General Issues in Measurement

Armin Zareiyan-PhD

a.zareian@ajaums.ac.ir

"Many people do ***not take instrument development as seriously*** as they should, perhaps because they themselves have completed so many poor-quality scales that they think this is standard or perhaps because **they think it is easy to piece together a collection of questions."**

So It May Seem Simple.

It is not!!

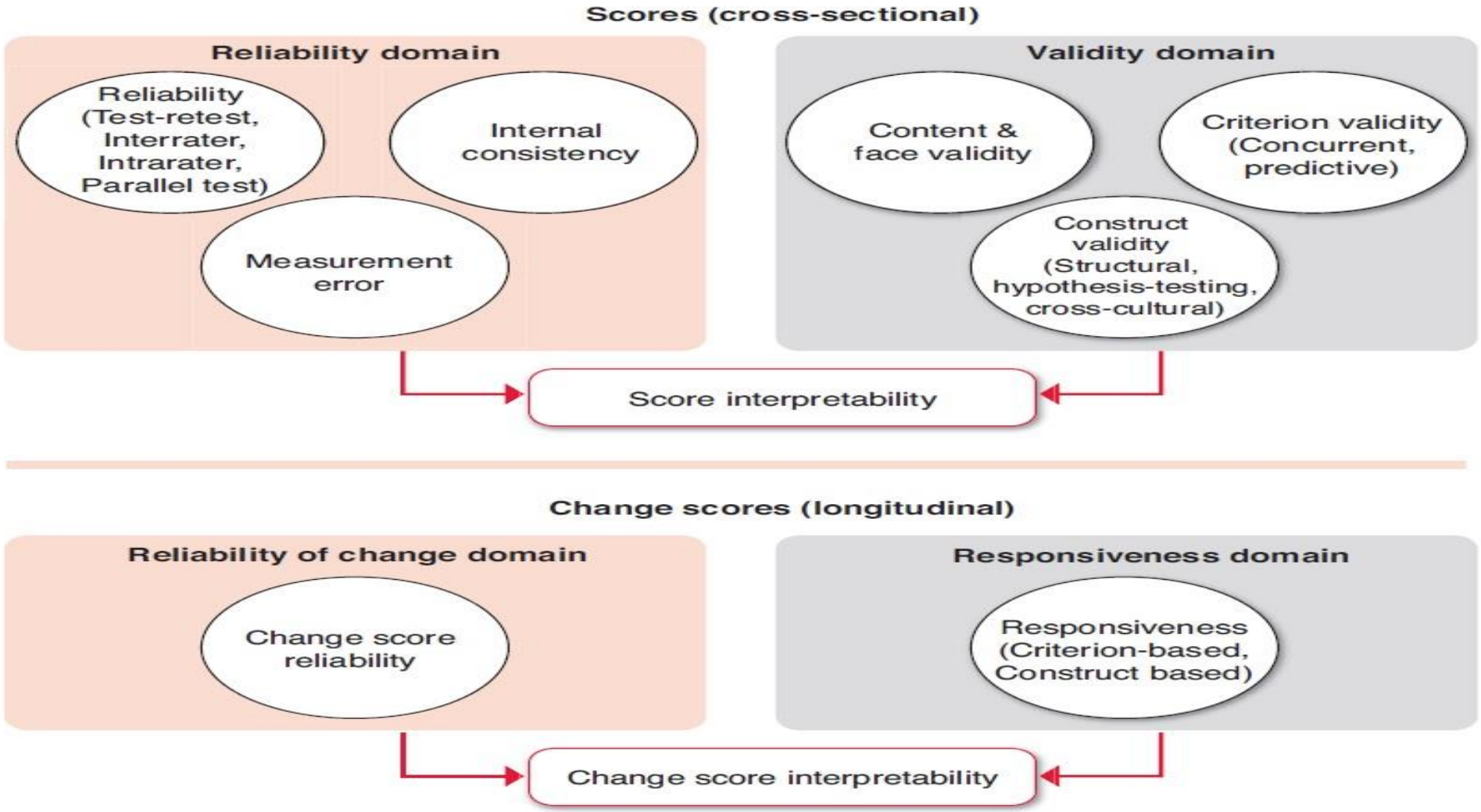
Polit & Yang,2016)

Measurement Terminology

- Psychometrics Properties
- Psychometrics Analysis
- Psychometrics Evaluation
- Validation

and… Clinimetrics

Taxonomy of Measurement Properties For an Instrument's Scores & Change Scores



Six Important Questions

1. Is the score of 80 at baseline the right score value for this patient? Is it possible that the score ***really*** should be **78?** Or **83?**
2. Is the scale truly measuring the construct *functional ability*? Or is it measuring something else?
3. What does a score of 80 *mean*? How can it be **interpreted?** Is it high or low for patients with spinal cord injury?
4. Does the change from 80 to 75 reflect a **true change**, or does it merely reflect a random fluctuation in measurement?
5. Does the change from 80 to 75 correspond to an improvement in the patient's degree of functional ability?
6. What does a 5-point change ***mean?*** Is the improvement large enough to be considered clinically significant?

Purposes of Measurement

- Discriminative:
  - Used to examine differences between individuals and groups
  - Used in cross-sectional studies
- Predictive:
  - Used to predict a health outcome
  - Screening and diagnostic instruments typically have primarily a predictive purpose
- Evaluative:
  - Used to assess the benefits and outcomes of a health treatment or clinical regimen
  - Such measures are useful in CLINICAL TRIALS and also in everyday clinical situations to monitor stability, improvements, and deterioration

Measurement Sources

- Measurement of health-related phenomena can be derived from many different sources.
- Least problematic measurements:
  - Biophysiologic equipment
  - Lab Analysis
- ***BUT***
- **Many important health constructs require different approaches.**

Measurement Sources…

Patient-Reported Outcomes (PROs or PROMs)

- PROs are often associated with the measurement of subjective states
  - (how patients ***feel***)
- Direct questioning can also be used to measure outcomes that would be more difficult or expensive to measure in other ways. such as:
  - Nutritional intake,
  - Sleep habits,
  - Smoking behavior,
  - Physical functioning
  - …

Measurement Sources…

Proxy-Reports (گزارش های وکالتی)

- For patients with cognitive or other impairments, **proxy reports** from caretakers or clinicians are sometimes used to measure patient outcomes.
- Some scales have been developed for either self-administration or proxy-administration
- for example:

World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0, 2010)

Measurement Sources…
Observation

- Observation Checklist
  - To indicate the *presence* or *absence* (or frequency of occurrence) of certain trait or behaviors
- Observational Rating Scale
  - Observer indicate the *intensity* of phenomena under scrutiny.
- یک مشاهده گر با استفاده از ابزار تحریکی-آرامبخشی ریچموند (Richmond Agitation-Sedation Scale) وضعیت بیمار تحت مراقبت های ویژه را در یک طیف ۱۰ نمره از امتیاز ۴+ (تحریک پذیر) تا صفر(*آرام و آگاه*) وتا ۵- (*غیر قابل تحریک پذیر*) در نظر می گیرد. (Sessler et al., 2002)

Health Measurements Variants

- Complexity
- Generality
- Adaptability

Measurement Complexity
Simple Measure

- Directly measure the construct and yield score value
  - Physiological: Thermometer or Visual Analog Scale (VAS)
- When it is important to **minimize error** it can be useful to use ***averaged values*** as the scores.
  - For example, ***clinical guidelines recommend that patients' blood pressure*** be measured two or more times on each visit and **then averaged** because single measurements have been found to misclassify a sizeable minority of patients into hypertensive categories (Handler, Zhao, & Egan, 2012).

Measurement Complexity
Composite Measure

- Several individual measurement to be combined: Self efficacy
  - Van der Ven and colleagues (2003) developed a composite scale(Confidence in Diabetes Self-Care Scale (CIDS)) with 20 items to measure self-efficacy relating to diabetes self-care.
- The 20 item scores are summed, and higher total scores reflect greater self-efficacy.

- مثلاً:

1. "من می توانم میزان انسولینم را در زمان مسافرت تنظیم نمایم".
2. "من می توانم میزان انسولینم را در زمان ورزش تنظیم نمایم".
3. "من می توانم میزان انسولینم را در زمان میهمانی تنظیم نمایم".

- **بیماران به هر سوال ۵ گزینه ای از امتیاز یک (نه؛ من مطمئنم که نمی توانم) تا امتیاز۵ (بله؛ من مطمئنم که می توانم) پاسخ می دهند.**

نکاتی مهم در خصوص پیچیدگی ابزار

ابزارهای چند گویه ای اغلب بر ابزارهای تک گویه ای ترجیح داده می شود.

- چه آنکه بسیاری از سازه ها آنقدر پیچیده هستند که تنها یک گویه نمی تواند طیف گسترده ای از اطلاعات مربوطه را اندازه گیری کند.
- ابزارهای تک آیتمی پایایی کمتری نسبت به ابزارهای چند آیتمی دارد:

چون وجود آیتم های متعدد خطای تصادفی را کاهش می دهد.

ابزار چند آیتمی تمایز بین افراد را بهتر نشان می دهد

نکاتی در خصوص نمره دهی ابزار چند آیتمی

- Sometimes, however, it is **attractive** to have differential weighting of items to reflect the items' contribution.
- اگرچه ممکن است دربعضی موقعیت ها وزن دهی سبب **بهبود روایی پیش بین** شود، لیکن گاهی ممکن است باعث پیچیدگی ابزار شده و به این ترتیب **احتمال خطا** را افزایش دهد.
- علاوه بر این، وزن دهی در ابزاری که برای یک جمعیت خاص ساخته شده، ممکن است در زمان استفاده بر روی جمعیت های متفاوت، **نامناسب** باشد.
- بنابراین وزن دهی یکسان سئوال ها، برای ابزارهای ترکیبی امری معمول و متداول است(پولیت و یانگ ۲۰۱۶)

Generic and Specific Measurement

- A **generic scale:**
- Is a measure of a construct that is broadly applicable across different clinical (and sometimes nonclinical) populations.
  - The ***SF-36 is a good example of a generic scale***.
- **Specific scales:**
- The most common are ***disease-specific scales*** that are designed for use with people who have a particular disease or condition.
  - For example, there are disease-specific quality of life scales for patients with stroke and aphasia, cancer, HIV/AIDS, and kidney disease.

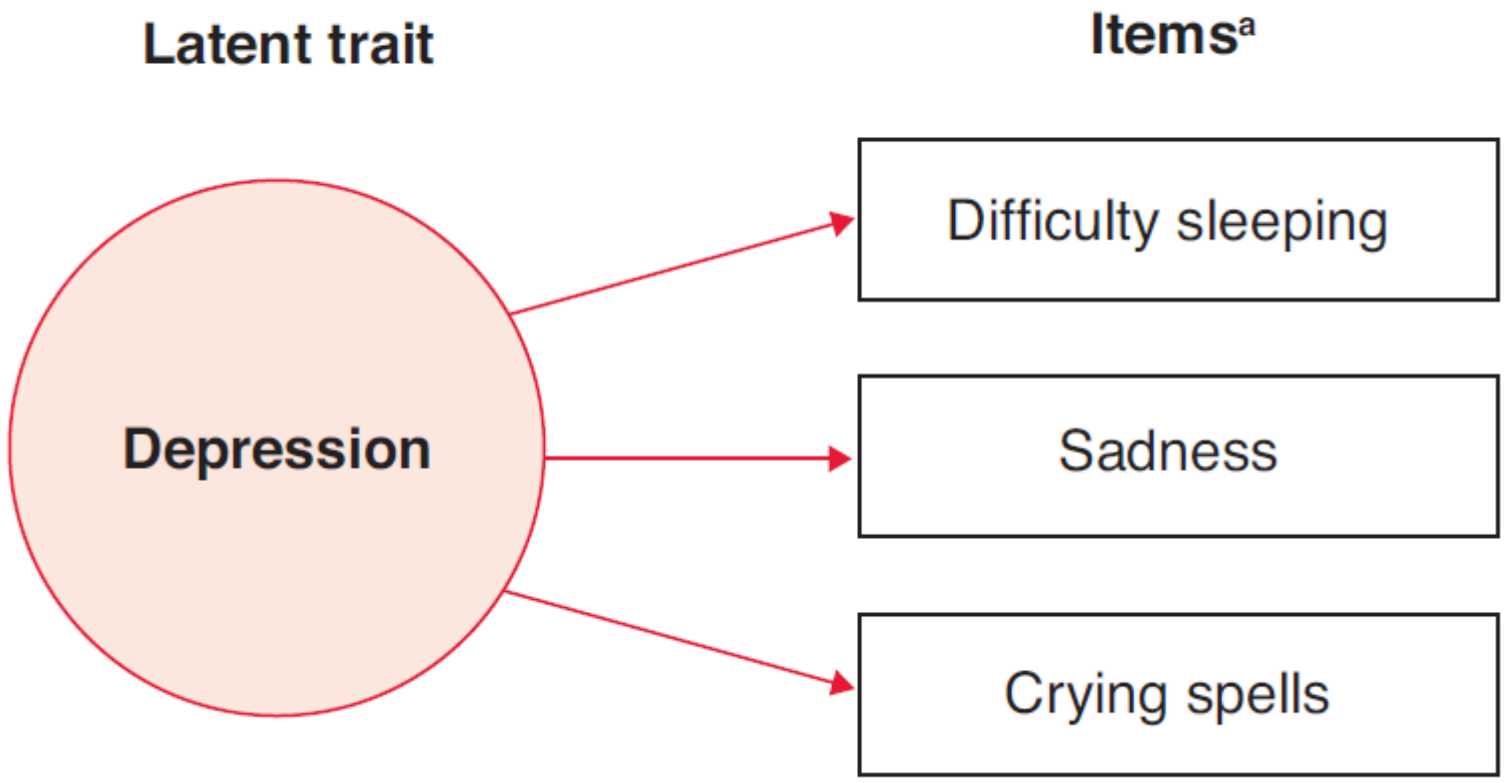
Static and Adaptive Measure

- A **static measure** is a fixed-length measure that is administered in a comparable manner for everyone who is measured.
- For static composite scales, people are asked to complete an **entire set of items** and then a **summary score** is developed on the basis of responses.
- An **adaptive measure** uses information from responses to early questions to guide the selection of subsequent questions.

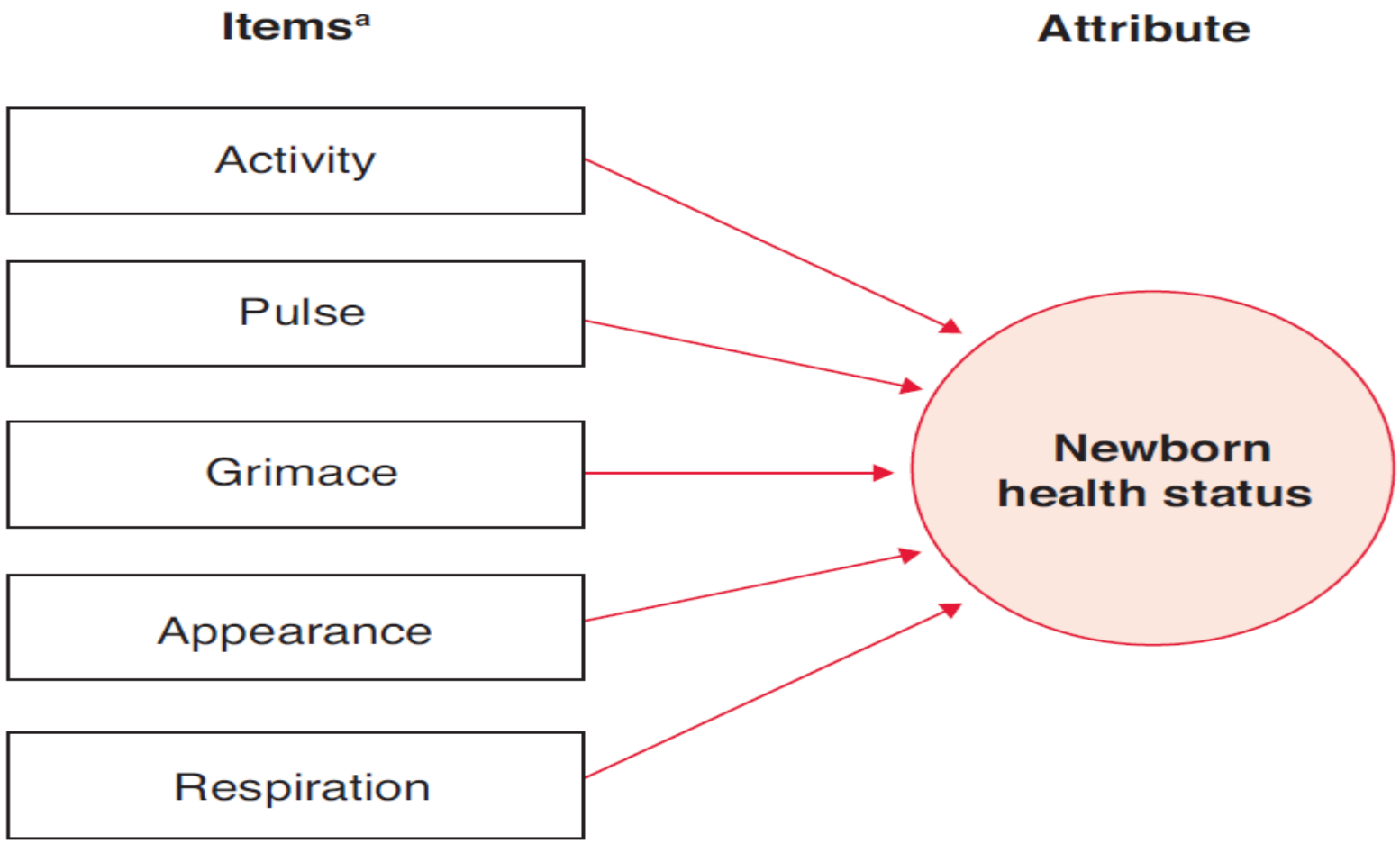
Reflective Scales and Formative Indexes

- Constructs are not directly observable ➔ Responses to Items
- Psychometric scales ➔ reflective scales because the items are viewed as *reflections* of the construct
- *The items on a reflective scale share a common cause ➔ fairly highly intercorrelated*
- **Formative Measures:**
  - A multi-item measure can be conceptualized as having items that "cause" or define the attribute (rather than being the effect of the attribute)

Reflective Scale to Measure Depressive Symptoms



Formative Index to Measure Newborn Health Status



Challenges in Scale Development

Decisions About Item Features

Number of Items

- There is ***no magic formula*** for the number of items that should be developed for a scale
- BUT: It is a good idea to generate a large pool of items for each dimension of the construct.
- Longer scales → More internally consistent, **BUT** Long scales → burdensome and noncompletion
- **Longer scales → Inappropriate in clinical settings**.
- DeVellis (2012) recommends starting with 3 to 4 times as many items as you think you will have in your final scale (30 to 40 items for a 10-item scale)
- BUT at a minimum, there should be 50% more (*15 items for an anticipated 10-item scale*)

Number of Response Options

- Traditional Likert scales use response options on a continuum of agreement, but other continua are possible, such as frequency (never/always) or importance (very important/totally unimportant).
- There is no simple answer to the question of how many response options there should be, BUT **variability is essential**.
- Most Likert scales have **five to seven** options.

Odd Versus Even Response Options

- When questions are not dichotomous, developers make a decision about having an odd or even number of response options.
  - یک **عدد فرد** به افراد این شانس را می دهد که **خنثی یا بینابینی** پاسخ دهند- یعنی انتخاب یک نقطه حد وسط.
  - برخی سازندگان مقیاس یک عدد زوج را برای جلوگیری از فرار افراد از پاسخ دهی (fence-sitting) ترجیح می دهند **(مثل ۴ یا ۶)**.
- The midpoint can be labeled with such phrases as:

1. "neither agree nor disagree"
2. "undecided"
3. "neutral"
4. or simply "?"

جعبه ۴-۱ انواع طیف‌های پاسخدهی (بدون لحاظ نقطه حد وسط)

- کاملاً مخالف، مخالف، موافق، کاملاً موافق
- کاملاً مخالف، بطور متوسط مخالف، اندکی مخالف، اندکی موافق، بطور متوسط موافق، کاملاً موافق
- هرگز، تقریباً هرگز (یا بندرت)، گاهی اوقات (یا گاه به گاه) غالباً (یا مکرراً)، تقریباً همیشه (یا همیشه)
- هرگز، یک بار، دو بار، ۳-۴ بار، ۵ بار و بیشتر
- هیچ زمانی، در زمان اندکی، در برخی زمان‌ها، در زمان نسبتاً مناسبی، در بیشتر زمان‌ها، در همه زمان‌ها
- کاملاً مهم، مهم، تا اندازه‌ای مهم، اندکی مهم، غیر مهم، کاملاً غیر مهم
- قطعاً خیر، احتمالاً خیر، ممکنه (امکان دارد)، احتمالاً، خیلی محتمل، قطعاً بله
- ابداً، تا حدی، متوسط، نسبتاً زیاد، شدیداً
- هیچ، خیلی خفیف، خفیف، متوسط، شدید، کاملاً شدید
- بدون ناراحتی، ناراحتی اندک، تا اندازه‌ای ناراحت، ناراحتی به میزان زیاد، ناتوان در انجام کار

Positive and Negative Stems

- **A generation ago**, psychometricians advised scale developers to include both positively and negatively worded items and to reverse-score the negative ones.
- As an example, consider two items for a depression scale:

1. I frequently feel depressed
2. I don't feel sad very often

Positive and Negative Stems…

- Many experts **currently** advise against including negative and positive items on a scale because some people are confused by reversing polarities-especially if there are negative words in the response options (***never***).
- شواهد زیادی وجود دارد که وارد نمودن هر دو نوع از گویه ها (مثبت و منفی) در یک ابزار می تواند منجر به **"ابعاد ساختگی"** در تحلیل عاملی شود (Hankins, 2008; Marsh, 1996; Motl & DiStefano, 2002)
- Answering negative item is difficult cognitive task.

In general, negatively worded items should be avoided

Wording of the Items

1. ***Clarity:*** Scale developers should strive for items that are clear and unambiguous.
2. ***Jargon:*** Jargon should be avoided (not familiar to the average person)
3. ***Length:*** Avoid long sentences or phrases(eliminate unnecessary words)
4. ***Double negatives:*** It is preferable to word positively (I am usually happy) than negatively (I am not usually sad), but double negatives should always be avoided (I am *not* usually *un*happy).
5. ***Double-barreled items:*** Avoid putting two or more ideas in a single item.
6. ***Bias:*** Items should not suggest a "right" answer.

Floor and Ceiling Effects

- Variability in scores on an attribute is restricted either at the lower end of a continuum (**floor effects**) or at the upper end (**ceiling effects**).
- Avoiding floor and ceiling effects also requires careful thought about how and with whom the scale will be used. A scale may have no ceiling effects when administered to adults, for example, and yet be too "difficult" for adolescents.
- Th e issue of floor and ceiling effects has been an impetus for the development of specific rather than generic scales.

Acceptable range of Floor and Ceiling Effects: **15%**

The Challenge of Missing Values

- Missing answers could signal problems of interpretation or comprehension, discomfort with the question, inapplicability of the item.
- Th e first line of defense:
  - looking at the frequency of missing values for each item and by using cognitive interviews to identify why people might struggle with their answers to certain questions.
- It has been suggested that the maximum number of missing items for a given case should not exceed **20%** (Downey & King, 1998)
- **Handling Missing value:**
- For brief overviews, see Donders, Van der Heijden, Stijnen, & Moons, 2006; Fox-Wasylyshyn &El-Masri, 2005; Haukoos & Newgard, 2007

Telescoping

- به این مفهوم که افراد وقایع را نزدیک‌تر از آنچه هست، به یاد می‌آورند.
- توصیه می‌شود زمان ***دو هفته تا یک ماه***، برای پرسشگری پیرامون رفتارها در نظر گرفته شود.

سؤالات زمینه ای و دموگرافیک

- دقت نمایید که سؤالات دموگرافیک در صورتی پرسیده شود که در تحلیل نهایی مورد استفاده قرار گیرند. چرا که افزودن سؤال اضافی نه تنها سبب خستگی پاسخگو می‌شود بلکه می‌تواند در وی ایجاد حساسیت نماید و از خود بپرسد «چرا این سؤالات را از من می‌پرسند؟» و حتی نزد خود چنین نتیجه گیری نماید که «ممکن است از روی سوالات زمینه ای من را شناسایی کنند».

شرایط اجرای آزمون:

مدت اجرا

- بستگی به تعداد گویه ها و میزان دشواری درک مفهوم گویه ها دارد.
- **قانون سرانگشتی:**

به ازای هر گویه ۳۰ ثانیه در نظر گرفته شود