

# Concepts in Item and Scale Construction

- Today's topics:
  - Examples of problems in measurement
  - Constructs, Measurement, Instruments
  - 4 Components of Instrument Construction
  - Item Design and Response Format
  - Wrapping up...

# Practical Problems in Measurement

- To demonstrate the types of issues we will discuss related to test development and evaluation, consider the following three examples of measurement:
  1. A teacher wishing to evaluate student knowledge of math
  2. A psychologist wishing to measure a psychological disorder
  3. A researcher wishing to study attitudes toward gun control
- Note the common denominator here is not topic, but rather than each person is trying to assess a **latent trait** – these concerns apply any time you are trying to do that, regardless of what the trait is

## Example #1 – The Math Teacher

- A teacher constructs 20 pass/fail items for a math test that covers algebra and geometry, administers the test, and adds up the number of correct items to use as the math score for each student.
- In doing so, the teacher wonders...
  - Should there be one score or two scores for math ability?
    - One score for geometry items AND one score for algebra items?
    - If so, what about items that require both algebra and geometry?
  - If one score is sufficient...
    - How accurate is that single score as a measure of math ability?
    - How accurate would two scores be?
  - Are 20 items sufficient to give a reasonably accurate determination of each student's knowledge?
    - Should more be used? Could fewer have been used?

## Questions about Questions...

- Are all items good measures of math ability or are some items better than others? Are there other ways of getting the right answer besides ability?
- If different items had been used, would they have measured the same thing?
  - Equally well? Can two tests be made (with different items) so that the scores are interchangeable? Could a computer be used to administer the test adaptively?
- Are students who have low scores measured as accurately as students scoring highly or in the middle?
  - Test floor? Test ceiling?
- Are the items free from bias when given to students of different cultural backgrounds? In different languages?
  - Could some students have irrelevant problems with certain items because of differences in their background and experience?
  - How would we be able to know?

## Example #2 – The Clinical Psychologist

- A clinical psychologist writes a set true/false of items such as:
  - “I have difficulty sleeping.”
  - “I am afraid of heights.”
  - “I get tired easily.”
  - “I often have bad dreams.”
- The clinical psychologist has similar questions about his or her questions as the teacher did...
  - Dimensionality of traits to be measured?
  - Overall accuracy and efficiency of measurement?
  - Item quality, exchangeability, and bias?
  - Reliability across trait levels?

## Example #3 – The Survey Researcher

- A researcher creates a series of items to study attitudes about gun control, with 5 response options ranging from *strongly agree* to *strongly disagree*:
  - “Assault weapons do not belong in private hands.”
  - “All hand guns should be licensed.”
  - “Government interference with the right to bear arms is an infringement of liberties.”
  - “From my cold, dead hands...”
- More questions about questions...
  - Do positively and negatively worded items measure same trait?
  - Are all ‘strongly agrees’ created equal?

## A Non-Exhaustive List of Potential Worries in Test Construction...

- Dimensionality of traits and items:
  - How many traits are you measuring?
- Overall test accuracy vs. efficiency
  - Do you need to add or remove items?
  - Add or remove response options?
  - Just any items? Or targeted items?
- Reliability across trait levels
  - Avoid ceiling and floor effects
  - Customize test for specific measurement purposes
- Bias and generalizability across populations:  
Does your test ‘work’ for different groups?
  - Sufficiently unbiased?
  - Sufficiently sensitive for groups with different ability levels?

## Defining Constructs

(adapted from *Constructing Measures*, Wilson, 2005)

- Purpose of measurement:
  - Provide a reasonable and consistent way to summarize the responses that people make to express their abilities, attitudes, etc. through tests, questionnaires, or other types of scales
- Classical definition of measurement:
  - “process of assigning numbers to attributes”
  - But important steps precede and follow this part!
- All measurement begins with a *construct*, or unobserved (latent) trait, ability, or attribute that is the focus of study
  - i.e., the ‘true score’ in CTT, ‘factor’ in CFA, or ‘theta’ in IRT

## Defining Constructs, continued

- The models we'll utilize each assume the construct to be a *unidimensional* and *continuous* latent variable
  - Wilson (2005) calls this a 'construct map'
  - If not strictly unidimensional, try to think of sub-constructs that would be unidimensional, and focus efforts on each one of those
  - Qualitative distinctions (benchmarks) are ok as a means of *description*, but should be continuous in between those points
- Constructs made up of categorical latent 'types' instead? You may need another kind of measurement model:
  - Diagnostic Classification Models (e.g., Templin & Henson, 2006)
    - Goal is measurement of discrete attributes or skills, not traits
    - Useful when classification is the goal of measurement

## Construct Maps should include...

- Coherent, substantive definition of the construct
- An underlying continuum that can be manifested 2 ways:
  - *Ordering of persons to be measured (low to high)*
    - Could include descriptive labels for 'types of people'
    - Could include other characteristics (e.g., age, disease state)
  - *Ordering of item responses (low to high)*
    - Behaviors (e.g., 'sits quietly'.... 'kicks and screams on the floor')
    - Item options ('no problems', 'some problems', 'many problems')
  - Key idea: Responses have to *orderable*
- Some examples of construct maps...

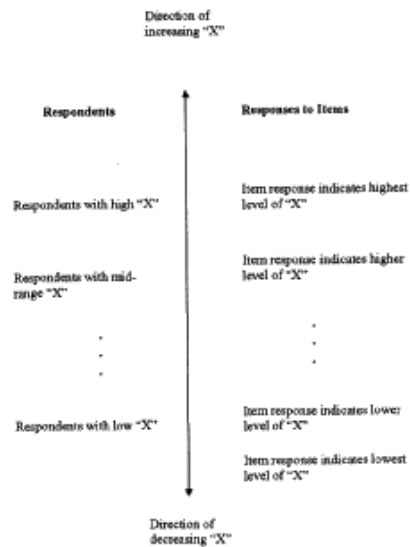
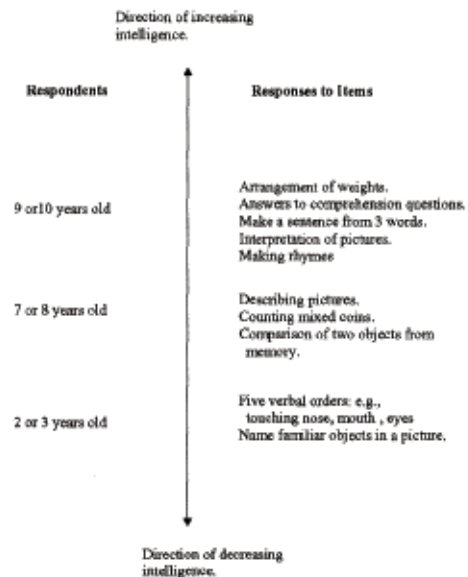


FIG. 2.1 A generic construct map in construct "X."  
From Wilson (2005)

## Template for a Construct Map

Left = PERSONS  
qualities  
characteristics

Right = ITEMS  
responses  
behaviors



From Wilson (2005) – originally adapted from Binet and Simon's (1905) Measuring Scale of Intelligence

## Example Construct Map for Intelligence

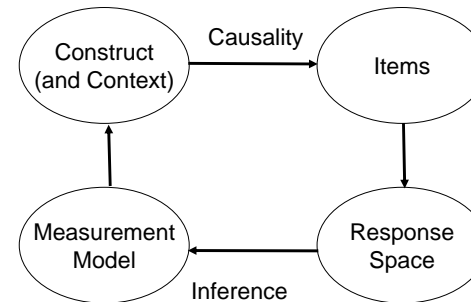
Left = PERSONS  
ages

Right = ITEMS  
skills  
behaviors

# Instrument Construction

- Once your construct is mapped in terms of ordering of persons and responses, next is instrument construction
- Instrument → Measurement method through which observable responses or behaviors in the real world are related to a construct that exists only as part of a theory
- 4 components of instrument construction:
  - Construct (and Context)
  - Item Generation
  - Response (Outcome) Space
  - Measurement Model

# 4 Instrument Building Blocks



Direction of causality: The construct determines which items are relevant (to represent the construct), the content of the items then causes a response, and *the response format then directs which measurement model to use.*

We then use the measurement model to make inferences about people's standing on the latent construct (trait as measured in a given context).

# Construct and Context

- Instruments should be secondary – they are created:
  - For the purpose of measuring a pre-existing latent **construct**
  - Within a specific **context** in which that measurement is needed
- Instruments should be seen as **logical arguments**:
  - Can the results be used to make the intended decision regarding a person's level of a construct in that context?
  - Constructive logic: Build instrument purposively with this in mind
  - Reflective logic: Pay attention to information gathered after-the-fact as to how well it is working (see list of potential worries)
- Instruments are created from items, which have 2 parts:
  - **Construct** component: Location on the construct map?
    - Want to include both hard and easy items to measure full range
  - **Descriptive** component: Other relevant item characteristics
    - Language? Context? Method of administration? Reporter/rater?

# Steps to Item Design

- Do your homework:
  - Literature review
    - What's been done before...And what's wrong with it?
  - Ask relevant people (participants, professionals):
    - What should we be focusing on? How should we ask the questions?
- Design the instrument:
  - Item design (construct and descriptive components)
  - Response format (location on 'openness' continuum)
- Get feedback from participants:
  - 'Think aloud' while solving problems
  - Exit interview

# (Good) Item Generation

- Ideally, items are *realizations* of existing constructs
  - Hmm...How do I measure this construct? (proceed to write item 1, 2, 3...)
  - In reality, this is an iterative process...
- Items should be unambiguous
  - Cover a single concept (no 'ands')
  - Specific with a clear referent
- Items should be simple to process
  - Short, common vocabulary
  - Negatives can be harder to process – and recent research has suggested negatively-worded (reverse-coded) items to be less discriminating
- Good items should span the full range of construct... but without going too narrow or too broad

# Actual (Not so Good) Items...

- *How important to you is it that...*
  - My family members have good relationships with extended family members (grandparents, in-laws, etc.).
  - My family is physically healthy.
- Assess the quality of the relationship that you have with your children?  
\_\_\_excellent \_\_\_very good \_\_\_good \_\_\_fair \_\_\_poor
- To what extent did others make it difficult for you to engage in various activities before your imprisonment?  
\_\_\_ 1. never \_\_\_ 2. rarely \_\_\_ 3. often \_\_\_ 4. most of the time

# Response (Outcome) Space

- **Outcome space = response format** → varies in flexibility
  - Most flexible: Open-ended response
    - e.g., essay, performance
    - Less work at beginning; more work at the end
  - Least flexible: Fixed format
    - e.g., multiple choice or likert scales
    - More work at beginning; less work at the end
- Ideally, instrument development **would start by seeking open-ended responses**, from which representative fixed format options would be created that are:
  - Research-based, well-defined, and context-specific
  - Finite and exhaustive (orderable responses; include n/a)

# Common Types of Fixed Response Formats

1. Completion ("Fill in the blank")  
e.g., The capital of New Jersey is: \_\_\_\_\_
  - Scoring method: 0/1 for wrong/right
  - Eliminates guessing... but can be difficult to define all possible correct answers (or what's 'close enough')
  - No partial credit available (thus, less info about ability)
2. Two-Choice or Multiple Choice  
(one right answer, varying number of distractors)
  - Scoring method: 0/1 for wrong/right
    - Potentially partial credit for selecting 'best' wrong answer
  - Understandable to examinees... but susceptible to guessing
  - Less amenable to non-cognitive items

## Common Types of Fixed Response Formats

### 3. Checklist ('check all that apply')

- How do you feel right now? (List of adjectives follow)
  - Which of the following are true? (List of statements follow)
  - Which of the following activities have you participated in?
- Usual scoring method: Sum of checked items
  - Are all options exchangeable?
    - More than one way to get the same total score
    - Some options should 'count' more than others (\*Ahem\*, IRT)

## Common Types of Fixed Response Formats

### 4. Ordered Category

- At least three ordered (or 'graded response') options
  - Likert (pronounced 'lick-ert') scales
- Scoring method: Integer assigned to category (1-2-3-4-5)
  - To 'neutral' or not to 'neutral'?
  - Ironically enough, a true 'likert scale' is NOT an integer-valued set of responses
    - In 1932 paper, Likert showed how his sophisticated scaling system was not really an improvement over simpler, integer-valued ratings... and the term 'likert' has been used ever since

## Special Type of Ordered Category: Guttman Scale

- Idea is that the ordering of categories dictates possible response patterns – if get hard item correct, must have gotten easier items correct (or endorsed 'easier' items)

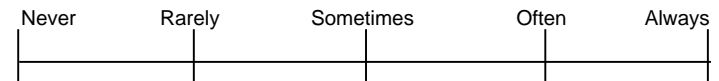
Item 1	Item 2	Item 3	Item 4	Score
agree	agree	agree	agree	4
agree	agree	agree	<b>disagree</b>	3
agree	agree	<b>disagree</b>	disagree	2
agree	<b>disagree</b>	disagree	disagree	1
<b>disagree</b>	disagree	disagree	disagree	0

- But people don't always cooperate....some kind of probabilistic model would pry be more realistic (\*ahem\*, IRT)

## Common Types of Fixed Response Formats

### 5. Graphic rating scale as ordered categories

- Mark the line where you fall
- Sometimes with qualitative anchors



- Scoring method: Actual distance on line
- Good idea in theory... but little evidence that having more choices helps (beyond 7-9 categories, anyway)
- Resultant distribution is harder to deal with, as it is likely going to be "bunched normal"

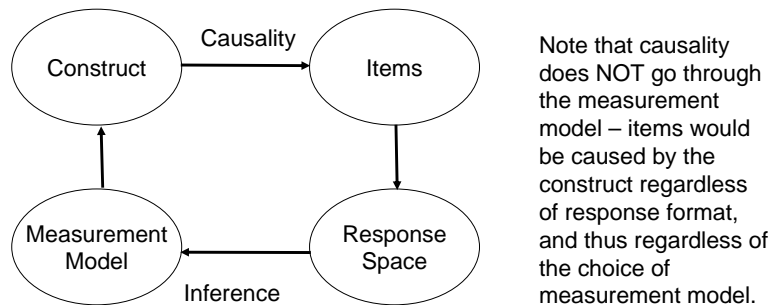
## Common Types of Fixed Response Formats

6. Forced choice
    - Do you prefer espresso-flavored or mint oreo cream?
    - Are you angry or sad?
  - Scoring method: ??? Preference does not necessarily imply 'choice' in an absolute sense
  - Possibility of logical inconsistency across items when responding to all possible combinations in a series
7. Rankings
    - Rank these ice creams in order of preference:
      - Vanilla, strawberry, chocolate, mint oreo
  - Scoring method: ??? Preference STILL does not necessarily imply 'choice' in an absolute sense

## Item-Level Measurement Models

- Type of response format will generally lend itself to an appropriate measurement model
  - Dichotomous (binary) item? (yes/no, MC → correct/not)
    - Logistic/probit model (IRT)
    - Normal approximation (CFA) pry won't work very well
  - Polytomous (quantitative) item? A few IRT options...
    - Graded response model
    - Partial credit model
    - Normal approximation (CFA) \*may\* not be too bad...
  - Unordered categorical item? Only one IRT option:
    - Nominal model (way hard to estimate)
  - No clear measurement model for many other types of item choices (i.e., forced choice, rankings)

## 4 Instrument Building Blocks



- Process of Inference:
  - Relate responses to construct via measurement model
  - In other words, *translate scores to locations on construct map*

## Wrapping Up...

- Instruments are created to measure pre-existing latent constructs: latent traits within desired contexts
  - Item construction is part art, part science
  - Seek as much info as possible before and after about your items
- Response options should be carefully considered:
  - Start with open-ended responses
  - Come up with fixed response categories eventually
- Measurement models provide basis for inference back to a person's position on the latent construct:
  - Specific model chosen on the basis of response format
  - The ones we'll use assume continuous underlying latent variable on which BOTH persons and items can be ordered