcome measures for the clinical research questions the measures were selected to examine.

Example of the Issues

Suppose that researchers have conducted an RCT to compare the effectiveness of two instructional approaches to the rehabilitation of military personnel who have sustained a brain injury (all examples in this article are fictional and were created for illustrative purposes only). The two interventions target cognitive skills in attention, memory, executive functions, and pragmatic communication, with the
The ultimate goal being return to employment. The researchers selected two outcome measures: (1) scores on the FIM™ Cognitive scale (Uniform Data Set for Medical Rehabilitation, 1997) and (2) whether the person worked for 30 or more hours per week in the past month. The researchers report that both measures have strong reported reliability and that the FIM Cognitive scale’s validity has been demonstrated. Does this evidence of satisfactory psychometric properties establish that these outcome measures are appropriate for the study? No, it does not; A number of important issues still need to be examined.

First, does the FIM cognitive scale sample the relevant areas the researchers expect will be influenced by either intervention? An attempt to match the areas of cognitive function that are the focus of the intervention to the FIM items reveals several gaps: No item specifically asks about attention; the problem-solving item may be sensitive to some aspects of executive function but does not cover all aspects typically included in this domain; and pragmatic communication may affect scores on the expressive communication item, but only if difficulties in this area are readily identifiable during ordinary interactions. Therefore, it is possible that this scale may not pick up changes in several of the areas being targeted by the interventions because this content is not well represented in the outcome measure.

A second important question is whether the selected measure is sensitive to the degree of change expected from the intervention. The FIM Cognitive scale has been shown to be sensitive to change in function from admission to discharge in people hospitalized after a stroke (Dromerick, Edwards, & Diringer, 2003); however, it has also demonstrated a significant ceiling effect in people who have returned to community living, especially if their remaining cognitive difficulties are in the milder range (Schepers, Ketelaar, Visser-Meily, Dekker, & Lindeman, 2006; van der Putten, Hobart, Freeman, & Thompson, 1999). The researchers in our example will examine change at 6 and 12 months postintervention. Thus, it is quite possible that a substantial number of participants will be at the ceiling, which will significantly reduce the study’s ability to detect differences in outcome between groups.

Several questions about the employment outcome are important to ask. Perhaps most important is the question of whether evidence exists to support the hypothesis that cognitive remediation interventions will significantly affect such a complex and multidetermined outcome as employment. In addition, one might ask whether the best way to capture “success” is to ask whether the person had sustained employment during the previous month. This measurement approach would miss participants who had sustained employment over a longer period but were not employed more recently (e.g., they had seasonal employment during the summer). Is full-time employment (more than 30 hr) the only useful indicator of the impact of the intervention for this population? Might measuring several different aspects of employment (e.g., average number of hours per week, total number of weeks worked) provide a more complete picture of the outcome?

As this brief example illustrates, selecting the best outcome measure requires consideration of issues far beyond the usual questions about reliability and validity. The ultimate value of a clinical trial or outcome study will be directly tied to how well the selected outcome measure matches the researcher’s understanding of what he or she expects to change, to what degree it is expected to change, over what period of time this change will happen, and how that change can best be identified.

The remainder of this article proposes a set of guidelines for thinking through these questions. Appendix 1 organizes the key questions into a structured format that can be used in study design or evaluation. Further explication of some of the ideas presented here can be found in Coster (2006).

**Question 1: What Is the Appropriate Outcome to Measure in This Study?**

When first beginning to plan a study, the researcher may identify the outcome he or she believes is most relevant in very broad terms such as participation, performance of daily activities, quality of life, or work performance. Before the target outcome can be operationalized (i.e., before an appropriate measure can be selected), however, it must be articulated much more clearly. One excellent way to do so is to create and refer to a well-specified causal model of the intervention process being tested. The causal model makes explicit the researcher’s thinking about how the intervention is expected to achieve its results—that is, what the hypothesized mechanism of change is and in which aspects of the person’s life changes are most likely to be evident. Illustrating this model with a visual diagram can be extremely helpful.

Figure 1 presents the type of simple model that one might construct to represent the early outline of a study of intervention for fine motor skills (again, this example is fictional). Examination of this model shows a significant difference in level of specificity between the focus of the intervention—discrete skills—and the focus of the
outcome measure—participation. This difference suggests that some intermediate steps need to be added to illustrate more precisely how changes in specific hand skills would lead to a change in the broad and complex outcome of participation. The literature suggests that the impact of skill development on a more distal outcome such as participation is indirect, through influence on the ability to perform relevant activities (e.g., Lawrence & Jette, 1996). If so, then the intervention is most likely to show an impact on activity performance first. Inclusion of an activity measure might enhance the study’s ability to detect changes as a result of the intervention. By also including a measure that focuses on participation in leisure activities that require good hand skills, the researcher could examine whether improved activity performance enhanced participation in related life situations. These refinements have been incorporated into Figure 2.

This simple example illustrates how explicating the causal model underlying a study’s design helps guide selection of appropriate outcome measures. Because most intervention and outcome studies are far more complex than this example, careful specification and evaluation of the causal model is an even more critical first step. See Whyte and Barrett (2012) for a thought-provoking discussion of theory development in rehabilitation science.

Question 2: How Can This Outcome Be Measured?

Explicating the causal model of the intervention process helps clarify the construct that will be the focus of the outcome measurement. Once the measurement construct has been identified, the researcher can begin to evaluate available options to measure that construct in the context of a particular research study. This step includes examining measures, instruments, and concrete indicators to identify the best match in four important areas: (1) how the measure operationalizes the construct, (2) potential item bias, (3) definition of the measurement dimension, and (4) whether the instrument is likely to capture change.

Examing How the Measure Operationalizes the Construct

The first area to consider is whether the construct as operationalized by a particular instrument matches the researcher’s definition of the target outcome. For example, many clinical measures are described as measures of function. These instruments often operationalize the construct in different ways, however: Some focus on performance of physical tasks such as walking or climbing stairs, and others include a broad spectrum of tasks ranging from sitting up to grooming to shopping. Some measures include social interaction skills as part of their definition of function, whereas others focus strictly on performance of concrete tasks.

The researcher needs to decide whether the contents of a particular measure match the specific domains of behavior that the intervention will address. In addition, it is important to examine whether the target domain or subdomain is a major focus of the instrument or whether this content is embedded in a set of items that represent many other less relevant subdomains. In the latter case,
although changes might be seen on a few items in the instrument, the impact on the overall score may be too small to show whether there are real differences related to the intervention. The earlier examples in Figures 1 and 2 illustrated this point by making the distinction between a broad measure of participation versus one that specifically examines participation in activities that depend on skilled hand use.

Examining Potential Item Bias

Another reason it is important to examine carefully the actual items in an instrument or outcome measure is to determine whether any features of the items could be problematic (biased) for the population in the study. For example, an item in the SF–36 (Ware, Kosinski, & Keller, 1994), a well known and frequently used measure of health-related quality of life, asks whether the person can “walk a mile.” Because of this feature, people using wheeled mobility can never obtain the maximum score for physical function, regardless of their state of health (Meyers & Andresen, 2000). If a researcher used the SF–36 to compare quality of life of people with spinal cord injury with that of people with traumatic brain injury, how should any group differences that are found be interpreted? Do the results represent new knowledge about quality of life, or are they simply reflecting what was already known: that people with spinal cord injury who use wheeled mobility generally cannot walk?

Such examples are more common than one might think. They often arise because instruments were not originally developed with the unique features of a particular population in mind. For example, the Vineland Adaptive Behavior Scales (VABS–II; Sparrow, Cicchetti, & Balla, 2005) were originally developed to measure daily living skills of people with intellectual disabilities, with an emphasis on the extent to which they can perform the activities in the expected manner. Therefore, the definitions of many of the items describe the typical means of demonstrating the skill and do not accommodate people who use assistive technology or alternative methods to accomplish the task. For example, an item in the socialization domain asks whether the child moves about looking for a parent nearby. No provision exists for the caregiver. This feature does not negate the value of the VABS–II in general, but it does illustrate why this instrument might not be a good outcome measure to use in a study of the effects of intervention for children with cerebral palsy.

Examining the Definition of the Measurement Dimension

Item content and wording is one way that the target construct is operationalized. A second important way is through the definition of the measurement dimension. Human behavior is multidimensional, which means that one could focus on many different aspects in measurement. It is important to consider whether the dimension used by the instrument is the most relevant, given the conceptual model of the intervention and its effects.

For example, consider an intervention designed to help older adults maintain or improve their ability to engage in usual daily activities. Is an instrument that measures the amount of assistance needed for daily activities the best way to capture this outcome? If the participants in the study are older women living alone in the community, they may already, by necessity, be taking care of daily activities independently. An instrument that measures function through the amount of assistance required is not very well matched to this context. A self-report instrument that asks about the degree of difficulty experienced while performing the activities would focus on a more relevant dimension of function for this group.

Variations in how a construct has been operationalized (i.e., the choice of outcome measure) may help explain apparent discrepancies in results across studies that appear, at first glance, to be very similar. This lack of comparability was an important impetus behind recent efforts such as the Patient Reported Outcomes Measurement Information System (PROMIS) and the Quality of Life in Neurological Disorders (Neuro-QOL) initiatives by the National Institutes of Health (Tulsky, Carlozzi, & Cella, 2011). These projects have developed linked sets of measures that, when included in future clinical trials and outcome studies, will ensure that results can be compared more readily.

Examining Whether the Instrument Is Likely to Capture Change

Instruments differ not only in their choice of measurement dimension but also in how the various points along the dimension are scaled. Therefore, it is important to consider whether the scale measures degrees of difference that are within the range expected in the population of interest. A scale might be excellent at discriminating differences in performance among people experiencing moderate degrees of functional limitation but not be able to detect differences among people with more significant limitations whose scores are closer to the bottom of the scale. To illustrate, on the Scales of Independent Behavior–Revised (Bruininks, Woodcock, Weatherman, & Hill, 1996),
a person must perform an item without any physical assistance to obtain a score above zero. This scale might be a good choice to evaluate improvement in the amount of verbal cuing or instruction a person with intellectual disabilities needs to complete activities; however, it is not well matched for a study designed to teach adolescents with quadriplegic cerebral palsy to participate as much as possible in their morning routine.

How a measure is scaled plays an important role in determining whether the instrument is likely to capture real change if it occurs. This issue relates to the properties of sensitivity and responsiveness. These terms are often used interchangeably, although they address related but different questions. Sensitivity is an indicator of the instrument’s ability to detect change in a population beyond measurement error. For example, is the instrument likely to show significant differences in the scores from admission to discharge in people hospitalized immediately after a stroke? This property, also referred to as longitudinal validity (Liang, 2000), is often calculated as a standard effect size for the amount of change measured across a clinically meaningful interval (e.g., the admission to discharge period in the example) or using a standardized response mean. A variety of additional methods are described in the literature (e.g., see Lehman & Velozo, 2010).

Multiple factors can influence a measure’s sensitivity to change in a particular research context. One of the most important factors is the presence of ceiling effects. If a large proportion of participants are already performing well on a measure at baseline (i.e., their scores are close to the top of the scale), then the scale will not be able to capture much change in that group. A related problem, known as restriction of range, arises when little variability exists in the scores within a sample. Because many statistical analyses use sample variability in their calculations, results of analyses could be seriously misleading if very little variability is found among participants. Both ceiling effects and restriction of range are the result of a poor match between the measurement capabilities of the instrument and the characteristics of the sample.

Sensitivity analysis evaluates whether an instrument is likely to detect differences across time. A statistically significant difference may have limited clinical importance, however. Responsiveness provides an indicator of the degree of change on a measure that represents real-world, meaningful difference. Responsiveness is typically evaluated by showing a relation between change on the measure and change as measured by some external criterion or “gold standard.” This requirement is difficult to meet in rehabilitation research because gold standards have not been established for most outcomes. Alternative methods for deriving the external standard include asking others who have observed the participant over the relevant period of time to rate the amount of change observed or having blind raters score videotapes of participants’ performance at the two time periods. Measures that have been created using Rasch or item response theory methods may also use item maps to identify the amount of change that would reflect a meaningful increase in skill or independent performance (Coster, Ludlow, & Mancini, 1999). Because the item map approach is less common, an illustration is provided in Figure 3.

Responsiveness is evaluated in a specific context—that is, in a particular population across a particular time period. Therefore, the applicability of information about an instrument’s responsiveness will depend on the similarity between the research context and the context in which responsiveness was evaluated. Two questions about the context are particularly important: (1) Are the populations similar and therefore may be expected to show similar degrees of change across time? (2) Is the interval of time over which responsiveness was measured similar to

![Participation Scale](image)

**Figure 3. Example showing use of an item map to identify meaningful change.** Meaningful change can be defined on the basis of the definitions of the ratings (1–5). For example, change in scores from 2 to 3 in at least 3 items represents a change from very limited to more complete participation.
the interval that will be used for follow-up in the clinical trial or outcome study? For example, a measure shown to be responsive to functional recovery in the month after a stroke may not be responsive to change in the month after the person has returned home. Recovery in the first month may be relatively rapid and captured well by measures of basic activities of daily living. Once at home, the person may still be recovering, although at a slower pace, and trying to resume instrumental activities that are not well sampled on the outcome measure.

Capturing Change in Other Types of Outcomes

Up to this point, the discussion about measuring change has assumed that the outcome is a continuous measure and, therefore, that the amount of positive change is the most relevant way to evaluate outcomes. In some situations, however, alternatives should be considered. For example, if employment is the target outcome, comparing the number of participants who worked continuously for at least 2 months might be more meaningful than comparing mean number of days worked. In the first approach, the researcher determines a priori what constitutes meaningful change and then compares how many people in each group actually reached the criterion. This information may be important if the researcher is interested in knowing how many clients are likely to benefit from the intervention and may suggest directions for exploration to understand why some of the participants did not reach the criterion. In contrast, comparing average days worked yields an overall conclusion about whether, on average, participants in one group benefited more than participants in the other, but it does not provide a direct indication about whether the average change was meaningful or how many participants actually achieved average change or better.

Although alternatives such as dichotomizing an outcome variable may be useful in some situations, this approach also has its problems. In the employment example above, the criterion for success was continuous employment for 2 months. How was that criterion selected? The researcher needs to justify the criterion and, ideally, provide evidence to support its validity. For example, input from consumers or professionals about what is perceived as a meaningful outcome may be obtained from focus groups. Alternatively, the point on a scale that represents a transition from moderate difficulty to minimal difficulty with most activities may be used as the cutpoint to define groups. Problems arise, however, when there is no obvious score to set as a cutpoint. This situation has led researchers to apply a variety of different cutpoints on the same instrument, which then makes comparison of results extremely difficult because the criterion to identify the participants with positive outcomes varies across studies (Duncan, Jorgensen, & Wade, 2000; Duncan, Lai, & Keighley, 2000).

Question 3: Who Should (or Could) Provide the Relevant Outcome Information?

Outcome measures are usually designed with a particular respondent or examinee in mind. Therefore, the next important question is whether the type and source of information suggested as most appropriate by the original causal model matches the instrument being considered. Three major sources of information may be considered. First, the outcome of interest may be a participant’s capability to perform specific behaviors under very specific conditions. In this case, the researcher needs a performance-based measure of the behaviors of interest. A second possibility is that the outcome of interest is the person’s typical performance in daily context as perceived by the person himself or herself. This situation calls for a self-report or patient-reported outcome (PRO) measure. Finally, the outcome of interest may be the person’s typical performance; however, the researcher wants the perspective of another party, such as a caregiver or professional. Each of these conditions has considerations.

Performance-Based Measures

If the researcher’s choice is a performance-based measure, the most obvious consideration is whether the participants are likely to be able to perform under the conditions specified. For example, will the participants have the general cognitive ability to understand the directions, and are they expected to have the physical, sensory, or communicative abilities to undertake the required activities? If the researcher proposes to assess people who are acutely ill, what is a reasonable length of time for the assessment to last? If such concerns exist, adaptations may be possible, but these may represent a deviation from standard procedure, and therefore their reliability and validity may need to be examined before the adapted measure is used in a study.

Measures of Typical Performance

Similar considerations apply if the choice of measure is a self-report. If the study design calls for participants to complete self-reports independently, the researcher needs to determine that the literacy and other demands of the instrument are appropriate for the intended respondents. The same is true if the respondent is a third party,
particularly if the respondent is a caregiver and not someone with a professional education. If the population of interest is aging adults, one needs to consider the possibility that the caregiver may also have cognitive, sensory, or literacy limitations. In addition, it is important to know whether the caregiver is likely to be available and knowledgeable about the participant’s functioning throughout the study period. For example, an adult daughter who is helping out immediately after her mother’s discharge following a hip fracture may no longer be present to serve as an informant 6 months later. Similarly, professional staff may be able to complete an assessment while the participant is hospitalized, but the same staff usually will not be available to provide follow-up information. A change in respondents across a follow-up period makes it difficult to determine whether a change in assessment scores is a true change or reflects the different perspectives of two different respondents.

**Question 4: When Should the Outcomes Be Measured?**

As noted earlier in the discussion of sensitivity and responsiveness, temporal considerations are important in outcome selection. The selection of one or more assessment points has to include consideration of measurement issues to avoid a mismatch between the study plan and the capacities of the measuring instrument. For example, it was noted earlier that the researcher needs to be sure that the instrument selected is responsive over the period of time between delivery of the intervention and measurement of outcomes. Evidence regarding the likely time course of change must be carefully evaluated to determine the optimal assessment point. For example, it may take longer for a person to return to community activities than to resume some home-based activities. If community participation is the outcome of interest, then a study design must incorporate a longer follow-up period and an instrument that is sensitive to change in this domain of function.

Another temporal consideration is the expected trajectory of change in the study population. For example, between admission to acute care and discharge home, people generally are expected to have a positive trajectory. People followed up 1 year after receiving rehabilitation services, however, were found to follow a variety of trajectories that included periods of decline, recovery to previous level, or even late-occurring recovery after a plateau (Prvu Bettger, Coster, Latham, & Keysor, 2008). Other conditions such as multiple sclerosis or rheumatoid arthritis also may have a variable impact on function. To accurately capture these patterns, researchers need to assess outcomes at intervals that make clinical sense given knowledge of the condition and need to use instruments that ask about the person’s function over a relevant time period. For example, it may make more clinical sense to ask people with multiple sclerosis, “Over the past month, how much of the time were you able to walk at least 3 blocks without resting?” rather than to ask the same question in reference to “the past year.” Researchers also need to anticipate the possibility of negative change and ensure that they have planned the appropriate statistical analyses to account for both directions of change (Coster, Haley, & Jette, 2006). When both types of change are present, analyses that simply use the average amount of change (calculated as posttest score minus pretest score) will be using an average of positive and negative numbers that may not accurately represent the status of participants in either group.

**Conclusion**

Selection of an appropriate outcome measure is a key factor that affects the ultimate value of results from clinical studies as a guide for improving the lives of people with health and disability challenges. To date, selection of measures has focused primarily on the basic psychometric properties of available instruments, but this focus fails to address important questions about the suitability of the measures for their intended use. This article has provided an initial set of guidelines to encourage researchers and consumers of research to expand their thinking about how to select appropriate outcome measures and to help ensure that future studies are optimally designed to advance knowledge about the effects of intervention and rehabilitation programs. To summarize,

- Evaluation of basic psychometric properties is only the first step in determining whether a particular instrument is appropriate to measure the outcomes of a clinical intervention or program.
- Different instruments may capture different aspects of complex phenomena, such as function or participation, and may not be equally valid for all people.
- A good match between the measure and what the researcher expects to change as a result of intervention is needed to ensure a valid picture of the outcomes. ▲

**Acknowledgments**

This article was supported in part by a grant from the National Institutes of Health (NIH), Eunice Kennedy Shriver National Institute for Child Health and Human Development (NICHD), to Boston University as part of...
the Medical Rehabilitation Infrastructure Network, NIH/NICHD 1R24HD065688-02, Alan Jette, principal investigator. An earlier version of this article was presented at the Advancing Clinical Trials and Outcomes Research (ACTOR) Conference, Fairfax, VA, December 2, 2011.

References


Appendix 1. Guiding Questions for Selecting Outcome Measures

Focus

What: Specification of the construct
1. Is there a well-specified explanatory model showing how the intervention links to the outcome of interest?
2. Have the most relevant dimensions or aspects of the outcome been specified clearly?

How: Rationale for selecting the measure
1. Does the measurement construct of the instrument match the study’s target outcome (as specified by the model)?
2. Does the instrument address the relevant domains of greatest importance?
3. Do the items sample the domain at the desired or appropriate level of specificity?
4. Are the items well suited to the characteristics of the population (i.e., are they free from bias)?
5. Does the measurement dimension reflect the type of change expected from the intervention?
6. Do points on the scale match the degrees of variation expected in the sample?
7. Are item and scale wording appropriate (i.e., meaningful, understandable) for this population?
8. Does evidence exist that the measure is sensitive to degrees of change expected in this population?
9. Does evidence exist supporting the ability of the measure to identify meaningful change?

Who: Determination of the most appropriate source of outcome information
1. Do the potential providers of outcome information (e.g., professional, caregiver) match the qualifications criteria of the instrument being considered?
2. If someone other than a professional will be the respondent, is it probable that the respondent will be able to complete the assessment (i.e., has the necessary sensory, literacy, cognitive, physical, and communication abilities)?
3. Can the measure be adapted if needed to accommodate functional limitations of the respondent?
4. Will the identified respondents be available throughout the study period (i.e., for all measurement points)?

When: Determination of when outcomes should be measured
1. Does the length of time between assessments match the time period over which this instrument is likely to show effects?
2. Can the measure be administered as often as required by the study design?