

DEVISING ITEMS

Lecture 3 Notes From:

McDowell I. *Measuring Health: A guide to rating scales and questionnaires, Third Edition*. Oxford: Oxford University Press, 2006.

Streiner DL, Norman GR. *Health Measurement Scales: A practical guide to their development and use, Fourth Edition*. Oxford: Oxford University Press, 2015.

- Streiner says: “The first step in writing a scale or questionnaire is, naturally, devising the items themselves. This is far from a trivial task, since no amount of statistical manipulation after the fact can compensate for poorly chosen questions; those that are badly worded, ambiguous, irrelevant, or even worse, not present.”
- Sell says: “The first step in writing a scale or questionnaire is, naturally, clearly defining the construct you intend to measure.”
- Once you’ve defined the construct you need to look at what others have done in the past to measure the same or similar constructs. Instruments rarely are developed totally new.
- There are a number of reasons that items are repeated from previous measures.
 - First, it saves work,
 - Second, the items have usually gone through repeated processes of testing so that they have proven themselves to be useful and psychometrically sound, and
 - Third, there are only a limited number of ways to ask about a specific construct.
- According to Streiner new items can come from five different sources:
 - the patients or subjects themselves, (community based participation!) (A point often overlooked in measure development).
 - clinical observation,
 - theory,
 - research, and
 - expert opinion.
- **Focus Groups:** A way to get items from patients or subjects, or clinicians or experts.
 - A focus group is "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.”
 - A focus group would first be used to discuss the construct and help the investigator define it. Later focus groups can be used to test the actual items.
- **Clinical Observation/Key Informant Interviews/Expert Opinion:** A way to get information directly from the professionals that deal with the construct that is being measured (who may ultimately also be the users of the measure), and are important sources of items for a measure. (For many measures individuals other than clinicians are best.... listing “clinical observation” in the text shows Streiner’s medical bias.) Key informants of course might be people with the construct (once again CBPR). The advantage of this approach is that if the informants/experts are chosen carefully, they probably represent the most recent thinking in an area. Without much effort, the scale developer has access to the accumulated knowledge and experience of others who have worked in the field.
- **Theory:** Individual laboratory results or physical findings convey far more information if they are components of a more global theory of an illness or behavior. The term *theory* can be used to describe formal models of how things relate to one another. I (Randy) use the term theory often to describe a definition or detailed description of the construct to be measured.
 - For example, if we wanted to devise a scale that could predict those post-MI patients who would comply with an exercise regimen, our task would be made easier if we had some model or theory

of compliance. The Health Belief Model postulates that compliance is a function of numerous factors, including the patient's perception of the severity of the disorder and his susceptibility to it, and his belief in the effectiveness of the proposed therapy, as well as external cues to comply and barriers which may impede compliance. If this model has any validity, then it would make sense for the scale to include items from each of these areas.

- **Research:** *Research findings* can be a source of items and subscales for measures. For the purposes of scale construction, research can be of two types: a literature review of studies that 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 the characteristics of a group of people or which differentiate them from other people.
 - Point one: It is sometimes necessary to perform research prior to the construction of the scale itself in order to determine the key aspects of the domain under investigation.
 - Point two: the initial item pool is often much larger than the final set of items.
- **Content Validity (Face Validity)** – (this is cut and pasted from last week's notes.... think about how this compares to the use of expert opinions to create scale items). A judgment that a scale looks reasonable and thoroughly measures a desired construct. Face validity simply indicates whether, on the face of it, the instrument appears to be assessing the desired qualities, that is, whether the scale appears appropriate for the intended purpose. Content validity is rarely tested empirically, but rather tested using face validity. Content validity is not sufficient for showing validity.
- **Generic Versus Specific Scales:**
 - The argument in favor of disease-specific and patient-specific questionnaires is two-fold. The first consideration is that if an instrument has to cover a wide range of disorders, many of the questions may be inappropriate or irrelevant for any one specific problem. The second reason for using disease- or patient-specific scales follows directly from the first problem. In order to keep the length of a generic questionnaire manageable, there cannot be very many items in each of the areas tapped. Thus, there will be fewer relevant questions to detect real changes within patients, or differences among them.
 - The argument against disease-specific and patient-specific questionnaires is also two-fold. First, the cost of the greater degree of specificity is a reduction in generalizability. That is, generic scales allow comparisons across different disorders, severities of disease, interventions, and perhaps even demographic and cultural groups. Second, since any one generic scale tends to be used more frequently than a given specific instrument, there are usually more data available regarding its reliability and validity.
- **Translating Measures:** Don't attempt on your own!