Highlights from Week #6

- Gave definitions and provided examples of measurement, phenomena, domains, concepts/conceptual models/conceptual frameworks
- Introduced classical measurement theory (CTT)
  - Discussed considerable limitations of CTT
- Gave examples of modern measurement theories (Generalizability Theory, Item Response Theory)

Highlights from Week #6

Validity

- Discussed traditional definitions of validity of instruments (three C's: content, criterion and construct)
- Described the meaning of measurement validity and recommendations for its estimation using quantitative methods
- Introduced consequential validity and discuss its importance for health research
  - No such thing as a valid instrument but rather an instrument ***valid for a given situation***
- Described the meaning of measurement validity and recommendations for its estimation
  1. Use empirical evidence
  2. An ongoing process
  3. Is not an all-or-none property of a measurement method
- Review the relationship between reliability and validity
  - All "valid" measures ***are reliable*** —but not all reliable measures are valid
  - Reliability is a necessary but ***NOT sufficient*** condition for validity

3

UCSF

Highlights from Week #6

- **Quantitative approaches to validity**
  - Content validity
    - Content validity index (CVI)
    - Index of Item-Objective Congruence (IOC)
  - Criterion-related validity
    - None
  - Construct validity
    - Multitrait-Multimethod Matrix (MTMM)
    - Factor analysis

4

UCSF

Highlights from Week #6

Reliability

- **Define main types of reliability measurements and associated statistical tests:**
  1. Internal consistency
     - Coefficient alpha (Cronbach's alpha, KR20 and KR21)
  2. Parallel/ alternate forms
     - Coefficient of equivalence and stability (Pearson correlation, Spearman's rho)
  3. Inter-rater agreement
     - Cohen's kappa
     - ICC
  4. Test-retest
     - Norm-referenced measures: Pearson correlation, Spearman's rho)
     - Criterion-referenced measures: Cohen's kappa

5 UCSF

Learning objectives

- **Theory:**
  - Review statistical underpinnings of scale measurement
  - Describe principles of instrument development
  - Define steps in designing new measures (norm- and criterion-referenced)
- **Practice:**
  - Methods of test development I: 1) Item analysis (based on the Classical Test Theory (CTT)); 2) Item response theory (IRT); 3) From items to scales: a) Item weighing and other methods; b) ROC and AUC; 4) Factor Analysis/ Principal Component Analysis.
  - Methods of test development II: Instrument design, construction and interpretation in cross-cultural situations

6 UCSF

Review

Statistical underpinnings of scale measurement

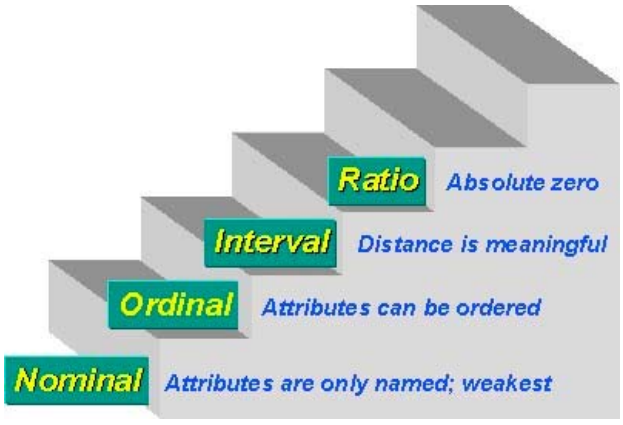
Types of Scales

- **Ratio Scale**
  - Any scale of measurement possessing magnitude, equal intervals and an absolute zero
  - Examples: Weight, Height
- **Interval Scale**
  - Any scale of measurement possessing magnitude and equal intervals, but not an absolute zero
  - Example: Temperature
- **Ordinal Scale**
  - Any scale that reflects only magnitude but does not contain equa intervals or an absolute zero
  - Example: Ratings or rank ordering
- **Nominal Scale**
  - Any scale that contains no magnitude
  - Examples: Classifications or Descriptions

7 Smith Stevens 1951

UCSF

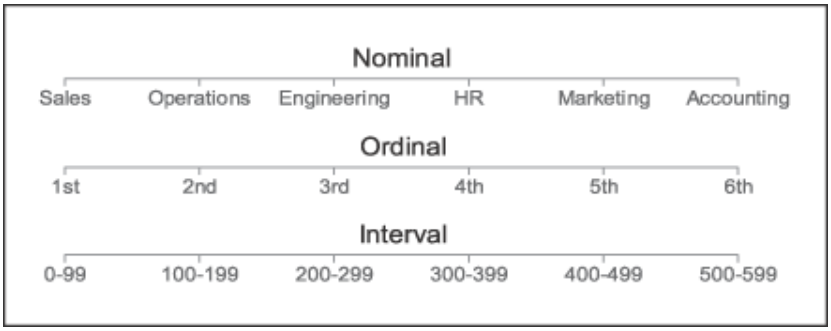
Types of scales



8

UCSF

Scale examples



UCSF

Relationship between Measurement and Statistics

Table 1. Examples of statistical measures appropriate to measurements made on various types of scales. The scale type is defined by the manner in which scale numbers can be transformed without the loss of empirical information. The statistical measures listed are those that remain invariant, as regards either value or reference, under the transformations allowed by the scale type.

Scale type	Measures of location	Dispersion	Association or correlation	Significance tests
Nominal	Mode	Information H	Information transmitted T	Chi square Fisher's exact test
Ordinal	Median	Percentiles	Rank correlation	Sign test Run test
Interval	Arithmetic mean	Standard deviation Average deviation	Product-moment correlation Correlation ratio	t test F test
Ratio	Geometric mean Harmonic mean	Percent variation Decilog dispersion		

 Stevens 1968

UCSF

Common measurement approaches: Scales

- A measuring tool or method composed of
  - A stem
  - A series of scale steps
  - Anchors that define the scale steps
- **Example:**

Indicate your degree of agreement with the following statement:

STEM: Noncompliance on the part of the patient indicates a need for additional attention to be directed toward the quality of care received by the patient.

SCALE STEPS: /___/___/___/___/___/___/___/___/___/___/___/___/

1 2 3 4 5 6 7 8 9 10 11 12

ANCHORS:

Completely Disagree — Completely Agree

UCSF

Scaling approaches

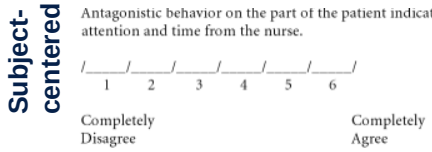
- Subject-centered
- Stimulus-centered (e.g., quality of service offered by different clinical departments)
- Response-centered (e.g., to assess the relative standings of individuals)
- **Types of scaling approaches:**
  - Summated ratings scales
    - Likert scaling (1932) to measure opinions, beliefs and attitudes
    - Semantic differential scales
  - Guttman scaling

UCSF

Scale variations

- **Summated rating scale** – scores for all scales in a set are summed or averaged to place an individual somewhere on a continuum of the attitude in question - Rensis Likert (1932)

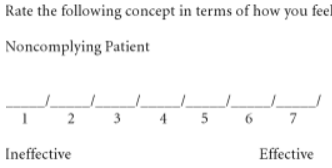
Subject-centered



1. a scale must contain multiple items
2. measure something that has a quantitative measurement continuum
3. each item has no "right" answer
4. each item in a scale is a statement

- **Semantic differential scale** – a specific concept is rated on a scale with bipolar adjectives anchoring the scale

Response-centered



used for measuring the *meaning* of things and concepts

UCSF

Example of Guttman Scale

- Place a checkmark against all statements with which you agree:
  - I like eating out
  - I like going to restaurants
  - I like going to themed restaurants
  - I like going to Chinese restaurants
  - I like going to Beijing-style Chinese restaurants

UCSF

Visual analog scales

- To measure intensity, strength, or magnitude of sensations and subjective feelings
- Employ a drawn, printed or computer-generated straight line of a specific length, with verbal anchors
- Subjects put a mark on the self-report of intensity, perception or sensation

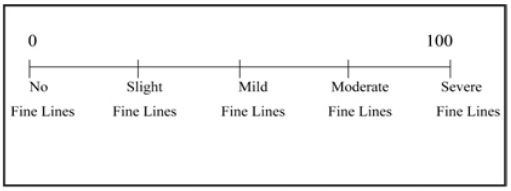
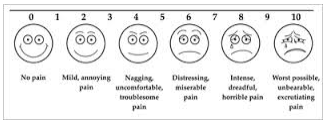


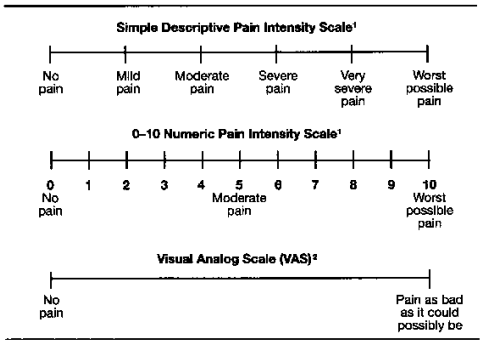
Figure 1. The modified 100mm Visual Analog Scale used by the grader.

UCSF

Visual analog scales
Example: Pain scales



- Reliability could be assessed by test-retest, but would not be very accurate because phenomena are very dynamic
- Validity could be assessed by comparing with other instruments (convergent validity)



¹If used as a graphic rating scale, a 10 cm baseline is recommended.
²A 10-cm baseline is recommended for VAS scales.

UCSF

Magnitude estimation scaling

- Subjects rate the magnitude of a stimulus relative to one of standard strength

- Not easy to use and requires abstract and complex thinking
- Reliability assessed by test-retest
- Validity is assessed by comparing MES with direct physical measurements of stimuli

17 UCSF

Review
Field methods

➢**Procedures for developing questionnaires**

- o Protocols for interview administration
- o Operations Manuals
  - o Staff training and re-training, certification
  - o Quality control measures
  - o Assessment of progress
- o Problems with recruitment/retention

18 UCSF

Principles of instrument development (#1)

- **Identify Research Aims/Questions**
  - What are the goals of your study?
  - Are you exploring:
    - Patterns over time?
    - Within subject changes?
    - Associations or Predictions?
  - Are your measures addressing these aims?
- **Identify Conclusions, Discussion Points, & Implications**
  - What do you hope to conclude?
  - Will your instrument and measures provide you with the information to make such conclusions?
- **Know Your Research Design**
  - Quantitative, qualitative, mixed-methods?
  - Cross-sectional?
  - Repeated Measures?
  - Longitudinal?
    - Interval between measurement waves?

Principles of instrument development (#2)

- **Know Your Research Subjects**
  - Children? Adolescents? Young adults?
  - Age comparisons?
  - Clinic sample?
  - At-risk sample?
  - School-based sample?
  - Reading level?
- **Identify & Define Constructs**
  - What are you trying to measure?
  - Do constructs address research aims/questions?
  - Can respondents answer questions?
  - Will venue allow you to ask these questions?
  - Are there existing measures for constructs? OR Need to develop new measures?

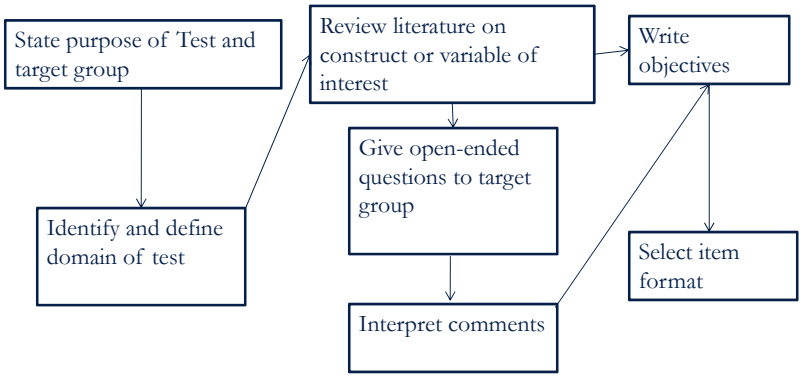
Principles of instrument development (#3)

- **Consider the Length of the Instrument**
  - Depends on the sample
  - Depends on the administration method
  - Depends on time and level of permission given

- **Consider the Instrument Format**
  - Paper/pencil
  - Web-based
  - ACASI
  - IPad

UCSF

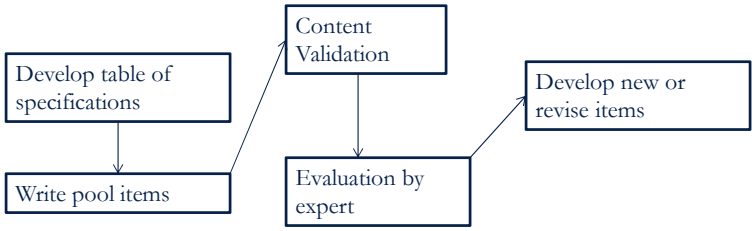
Instrument development: Phase 1: Planning



Benson et 1982.
A guide for instrument development and validation. AJOT36, 789-800

UCSF

Instrument development: Phase 2: Construction



UCSF

Respondents are able to respond

- **Items are specific, not ambiguous.**
- **Respondents know and remember information.**
- **Minimize burden of recall.**
  - If recall is needed, help respondent:
  - Place events or behavior in time
  - Use memory aids
- **Examples:**
- **Factual Data:**
  - SES, Parent Education, Income, Medical Data.
  - Can ask the source (e.g., parents, teachers, chart reviews)
- **Event-Specific Data:**
  - Since high school graduation…
  - Since we last surveyed you…
  - Provide calendars…

UCSF

Respondents are willing to respond

- Items should be written such that respondent does not feel the need to respond inaccurately.
- Assure confidentiality.
- **Handling "Don't Know" and "Not Sure" Responses:**
  - Can provide a screening question
  - Can provide an option for N/A, None, Don't Know
  - Don't assume or judge

Review
Measurement frameworks

- **Norm-referenced – evaluate a subject relative to other subjects in a well-defined norm or comparison group**
  - Eg., The Stress of Discharge Assessment Tool (SDAT-2) for patients with acute MI discharged from the hospital. Patient's score is compared to the scores obtained by other patients who responded to the same tool.
- **Criterion-referenced – determine whether a subject has acquired a predetermined set of characteristics or behaviors**
  - Eg., standards defined by the Guidelines for the Management of Patients with Chronic Stable Angina, by the AHA
  - Distribution has less variance and is typically skewed
- **Could be quantitative or qualitative**

Steps in designing a norm-referenced measure

How does the respondent compares to others?

1. Determine a phenomenon of interest
2. Select a conceptual model for delineating the nursing or health aspects of the measurement process
3. Explicitly define objectives of the measure (purpose?)
4. Develop a blueprint of the measure
5. Construct a measure, including:
   - Administration procedures (Operation's Manual)
   - An item set (questionnaire)
   - Scoring rules and procedures

3. Explicitly define objectives of the measure (purpose?)

What is taxonomy?

- A useful mechanism for defining a critical behavior to be assessed by an objective in such a manner that all who use the same taxonomy or classifying scheme will assess the same behavior in the same way, thus increasing ***reliability*** and ***validity*** of measurement.
- Taxonomy is specific to various domains

4. Develop a blueprint of the measure

- Show the constructed instrument to content experts to have them judge the following:
  - Does the instrument capture all domains of the conceptual framework?
  - Does the instrument solicits the same information from all subjects?
  - How well does the blueprint fit the measure objectives?

5. Construct a measure, including: Administration procedures

Operation's Manual, Questionnaire, Rules and Procedures

- WHO Drug Injecting Study Phase II
  - WHO Drug Injecting Study Operations Manual
  - WHO Drug Injecting Study Questionnaire
  - WHO Drug Injection Study Master Codebook

Steps in the development of criterion-referenced measure

What skills or abilities the respondent has?

1. Determine a phenomenon of interest
2. Clearly define the content of the domain tested
3. Include a relatively homogeneous collection of items or tasks that accurately assess the content domain
4. Determine criteria or performance standards (test specifications)

Methods of test development I:

Item analysis/ item response theory

Methods of test development I:
1) Item analysis (based on the CTT)

- The examination of any statistical property of examinees' responses to an individual test item.
- Used to assess the quality of these items and of the test as a whole
- Three general categories for item analysis
  - Indices that describe the distribution of responses to a singular item (mean, SD)
  - Indices that describe the degree of relationship between response to the item and some other criterion (item difficulty)
  - Indices that are a function of both item variance and relationship to a criterion (index of discrimination)

Item statistics

- Mean, variance and item difficulty
- Item difficulty (*p*): the proportion of participants who answer an item correctly.
  - P0 = 0.50 + 0.50/m
    - m = number of choices

Item discrimination

- Refers to the ability of an item to differentiate among students on the basis of how well they know the material being tested
  - To identify items for which high scoring participants have a high probability of answering correctly and low-scoring participant have a low probability of answering correctly.

UCSF

Index of Discrimination (D)

- For dichotomously scored items.
- Designating cut scores to identify upper and lower groups(50%, 33%, 27%)
  - D= Pμ - Pl
- Pμ - proportion in the upper group who answered the item correctly
- Pl– proportion in the lower group who answered the item correctly.
- Range from -.100 to 1.00

UCSF

Interpretation of D value

- Ebel (1965) provided the following guidelines:
  - D ≥ .40 Satisfactory
  - 0.30 ≥ D ≤ .39 Little or no revision
  - 0.20 ≥ D ≤ .29 Marginal and needs revision
  - D ≤ .19 Eliminated or completely revised

UCSF

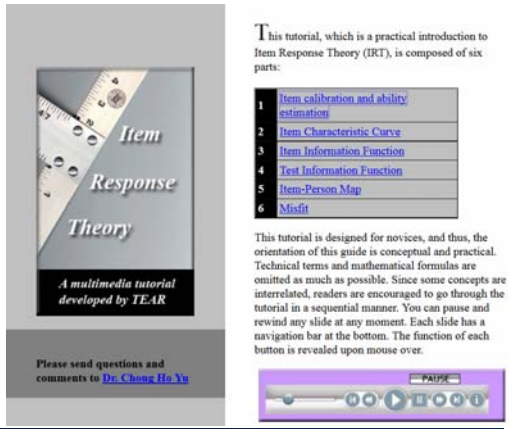
Methods of test development I:
2) Item Response Theory (IRT)

- IRT is the study of test AND item scores
  - IRT is a model-based measurement in which trait level estimates (e.g., level of physical functioning or level of depression) depend on both persons' responses AND on the properties of the questions that were administered
- Assumptions are based on the mathematical relationship between abilities (or other hypothesized traits) and item responses
- **Other names and subsets include**: *Item Characteristic Curve Theory, Latent Trait Theory, Rasch Model, 2PL Model, 3PL model and the Birnbaum model*

UCSF

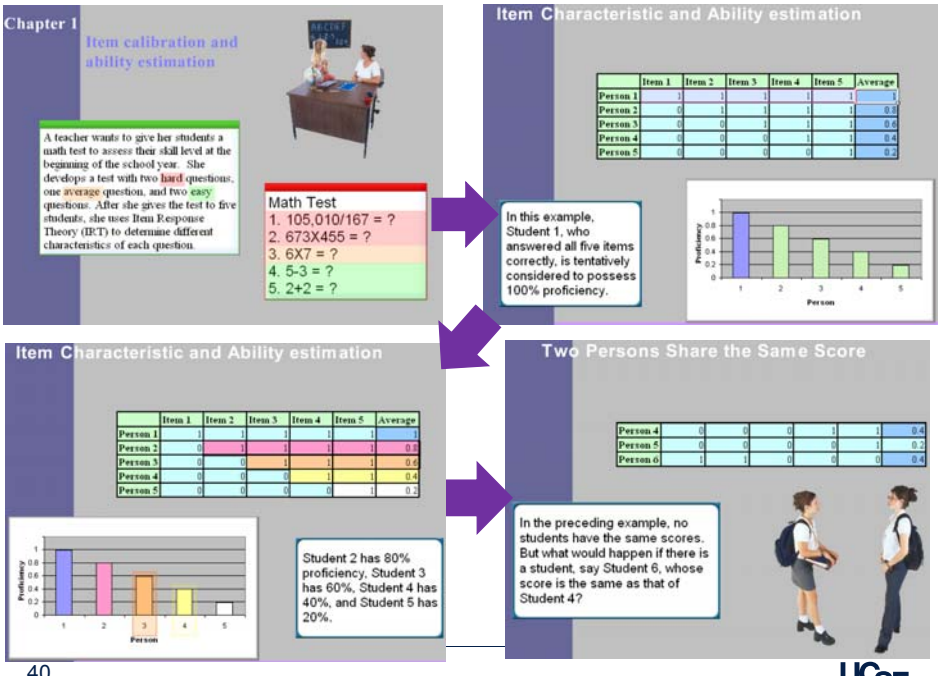
http://www.creative-wisdom.com/multimedia/IRTTHA.htm

- IRT multi-media tutorial



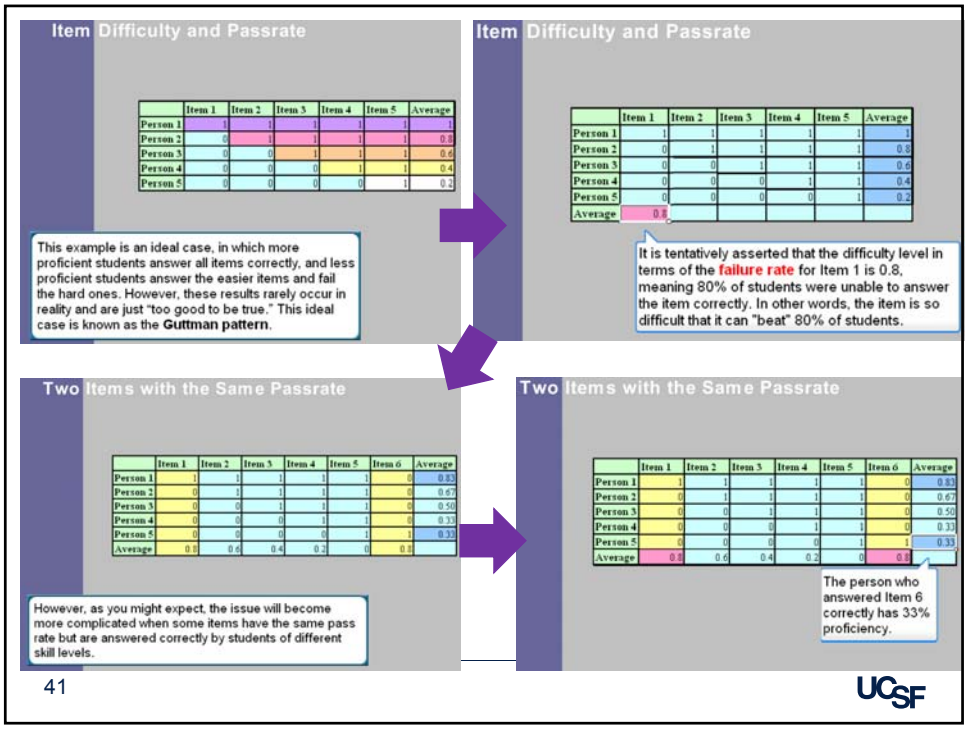
39

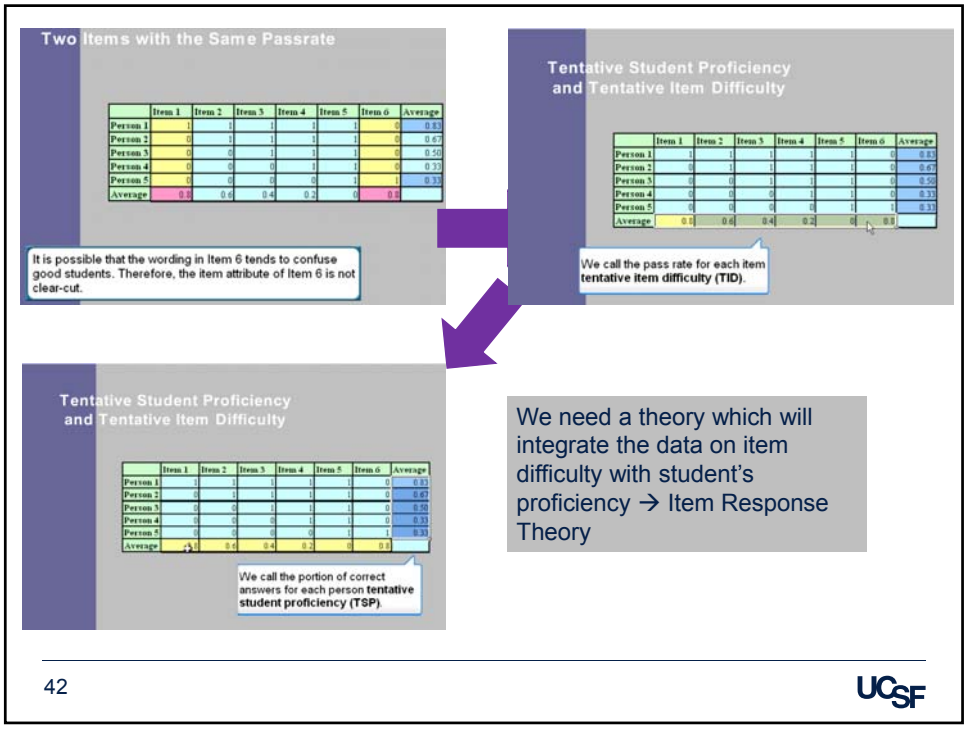
UCSF



40

UCSF


Item Difficulty and Passrate
This example is an ideal case, in which more proficient students answer all items correctly, and less proficient students answer the easier items and fail the hard ones. However, these results rarely occur in reality and are just "too good to be true." This ideal case is known as the Guttman pattern.
Item Difficulty and Passrate
It is tentatively asserted that the difficulty level in terms of the failure rate for Item 1 is 0.8, meaning 80% of students were unable to answer the item correctly. In other words, the item is so difficult that it can "beat" 80% of students.
Two Items with the Same Passrate
However, as you might expect, the issue will become more complicated when some items have the same pass rate but are answered correctly by students of different skill levels.
Two Items with the Same Passrate
The person who answered Item 6 correctly has 33% proficiency.
41
UCSF


Two Items with the Same Passrate
It is possible that the wording in Item 6 tends to confuse good students. Therefore, the item attribute of Item 6 is not clear-cut.
Tentative Student Proficiency and Tentative Item Difficulty
We call the pass rate for each item tentative item difficulty (TID).
Tentative Student Proficiency and Tentative Item Difficulty
We call the portion of correct answers for each person tentative student proficiency (TSP).
We need a theory which will integrate the data on item difficulty with student's proficiency → Item Response Theory
42
UCSF

IRT-based instruments

- Because the expected participant's scale score is computed from their responses to each item (that is characterized by a set of properties), the IRT estimated score is sensitive to differences among individual response patterns and is a better estimate of the individual's true level on the trait continuum than CTT's summed scale score

UCSF

IRT basics

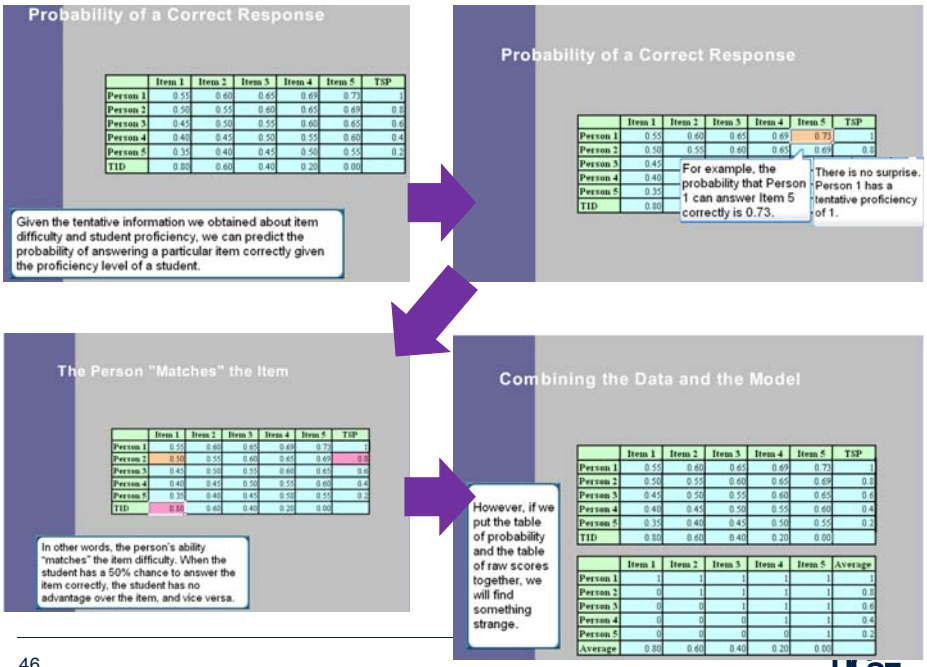
- IRT is a model for expressing the association between an individual's response to an item AND the underlying latent variable (often called "ability" or "trait") being measured by the instrument.
- The underlying latent variable in health research may be any measurable construct such as physical functioning, risk for cancer, or depression. The latent variable, expressed as theta (θ), is a continuous unidimensional construct that explains the covariance among item responses (Steinberg & Thissen, 1995).

UCSF

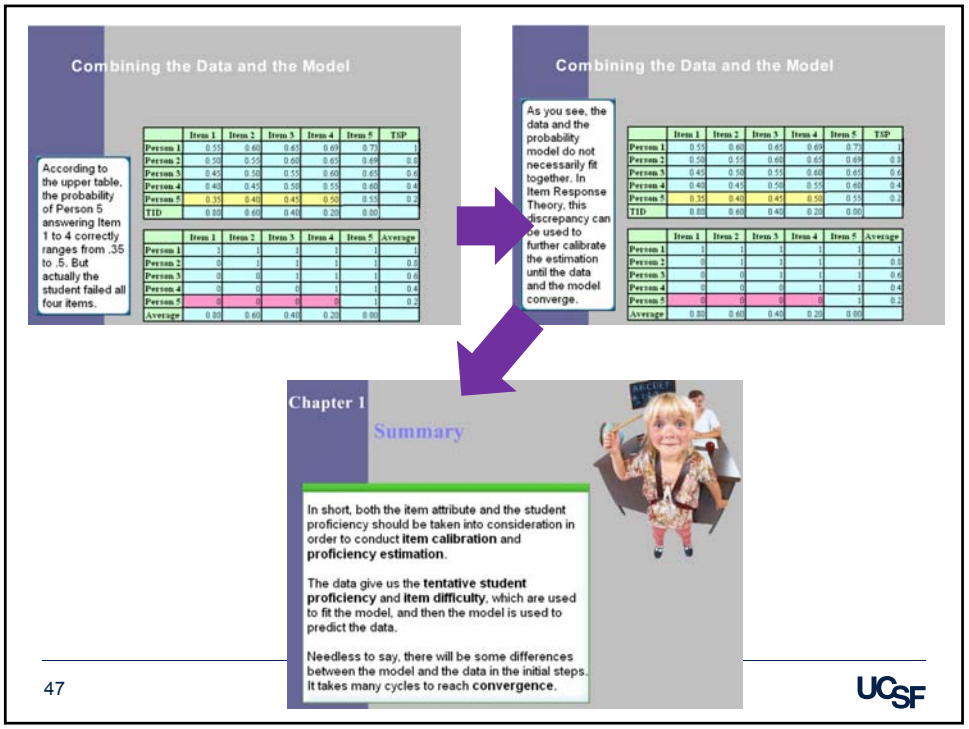
Assumptions of Item Response Theory Models

- People at higher levels of θ have a higher probability of responding correctly or endorsing an item.
- The items are measuring a single continuous latent variable θ ranging from -∞ to +∞
- The item responses are assumed to be independent of one another: the assumption of local independence (the only relationship among the items is explained by the conditional relationship with the latent variable θ)

UCSF



UCSF



IRT basic premises:

1. An examinees test performance can be predicted or explained by a set of factors referred to as traits, latent traits, abilities, or proficiencies
2. The relationship between the examinee's item performance and the set of abilities underlying item performance can be described by a monotonically increasing function, *item characteristic curve (ICC)*

- The probability of a person correctly answering a test item:
  - Person's attributes: abilities, aptitude
  - Item attributes: difficulty level

➢ **Conduct a logistic regression and choose a model that fits the data better than other models and with fewer parameters**

48

UCSF

IRT basics

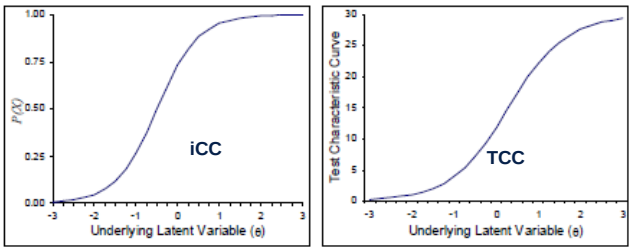


Figure 1: Left figure is an IRT item characteristic curve (trace line) for one item. Right figure is test characteristic curve for thirty items.

TCC (the test characteristic curve): the sum of the probabilities of the correct response of the item trace lines yields the test characteristic curve

The higher a person's trait level (moving from left to right along the θ scale), the greater the probability that the person will endorse the item.

- **E.g.,** "Are you unhappy most of the day?", then the left graph of Figure 1 shows that people with higher levels of depression (θ) will have higher probabilities for answering yes to the question.

UCSF

IRT uses in test development

- IRT provides direction for a method of test construction that takes into account item analysis and selection, and a measurement scale for reporting scores
- Models appropriate for dichotomous item and response data

➢One parameter → item difficulty

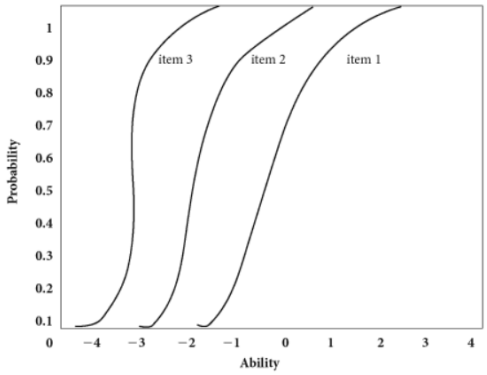


FIGURE 3.15 One-parameter ICC for three items.

UCSF

The Rasch Simple Logistic Model

- This model shows the dependent variable, the probability of endorsing an item, as a function of the difference between two independent variables, the person's level on the underlying trait θ and the item threshold b (difficulty).

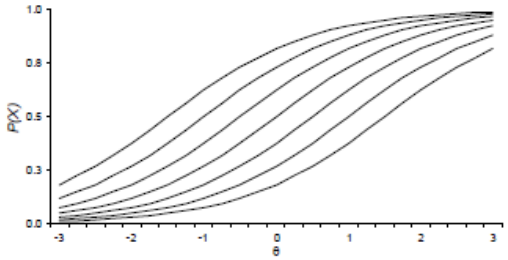


Figure 4 Rasch / One Parameter Logistic IRT Model (7 items)

51

UCSF

IRT uses in test development

- Two parameters → item difficulty and discrimination
- Item discrimination is defined by the slope of the ICC
  - Items #1 and #3 are better able to discriminate among examinees compared to item #2

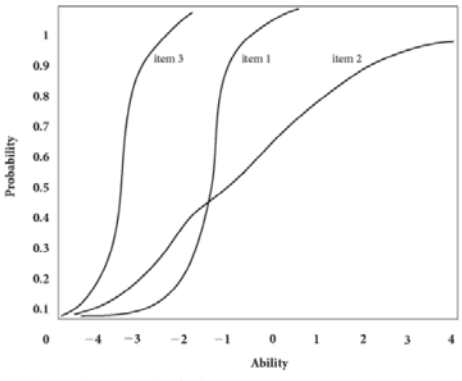


FIGURE 3.16 Two-parameter ICC for three items.

52

UCSF

The two-parameter logistic model (2PL; Birnbaum, 1968)

- allows the slope or discrimination parameter a to vary across items instead of being constrained to be equal as in the one-parameter logistic or Rasch model. The relative importance of the difference between a person's trait level and item threshold is determined by the magnitude of the discriminating power of the item

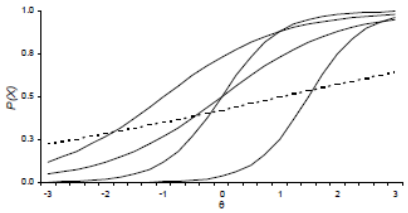


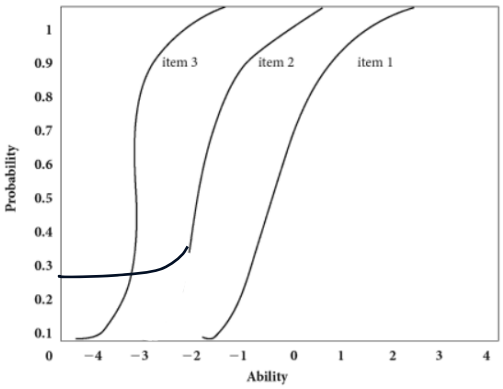
Figure 5 Two Parameter Logistic Trace Lines (5 items)

53

UCSF

IRT uses in test development

- Thee parameters → item difficulty, discrimination and proneness to guessing the item
- Item discrimination is defined by the slope of the ICC
  - Item #2 is guessed correctly by examinees with low abilities



54

UCSF

The three-parameter logistic model (3PL; Lord, 1980)

- developed in educational testing to extend the application of item response theory to multiple choice items that may elicit guessing.

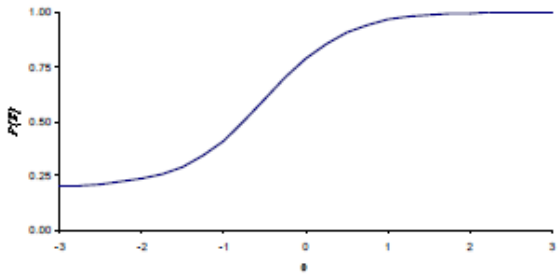


Figure 6 Three Parameter Logistic Model (1 item)

55

UCSF

Example

- Rasch or 1PL IRT model

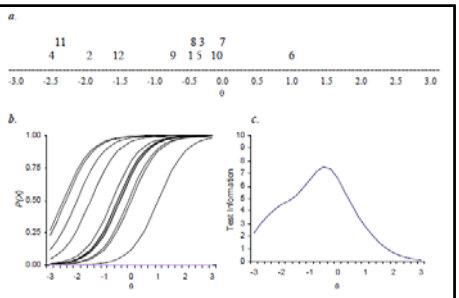


Figure 10 a. Item map for twelve items along the latent trait θ. b. Item characteristic curves for the same twelve items. c. Test information curve for the twelve-item scale.

a. presents an item map for twelve items. Except for item #6, the items are located at the lower end of the latent scale, meaning that the items are measuring low levels of the trait θ being measured by the instrument. Items # 1, 3, 5, and 8 measure people at similar trait levels, suggesting that an instrument developer may wish to remove one or two of the items because they provide redundant information.

b.displays the item characteristic curves (trace lines) for the same set of twelve items. Items # 1, 3, 5, and 8 are bunched together around theta θ levels of –0.47.

c.displays the precision of measurement of respondents at different levels of the underlying latent construct θ.

Conclusion: The instrument lacks information for measuring respondents with high levels of the trait θ.

56

UCSF

IRT vs. CTT

- An approach to constructive cognitive and psychological measures that usually serves as the basis for ***computer-adapted testing***
- Under the CTT, an attribute is assessed on the basis of responses of a sample of subjects or examinees to the set of items on a measure → focus on test performance
- Under the IRT, a set of attributes are assessed to understand how well an examinee or a group of examinees might perform when presented with a given item from the measure → focus on item performance and on the examinee's ability

UCSF

IRT vs. CTT

- IRT item parameters are not dependent on the sample used to generate the parameters, and are assumed to be invariant across divergent groups within a research population and across populations.
- IRT models measure scale precision across the underlying latent variable being measured by the instrument
- IRT-estimated person's trait level is independent of the questions being used

- CTT statistics:
  - item difficulty (proportion of correct responses)
  - item discrimination (corrected item-total correlation)
  - reliability

  are contingent on the sample of respondents to whom the questions were administered → that's why we always need to pre-test instruments!
- CTT yields only a single estimate of reliability and corresponding standard error of measurement
- CTT-based participant's score depends on the set of questions used for analysis

UCSF

Summary: CTT vs. GT vs. IRT

- Classical Test Theory (CTT) is used for assessment of random measurement error
- Generalizability Theory (GT) is an extension of the CTT and takes into account multiple sources of error. G-coefficient indicates how accurately one can generalize the results of measurement to a specified universe
- Item Response Theory (IRT) is an approach to constructing cognitive and psychological measures, which focuses on item performance and examinee's ability, unlike CTT, which focuses on test performance

- Develops computer-adapted testing (CATs) for various concepts across various chronic illnesses
- System for gathering and analyzing data
- The Domain framework has 3 sections: Physical Health, Mental Health and Social Health

Example:

ELSEVIER
Journal of Clinical Epidemiology

Construction of the eight-item patient-reported outcomes measurement information system pediatric physical function scales: built using item response theory

Esi Morgan DeWitt[a,*], Brian D. Stucky[b], David Thissen[b], Debra E. Irwin[c], Michelle Langer[a], James W. Varni[a,f], Jin-Shei Lai[g], Karin B. Yeatts[c], Darren A. DeWalt[h]

Abstract

Objective: To create self-report physical function (PF) measures for children using modern psychometric methods for item analysis as part of patient-reported outcomes measurement information system (PROMIS).

Study Design and Setting: PROMIS qualitative methodology was applied to develop two PF item pools that comprised 32 mobility and 38 upper extremity items. Items were computer administered to subjects aged 8–17 years. Scale dimensionality and sources of local dependence (LD) were evaluated with factor analysis. Items were analyzed for differential item functioning (DIF) between genders. Items with LD, DIF, or low discrimination were considered for removal. Computerized adaptive testing performance was simulated, and short forms were constructed.

Results: Three thousand forty-eight children (51.8% female, 40% nonwhite, and 22.7% chronically ill) participated. At least 754 respondents answered each item. Factor analytical results confirmed two dimensions of PF. Fifty-two of 70 items tested were retained. A 23-item mobility bank and a 29-item upper extremity bank resulted, and an eight-item short forms were created. The item banks have high information from the population mean to three standard deviations below.

Conclusions: PROMIS pediatric PF item banks and eight-item short forms assess two dimensions, mobility, and upper extremity function and show good psychometric characteristics after large-scale testing.

61 UCSF

Limitations of IRT, Waltz et al. 2010

There are other factors related to IRT that are identified as potentially limiting factors in its use for CAT related to measurement of generic health outcomes. Among these are the following: (1) many generic concepts are not unidimensional; (2) the ordered continuum of IRT items is very different from the more traditional Likert-type response format; (3) time and complexity required in test development is large; (4) there is a large number of respondents required for modeling of the test; (5) there is the requirement for a large item pool; and (6) the resulting tests created would be generic as each would be individualized to the respondent as opposed to the more proprietary approach common today (McHorney, 1997). Extensive resource commitment is needed to support the development of CAT such as is seen in the complex processes of the PROMIS initiative. Nunnally and Bernstein (1994) note that clear advantages to classical testing methods, which are robust, remain.

62 UCSF

Methods of test development I:
3) From items to scales

- To evaluate subject's performance on a scale, we usually add the scores from individual items
  - some items may be more important than others, and perhaps should make a larger contribution to the total score
  - Different items could use different metrics
    - ➢ **Item weighing**
- Each new scale is reported on a different metric, making comparisons among scales difficult
    - ➢ **Convert row scores to percentiles –> depends on the sample, percentiles are ordinal data, cannot use means and standard deviations**
    - ➢ **Standardize, normalize scores**
    - ➢ **Establish cut-points, calculate ROC and AUC**

UCSF

From items to scales ROC, AUC

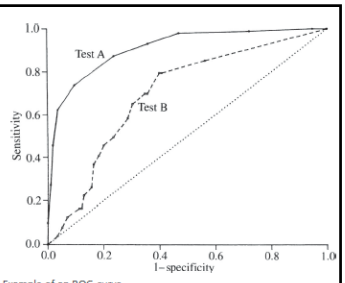


Fig. 7.3 Example of an ROC curve.

- The better a test is in dividing cases from non-cases, the closer it will approach the upper left hand corner
- The cut point, which minimizes the overall number of errors, is the score that is closest to that corner - the Youden index, or Youden's *J*, defined as (Sensitivity + Specificity) – 1 (Youden 1950)
- A non-discriminating test (one which falls on the diagonal) has an area under the curve (AUC) of 0.5, and a perfect test has an area of 1.0. If a test falls below the line, and has an AUC<0.5, then it is performing worse than chance
- Estimate AUC by doing a definite integral (could do in Excel)
- Interpretation of AUC: If AUC for Test A is 91.1%, then if we randomly select two people, one who is a case and one who is a non-case, the probability is 91.1 per cent that the former will have a lower score (in this situation, where a lower score is worse) than the former.

UCSF

Methods of test development I:
4) Factor Analysis/ Principal Component Analysis

- We will cover this in computer lab

UCSF

Methods of test development II:

Instrument design, construction and interpretation in cross-cultural situations

UCSF

Collecting sensitive information

- Recent legislations to protect sensitive information:
  - Genetic Information Nondiscrimination Act (GINA)
  - Health Insurance Portability and Accountability Act (HIPAA)
- Research on sensitive information could be perceived as embarrassing or threatening and could lead to subjects refusing to participate in the research study

Methods of collecting sensitive information

- Randomized Response Technique (RRT)
  - Subjects are assured that their response cannot be directly attributed to them
  - Each subject is given a randomization device (coin, die, etc.)
  - The answer to the question could be true or could be due to flipping a coin (all questions are dichotomous)
  - An aggregate response from a sample of 100 subjects could be due to chance (50%) or due to some characteristic (60-50=10%)
  - Prevalence of condition is 10/50=20%

Cultural Issues in Measurement

- U.S. population is becoming increasingly more diverse.
- Growing research on health disparities to:
  - Describe health disparities or differences in health across groups;
  - Identify determinants of health disparities (individual level vs. environmental level); and
  - Intervene to reduce health disparities.

Validation Approaches for Translated Instruments

- Back translation procedures
- Administration of both versions to fluent
- bilingual individuals
  - Total score agreement (or correlation)
  - Item agreement (or correlation)

Issues in Translation

- **Literal translation**
  - Asymmetric -- loyalty to one language (source)
  - Use: for operational study goal
  - Criterion-referenced
- **Cultural equivalence**
  - Symmetric -- equal loyalty of meaning and familiarity for source and target language
  - Use: for comparative study goal
  - Construct-referenced

71 UCSF

Five Dimensions of Cross-Cultural Equivalence

- Content equivalence / Content
- Semantic equivalence / Language
- Technical equivalence / Measurement
- Criterion equivalence / Norms
- Conceptual equivalence / Concept or theoretical construct

72 Flaherty et al., 1988 UCSF


Postpartum Depression Screening Scale

Spanish Version

Cheryl Tatano Beck ▾ Robert K. Gable

Screening Performance of the Postpartum Depression Screening Scale—Spanish Version

CHERYL T. BECK, DNSc, CNM, FAAN
University of Connecticut
ROBERT K. GABLE, EdD
Johnson and Wales University

J Immigrant Minority Health (2010) 12:249–258
DOI 10.1007/s10903-009-9260-9

ORIGINAL PAPER

The Postpartum Depression Screening Scale-Spanish Version: Examining the Psychometric Properties and Prevalence of Risk for Postpartum Depression

Huynh-Nhu Le · Deborah F. Perry · Glorimar Ortiz

Example: Semantic Equivalence of Spanish Version

- Postpartum Depression Screening Scale (PDSS)
- Methods used:
  - Back translation (2 each from 4 countries)
  - Committee approach
  - Pretest technique
  - Alternate forms equivalence (with bilingual subjects)

73 Beck CT, Bernal H, Froman RD. Methods to document semantic equivalence of a translated scale. Res Nurs Health. 2003 Feb;26(1):64-73.

UCSF

Description of PDSS

- A 35-item Likert-type self report scale
- 7 dimensions with 5 items each
  - Sleeping/eating disturbances
  - Anxiety/insecurity
  - Emotional liability
  - Cognitive impairment
  - Loss of self
  - Guilt/shame
  - Contemplating harming oneself

74

UCSF

Psychometric Results with PDSS

- Internal consistency .83 to .94
- Strong correlation with 2 established
- depression scales (.79-.81)
- **ROC curves showed:**
  - Sensitivity 94% and specificity 98% for a cutoff score of 80 for major postpartum depression
  - Sensitivity 91% and specificity 72% for a cutoff score of 60 for major/minor postpartum depression screening

Back Translation Procedures

- Requires a minimum of 2 translators who
- work independently
- 8 used in this study: 2 each for the 4
- predominant Hispanic groups in the US
- Translators -- bilingual and bicultural
- Decentering -- considers original and
- translated language equally important

Example from Translation: 1

- Two or more possible translations for an item: “I felt like my emotions were on a roller coaster.”
  - 1. “I felt like my emotions have been on a roller coaster.”
  - 2. “Felt that my emotions were up and down.”
  - 3. “Felt emotional instability.”
- Resolved by Committee Approach

Example from Translation: 2

- **Variations among sub-groups**
  - English item: I was scared that I would never be happy again
  - Cuban: I feared I would never be happy again
  - Puerto Rican: I felt afraid I would never be happy again
  - Mexican: I was frightened I would never be happy again
  - Other countries: I was afraid I would never be happy again

Committee Approach

- Goal -- group consensus on problematic items
- Typical input for discussion and resolution:
  - The Puerto Rican women will not understand that term. We need to use another word.
  - A different word is needed because women in my group may have only a 6th grade education.
- Final review for errors in syntax and grammar

UCSF

Anxiety/insecurity

felt all alone (2)
felt really overwhelmed (9)
felt like I was jumping out of my skin (16)
got anxious over even the littlest things that concerned my baby (23)
felt like I had to keep moving or pacing (30)

J Immigrant Minority Health (2010) 12:249–258
DOI 10.1007/s10903-009-9260-9

ORIGINAL PAPER

The Postpartum Depression Screening Scale-Spanish Version: Examining the Psychometric Properties and Prevalence of Risk for Postpartum Depression

Huynh-Nhu Le · Deborah F. Perry · Glorimar Ortiz

Table 3 The internal consistencies of the PDSS-Spanish version by group and total sample

PDSS dimensions (# items)	El Salvador ($n = 91$)	Other Cent Am ($n = 40$)	Mexico ($n = 24$)	Total ($N = 155$)
Sleeping/Eating Disturbances [5]	0.85	0.85	0.83	0.84
Anxiety/Insecurity [5]	0.68	0.82	0.56	0.72
Emotional Lability [5]	0.76	0.89	0.76	0.80
Cognitive Impairment [5]	0.86	0.88	0.81	0.86
Loss of Self [5]	0.90	0.88	0.85	0.89
Guilt/Shame [5]	0.88	0.91	0.91	0.89
Contemplating Harming Oneself [5]	0.97	0.97	0.99	0.97
PDSS total score long version [35]	0.97	0.97	0.97	0.97
PDSS total score short form [7]	0.85	0.83	0.77	0.83

Note: *Other Cent Am* other Central American countries

Results:

- coefficients for the Anxiety/Insecurity subscale were lower than the other dimensions for all three groups and the entire sample in the study
- Beck and Gable's findings for this subscale ranged from 0.71 to 0.81
- Explanation: Mexican women chose either feeling lonely (33%) or anxiety about the baby (44%) but not the other 3 responses.

UCSF

The Measurement Problem

- Most self-report measures have been developed and tested in mainstream, well-educated groups.
- **Measurement goal:**
  - Identify measures that can be used across all groups
  - Sensitive to diversity
  - Minimal bias between groups

Cultural Issues

- Observed mean differences between groups may be due to:
  - Culturally-mediated differences in “true score” (true differences).
  - Bias-systematic differences between group observed scores that are unrelated to “true” scores.

Cultural Relevance

- Extension of the instrument to a new group that differs from the original group used in standardizing the instrument.
  - Age, gender, socio-economic status, culture/ethnicity, language, etc.
- When crossing cultures, initial work is needed to determine whether the construct exists in the same way as in the original culture.

Conceptual Adequacy

- Is the concept meaningful, relevant, and acceptable?
- **Traditional research:**
  - Conceptual adequacy = merely defining a concept.
  - Mainstream population "assumed."
- **Cross-cultural research:**
  - Mainstream concepts may be irrelevant and/or non-existent.

Psychometric Adequacy

- **Minimal standards:**
  - Sufficient variability
  - Minimal missing data
  - Adequate reliability and evidence of construct validity
  - Evidence of responsiveness to change

UCSF

Example: Testing the Conceptual Equivalency of the Chinese Family Assessment Device (FAD)

- Assessed the FAD in a Chinese population of hospitalized children (n=313) compared to families with healthy children (n=29) in Hong Kong and Chinese Mainland.
- Compared to U.S. studies, the Chinese FAD has a different factor structure, reliabilities, and mean scores on several subscales.
- Bilingual and bicultural experts reviewed all 60 items and some items had low cultural relevance to family functioning.

Chen et al. (2003). Culturally Appropriate Family Assessment: Analysis of the Family Assessment Device in a Pediatric Chinese Population. Journal of Nursing Measurement, 11, 41-59.

UCSF

Example of Nonequivalent Concept: Depression

- **Depression (concept):**
  - In mainstream U.S. culture, depression is usually expressed and reported via affect, somatic symptoms, behavior, and thought patterns.
- **In some Asian cultures:**
  - Public expression of self-reflection may be discouraged.
  - "Saving face" and "self-sacrifice" may serve as powerful forces in molding expression and behavior.

Validation Approaches for Translated Instruments

- Back translation procedures
- Administration of both versions to fluent bilingual individuals and examine:
  - Total score agreement (or correlation)
  - Item agreement (or correlation)

Steps in Developing Cross-Cultural Research Instruments

- Step 1-Defining the Concept in Emic Terms
- Step 2-Obtaining Cultural Consensus
- Step 3-Developing Measurement Items
- Step 4-Peer Validation
- Step 5-Pretesting Items with Prospective Study Participants
- Step 6-Seeking Additional Peer Input for Revising Items
- Step 7-Pilot Testing
- Step 8-Psychometric Evaluation
- Step 9-Evaluating Efficacy

89 UCSF

Steps in Developing Cross-Cultural Research Instruments from

- **Step 1-Defining the Concept in Emic Terms:**
  - Researcher begins with a concept that is understood well by the scientific community in the original culture.
  - The meaning of the concept must be established with the second ethnocultural group.
  - A preliminary technique for exploring cultural meaning (emic perspective) of a concept is to conduct in-depth qualitative interviews.

90 Tran & Aroian (2000) UCSF

Step 2-Obtaining Cultural Consensus:

- Conducting focus groups with representative members of the second ethnocultural group to reach consensus about the essential elements of the concept.
- Moderate-sized groups of 6-10 people are effective at facilitating group interaction.
- The unit of analysis is the group (rather than the individual).

Step 3-Developing Measurement Items:

- Generated items will serve as a pool of potential items that will undergo further evaluation.
- Goal is to develop more than a sufficient number of items, since some of these items will be revised or dropped during subsequent steps.

Step 4-Peer Validation:

- Item pool undergoes systematic peer review by a panel of experts for content relevance.
- At least 3 experts needed to assess the items.
- Revised items are re-evaluated as needed for content validity.

Step 5-Pretesting Items withProspective Study Participants:

- Pretesting is essential for measures being used on a new population.
- Involves administering the measure to a small sample of people (e.g., 25-50) from the second ethnocultural group.
- Pretesting allows one to detect any problems with the method of administration, respondent burden, procedures, and comprehension of items.

Step 6-Seeking Additional Peer Input for Revising Items:

- Obtaining input from representative members of the second ethnocultural group about revising any questionable, problematic, and misleading items.
- Conducting focus groups is a useful method for seeking input about revising items that are questionable.
- Revised items are then subjected to peer review and pretesting procedures as outlined in Steps 4 and 5.

Step 7-Pilot Testing:

- Purpose is to time the length of administering the measure and examine missing data and item frequencies.
- Missing data and items with little variance may suggest that the items are not clearly measuring the concept.
- A lot of missing data could suggest that the items elicit information that respondents find too threatening, personal, or irrelevant.
- It is preferable (if time permits) to pilot-test different methods of data collection, such as collecting data in-person or over the phone.

Step 8-Psychometric Evaluation:

- Collect data for thorough psychometric evaluation.

UCSF

Step 9-Evaluating Efficacy:

- Assessing the instrument's ability to correctly classify "cases" and "noncases."
- Requires the researcher to collect data from two different populations, e.g., a clinically depressed patient population and a community-based non-depressed population.
- Sensitivity and specificity are two commonly used approaches to assess efficacy.

UCSF

Some Strategies for Low Literacy or Unfamiliarity with Format

- Using interview methods in samples with a high proportion of low literacy.
- **Unfamiliarity with Likert response scales:**
  - Visual aids for response scale points.
    - Use of color blocks of increasing intensity to match the scaling anchors.
    - Use of faces (smiling or frowning) to match the scaling anchors.

UCSF

Conclusion

- Measurement studies in culturally/ethnically diverse populations are recently emerging.
  - Few guidelines.
  - We need more studies in this area!
- Developing cultural equivalent research instruments is a lengthy/time-consuming process but critical first step before using the instruments to evaluate clinical interventions.
- Pretest, pretest, pretest!!!

UCSF

Summary of Measurement Theory and Principles

Principles of measurement development:

- Measures need to provide answers containing important information that directly addresses research aims
  - Go back to research aims and questions; develop conceptual and measurement frameworks
- Measures should provide comparable information about people, events or behavior
  - All respondents must be answering the same question.
  - All respondents must understand each item and attribute the same meaning to it.
  - Assure that researchers and respondents are using the same definition.
  - Will items produce meaningful information from adolescents and adults?
    - Cannot conclude age or developmental differences if it is really just a reflection of interpretation.
- Measures should provide reliable and valid information
- Most instruments are valid only in the population for which they have been developed → need to conduct validation studies
  - Instruments developed using IRT, e.g. CAT questionnaires, could be used in new studies without prior validation studies

UCSF