Introduction to Latent Trait Measurement Models (LTMM)

- Today's Class:
  - Test Theory—definitions and historical context
  - Latent trait measurement models (LTMM)
    - Confirmatory Factor Analysis (CFA)
    - Item Response Theory (IRT) and Item Factor Analysis (IFA)
  - Advantages and disadvantages of LTMM framework
  - Item and scale construction

Test Theory

- Test theory is an abbreviated expression for:
  - "Theory of Psychological Tests and Measurements"
  - Or "Psychometric Theory" (even when not used in Psychology)

- Test theory is a general collection of **statistical models** for evaluating the development and use of instruments
  - **Operationalize** practical problems in measurement
  - **Provide answers to** practical problems in measurement

- 3 'branches' of measurement models for latent traits that are inter-related… you actually know one of these already

Classical Test Theory (CTT)

- What you have learned about measurement so far pry* falls under the category of CTT:
  - Writing items and building scales
  - Item analysis
  - Score interpretation
  - Evaluating reliability and construct validity

- Big picture: We will view CTT as model with a restrictive set of assumptions within a more general family of latent trait measurement models

**pry = "probably" in my midwestern vernacular*

What is a ‘latent trait’?

- **Latent trait** = Unobservable ability or trait
  - e.g., “Intelligence”, “Extroversion”, “Depression”

- But how can we measure something unobservable?
  - Build **measurement models**!

- Big picture: Latent traits can be measured using observed behaviors or responses (“indicators”)
  - Common part of the variance across items that measure the same thing is supposed to measure the latent trait
  - But not all constructs should use latent trait measurement models! (e.g., formative vs. **reflective indicators**)

Differences Among Measurement Models

- What is the **name of the latent trait** measured by a test?
  - Classical Test Theory (CTT) → "True Score" (T)
  - Confirmatory Factor Analysis (CFA) → "Factor Score" (F)
  - Item Factor Analysis (IFA) → "Factor Score" (F)
  - Item Response Theory (IRT) → "Theta" (θ)
- Fundamental difference in approach:
  - **CTT → unit of analysis is the WHOLE TEST** (item sum or mean)
    - **Sum = latent trait**, and the sum doesn't care how it was created
    - Only using the sum requires restrictive assumptions about the items
  - **CFA, IFA, IRT, and other LTMM → unit of analysis is the ITEM**
    - Model of how item response relates to an **estimated latent trait**
    - Different names of models for differing item response formats
    - Provides a framework for testing adequacy of measurement models

Latent Trait Measurement Models (LTMM)

- Families of latent trait measurement models are labeled based on their indicators' response format:
  - Continuous responses? → Confirmatory Factor Models
  - Categorical responses? → Item Response Theory or Item Factor Models
  - Measurement models for other responses exit too (like counts), but they don't necessarily have special names

- Other relevant, related terms:
  - "Structural Equation Modeling" (SEM) is correlation or regression among the latent traits defined by the measurement models
    - Things that can go wrong in SEM most often reflect problems with the measurement models—that is why we spend most of the semester on this!
  - "Path Analysis" is just regression among observed variables only
  - "Mediation" is regression with a better marketing campaign
  - "Moderation" is an interaction term with a better marking campaign

Problems in Test Theory…

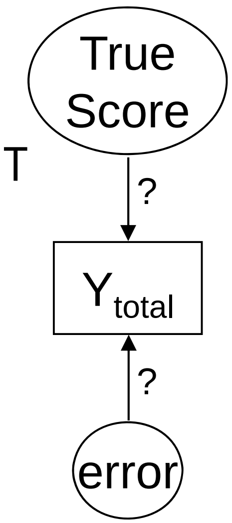
- Motivated by problems in education and psychology
  - Education → Assessment, Psychology → Understand structure
  - Several independent groups → Piecemeal approach
- Theory developed largely before mainstream availability of high-powered computers
  - Had rationale, but couldn't be solved computationally
  - That's how many 'approximations' or 'devices' came about (many of which are still used today, for better or for worse)
- Then mathematicians tried to help…
  - But published works that were written for other mathematicians
- A little historical context…

History of Test Theory

- Begins about the mid 19th century in psychophysical laboratories considered with measuring intelligence
- 1904: Charles Spearman published 2 seminal papers
  - One showed how to estimate amount of error in test scores
    - Gave rise to classical true score theory (**classical test theory**)
  - One showed how to recognize from test data that the tests measure just one psychological attribute in common
    - How to measure Spearman's "G"
    - Led to development of **common factor theory** (CFA)

Classical Test Theory (CTT)

- In CTT, the **TEST** is the unit of analysis: $\mathbf{Y_{total} = T + e}$
  - **True score T:**
    - Best estimate of 'latent trait': Mean over infinite replications
  - **Error e:**
    - Expected value (mean) of 0, expected to be uncorrelated with T
    - e's are supposed to wash out over repeated observations
  - **So the expected value of T is $\mathbf{Y_{total}}$**
  - In terms of observed variance of the test scores:
    - Observed variance = true variance + error variance



- Goal is to quantify ***reliability***
    - Reliability = true variance / (true variance + error variance)

- Because the CTT model does not include individual items, **items must be assumed exchangeable** (and more items is better)

Classical Test Theory, continued

- CTT unit of analysis is the WHOLE TEST (sum of items)
  - Want to ascertain how much of observed test score variance is due to 'true score' variance versus 'error' variance
  - Quantify 'error variance' in various ways
  - 'Error' is a unitary construct in CTT (and error is 'bad')
  - Goal is then to reduce 'error' variance as much as possible
    - Standardization of testing conditions (make confounds constants)
    - Aggregation → more items are better (errors should cancel out)
  - Items are exchangeable; item properties are NOT taken into account in indicating the latent trait of a given person (which is just the sum)
- Followed by *generalizability theory* to decompose error
  - e.g., rater variance, person variance, time variance

Classical Test Theory, continued

- Brief history of solutions for quantifying reliability:
  - 1904: Spearman: from alternate forms or test-retest
  - 1945: Guttman: from the relations between the items within a test (i.e., coefficient alpha)
  - 1951: Cronbach further developed Guttman's work → "Cronbach's alpha"
    - Called "Guttman-Cronbach alpha" by McDonald (and no one else)
    - Cronbach's work further elaborated into generalizability theory
  - 1950: Gulliksen classic text for CTT
    - See also Nunnally's texts from the 1970's - 1990's
- More CTT specifics in upcoming classes…
- *Next, tracing the other contribution of Spearman…*

Confirmatory Factor Analysis (CFA) Models

- Main idea: Build a measurement model of which response variables should 'go together' to measure the same thing
  - **CFA = Linear regression model** predicting each continuous observed outcome variable (item, subscale) from a latent trait predictor variable(s)

- Differs from exploratory factor analysis (is NOT a model):
  - YOU impose the number and content of factors
  - Alternative models are COMPARABLE and TESTABLE

- Uses of confirmatory factor analysis models:
  - Analyze relationships among subscales that have normal, continuous distributions (or "incorrectly" to analyze item-level data)
  - Any LTMM provides comparability across persons, items, and occasions

Confirmatory Factor Analysis (CFA)

- **The CFA unit of analysis is the ITEM (as in any LTMM):**

  $\mathbf{y_{is} = \mu_i + \lambda_i F_s + e_{is}}$ → **both items AND subjects matter**

  - Observed response for item *i* and subject *s*
    = intercept of item *i* (μ)
    + subject *s*'s latent trait/factor (F), item-weighted by λ
    + error (e) of item *i* and subject *s*

Should look familiar…
$y_{is} = \beta_{0i} + \beta_{1i} X_s + e_{is}$

- **Dimensionality** is assumed known (usually 1 latent trait per item)
  - Local Independence → e's are independent after controlling for factor(s)
  - The factor is the reason why item responses were correlated in the first place

- **Linear model** → a one-unit change in latent trait/factor F has same increase in expected response Y at all points of Y
  - Won't work well for binary or categorical data… thus, we need another LTMM

- Items are allowed to differ from each other in how much they relate to the latent trait, *but a good item is equally good for everybody*

"Common Factor Theory" (→CFA)

- 1900's: Spearman's G
  - Went looking for single-factor model… and "found" it
  - Led to development of other IQ tests (Stanford-Binet, Wechsler)
- 1930's and 1940's: Thurstone elaborated Spearman's model into a "multiple factor" model
  - Beginnings of exploratory factor analysis to do so
  - Later applied in other personality tests (e.g., MMPI)
- 1940's and 1950's: Guttman's work
  - Factor analysis and test development is about generalizing from measures we have created to more measures of the same kind
  - Thus, need to think about measurement structure before-hand

Common Factor Theory, continued

- 1940's: Lawley → rigorous foundation for statistical treatment of common factor analysis
  - But had to wait for better computers to be able to do it!
- 1952: Lawley → beginnings of confirmatory factor model
  - Later extended by Howe and Bargmann (1950's)
  - Further extended by Jöreskog (the King of LISREL – 1970's)
- But this linear model pry should not be applied to binary, ordinal, or other not-normal responses…
  - Probability of correct response will go out of bounds
  - Errors can't be normally distributed with constant variance
- So then what? Item Response Theory to the rescue…
  - *aka*, LTMM for generalized response formats

Item Response Theory (IRT) Models

- IRT resulted from combination of ideas from factor analysis and phi-gamma law of psychophysics
  - When detecting stimuli of varying intensity (e.g., light), the response follows a smooth, S-shaped curve that can be represented by the cumulative normal distribution
  - That response function also works to model probability of a correct response given (1 to 4) model parameters

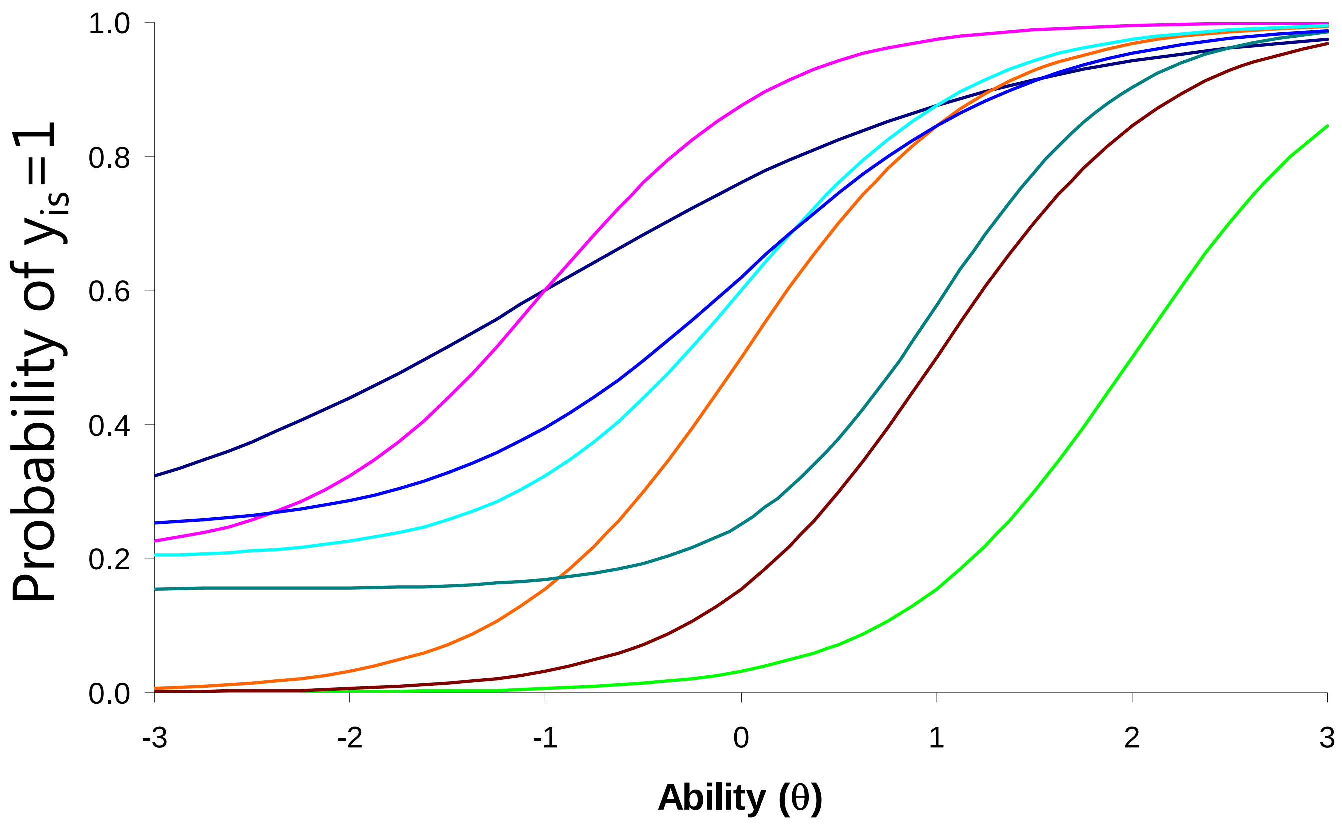
- 1950: Lazarsfeld: Introduced "latent structure analysis" → factor analysis for binary item responses
  - Beginnings of item response theory (which is not a theory per se, but a set of latent trait measurement models)

Item Response Theory (IRT) Models

- Linear regression is to confirmatory factor models as to:
  - Logistic regression is to binary IRT models
  - Ordinal/nominal regression is to "polytomous" IRT models
  - IRT = linear model predicting each categorical observed outcome variable from latent variables using link functions
- A "Rasch model" is a restricted subset of an IRT model (but don't let any Rasch people hear you saying that)
- Uses of IRT models:
  - *Correctly* analyze item-level data (binary items, Likert scales)
  - Examine sensitivity of measurement across range of latent trait
  - Any LTMM provides comparability across persons, items, and occasions

Item “Characteristic” Curves

a = Discrimination = slope of ‘line’
b = Difficulty = location of ‘line’
c = Lower Asymptote of ‘line’
d = Upper Asymptote of ‘line’



Item Response Theory, continued

- The **IRT** unit of analysis is the individual **ITEM** (as in any LLTM) $\mathbf{Logit(y_{is}) = 1.7a_i(\theta_s - b_i)}$ → both items AND subjects matter
  - Items and persons are located on the same latent metric
  - Probability of getting an item right depends (at least) on the subject's ability ($\boldsymbol{\theta_s}$ **= "Theta"**) and the item's difficulty ($\mathbf{b_i}$), weighted by its discrimination ($\mathbf{a_i}$, how related the item is to the latent trait)
  - **"Item factor analysis" (IFA)** re-arranges the model into something that looks more like CFA (and usually uses limited information estimation)

- All items are NOT created equal (not exchangeable)
  - Having items that differ in their properties is a GOOD THING, because you can customize tests for different groups or purposes
  - Reliability ("information") varies across ability level, and depends specifically on how well the items' difficulty matches subjects' ability

Item Response Theory, continued

- 1952: Lord's seminal paper: Spearman's single-factor model can be applied to dichotomous items
  - Binary responses modeled by normal ogive function ("probit")
  - Later work used logit link instead (logit ≈ probit*.17)
  - Elaborated in 1960's by Birnbaum

- 1968: Lord & Novick → first CTT text to also include IRT
  - Well-connected to emerging scholars in both educational testing and psychometric methods… and BOOM…

- 1960: Separate work by Rasch (common 'a' parameter)
  - Restricted IRT model, but with highly desirable properties…
  - … and different philosophical viewpoint

Unified View of Test Theory

- Classical test theory can be viewed as a restricted form of the common factor model, but the focus is the TEST…
  - Originated by Spearman, elaborated by Thurstone, formalized by Lawley, and made practical by Jöreskog
- Item response (and Rasch) models are common factor models used for binary/ordinal responses…
  - Developed by Lord, Birnbaum, and Rasch and their students
- Common factor models (CFA) are for continuous data…
  - Approximation for ordinal data with varying degrees of success
- Latent traits can also be indicated by other kinds of non-normal item responses (count, zero-inflated, two-part)….
  - But they don't have special names (I'd say "generalized SEM"?)

Advantages of LTMM Framework (CFA, IRT, IFA, and beyond)

- Explicit, testable models of dimensionality
- Concrete guidelines for selecting items to builds scales
- Assess measurement sensitivity across range of latent trait (i.e., know where the 'holes' are)
- Provide comparability across persons, items (different forms scales or different scales), and occasions
- Examine comparability across groups or repeated measures
  - Confirmatory factor analysis → "Measurement invariance"
  - Item response theory → "Differential item functioning"
- Internal and external evidence for construct validity
- Generalized measurement models can even accommodate different response formats within the same instrument

Disadvantages of LTMM Framework

- Primary: Required sample size
  - Casts of 100s for sure, and preferably 1000s
  - Uses maximum likelihood (Limited info WLSMV estimator in Mplus can also be used for multidimensional IRT models)

- Technical difficulties
  - Estimation is harder, especially in multidimensional IRT
  - References written in Greek (literally)
    - Except your textbook and selected readings, so please read them!

- Misnomers about what LTMM (SEM) can do…
  - Bad items are still bad items, no matter what model is used

Summary: LTMM Introduction

- Test Theory = a set of statistical models used to evaluate how well an instrument measures a trait(s)
- The branches we will cover:
  - "Classical Test Theory" (CTT)
    - Just add the items up: Focus on TEST as unit of analysis
    - Simple, yet very restrictive; requires belief instead of evidence
    - (We'll stop by EFA just for a point of reference, too)
  - "Latent Trait Models" (CFA, IRT, IFA… and beyond)
    - Estimate a latent trait; Focus on ITEM as unit of analysis
    - Flexible models that differ by response format of items
    - More complex, but more powerful and useful

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:
  - A teacher wishing to evaluate student knowledge of math
  - A psychologist wishing to measure depression

- Note the common denominator here is not the topic, but rather than each example 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 a math ability
- 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 equally good measures of math ability or are some items better than others?
- Are there other ways of getting the right answer besides ability?
- Would different items 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 give 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 floors and ceilings necessarily bad things?
- 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?

Example #2 – The Psychologist

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

- The psychologist may have similar measurement questions…
  - 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: How many traits do these items measure?
  - Here's a tip: if the trait name has a slash or an and, it's not a single trait!
- Overall test accuracy vs. efficiency: Add or remove response options?
  - Do you need to add or remove items? Just any items? Or targeted items?
- Reliability across trait levels: How is the trait distributed?
  - How to write enough items to avoid ceiling and floor effects?
  - How to customize test for specific measurement purposes?
- Bias and generalizability: 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 pry need another kind of measurement model:
  - Diagnostic Classification Models (Rupp, Templin & Henson, 2010)
    - Measure categorical attributes or skills, not continuous traits
    - Useful when *classification* is the goal of measurement (not trait amount)

Construct Maps should include…

- Coherent, substantive definition of the construct
- An underlying continuum is manifested in two 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…

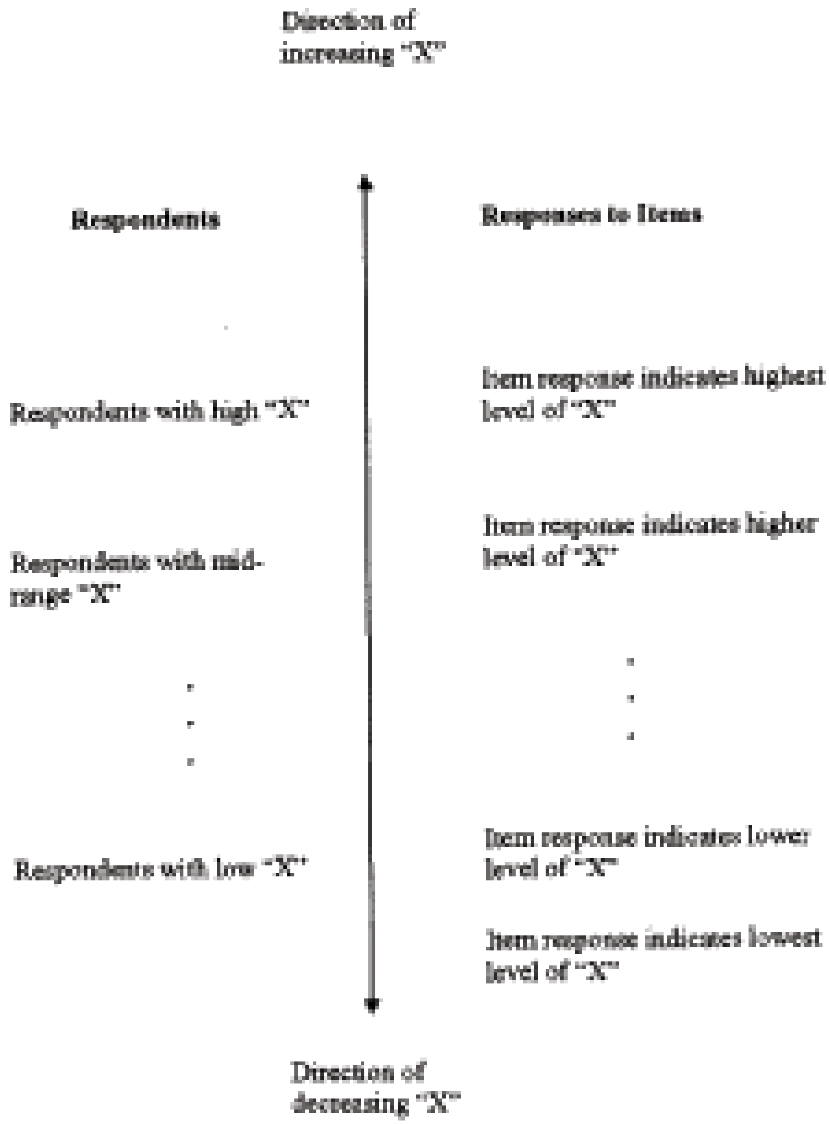


FIG. 2.1 A generic construct map in construct "X."

From Wilson (2005)

Template for a Construct Map

<u>Left = PERSONS</u>
- qualities
- characteristics

<u>Right = ITEMS</u>
- 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

Construct Map for Standardized Interviewing

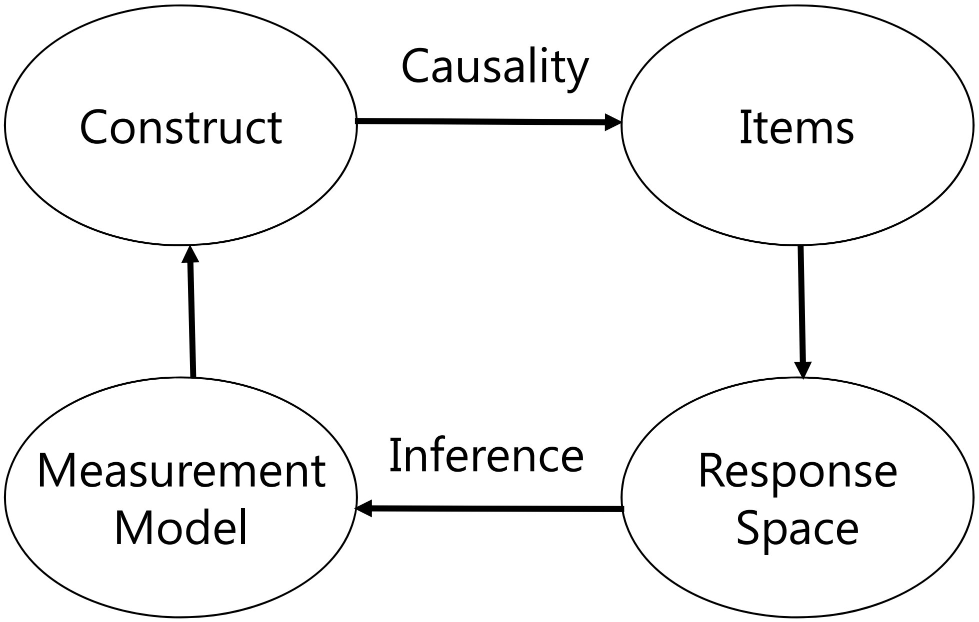
Types of people	**Item response options**
ATA-certified SLLs specifically trained to work with surveys	Can translate survey questions, maintaining standardization of question wording
SLLs who are certified by the American Translators Association (ATA)	Can translate documents from second language into first language
SLLs who have studied both languages and have studied translation theory	Can revise translated documents
SLLs with at least 5 years of language study	Can write in the first and second language
SLLs with at least 3 years of language study	Can speak in the first and second language
SLLs with at least 1 year of language study	Can read in the first and second language
An individual with at least 10 years of educ	Can write in at least one language
Any literate individual	Can read in at least one language
Anyone over the age of two who has not been raised in isolation	Can speak at least one language

SLl = Second Language Learners

Instrument Construction

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

4 Instrument Building Blocks



The direction of 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.

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, fraught with trial and error…

- Items should be unambiguous
  - Cover a single concept (no 'ands') with a clear referent

- Items should be simple to process (short, simple)
  - Negatives can be harder to process; research has suggested negatively-worded (reverse-coded) items are less discriminating
  - Do NOT confound item stem/valence with construct!

- Good items should span the full range of construct… but not be too narrow ("bloated specific") 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

Example: Confounded Valence and Construct

Nonacceptance of Emotional Responses
11. When I'm upset, I become angry with myself for feeling that way.
12. When I'm upset, I become embarrassed for feeling that way.
21. When I'm upset, I feel ashamed with myself for feeling that way.
23. When I'm upset, I feel like I am weak.
25. When I'm upset, I feel guilty for feeling that way.
29. When I'm upset, I become irritated with myself for feeling that way.

Difficulties in Engaging in Goal-Directed Behaviors
13. When I'm upset, I have difficulty getting work done.
18. When I'm upset, I have difficulty focusing on other things.
20. When I'm upset, I can still get things done. (R)
26. When I'm upset, I have difficulty concentrating.
33. When I'm upset, I have difficulty thinking about anything else.

Impulse Control Difficulties
3. I experience my emotions as overwhelming and out of control.
14. When I'm upset, I become out of control.
19. When I'm upset, I feel out of control.
24. When I'm upset, I feel like I can remain in control of my behaviors. (R)
27. When I'm upset, I have difficulty controlling my behaviors.
32. When I'm upset, I lose control over my behaviors.

Lack of Emotional Awareness
2. I pay attention to how I feel. (R)
6. I am attentive to my feelings. (R)
8. I care about what I am feeling. (R)
10. When I'm upset, I acknowledge my emotions. (R)
17. When I'm upset, I believe that my feelings are valid and important. (R)
34. When I'm upset, I take time to figure out what I'm really feeling. (R)

Limited Access to Emotion Regulation Strategies
15. When I'm upset, I believe that I will remain that way for a long time.
16. When I'm upset, I believe that I'll end up feeling very depressed.
22. When I'm upset, I know that I can find a way to eventually feel better. (R)
28. When I'm upset, I believe there is nothing I can do to make myself feel better.
30. When I'm upset, I start to feel very bad about myself.
31. When I'm upset, I believe that wallowing in it is all I can do.
35. When I'm upset, it takes me a long time to feel better.
36. When I'm upset, my emotions feel overwhelming.

Lack of Emotional Clarity
1. I am clear about my feelings. (R)
4. I have no idea how I am feeling.
5. I have difficulty making sense out of my feelings.
7. I know exactly how I am feeling. (R)
9. I am confused about how I feel.

The "lack of emotional awareness" scale has only reverse-coded items. So these items could be correlated (i.e., seem to indicate a trait) due to their content OR their valence.

In addition, the first items on the scale do not have the referent "when I'm upset"—this could cause them to be responded to differently than the rest of the later scale items that have a different, more specific, referent.

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 if relevant)

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 an LTMM. 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"

- More response options are only better if the categories stay distinguishable! Including more items instead will result in more information.
- Also, if you don't know what to call the middle categories, how are people supposed to know when to use them??

Item-Level Measurement Models

- Type of response format will generally lend itself to an appropriate latent trait measurement model
  - Binary item? (yes/no, MC → correct/not)
    - Logistic/probit model (IRT; IFA)
    - Normal approximation (CFA) pry won't work very well
  - Polytomous (quantitative) item? A few IRT options…
    - Graded response or 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 easy measurement model for many other types of item choices (i.e., forced choice, rankings)
    - Avoid ipsative response formats (e.g., rankings) if you can!

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 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