JRRD Guest Editorial

Kathryn E. Roach, PhD, PT; Alison A. Linberg, DPT, ATC; Michele A. Raya, PhD, PT, SCS, ATC

Developing and testing a new outcome measure

Health outcome measures are used to assess treatment effectiveness. Historically, survival was the most important health outcome. Treatments were assessed based on whether the patient lived or died. Now in the 21st century, treatments for many injuries and illnesses are so effective that many, if not most, patients survive. If patients survive, the quality of their survival becomes an important health outcome. The severity of impairments in body structure (limb loss) and function (weakness) are one way to assess the quality of survival. However, from the perspective of the patient, discomfort and disability are probably more important health outcomes. Limitation in mobility activities is an important component of disability, particularly for individuals with lower-limb loss (LLL). The treatment effectiveness of the rehabilitative care provided to servicemembers with LLL should be assessed by examining its effect on mobility limitations. Unfortunately, it is much more difficult to accurately measure mobility limitations than it is to measure survival.

Measurement requires developing a set of rules to assign numbers to represent a concept or health outcome. When outcome measures contain multiple items, rules must also be developed to combine item scores to generate total and subscale scores. Determining the set of rules that will best represent a particular health outcome is affected by both the reason for measuring the outcome and the types of individuals being measured. Outcome measures can be used to examine small changes resulting from a treatment or to place individuals into broad categories. The purpose for using the outcome measure will dictate the types of items selected and the measurement dimension attached to these items. Measures of mobility activity limitations can be either performance-based or self-report. The types of individuals being measured will often determine which approach is best. A performance-based outcome measure requires both a set of rules for performing the test and a set of rules for scoring the test.

It is not enough to simply create an outcome measure. It is essential to determine whether the rules used to create the outcome measure work to consistently and accurately represent the concept being measured.

RELIABILITY

For an outcome measure to be useful, it must be reliable in that it produces consistent findings if no real change has occurred. Performance-based outcome measures use raters. Raters must be trained to follow a standard set of rules to administer and score the measure. If raters do not adhere to rules, measurement errors may occur that adversely affect the reliability of the

measure. Two types of rater reliability can be examined. Intrarater reliability indicates how consistently a rater administers and scores an outcome measure. Interrater reliability indicates how well two raters agree in the way they administer and score an outcome measure. To evaluate a measure's ability to detect real change, we must also examine score consistency over short periods in which no real change should occur. This is called test-retest reliability. An outcome measure used to evaluate progress over time must also be responsive in its ability to detect real change. Test-retest reliability is a critical factor in determining how well an outcome measure will detect real change.

Reliability can be examined experimentally by testing how well scores agree between raters and time periods. Agreement is expressed mathematically by calculating a reliability coefficient representing the ratio of true score variance divided by true score variance plus error variance. A reliability coefficient of 1.0 represents perfect reliability, indicating that all of the differences between scores represent real differences between individuals. A reliability coefficient of 0.43 indicates that 43 percent of the variance is due to true score and 57 percent of the variance is due to measurement error. In general, reliability coefficients below 0.50 are considered poor and above 0.75 are considered good.

VALIDITY

Minimally, an outcome measure must be reliable. However, reliability is not enough. To be useful, an outcome measure must accurately represent the phenomenon of interest in a particular group of individuals. The degree to which a measure represents a particular concept is called validity. To test validity, researchers make a series of assumptions about how scores should behave if the instrument measures what it is supposed to measure. There are many types of validity. Criterion validity is the most straightforward type of validity. The criterion validity of an outcome measure is tested by comparing the results of the new measure to a gold standard or criterion test. If the new test measures what it is intended to measure, then its results should agree with the results of the gold standard criterion test. Often new measures are developed because there are no existing measures that can be used for a particular purpose with a particular group of individuals. In these circumstances, there is no gold standard and criterion validity cannot be tested.

In the absence of a gold standard, it is still possible to validate a test. Convergent validity can be used to determine whether a test is valid. Convergent validity is demonstrated when scores on the test being examined are highly correlated with scores on a test thought to measure similar or related concepts. Convergent validity is examined experimentally by administering multiple measures to the same group of individuals and calculating correlation coefficients. If the correlation coefficients are of the magnitude and direction theorized, the validity of the measure is supported.

The known-groups method can also be used to determine construct validity. This approach is based on the assumption that if you give an outcome measure to groups of individuals that you know differ on the phenomenon you are measuring, the scores of the groups should differ if the outcome measure accurately measures what it is intended to measure.

RESPONSIVENESS

If an outcome measure is used to evaluate changes in patients over time, the measure must be able to detect this change. Responsiveness has been defined as the ability of an instrument to accurately detect change when it has occurred. Responsiveness is typically examined by administering a measure before and after a treatment that is known to be effective. Reliability is a critical component of responsiveness. Measures with poor reliability will have difficulty detecting real change because the noise introduced by measurement error will obscure any real change that has occurred. The minimal detectable change incorporates both reliability and subject variability to determine the smallest change that exceeds measurement error.

For a measure to be responsive, it is also important that most respondents do not initially achieve the highest possible score on that measure, also know as a ceiling effect. For a measure to be responsive, there must be a potential for the scores to improve after most respondents have been administered the measure. This is one of the reasons why an outcome measure must be examined for a particular purpose in a particular group of individuals. An outcome measure could work well for one group of individuals but demonstrate a ceiling effect for another.

SUMMARY

Measuring an activity limitation health outcome is much more challenging than measuring survival. Activity limitation outcome measures must be tested for reliability, validity, sensitivity to change, and responsiveness in the context of a particular purpose for a particular group of individuals before they can be used.

Kathryn E. Roach, PhD, PT;[1*] Alison A. Linberg, DPT, ATC;[2] Michele A. Raya, PhD, PT, SCS, ATC[1]

[1]Department of Physical Therapy, Miller School of Medicine, University of Miami, Coral Gables, FL; [2]Military Advanced Training Center, Walter Reed Army Medical Center, Washington, DC

[*]Email: keroach@miami.edu

This editorial and any supplemental material should be cited as follows:
Roach KE, Linberg AA, Raya MA. Developing and testing a new outcome measure. J Rehabil Res Dev. 2013;50(7):xvii–xx.
http://dx.doi.org/10.1682/JRRD.2013.05.0112