

EVERYDAY PSYCHOMETRICS

PAIN AND DISABILITY ACROSS THE LIFESPAN
OCTOBER 7-8, 2009
MINNEAPOLIS

Lynn Breau, Ph.D., Registered Psychologist

Departments of Nursing, Pediatrics & Psychology, Dalhousie University;
Complex Pain Team & Centre for Pediatric Pain Research, IWK Health Centre;
Halifax, Nova Scotia, Canada

DEFINING A “MEASUREMENT TOOL”

- ◎ To be considered a “measure” a checklist / scale / tool ***MUST***:
 - ◎ Use numbers to describe something.
 - ◎ Have the same items for every time it is used.
 - ◎ Scales that have “fill in the blanks” cannot be summed because essentially a “new” scale is used for each person. This means ***any statistical analyses of them are invalid.***

THE BASIC PROPERTIES OF A GOOD OBSERVATIONAL TOOL

- ⊙ Validity

- ⊙ Does it measure what I think it does?

- ⊙ Reliability

- ⊙ Does it measure the same way each time?

QUALITY CONTROL FOR ASSESSMENT TOOLS

- ◎ If the tools you use are not accurate and reliable, they will not give you a true picture of what you are trying to assess.
 - ◎ You may not know they aren't working,
 - ◎ You should always evaluate how a tool was developed before using it clinically,
- ◎ Lab tests, equipment; we expect these must be calibrated,
- ◎ Surveys/questionnaires/checklists are no different.

PRACTICALITY VERSUS PSYCHOMETRICS

- ◎ To be useful clinically, a tool must be as ***practical as possible***, easy to understand, use, add up.
- ◎ But, if the tool is designed for practicality first, it may not be valid or reliable.
- ◎ Many tools that appear “user-friendly” have weaknesses that make them invalid:
 - ◎ Too short, items that were important were missed
 - ◎ Item wording is vague, unclear, can be interpreted several ways, items have overlapping components

WORDING ITEMS

- ⊙ Bad: Nurse judges child to be uncooperative.
 - Relies on observer to decide, no guidance.
- ⊙ *Better: Child will not follow nurse directions.*

- ⊙ Bad: Child is sad.
 - Relies on judgement of non-observable behaviours.
- ⊙ *Better: Child has downturned mouth.*

- ⊙ Child is irritable and crying.
 - Includes 2 behaviours.
- ⊙ Better: Child is complaining, will not follow directions (Item 1); Child is crying (Item 2).

UNDERSTANDING YOUR CONSTRUCT

- ◎ Researchers too often start developing a scale when they do not fully understand their construct (e.g. pain versus behavioural reflections of pain). This can be done, but:
 - ◎ They **must** use a large number of items to capture **all possible nuances** of construct. This is because few developers will ever go back to add items after a scale has been tested.
 - ◎ They **must** be willing to omit items after initial studies show they are not valid
 - ◎ You **may** include “write-in” responses to allow for aspects you did not predict, but this is part of development.

UNDERSTANDING YOUR CONSTRUCT

- ⊙ Researchers too often start to develop a scale, thinking they know what their construct looks like, but:
 - ⊙ Omit items because they believed they were not critically important and they want to limit scale length
 - ⊙ Merge two aspects of a construct into one item to limit length
 - ⊙ Base item-selection on the current sample they intend to study, neglecting possible items of relevance to the construct as it occurs in the target population
 - ⊙ Believe that if the psychometrics that result from a first study are weak that the scale is the problem, instead their sense of the construct is wrong or their study was at fault.

VALIDITY

- ◎ **Face validity**
- ◎ **Content validity**
 - ◎ Internal consistency (reliability)
- ◎ **Construct validity**
 - ◎ Criterion validity
 - ◎ Concurrent validity
 - ◎ Discriminative validity
 - ◎ Predictive utility

VALIDITY

◎ **Face validity**

- ◎ Do others think the scale asks the right questions?
 - ◎ Do the questions seem appropriate?
 - ◎ Can they be answered?
 - ◎ Is the format easy to follow?
- ◎ These issues may not be under your control if you are using an established questionnaire. But , they should be considered when choosing a questionnaire.

VALIDITY

⊙ Content validity versus Construct validity

⊙ Content:

- Are all the parts that are there, supposed to be there?

⊙ Construct:

- Does the thing look like I think it should?
- Does it have the characteristics I expected?
- Are the parts in the right places?
- Is it related to other things the way I expect?
- Does it act how I expected it to act?

DIAGRAM LEGEND

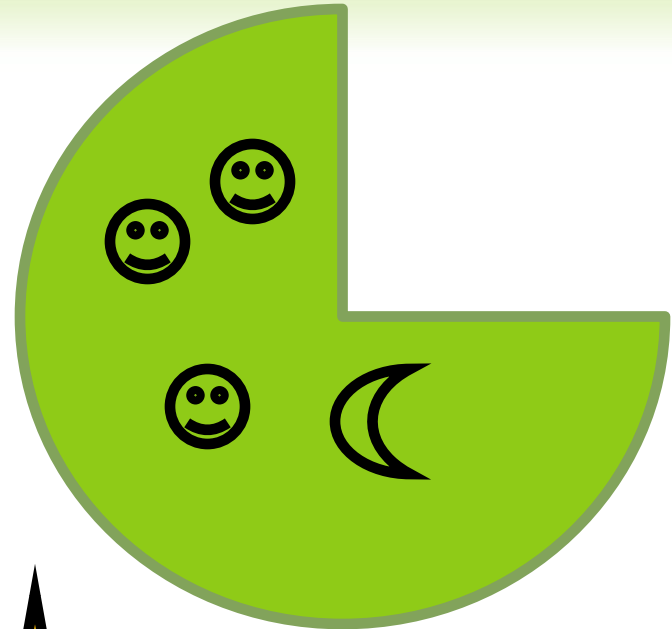
- ◎ The following diagrams are an attempt to visualize the different types of validity.
- ◎ In each diagram, the objects on the left reflect what was expected. That is, the construct the tool is attempting to measure.
- ◎ The objects on the left reflect what was actually found. That is, the construct as it exists in the real world.
- ◎ The green shapes represent the “construct”, such as pain or depression.
- ◎ The star and square reflect other constructs that may be similar to the one you are interested in. How close they are to your construct reflects how strong the relation between them is. For example, perhaps you want to measure pain, the star reflects distress, and the square reflects disability.
- ◎ The happy faces within the shapes reflect the parts of the construct. Their placement reflects their relationship to each other. On a tool, these would be items.

PROBLEMS WITH CONTENT VALIDITY

You expected...



You found...



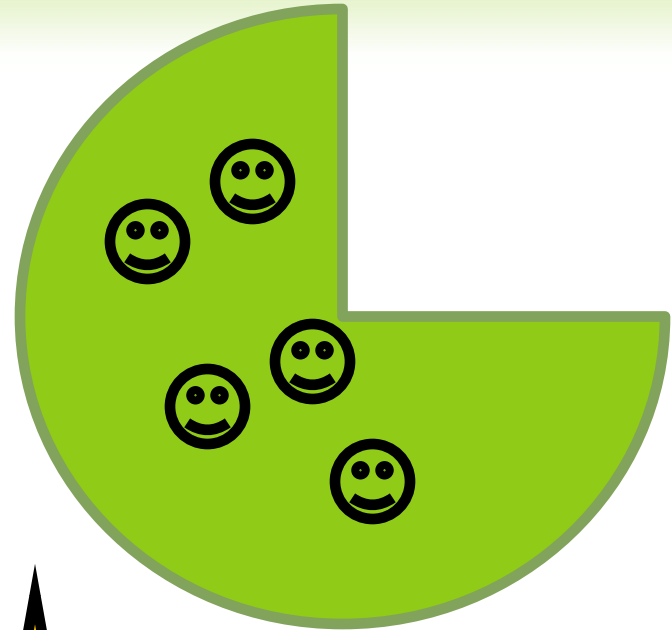
A moon was found, which was not expected. Must be removed.

PROBLEMS WITH CONTENT VALIDITY

You expected...



You found...



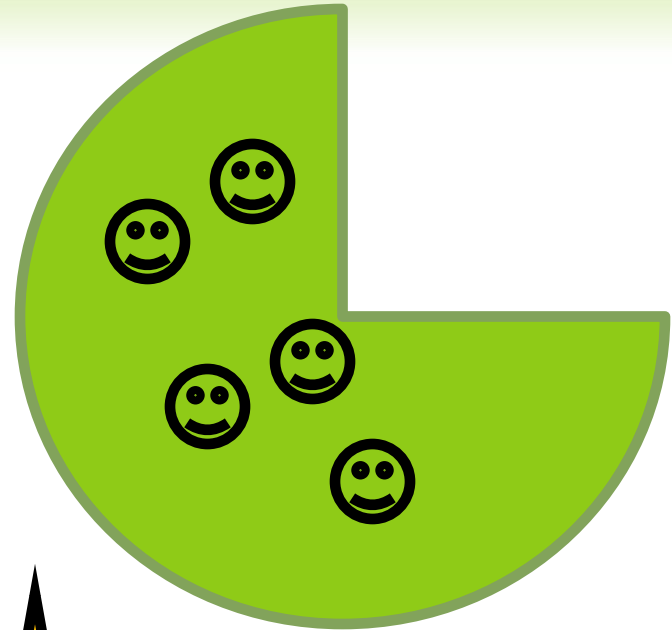
A part of the construct was not expected. Your tool was missing items that are part of the construct. ***Caution: This cannot be fixed without adding items, something researchers rarely do!***

PROBLEMS WITH CONTENT VALIDITY

You expected...



You found...



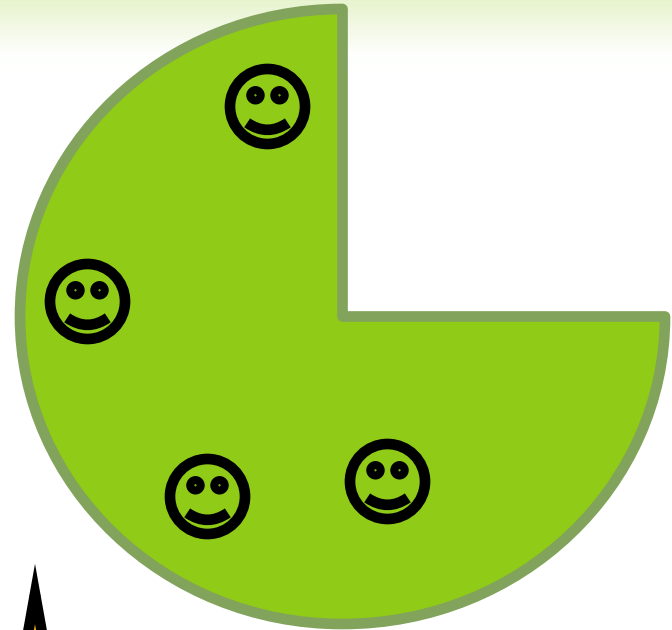
The construct contains more than expected. Your tool had extra items that are not part of the construct. This can be fixed by removing items.

PROBLEMS WITH CONSTRUCT VALIDITY

You expected...



You found...



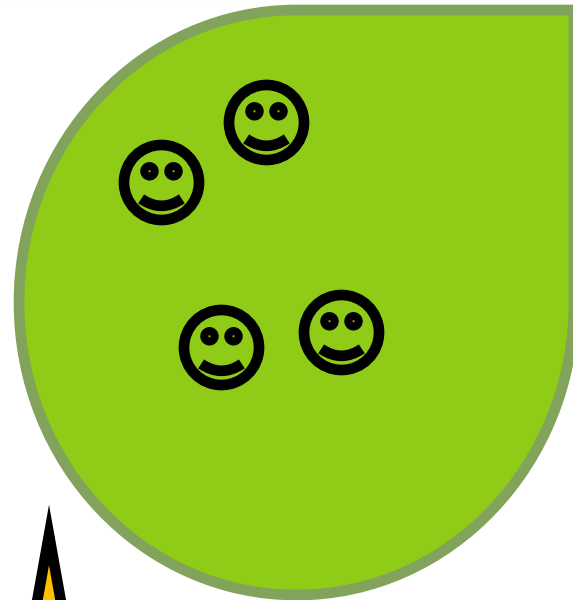
Relationships among items (distance) was not as expected. E.g. Crying was not related as closely to tears as expected during pain. Need to redefine construct or change items to capture relationship better.

PROBLEMS WITH CONSTRUCT VALIDITY

You expected...



You found...



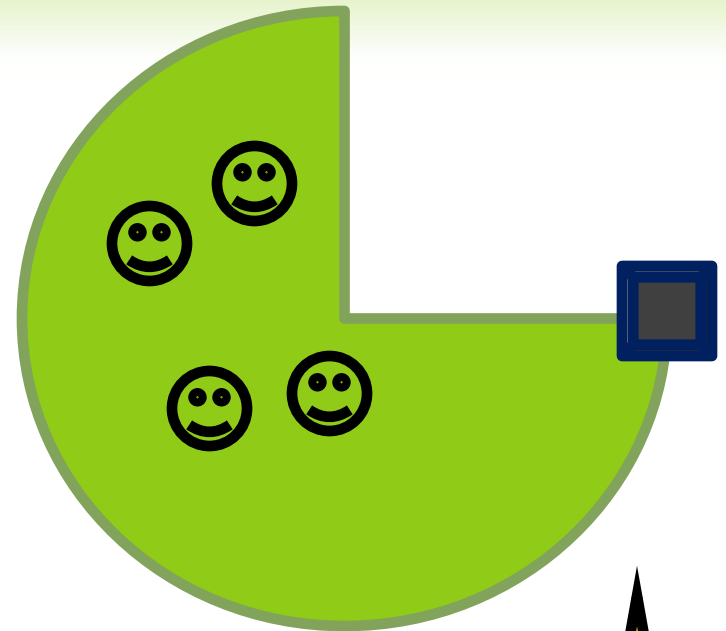
The shape of the actual construct was not as expected. Are you sure you are measuring what you thought? Are you measuring distress, not pain?

PROBLEMS WITH CONSTRUCT VALIDITY

You expected...



You found...



The relationship between your construct of interest and another construct was not as expected. They are more/less closely related than thought; one overlaps with your construct (poor discrimination).
E.g. Some parts of your tool are measuring distress, not pain.

LOOKING MORE CLOSELY AT THE TYPES OF VALIDITY AND RELIABILITY

CONTENT VALIDITY

- ⊙ Internal reliability (also called “internal consistency”)
 - Do the items belong on the scale?
- ⊙ Cronbach’s alpha (average correlation between items)
- ⊙ corrected item-total correlations
 - If $< .20$, item is not related to others
 - If $> .80$ item may be redundant
- ⊙ If the scale has subscales; compute for subscales
- ⊙ More items = higher alpha possible
- ⊙ This means:
 - alphas should be higher for subscales than a scale total
 - Alphas can be higher for a long scale than a short one

CONSTRUCT VALIDITY

⊙ **Criterion validity:**

- How strongly is the scale related to the best other measure of the same thing? (Gold standard). Only possible if gold standard is available.
- Correlations, measures of association

⊙ **Concurrent validity:**

- How strongly is the scale related to other measures of the same thing? These may or may not be behavioural and may not be “gold standards”.
- Correlations, measures of association

CONSTRUCT VALIDITY

◎ Discriminative validity (Sensitivity and Specificity)

- How weakly is the scale related to measures of other constructs.
- How well can the scale tell when the thing is present/absent?
- Develop cut-off scores and assess sensitivity / specificity (should be above .80)
- Compare means (t-tests, ANOVA's, etc)

◎ Predictive utility:

- If I administer the scale now, how well does it predict the thing later?
- Correlation, multiple regression, logistic regression, odds ratios, etc
- Must have two time points

RELIABILITY

⊙ Internal reliability (also called “internal consistency”)

⊙ How cohesive are the items?

- Cronbach’s alpha (average correlation between items)
 - ⊙ corrected item-total correlations
 - If $< .20$, item is not related to others
 - If $> .80$ item may be redundant
- If the scale has subscales; compute for subscales
- More items = higher alpha

⊙ Inter-rater reliability

- ⊙ Do two (or more) people provide the same scores when rating/reporting about the same event/construct?
 - Kappa; Intra-class correlation coefficient; correct for chance-agreement
 - $K = .49 - .59$ (moderate); $K = .60 - .79$ (substantial); $K = .80$ (exceptional). (Landis & Koch, 1977)

RELIABILITY

⊙ Intra-rater reliability

- ⊙ Does the same person answer the same way across administrations?
- ⊙ “drift”; gradual change in ratings away from standard
 - Matched sample t-test (total scores)
 - Repeated measures analysis of variance (subscale scores or items)

⊙ Test-retest reliability

- ⊙ Does the score stay the same when there is no change in the condition/construct being measured?
 - Matched sample t-test (total scores)
 - Repeated measures analysis of variance (subscale scores or items).

HOW RELIABLE?

- ◎ Weiner & Stewart (1984): .85
- ◎ Kelley (1927): .94
- ◎ Streiner & Norman (1995)
 - ◎ Lower levels acceptable for research/groups than for tools used clinically to make decisions about care for an individual.
- ◎ Lynn says: .80 reasonable in most cases according to most sources

SAMPLES

- ⊙ Reported psychometrics only apply to new samples that are **the same as** the original sample, in the same setting, in the same manner.
- ⊙ A tool designed for use in hospital cannot be used at home and vice versa
- ⊙ A tool designed to be used by nurses cannot be used by parents / family / care workers and vice versa
- ⊙ A tool designed for people who speak cannot be used for people who cannot speak and vice versa

MISSING DATA, SUBGROUPS AND OTHER RED FLAGS

- ◎ Some studies report statistical tests for a full sample and then tell you they conducted a statistical test on a subgroup.
- ◎ Caution:
 - ◎ Ask yourself why? Was the data too hard to collect (poor feasibility of the tool) for some members?
 - ◎ Ask yourself whether the subgroup is different in any way from the full sample.
 - If cut-off scores are only conducted on a group receiving medication for pain in an inpatient study, that means patients with mild pain, or pain judged as not requiring medication by the physician.
 - This means the cut-off scores only apply to patients with more severe pain. The tool should not be used with patients who have mild pain or pain not judged as needing treatment by a physician.

TAKE HOME MESSAGE

- ◎ Psychometric properties are sample and situation specific.
- ◎ The generalizability of a scale's psychometrics must be shown over multiple studies, multiple settings, multiple situations if it is to be used as a general tool.
- ◎ Practicality follows after psychometrics. A hammer that is easy to use isn't really practical if it hasn't the weight to hammer in the nail!
- ◎ Be cautious of scales that are simplistic. ***Human behaviour is not!***