Measuring Quality and Outcomes in Sports Medicine

Joseph J. Ruzbarsky, MD, Niv Marom, MD, Robert G. Marx, MD*

KEYWORDS

• Patient-reported outcome measures • Psychometric properties • Sports medicine
• Computer-adaptive testing

KEY POINTS

- Sports medicine contains an ever increasing array of available patient-reported outcome measures, which are commonly used metrics for assessing outcomes.
- Most traditional patient-reported outcome measures are constructed using classic test theory and are administered to patients as a whole, either on paper or digitally, for completion and score tabulation.
- Modern patient-reported outcomes measures, like the Patient-Reported Outcomes Measurement Information System, are constructed using item response theory and are amenable to computer-adaptive testing, which can decrease the question burden for patients and provide accurate results.

INTRODUCTION

Over the past several decades there has been a dramatic increase in the number of patient-reported outcome measures (PROMs) within medicine and orthopedics, specifically, as we have moved into an era of patient-centric care. PROMs have become so plentiful that it is difficult to keep track of the details of each instrument: how it is scored, for what conditions it was developed, and its intrinsic test characteristics or metrics.

A basic understanding of both the questionnaire content and the population for which it has been designed and tested is critical when collecting data for research purposes or for interpreting the published literature. When gathering data on patients,

Disclosure Statement: Dr R.G. Marx discloses his role as the Deputy Editor of Sports Medicine at the *Journal of Bone and Joint Surgery* and royalties from books published by Springer and Demos Health.
Conflicts of Interest and Source of Funding: There are no funding sources to disclose.
Department of Sports Medicine, Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021, USA
* Corresponding author.
E-mail address: Marxr@hss.edu

Clin Sports Med 37 (2018) 463–482
https://doi.org/10.1016/j.csm.2018.03.001 **sportsmed.theclinics.com**

choosing the single or few most appropriate instruments is important to accurately measure the outcomes of interest while minimizing the test burden for the patients. Furthermore, there may soon be a financial incentive to better understand PROMs, as they may potentially be linked to performance-based pay models after surgical procedures in the future.

With these important characteristics of PROMs in mind, the goals of this article are first to review some of the psychometric properties of these instruments and then to focus on the most commonly used PROMs within the realm of sports medicine, specifically focusing on the shoulder, elbow, knee, and hip joints. For the purposes of this review, measures focusing exclusively on lower extremity arthritis and general health measures, such as the 36-Item Short Form Health Survey (SF-36), are excluded.

BACKGROUND

Before a discussion about specific instruments, it is important to discuss the metrics used to quantify their intrinsic characteristics, also known as their psychometric properties. In general, the term *validity* is an index of how well a test measures what it is supposed to measure.[1,2] There are several types of validity that are used to describe an instrument. *Criterion or construct validity* is assessed by correlating the scores of the tool with that of a gold standard measure.[2] *Face validity* is when an expert in a specific field reviews the questions in the instrument and confirms that they measure the concept.[2] Finally, *content validity* measures whether the scale includes representative samples of the concept that the investigator is attempting to measure.[2]

Reliability is a measure of consistency or degree of dependability. In other words, reliability is the random error of a measure or the extent to which the scores are reproducible.[3] Reliability testing involves administration of an instrument at 2 time periods (usually days to weeks[2]) to the same individual and then determining the similarity of those responses. Agreement is then reported numerically in the form of the intraclass correlation coefficient (ICC). ICC values range from −1 to +1, with a value of 0 indicating only a random correlation. Internal consistency, reported by the Cronbach alpha, is another measure of reliability.[2] Cronbach alpha values range from 0 to 1, with values of 1 representing perfect internal consistency, which is a measure of the interitem correlation of all items in the scale. Values of 0.7 are generally considered acceptable.

Responsiveness is the ability of an instrument to detect clinical change over time. This value is measured by several statistics, including the responsiveness index (mean change score/variability of scores among subjects' scores); standardized response mean (mean change in score divided by the standard deviation of the change scores); or the effect size (mean change score divided by the standard deviation of baseline scores).

Minimal clinically important difference (MCID) is defined as the smallest difference in score in the domain of interest which patients perceive as beneficial.[4,5] Although values of MCIDs may not always be available or reported, a systematic review determined that under many circumstances, when patients with a chronic disease are asked to identify minimal change, the estimates fall very close to half a standard deviation of the results.[6]

Classic test theory is an instrument development theory whereby all of the questions of an instrument in combination is validated; this is in contrast to *item response theory* (IRT) whereby each item is independently validated.[7] Independent question validation allows for questions to be mixed and matched in administration and is required for computer adaptive testing.