Commentary

Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures

Joseph C. Cappelleri, PhD[1]; J. Jason Lundy, PhD[2]; and Ron D. Hays, PhD[3]

[1]Pfizer Inc, Groton, Connecticut; [2]Critical Path Institute, Tucson, Arizona; and [3]Division of General Internal Medicine & Health Services Research, University of California at Los Angeles, Los Angeles, California

ABSTRACT

Background: The US Food and Drug Administration's guidance for industry document on patient-reported outcomes (PRO) defines *content validity* as "the extent to which the instrument measures the concept of interest" (FDA, 2009, p. 12). According to Strauss and Smith (2009), construct validity "is now generally viewed as a unifying form of validity for psychological measurements, subsuming both content and criterion validity" (p. 7). Hence, both qualitative and quantitative information are essential in evaluating the validity of measures.

Methods: We review classical test theory and item response theory (IRT) approaches to evaluating PRO measures, including frequency of responses to each category of the items in a multi-item scale, the distribution of scale scores, floor and ceiling effects, the relationship between item response options and the total score, and the extent to which hypothesized "difficulty" (severity) order of items is represented by observed responses.

Results: If a researcher has few qualitative data and wants to get preliminary information about the content validity of the instrument, then descriptive assessments using classical test theory should be the first step. As the sample size grows during subsequent stages of instrument development, confidence in the numerical estimates from Rasch and other IRT models (as well as those of classical test theory) would also grow.

Conclusion: Classical test theory and IRT can be useful in providing a quantitative assessment of items and scales during the content-validity phase of PRO-measure development. Depending on the particular type of measure and the specific circumstances, the classical test theory and/or the IRT should be considered to help maximize the content validity of PRO measures. (*Clin Ther*. ▮▮▮▮;▮:▮▮▮–▮▮▮) © 2014 Elsevier HS Journals, Inc. All rights reserved.

Key words: classical test theory, content validity, item response theory, patient-reported outcomes, scale development.

INTRODUCTION

The publication of the US Food and Drug Administration's guidance for industry on patient-reported outcomes (PRO)[1] has generated discussion and debate on the methods used for developing, and establishing the content validity of, PRO instruments. The guidance outlines the information that the FDA considers when evaluating a PRO measure as a primary or secondary end point to support a claim in medical product labeling. The PRO guidance highlights the importance of establishing evidence of *content validity*, defined as "the extent to which the instrument measures the concept of interest" (p. 12).[1]

Content validity is the extent to which an instrument covers the important concepts of the unobservable, or latent, attribute (eg, depression, anxiety, physical functioning, self-esteem) that the instrument purports to measure. It is the degree to which the content of a measurement instrument is an adequate reflection of the construct being measured. Hence, qualitative work with patients is essential to ensure

Accepted for publication April 9, 2014.
http://dx.doi.org/10.1016/j.clinthera.2014.04.006
0149-2918/$ - see front matter

© 2014 Elsevier HS Journals, Inc. All rights reserved.

that a PRO instrument captures all of the important aspects of the concept from the patient's perspective.

Two reports from the International Society of Pharmacoeconomics and Outcomes Research Good Research Practices Task Force[2,3] detail the qualitative methodology and 5 steps that should be employed to establish content validity of a PRO measure: (1) determine the context of use (eg, medical product labeling); (2) develop the research protocol for qualitative concept elicitation and analysis; (3) conduct the concept elicitation interviews and focus groups; 4) analyze the qualitative data; and (5) document concept development, elicitation methodology, and results. Essentially, the inclusion of the entire range of relevant issues in the target population embodies adequate content validity of a PRO instrument.

Although qualitative data from interviews and focus groups with the targeted patient sample are necessary to develop PRO measures, qualitative data alone are not sufficient to document the content validity of the measure. Along with qualitative methods, quantitative methods are needed to develop PRO measures with good measurement properties. Quantitative data gathered during earlier stages of instrument development can serve as: (1) a barometer to see how well items address the entire continuum of the targeted concept of interest; (2) a gauge of whether to go forward with psychometric testing; and (3) a meter to mitigate risk related to Phase III signal detection and interpretation.

Specifically, quantitative methods can support the development of PRO measures by addressing several core questions of content validity: What is the range of item responses relative to the sample (distribution of item responses and their endorsement)?; Are the response options used by patients as intended?; Does a higher response option imply a greater health problem than does a lower response option?; and What is the distance between response categories in terms of the underlying concept?

Also relevant is the extent to which the instrument reliably assesses the full range of the target population (scale-to-sample targeting), ceiling or floor effects, and the distribution of the total scores. Does the item order with respect to disease severity reflect the hypothesized item order? To what extent do item characteristics relate to how patients rank the items in terms of their importance or bother?

This article reviews the classical test theory and the item response theory (IRT) approaches to developing PRO measures and to addressing these questions. These content-based questions and the 2 quantitative approaches to addressing them are consistent with construct validity, now generally viewed as a unifying form of validity for psychological measurements, subsuming both content and criterion validity.[4] The use of quantitative methods early in instrument development is aimed at providing descriptive profiles and exploratory information about the content represented in a draft PRO instrument. Confirmatory psychometric evaluations, occurring at the later stages of instrument development, should be used to provide more definitive information regarding the measurement characteristics of the instrument.

CLASSICAL TEST THEORY

Classical test theory is a conventional quantitative approach to testing the reliability and validity of a scale based on its items. In the context of PRO measures, classical test theory assumes that each observed score (X) on a PRO instrument is a combination of an underlying true score (T) on the concept of interest and nonsystematic (ie, random) error (E). Classical test theory, also known as *true-score theory,* assumes that each person has a true score, T, that would be obtained if there were no errors in measurement. A person's *true score* is defined as the expected score over an infinite number of independent administrations of the scale. Scale users never observe a person's true score, only an observed score, X. It is assumed that observed score (X) = true score (T) + some error (E).

True scores quantify values on an *attribute of interest,* defined here as the underlying concept, construct, trait, or ability of interest (the "thing" intended to be measured). As values of the true score increase, responses to items representing the same concept should also increase (ie, there should be a monotonically increasing relationship between true scores and item scores), assuming that item responses are coded so that higher responses reflect more of the concept.

It is also assumed that random errors (ie, the difference between a true score and a set of observed scores in the same individual) found in observed scores are normally distributed and, therefore, that the expected value of such random fluctuations (ie, mean of the distribution of errors over a hypothetical infinite number of administrations in the same subject) is taken to be zero. In addition, random errors are assumed to

Download English Version:

https://daneshyari.com/en/article/5825506

Download Persian Version:

https://daneshyari.com/article/5825506

Daneshyari.com