Journal of Psychosomatic Research 68 (2010) 319–323

Introduction to health measurement scales

András P. Keszei[a], Márta Novak[a,b], David L. Streiner[c,*]

[a]Institute of Behavioral Sciences, Semmelweis University, Budapest, Hungary
[b]Department of Psychiatry, University Health Network and University of Toronto, Toronto, Canada
[c]Kunin-Lunenfeld Applied Research Unit Baycrest Centre, Toronto, Canada

Received 1 May 2009; received in revised form 14 January 2010; accepted 14 January 2010

Abstract

Both research and clinical decision making rely on measurement scales. These scales vary with regard to their psychometric properties, ease of administration, dimensions covered by the scale, and other properties. This article reviews the main psychometric characteristics of scales and assesses their utility.

Keywords: Rating scales; Reliability; Validity

Introduction

There is no single variable that can be used to describe health, and health cannot be measured directly. Health measurement requires several steps and involves the evaluation of several health-related indicators.

Rating scales are used in numerous settings to measure various aspects of health such as different symptoms or the presence of a particular trait. Health measurement scales can be classified in (at least) three ways, according to their function, description, and methodology. Functional classification focuses on the application of methods and how they are used, such as Bombardier and Tugwell's [1] classification of diagnostic, prognostic, and evaluative health measurements; however, others [2] have argued that this classification ignores the way scales are actually used in practice. Descriptive classification of health measurements is concerned with the range of topics covered by a particular measurement. For example, one might focus on a particular organ system, a diagnosis, or a broader concept such as anxiety or quality of life. Another distinction can be between broad classification of generic health measures and specific instruments. Specific instrument can be concerned with not only a particular disease, but also a particular target population, such as children. Methodological classification distinguishes among rating scales, questionnaires, indices, and subjective vs. objective measures.

Whether rating scales are to be used in a research project or to make clinical decisions, it is essential to evaluate how well they perform. By how well, we mean how much random error is present in the measurement (i.e., its reliability) and whether the scores give us meaningful information about the respondent (the validity of the instrument). A third measure of performance addresses the issue of whether it is feasible to use the instrument for a particular purpose. In this article, we will give an introduction to some of the properties of rating scales, the concept of validity and reliability. Those who are interested in the details of constructing measurement scales are referred to more comprehensive texts [2–4].

Scale development can be approached in two ways: questions may be chosen from an empirical or a theoretical viewpoint [5]. With the empirical approach, a large number of questions are tested and statistical procedures are used to select the ones that best predict the outcome of interest. However, the disadvantage of this method is that it is difficult to interpret why individuals answering a certain question in a certain way tend to have different outcomes. Questions in the Health Opinion Survey [6] were selected because they distinguished between those who do and do

* Corresponding author. Kunin-Lunenfeld Applied Research Unit, Baycrest Centre, 3560 Bathurst Street, Toronto, Ontario, Canada M6A 2E1. Tel.: +1 416 785 2500x2534; fax: +1 416 785 4230.
E-mail address: dstreiner@klaru-baycrest.on.ca (D.L. Streiner).

0022-3999/10/$ – see front matter
doi:10.1016/j.jpsychores.2010.01.006

not have psychiatric problems. However, debates over what exactly the scale measures are still continue. Scales developed entirely from an empirical stance may have clinical value, but they do not advance our understanding of the underlying phenomena. The alternative strategy is to select questions that are thought to be relevant from a standpoint of a particular theory, such as the McGill Pain Questionnaire [7]. In psychology, at least, the trend over the past 50 years has been a move toward theoretically derived instruments [2].

Items on a scale

Items on a scale can come from several different sources: existing scales, reports of individuals' subjective experiences, clinical observations, expert opinion, research findings, and theory. One should be aware of the strengths and weaknesses of each source when considering a scale for a particular use. The advantage of using existing items from older scales is that items have probably already gone through a rigorous process of assessment and are, therefore, more likely to be useful. It may thus save time and work rather than construct new items. However, outdated terminology may render some older items useless.

Patients experiencing a trait or disorder can be excellent sources of scale items, especially when the interest lies in the more subjective elements of the trait. *Focus groups* and *key informant interviews* are techniques that can be used to acquire patients' viewpoints in a systematic manner [8].

Clinical observation as a means of developing scale items can be useful, as these observations precede any theory, research, or expert opinion. Scales developed in this manner can be seen as a structured way of assembling clinical observations. The major disadvantage, however, is that the clinician developing the scale might have been wrong in his/her observation. If a scale is based on unreplicated findings, it leads to a useless scale. For example, a scale that is based on the erroneous observation that the incidence of epilepsy is lower in the schizophrenic population is destined to failure. Moreover, the clinician observes a particular phenomenon on a limited sample of patients and, therefore, may miss other relevant factors that would be apparent in another population. One way to overcome the problem of mistaken observations is to use the judgment of not just one but a panel of experts. The advantage of this approach is that an expert panel probably represents the most recent views in a topic. A note of caution is in order, however, as there are no rules on how and how many experts have to be chosen and how opinions have to be synthesized. If the selected experts all share the opinion and perhaps biases regarding the domain to be measured, then using a panel of experts does not provide additional advantage to using the views of just one person.

Research findings from literature reviews of previous studies in the area or new research carried out for developing a scale can be another source of items. An example is a subset of a scale developed to differentiate between irritable bowel syndrome and organic bowel disease [9]. It consists of laboratory values and clinical history that were chosen on the basis of previous research, indicating difference between irritable bowel syndrome and organic disease regarding these variables.

As mentioned previously, a set of clinical and/or laboratory observations forming a theory about differences in patients might also provide items. In this context, we should not only think of formal theories, but also more vague ideas such as the notion that patients believing in the efficacy of a treatment will be more compliant. The weakness of using "theories" in item selection is the possibility of using a wrong model, and this may only be apparent later when the validity of the scale is assessed.

Criteria to identify useful items

Not every item intended for a scale will perform well; therefore, several aspects of items have to be checked to decide which are likely to be useful.

It is important to use clear, comprehensible language. Very often, technical or jargon terms are used (e.g., *stool*, *shock*, or *cardiovascular*), which would be fine if the scale is to be used on health professionals but not if lay people are the intended respondents. Since people are different in their reading ability, items should not require more than very basic reading skills. The scales should be tested on the target group to verify that the used terms are understandable.

Another potential problem related to language that can result in unintended responses is ambiguity, which can be caused, for example, by using vague terms. The answer to the question, "Have you been in hospital recently?" depends on how the respondents interpret "recently" and if they differentiate among being an inpatient, outpatient, or even a visitor. We should also check for and avoid items that incorporate two questions (e.g., "I feel sad and lonely"), because some people would answer "yes" if one part of the question applies to them, while others will say "yes" only if both apply.

Terms that might offend or prejudice people should be avoided. Items such as "Do physicians make too much money" may indicate to the respondent what the desirable answer would be.

Items that are very likely (more than 90% of the time) answered in one way or the other are not very useful. If everyone answers "yes" to a question, then that item does not contribute to discriminating between individuals who have a certain characteristic and those who do not. Furthermore, such questions can introduce unnecessary measurement error caused by random responding or other reasons for not giving the true answer.

Reliability

Before one can start using an instrument, it should be established that it is measuring "something" in a