Tool development in health care research

P. J. Parameaswari
Department of Community Medicine, Sree Balaji Medical College & Hospital, Bharath University, Chromepet, Chennai-600044, Tamil Nadu, India.

Article history
Received 01 November 2012
Accepted 07 February 2013
Early online 25 February 2013
Print 28 February 2013

Abstract
Measures of Population Health would need to be compatible across nations and cultures to serve as measures for making global health policy. Health state valuations on the basis of utility instruments such as rating scales, ex. the Visual Analog Scale or methods such as personal trade-off, time trade-off and standard gamble, are one of the critical inputs that contribute to the calculation of summary measures of population health. These methods are supposed to assess an individual’s valuation of hypothetical health states in terms of a single number indicating the value placed on a health state relative to perfect health or death.

Key words: validity, reliability, measurement, Likert scale, qualitative, quantitative

In the light of scarcity of formal cross-cultural comparisons, we see two major research strategies applicable to the investigation of variation in health state valuations. First, we believe that qualitative research is needed to precede and accompany quantitative research in health state valuations. Several scholars have stressed the need to conduct preparatory qualitative research before developing measures of population health. Qualitative research is particularly useful in elucidating and contextualizing determinants of variance in health state valuations in extremely diverse cultural settings. Such qualitative research can address a number of needs. It would complement or precede psychometric and econometric work by assisting the definition, understanding and development of conceptual domains and culturally appropriate scale items (Kessler and Mroczek 1996; Krause and Jay 1994); it would lead to contextually sensitive and culturally relevant indicators (Nazarea et al 1998) and it would document and illuminate the large cross-cultural variability of attitudes and beliefs regarding health-related impairment and disability in a wide variety of settings.

A number of modern anthropological methods are available, particularly those methods bridging qualitative and quantitative dichotomies (Bernard 1994; Trotter and Sommerfeld 1996). Qualitative and ethnographic research can be a valuable complement to comparative epidemiological research, as it sheds light on the socio-cultural and emotional context of health-related quality of life and the consequences of ill-health (Heggenhougen and Shore 1986; Marshal 1990). In addition, such qualitative research yields knowledge on the background of variations in health state valuations across individuals, ex. Potential sources of variation based on age, sex or income. Qualitative research should not be limited, however, to preparatory or formative research. Well-designed qualitative research with similar questions and standardized instruments could yield important comparative data and thus contribute to comparative quantitative research.
Measurement
Quantification refers to the application of mathematical models to help the scientist to objectively establish the precise and meaningful interpretation of information i.e. measurement is the assignment of number to events or objects according to rules of number system. It is common to distinguish four types of measurement scales according to the rules by which numbers are assigned to objects or events. It is a widespread notion that the scale on which a variable is measured determines the types of statistics that can appropriately be performed on the data, the four types of measurement scales are nominal scale, ordinal scale, interval scale and ratio scale.

- The two basic types of measurement errors are *systematic error* and *random error*.
- Measurement tools must be *reliable and valid*

Reliability
Reliability is the *consistency* of scores obtained by the same persons when retested with the identical test or with an equivalent form of the test

Validity
Validity is the degree to which the test actually measures what it purports to measure. Types of validity of measurement are:*construct validity*, *face validity*, *content validity* and *criterion validity*

Standardization implies uniformity of procedure in administering and scoring the test.

For example, *Psychological test* as a 'set of standardized or controlled occasions for response presented to an individual with design to elicit a representative sample of his behavior, to measure objectively one or more aspects of a total personality by means of samples of verbal or non-verbal responses, or by means of other behaviors.

The two major classifications of tests are*Ability test* and *Personality test*. *Ability tests* focus on the question of what people can do when they are at their very best. *Intelligence test* refers to overall capacity for learning and problem solving. *Personality test* assesses underlying thought processes, emotional states, or behavioral patterns reflecting disorders. *Aptitude test* refers to the ability to learn a particular kind of skill required in a specific situation. No amount of statistical manipulation can compensate for poorly chosen questions.

More importantly, the motivation for developing a new tool is the investigator’s belief that the previous scales are inadequate for one reason or another, or do not completely cover the domain under study. At this point, new items can come from five different sources: the patients or subjects themselves, clinical observation, theory, research and expert opinion, although naturally the lines between these categories are not firm.

In the area of *scale development*, the participants would be patients who have the disorder, or subject’s representative of those, whose opinions would be elicited by the instrument. At first, their task would not be to generate the specific items, but rather to suggest general themes which the research team can use to phrase the items that have been written; focus groups can again be used to discuss whether these items are relevant, clear, unambiguous, written in terms that are understood by potential respondents, and if all the main themes have been covered. These groups are much more focused than during the theme generation stage, since there is a strong externally generated agenda; discussing the items themselves.

Source of items
A point often overlooked in scale development is the fact that patients and potential research subjects are an excellent source of items. Whereas clinicians may be the best observers of the outward manifestations of a trait or disorder, only those who have it can report on the more subjective elements. Over the years, a variety of techniques have been developed which can elicit these view points in a rigorous and systematic manner; these procedures are used primarily by 'qualitative' researchers, and are only now finding their way into more 'quantitative' types of studies.

Williams and Johnson (1993) describe a *focus group* as a discussion in which a small group of informants (six to twelve people) guided by a facilitator, talk freely and spontaneously about themes considered important to the investigation. The participants are selected from a target group whose opinions and ideas are of interest to the researcher. Sessions are usually tape recorded and an observer (recorder) also takes notes on the discussion.

*Key informant interviews*: These are *in-depth* interviews with a small number of people, chosen because of their unique knowledge. These can be patients who have or have had the disorder, for example, and who can articulate what they felt; or clinicians who have extensive experience with the patients and can explain it from their perspective. The interviews can range from informal or unstructured ones, which are almost indistinguishable
from spontaneous conversations, to highly structured ones, where the interviewer has a preplanned set of carefully worded questions. Generally, the less that is known about the area under study, the less structured the interview.

Clinical observation: It is one of the most fruitful sources of items. Indeed, it can be argued that observation, whether of patients or students, precedes theory, research or expert opinion. Scales are simply a way of gathering these clinical observations in a systematic fashion, so that all observers are ensured of looking for the same thing, or all subjects of responding to the same items. Kruise et al.(1984) devised a scale to try to differentiate between irritable bowel syndrome and organic bowel disease. The first part of their questionnaire consists of a number of items asked by the clinician of the patient- presence of abdominal pain and flatulence, alteration in bowel habits, duration of symptoms, type and intensity of pain, abnormality of the stools, and so forth. The choice of these items was predicated on the clinical experience of the authors, and their impressions of how irritable bowel syndrome patients’ symptomatology and presentation differ from other patients. Similarly, menstrual distress questionnaire (Moos 1984)\(^5\) consists of 47 symptoms, such as muscle stiffness, skin blemishes, fatigue, and feeling sad or blue, which have been reported clinically to be associated with premenstrual syndrome. However, there has been an increasing appreciation of the role that theory can play in scale and questionnaire development. This is seen most clearly when we are trying to assess attitudes, beliefs, or traits. From The Health belief model (Becker et al.1979)\(^6\) we realize that our task would be made easier if we had a model or theory (valid) of the items under study while devising a new scale.

Research findings can be a fruitful source of items and sub-scales. For the purposes of scale construction, research can be of two types; a literature review of studies which have been done in the area or new research carried out specifically for the purpose of developing the scale. In both cases, the scale or questionnaire would be comprised of items which have been shown empirically to be characteristic of a group of people, or which differentiate them from other people.

When entering into a new area, though, there may not be any research which can serve as the basis for items. Under these circumstances, it may be necessary for the scale developer to conduct some preliminary research, which can then be the source of items.

Brumback and Howell had highlighted two points; first, it is sometimes necessary to perform research prior to constructing the scale itself in order to determine key aspects of the domain under investigation. Second, the initial item pool is often much larger than the final set of items.

Information had to be gathered prior to the construction of the scale.

Expert’s opinion should be significantly considered. There are no hard and fast rules governing the use of expert judgments, because they probably represent the most recent thinking in an area. A scale may consist of items derived from some or all of these sources.

Translation

In most large studies, translation of scales and questionnaires into the languages most commonly used within the catchment area of the study is in practice, but there is little assurance that the psychometric properties of the scale (its reliability and validity) have remained constant. It is therefore necessary to revalidate that instrument, as if it were a new one.

The minimum number of categories used by raters should be in the region of five to seven while constructing a continuous scale, example, Likert scale, semantic differential scale, etc. One bias of rating scales is the halo effect; since items are frequently ordered in a single column on a page it is possible to rapidly rate all items on the basis of a global impression, paying little attention to the individual categories. We also rarely commit to the extreme categories on the scale, effectively reducing the precision of measurement.

Developing measurement tools

The technique of developing measurement tools involves four-stages

Stage 1: Concept development

The researcher should arrive at an understanding of the major concepts pertaining to his study. The concept development is more apparent in theoretical studies than in the more pragmatic research, where the fundamental concepts are often already established.

Identification of potential domains of further inquiry:

- what problems are there and which of them merit priority attention
- who is at risk from the problem or is exposed to or affected by it, where those at risk live and
work, their social class and the ‘catchment area’ where they can be reached.

- the important one is the extent to which it is intended to reduce the problem, and the target time for achieving the reduction

**Stage 2: Specification of concept dimensions**

The researcher should specify the dimensions of the concepts that he developed in the first stage. This task may either be accomplished by deduction i.e. by adopting a more or less intuitive approach or by empirical correlation of the individual dimensions with the total concept and/or the other concepts. For instance, one may think of several dimensions such as product reputation, customer treatment, corporate leadership, concern for individuals and sense of social responsibility when one is thinking about the image of a certain organization.

We search ‘psychological abstracts’ for psychological scales and ERIC for instruments designed for educational purposes. The volume entitled ‘measuring health: a guide to rating scales and questionnaires’ is a critical review of scales designed to measure a number of characteristics of interest to researchers in the health sciences; such as pain, illness behavior, and social support. Questions designed to assess sensitive areas are likely to be less obviously related to the underlying attitude or behaviour. The scale should measure something in reproducible fashion.

**Stage 3: Selection of indicators**

Once the dimensions of a concept have been specified, the researcher must develop indicators for measuring each concept element. Indicators are specific questions, scales, or other devices by which the respondents’ knowledge, opinion, expectation, etc. are measured. As there is seldom a perfect measure of concept, the researcher should consider several alternatives for the purpose. The use of more than one indicator gives stability to the scores and it also improves their validity.

The researcher must pay attention to the following points in constructing an appropriate and effective questionnaire or a schedule:

1. The researcher must keep in view the problem he is to study for it provides the starting point for developing the questionnaire/schedule. He must be clear about the various aspects of his research problem to be dealt with in the course of his research project.

2. Appropriate form of questions depends on the nature of information sought, the respondents and the kind of analysis intended. The researcher must decide whether to use closed or open-ended question. Questions should be simple and must be constructed with a view to their forming a logical part of a well thought-out tabulation plan. The units of enumeration should also be defined precisely so that they can ensure accurate and full information.

3. A rough draft of the questionnaire/schedule must be prepared, giving due thought to the appropriate sequence of questions. Questionnaires or schedules previously drafted (if available) may as well be looked into at this stage.

4. The researcher must invariably re-examine, and in case of need may revise the rough draft for a better one. Technical defects must be minutely scrutinized and removed.

5. A pilot study should be undertaken for pre-testing the questionnaire. The questionnaire may be edited in the light of the results of the pilot study.

6. The questionnaire must contain simple but straightforward directions for the respondents so that they may not feel any difficulty in answering the questions

**Steps involved in initial selection of items**

1. Pre-test the items to ensure that they:
   - are comprehensible to the target population
   - are unambiguous
   - ask only a single question

2. Eliminate or rewrite any items which do not meet these criteria, and pretest again.

3. Discard items endorsed by very few (or very many) subjects

4. Check for the **internal consistency** of the scale using:
   - Item-total correlation
     - correlate each item with the scale total, omitting that item
     - eliminate or rewrite any with Pearson r’s less than 0.20
   - rank order the remaining ones and select items starting with the highest correlation
or

ii. Coefficient α or KR-20
   a. Calculate α eliminating one item at a time
   b. Discard any item where α significantly increases

5. For multi-scale questionnaires, check that the item is in the ‘right’ scale by:

   a. Correlating with all the scales, eliminating items which correlate more highly on scales other than the one it belongs to; or
   b. Factor-analysing the questionnaire, eliminating items which load more highly on other factors than the one it should belong to.

Stage 4: Formation of index

Combine the various indicators into an index. When we have several dimensions of a concept or different measurement of a dimension, we may need to combine them into a single index. One simple way for getting an overall index is to provide scale values to the responses and then sum up the corresponding scores. Such an overall index would provide a better measurement tool than a single indicator because of the fact that an ‘individual indicator has only a probability relation to what we really want to know’. In this way we must obtain an overall index for the various concepts concerning the research study.

Acknowledgments: None

Conflict of interest: None

References