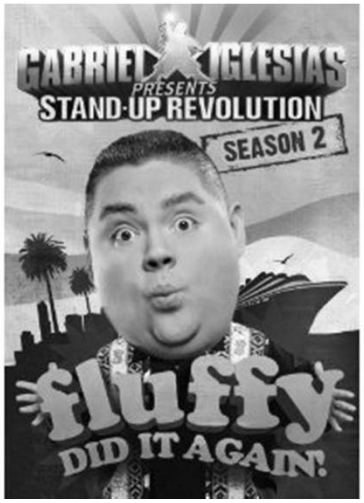
Concepts Underlying Scale Development

Latent Trait Measurement and
Structural Equation Models
Lecture #5

February 6, 2013

In this Lecture…

- The “fluffy” part of factors and scales…



CONCEPTS UNDERLYING SCALE DEVELOPMENT

Practical Problems in Measurement

- To demonstrate the types of issues we will discuss related to test development and evaluation, consider the following two examples of measurement:
  1. A teacher wishing to evaluate student knowledge of math
  2. A psychologist wishing to measure depression

- The common denominator here is not topic, but rather that 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 Psychologist

- A clinical psychologist writes a set of items to measure depression, with 5 options ranging from “rarely” to “almost always” such as:
  - “I have lots of energy.”
  - “I sometimes feel sad.”
  - “I think about ending my life.”
  - “I cry.”

- The psychologist may have similar questions about measurement…
  - Dimensionality of traits to be measured?
  - Overall accuracy and efficiency of measurement?
  - Item quality, exchangeability, and bias?
  - Reliability across trait levels?
  - Do positively and negatively worded items measure same trait?
  - Are all ‘almost always’ responses 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 single factor CFA model assumes 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?
  There are other kinds of measurement models:
  - Diagnostic Classification Models (e.g., Rupp, Templin & Henson, 2010)
    - 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…

Template for a Construct Map

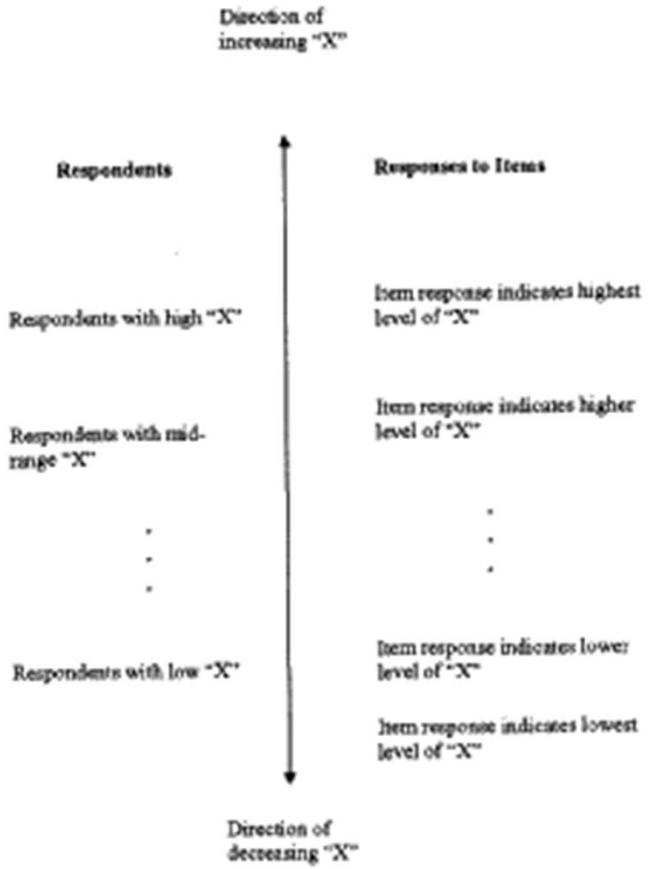


FIG. 2.1 A generic construct map in construct "X."

From Wilson (2005)

Left = PERSONS
- qualities
- characteristics

Right = ITEMS
- responses
- behaviors

Direction of increasing speech sound development for *girls*

Respondents	Responses to Items
9 ½ yrs.	All speech sounds are accurate
9 yr. olds	spr, thr, skr, str
8 yr. olds	r-, -er, pr, br, tr, dr, gr, kr, fr
7 yr, olds	-ng, s, z, th, sp, st, sk, sp, sm, sn, sw, sl, spl, skw
6 yr. olds	sh, ch, j, th, -l
5 ½ yr. olds	-f, v, pl, bl, kl, gl, fl
5 yr. olds	l-
4 yr. olds	y-, t, tw, kw
3 ½ yr. olds	n, g, k, f-
3 yr. olds	m, h, w, p, b, d
1 yr. olds	No accurate speech sounds

Direction of increasing speech sound development for *boys*

Respondents	Responses to Items
9 ½ yr. olds	All speech sounds are accurate
9 yr. olds	spr, thr, skr, str
8 yr. olds	th, \r-, -er, pr, br, tr, dr, gr, kr, fr
7 yr, olds	-ng, s, z, th, sp, st, sk, sp, sm, sn, sw, sl, spl, skw, -l, j, ch, sh
6 yr. olds	l-, pl, bl, kl, gl, fl
5 ½ yr. olds	-f, v, tw, kw
5 yr. olds	y-
4 yr. olds	g
3 ½ yr. olds	t, k, d, f-
3 yr. olds	m, h, n, w, p, b, d
1 yr. olds	No accurate speech sounds

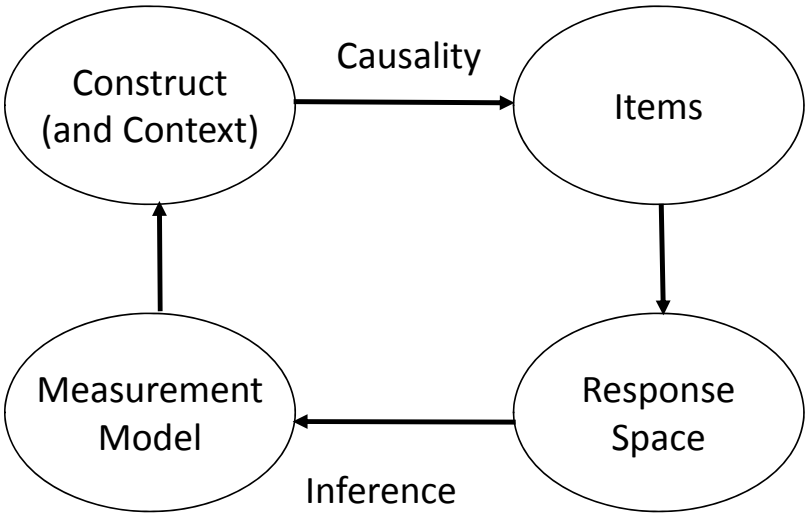
Instrument Construction

- 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

- Four components of instrument construction:
    1. Construct (and Context)
    2. Item Generation
    3. Response (Outcome) Space
    4. 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?
  - Build instrument purposively with this in mind, but pay attention to information gathered after-the-fact as to how well it is working

- 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? (write item 1, 2, 3…)
  - In reality, this is an iterative process…
- Items should be unambiguous
  - Cover a single concept (no 'ands') with a clear referent
- Items should be simple to process
  - Short, common vocabulary
  - Negatives can be harder to process – and 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)

Specificity of Response Space

Response options can be item-specific to maximize their utility:

Do you feel confident in explaining your religious beliefs to others?

_____ Not at all confident
_____ Mostly not confident
_____ Confident
_____ Very confident
_____ Totally confident

How often do you explain your religious beliefs to others?

_____ Never
_____ Once a year
_____ Every couple months
_____ Couple times a month
_____ Once a week,
_____ Couple times a week
_____ Everyday

How good are you at explaining your religious beliefs?

_____ I have no idea how to explain my beliefs
_____ I struggle a lot in explaining my beliefs
_____ I struggle a little in explaining my beliefs
_____ I am pretty good at explaining my beliefs
_____ I am very good at explaining my beliefs
_____ I am extremely good at explaining my beliefs

Item response formats DO NOT all have to be the same if you are using a latent trait model – you can and should customize them to be most informative for the question at hand.

Specificity of Response Space

Versus something like this:

- Sometimes I feel caught between wanting to buy things to make me look better in some way to others, when I really should be spending more money in ways that have more spiritual meaning.

_____ Strongly Disagree
_____ Disagree
_____ Somewhat Disagree
_____ Neither
_____ Somewhat Agree
_____ Agree
_____ Strongly Agree

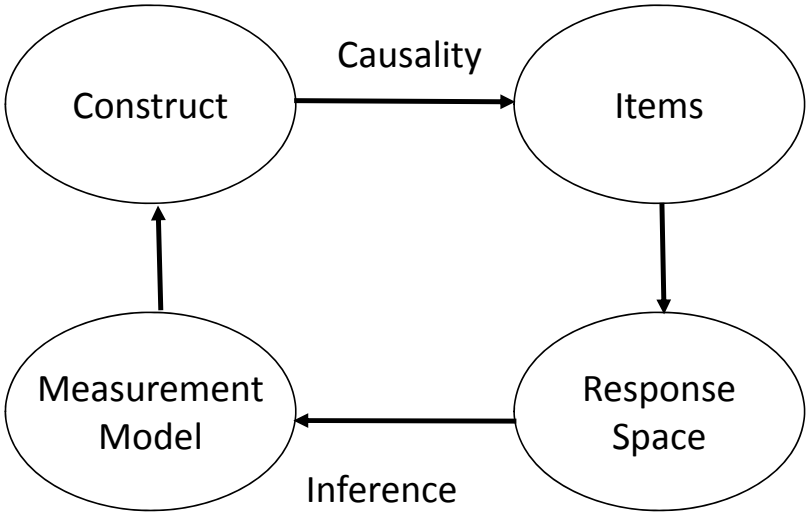
Another instance of what not to do:
unlabeled options:

1. "Never"
2. …
3. …
4. …
5. "Always"

Item-Level Measurement Models

- Type of response format will generally lend itself to an appropriate measurement model
  - Dichotomous (binary) item? (yes/no, Multiple Choice: correct/not)
    - Logistic/probit model (IRT)
    - Normal approximation (CFA) probably 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…**
      - The focus of this class
  - 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



Note that causality does NOT go through the measurement model – items would be caused by the construct regardless of response format, and thus regardless of the choice of measurement model.

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

Moving from Concept to Practice

- 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 flexible but 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

VALIDITY OF THE FACTOR

Scale Interpretation: Validity of Measure

- The interpretation of the scale starts to approach two very important topics in psychometrics, both centered around

- **Reliability:**
  - "Extent to which the instrument does what it is supposed to *with sufficient consistency* for its intended usage"
  - "Extent to which same results would be obtained from the instrument after repeated trials"

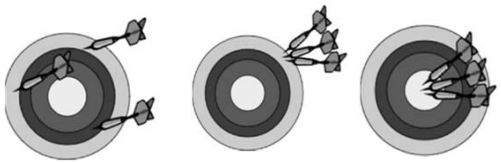
- **Validity:**
  - "Extent to which the instrument measures *what it is supposed to* (i.e., it does what it is intended to do)" or "Validity for WHAT?"
  - Is measure of degree, and depends on USAGE or INFERENCES
    - Scales are not "valid" or "invalid" – validity is NOT a scale property
    - e.g., Test of intelligence: Measure IQ? Predict future income?

Another Way to Think About Reliability and Validity

Factor score = true score + error (F = T + e)

- Error can be 'random'
  - Random error can be due to many sources (internal, external, instrument-specific issues, rater issues)
  - **Random error compromises reliability**
- Error can also be 'non-random'
  - Non-random error is due to constant source of variation that get measured consistently along with the construct (e.g., acquiescence)
  - **Non-random error compromises validity**

- In other words… reliability concerns how well you can hit the bulls-eye of the target…Validity concerns whether you hit the right target!



More about Validity

- The process of 'establishing' validity should be seen as building an argument:
  - To what extent can we use this instrument for its intended purpose (i.e., as a measure of construct X in this context)?

- Validity evidence can be gathered in two main ways:
  - Internal evidence
    - From construct map – does the empirical order of the items along the construct map match your expectations of their order?
  - External evidence
    - Most of CFA is focused on this kind of evidence
    - This will be our focus for now…

Historical Classification of Types of Validity

- In 1954, the American Psychological Association (APA) issued a set of standards for validity, defining 4 types
  - Predictive Validity
  - Concurrent Validity
  - Content Validity
  - Construct Validity

- Cronbach and Meehl (1955) then expanded (admittedly unofficially) on the logic of construct validity

Predictive and Concurrent Validity

- Predictive and concurrent validity are often categorized under 'criterion-related validity' (which makes it 3 kinds)
  - Predictive validity/utility: New scale relates to future criterion
  - Concurrent validity: New scale relates to simultaneous criterion

- Criterion-related validity implies that there is some known comparison (e.g., scale, performance, behavior, group membership) that is immediately and undeniably relevant
  - e.g., Does newer, shorter test 'work as well' as older, longer test?
  - e.g., Do SAT scores predict college success?
  - This requirement limits the usefulness of this kind of validity evidence, however…

Content Validity

- Content validity concerns how well a scale covers the plausible universe of the construct…
  - e.g., Construct: Spelling ability of 4th graders –
    Is the sample of words on this test representative of all the words they should know how to spell?

- 'Face validity' is sometimes mentioned in this context
  - Does the scale 'look like' it measures what it is supposed to?

- What might be some potential problems with 'establishing' these kinds of validity evidence?

The Big One: Construct Validity

- Extent to which scale can be interpreted as a measure of the latent construct (and for that context, too)
  - Involved whenever construct is not easily operationally defined…
  - Required whenever a ready comparison criterion is lacking…

- Depends on having a 'theoretical framework' from which to derive expectations…
  - The more elaborate the theoretical framework around your construct, the pickier you need to be…

Construct Validity: 3 Steps for Inference

- Predict relationships with related constructs
  - Convergent validity
    - Shows expected relationship (+/-) with other related constructs
    - Indicates “what it IS” (i.e., similar to, the opposite of…)
  - Divergent validity
    - Shows expected lack of relationship (0) with other constructs
    - Indicates “what it is NOT” (unrelated to…)
- Find those relationships in your sample
  - No small task…
- Explain why finding that relationship means you have shown something useful
  - Must argue based on ‘theoretical framework’

3 Ways to Mess Up a Construct Validity Study…

1. Is your instrument broken?
   - Did you do your homework, pilot testing, etc?
   - Did you measure something reliably in the first place? Reliability precedes validity, or at least examination of it does
   - Is that something the right something (evidence for validity)?

2. Wrong theoretical framework or statistical approach?
   - Relationships really wouldn't be there in a perfect world
   - Or you have the wrong kind of sample given your measures
   - Or you lack statistical power or proper statistical analysis
     - Watch out for discrepant EFA-based studies…

The 3rd Way to Mess Up a Construct Validity Study…

3. Did you fool yourself into thinking that once the study (or studies) are over, that your scale "has validity"?

- **SCALES ARE NEVER "VALIDATED"!**
- Are the items still temporally or culturally relevant?
- It is being used in the way that's intended, and is it working like it was supposed to in those cases?
- Has the theory of your construct evolved, such that you need to reconsider the dimensionality of your construct?
- Do the response anchors still apply?
- Can you make it shorter or adaptive to improve efficiency?

The Last Words I Will Utter About Validity

- Reliability is a precursor to validity…
  - You cannot have a valid scale if the scale is not measuring anything reliably
  - You will sometimes hear people state that to increase validity, you need to decrease reliability – this is dead wrong

- Most approaches to validity are largely external…
  - Depend on detecting expected relationships with other constructs, which can be found or not for many other reasons besides problems with validity
  - This kind of externally-oriented validity is "nomological span"

- In my opinion, arguing the validity of a test is a little like splitting hairs
  - The most absurd examples are the only ones that aren't valid (e.g., using a tape measure to get an estimate of happiness)
  - Most tests can be shown to be "valid" using some type of body of evidence

WRAPPING UP

Wrapping Up

- This lecture was about the non-technical aspects of building psychometric instruments
  - These are going to always be assumed as we go forward
  - The strength of the instrument depends on these features

- Up next are the technical aspects of factor analysis