
QUESTIONNAIRES 101:

Selecting, Evaluating, and Adapting Self-Reported Measures

Alison Huang, MD, MAS
Professor of Medicine, Urology,
and Epidemiology & Biostatistics

Special thanks to:

Anita Stewart, PhD, Professor,
Institute for Health & Aging,
UCSF School of Nursing



Miriam Kuppermann, PhD, MPH,
Professor, UCSF Department of
Obstetrics, Gynecology, &
Reproductive Sciences



Today's goals

- Examine different types of self-reported measures used in clinical research
- Discuss aspects of self-reported measures that affect the quality of the data they produce
- Outline procedures for selecting, evaluating, and adapting self-reported measures for a given study

Common approaches to collecting self-reported data in clinical research

- Structured interviews to screen a participant for eligibility
- Emailed surveys to assess respondents' health history
- Paper-based questionnaires measuring quality of life
- Participant diaries that document timing of symptoms
- Electronic logs to record health-related behaviors

Terminology

A *questionnaire* is a document containing one or more questions (*items*) to measure self-reported concepts

Questionnaires may combine multiple items to provide a more robust or multidimensional quantitative measure of a concept (*multi-item scales*)

A questionnaire or self-reported measure is sometimes referred to as a *study instrument*

Collection of self-reported data in a cross-sectional study can be called a *survey*

Single versus multi-item measures

Single-item measure of overall health:

Would you rate your health as excellent, very good, good, fair, or poor?

☐ Excellent ☐ Very good ☐ Good ☐ Fair ☐ Poor

PROMIS-10 global health scale:

10 questions assessing overall physical health, mental health, social health, pain, fatigue, and overall perceived quality of life

Uni- versus multidimensional measures

- Perceived Stress Scale (PSS) – 14 items examining the subjective experience of stress, designed to generate a single score from all items
- Multidimensional Behavioral Health Screen (MBHS)– 27 items examining multiple types of behavioral symptoms (anxiety, depressive, attention/cognitive-related)

Condition-specific versus generic measures

- Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire (PISQ)

Sample item: How much does the fear of leaking urine cause you to avoid sexual activity?

- Female sexual function index (FSFI)

Sample item: How satisfied have you been with your overall sexual life?

Goals of self-reported data collection = goals of physiologic data collection

- Obtain the most precise, sensitive, specific measurement possible
- Produce data that are robust and reproducible in future research studies
- Administer measures that minimize both systematic and random error

Sources of error from physiologic versus self-reported measures

Physical exam measure: Body weight	Self-reported measure: Depression symptoms
• Inconsistent weighing protocol • Variable weight of clothes • Miscalibration of scale • Staff or observer error	• Reluctance to tell interviewer • Confusing stems or response options • Culturally insensitive measures • Excessive burden of questionnaires

Steps in developing new measures

- Develop concept

- Create item pool

- Pretest/revise items

- Field testing of measure

- Psychometric analysis

- Finalization of measure

When should you develop a new self-reported measure?

- Does a robust self-reported instrument already exist for the characteristic you want to study? **NO**
- Can you adapt an existing but suboptimal instrument for your own research? **NO**
- Is the characteristic a key study variable (major predictor or outcome)? **YES**
- Do you have the time and resources to invest in developing a completely new instrument? **YES**

Steps in selecting existing measures

- 1) Define the concept you want to measure and the context or population in which you want to measure it
- 2) Search for existing self-reported measures described in prior publications or research repositories
- 3) Review measures for face validity, acceptability, practicality, and conceptual/psychometric adequacy
- 4) Pre-test candidates and choose the best one, or adapt measures after pre-testing

Defining the concept being measured

Vaginal health in postmenopausal women

```
graph TD; A[Vaginal health in postmenopausal women] --> B[Postmenopausal vaginal symptoms in women]; B --> C[Impact of postmenopausal vaginal symptoms on quality of life in women]; C --> D[Impact of postmenopausal vaginal symptoms on intimate relationships in Latina American women];
```

Postmenopausal vaginal symptoms in women

Impact of postmenopausal vaginal symptoms on
quality of life in women

Impact of postmenopausal vaginal symptoms on
intimate relationships in Latina American women

Where can you find existing measures?

PhenX toolkit

<https://www.phenxtoolkit.org/>

NIH toolbox

<https://www.healthmeasures.net/explore-measurement-systems/nih-toolbox>

PROMIS (Patient-Reported Outcomes Measurement Information System)

<https://www.healthmeasures.net/explore-measurement-systems/promis>

Science of Behavior Change Measures Repository

<https://scienceofbehaviorchange.org/Measures>

National Institutes of Health Common Data Element (CDE) Resource Portal

<https://www.nlm.nih.gov/cde>

REDCap HealthMeasures

<https://www.healthmeasures.net/implement-healthmeasures/administration-platforms/redcap>

Reviewing existing self-reported measures

- 1) Initial face validity— do the individual items or scales seem clear, relevant, and appropriate?
- 2) How easy or burdensome will it be to administer this measure in your own study context?
- 3) Does the measure show evidence of **conceptual and psychometric adequacy** for your study?

Evaluating self-reported measures: Are items too burdensome?

Please rank the following sources of information according to how helpful they were to you in making your prenatal testing decision.

- ☐ Your midwife or obstetrician
- ☐ A genetics counselor
- ☐ A group class taught by a health care professional
- ☐ Written material given to you by your healthcare providers
- ☐ Your friends and relatives
- ☐ The internet

Evaluating self-reported measures: Medical or scientific terms

Please indicate if you have had any of the following health conditions:

- ☐ Myocardial infarction
- ☐ Cerebrovascular accident
- ☐ Thrombophlebitis
- ☐ Pulmonary embolism

Evaluating self-reported measures: Double-barreled questions

How satisfied were you with the care you received from doctors and nurses in the hospital?

- ☐ Very satisfied
- ☐ Moderately satisfied
- ☐ Neither satisfied nor dissatisfied
- ☐ Moderately dissatisfied
- ☐ Very dissatisfied

Evaluating self-reported measures: Hidden assumptions

How often in the past week did you feel down or depressed even with help from your family?

- ☐ All of the time
- ☐ More than half of the time
- ☐ About half of the time
- ☐ Less than half of the time
- ☐ None of the time

Evaluating self-reported measures: Leading questions

In what ways has the current medical insurance system failed to meet its ethical mission? (check all that apply)

- ☐ Too many uninsured
- ☐ Insufficient catastrophic coverage
- ☐ Inability to access specialists
- ☐ Too difficult to obtain reimbursement
- ☐ Other: (please specify) _____

What does it mean for a questionnaire or other self-reported measure to be 'validated'?

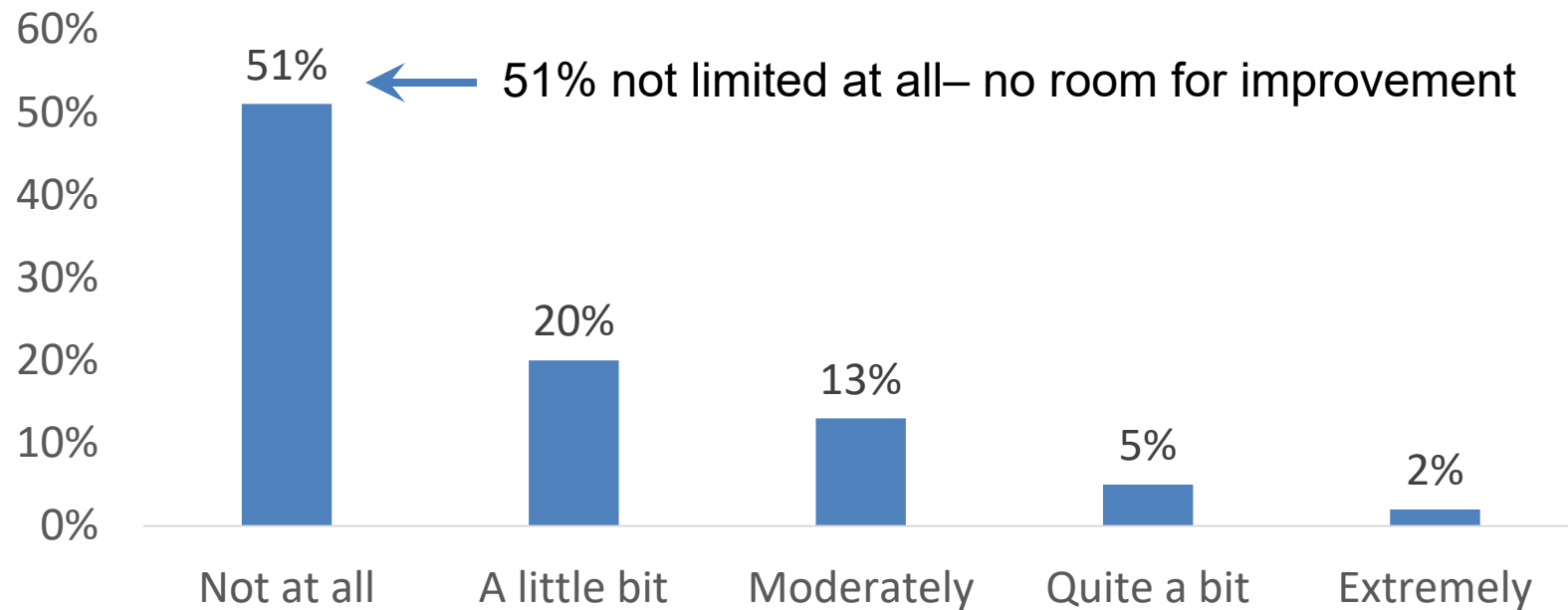
- Validation refers to the process of evaluating the psychometric properties of a measure
- A measure may be psychometrically robust in one domain or population but not another
- Common properties examined: variability, reliability, validity, interpretability

What makes a good participant-reported outcome measure?

- Good variability in responses or scores at baseline with room to improve
- Evidence of excellent reliability in the population under study
- Evidence that the measure produces data that are *valid* and also easy to *interpret*
- Sensitivity to change (for a prospective or interventional study)

Variability: appropriate representation of different response or score levels

“In the past four weeks, how much have vaginal symptoms made it uncomfortable to wear the clothing or underwear you want?”



Reliability: the extent to which a measure score is free from random error

- **Inter-rater reliability:** Are responses similar regardless of the person *who* administers the questionnaire?
- **Test-retest reliability:** Are responses similar *when* the measure is re-administered under the same conditions?
- **Internal consistency reliability:** How consistent responses are across items within a scale (Cronbach's alpha)?
 - Reliability coefficient– range of .00 to 1.00 – coefficient of .70 means that 30% of variance is due to error

Validity/interpretability: whether a measure reflects the construct it is designed to reflect

- **Face validity:** Does the measure “on the surface” or “on the face” seem like a good representation of the characteristic?
- **Construct validity:** Does the measure relate to other similar or different measures in expected ways?
- **Predictive validity:** Do scores accurately predict an outcome that is important or accepted?

Sensitivity to change

- Is a measure sensitive to true differences or changes in the characteristic being measured?
- Is it capable of capturing small differences in the characteristic that are still clinically important?
- “Minimal clinically important difference” (MCID) = the smallest difference in score that is clinically meaningful

Can you adapt an existing measure?

CONS:

- Potential change in meaning of items and overall scores
- Decreased ability to compare findings across studies

PROS:

- Still benefits from prior work to establish reliability and acceptability of specific items
- For measures with multiple subscales, one subscale can usually be administered separate from others

Adapting measures for diverse populations

Example: will Chinese American participants give different answers on a standard depression questionnaire due to:

- Differences in cultural norms about expressing negative feelings?
- Decreased comfort with being interviewed by strangers about emotional well-being?
- Nuances in the meaning of the closest term for “depression” in Chinese languages?

Decisions in administering measures

- Choosing paper-, telephone-, computer- or email-based administration
- Designing measures to be self-administered versus interviewer-administered
- Tradeoffs for ease of collecting, entering, and cleaning data (discussed by Michael Kohn)

Wrap-up