Introduction to this Course and to Latent Trait Measurement Models (LTMM)

- Topics:
  - What to expect this semester
  - 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
  - Advice about item and scale construction

What to Expect This Semester

- I believe that **everyone is capable** and **can significantly benefit**** from learning more quantitative methods!
  - Our focus is measurement models (which are indeed statistical models)

- **Philosophy:** Focus on accessibility + mastery learning
  - No anxiety-prone tasks (e.g., hand calculations, memorizing formulas)
  - No anxiety-prone methods of evaluation (e.g., timed tests)

- **Materials:** Unit = (wordy) lecture + example(s); 9 units planned
  - **Lecture** slides present concepts—the **what** and the **why**
  - **Example** documents: reinforce the concepts and demonstrate the **how using software**—Mplus (and to some extent R lavaan)
  - All available at the course website (hosted outside of ICON)

** ***Benefits*** *include but are not limited to: Better research, more authorship opportunities, and actual money*

Course Requirements

- **All course requirements are take-home, open-note, and untimed**
- Late* work will be accepted (−2 for HW; −1 for FA)
  - **Extensions will be granted if requested at least 2 weeks in advance*
  - HW due dates **may be pushed later** (to ensure approximately 1 week after covering the material before HW is due), but never sooner
- **5 formative assessments (FA, 10 points) in ICON:** Top-of-head questions and story problems for structured review (will discuss answers in next class)
  - 2 points for an honest attempt to complete each FA (mostly without feedback)
  - An opportunity for you to request topics for further clarification and review
- **6 homework assignments (HW, 90 points):** Practice doing data analysis
  - Based directly on examples given (no googling or ChatGPT should be needed)
  - Online versions (HW 2, 4, and 6) will have two parts:
    - **Computation** sections: Instant feedback, infinite attempts
    - **Results** sections (multiple-choice questions): Delayed feedback, single attempt (but repetition of concepts and vocabulary across the semester)
  - Written versions (HW 1, 3, and 5) will each have one opportunity for **revision**
    - Practicing fitting measurement models on item-level data (ideally that you care about)

Homework using Your Own Data

- HW 1, 3, and 5 will require individual-specific item-level data
  - At least 6 items thought to measure one latent trait or 8 items thought to measure 2 latent traits
  - Any response format (ordinal, slider, RT, etc)
  - No sample size requirement, but ideally >100 respondents
  - Preferable source: data from your own research area that you care about and want to do something with anyway
  - Otherwise: any publicly-available data you can find, such as through the International Personality Item Pool; archives at ICPSR, Berkley, or Harvard; Healthy Minds Data; or Early Childhood Data

- Not sure if the data you want to use will work? Ask me!
  - Describe psychometric evidence for your items: HW1 due 2/12
  - Do analyses with your items: HW3 due 3/25 and HW5 due 4/22

Our Other Responsibilities

- My job (besides providing materials and assignments):
  - **Answer questions** via email, in individual meetings, or in group-based zoom office hours—you can each work on homework during office hours and get (near) immediate assistance (and then keep working)
    - Email me first (but you can follow up with the TAs if they help you)

- Your job (in descending order of timely importance):
  - **Ask questions**—preferably in class, but any time is better than none
  - **Frequently review** the class material, focusing on mastering the vocabulary (words and symbols), logic, and procedural skills
  - Don't wait until the last minute to start homework, and don't be afraid to **ask for help if you get stuck** on one thing for more than 15 minutes
    - Please email me a screenshot of your code+error so I can respond easily
  - **Do the readings** for a broader perspective and additional examples (best after lecture; readings are for the whole unit, not just that day)
  - **Practice** using the software to implement the techniques you are learning **on data you care about**—this will help you so much more!

More About Your Experience in this Class

- **Attendance**: Strongly recommended but not required
  - **You choose each class**: In-person "roomer" or online "zoomer"
  - **Masks** are still welcome for in-person attendees
  - **Please do not attend in-person if you might be sick!**
  - You won't miss out: I will post YouTube-hosted **recordings** (audio + screenshare only) for each class at the course website
  - **Ask questions aloud or in the zoom chat window (+DM)** (even if you are attending class in-person as a "roomer")

- **Changes** will be sent via email by 9 am on class days
  - I will change to zoom-only if I am exposed to Covid!
  - I will change to zoom-only for dangerous weather
  - Nothing is more important than our health and safety…

Class-Sponsored Statistical Software

- I will show examples using **M*plus*** (currently v. 8.10)
  - Mplus is expensive to purchase, but it is available for free to course participants through the UIowa Virtual Desktop
  - Why? Because it's the only package that does everything I plan to cover, with the best integration into structural equation models
  - Also, Mplus syntax is (relatively) easy to follow and replicate

- That being said, M*plus* is not the only option:
  - R program lavaan can estimate some of the models covered, and can be used for some homework (canned or your own data)
  - STATA SEM or GSEM can be used to analyze your own data, but cannot be used for all homework (different missing data routines)
  - SAS CALIS and SPSS AMOS can only do models for continuous responses (as far as I know), so these won't work for our purposes

This Semester's Topics

- Section 1: Concepts and Old-School Techniques
  - Introduction to latent trait terminology; (dimensionality assessment via PCA and EFA); reliability assessment via CTT
- Section 2: Latent Trait Measurement Models
  - For continuous or continu-ish responses (CFA)
  - For other responses (IRT/IFA/Generalized SEM)
    - Please review binary/ordinal regression from PSQF 6270 first!
  - Invariance across groups or occasions either way
- Section 3:
  - Multidimensional measurement in practice—higher-order factor models, method factor models, and bifactor models
  - Path analysis and structural equation models for examining relationships among observed or latent variables (the finale!)

Latent Traits Need 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
  - So yes, measurement models are indeed statistical models!

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

Classical Test Theory (CTT)

- What you first learned about measurement pry* falls under the category of Classical Test Theory (CTT):
  - Writing items and building scales (or "tests")
  - Item analysis for differentiating "good" from "bad" items
  - Evaluating dimensionality underlying the items
  - Interpretating scale or test "scores"
  - Evaluating reliability and construct validity

- Big picture: We will view CTT as a 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 construct ("factor")
  - Many types of variables: ability, attitude, tendency, etc.
  - e.g., "Intelligence", "Extroversion", "Depression"

- But how can we measure something unobservable?
  - Build **measurement models** by which to represent them!

- Big picture: Latent traits can be measured using observed responses → "**items**" or "**indicators**"
  - A new latent variable is created from the common variance across indicators thought to measure the same construct
  - But not all constructs should use latent trait measurement models! (e.g., formative vs. **reflective indicators**)

Differences Among Latent Trait Measurement Models (LTMMs)

- What do we call **the latent trait** measured by the indicators?
  - 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**, so items and persons are inherently tied together → bad
    - Only using the sum requires restrictive assumptions about the items
  - **CFA, IFA, IRT, and other LTMMs → unit of analysis is the ITEM**
    - Model of how item response relates to a **separately estimated latent trait**
    - Provides way of separating item and person properties → good for flexibility
    - Different names of models are used for differing item response formats
    - Provides a framework for testing adequacy of measurement models

Latent Trait Measurement Models (LTMMs)

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

- 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 just regression with a better marketing campaign
  - "Moderation" is an interaction term with a better marking campaign

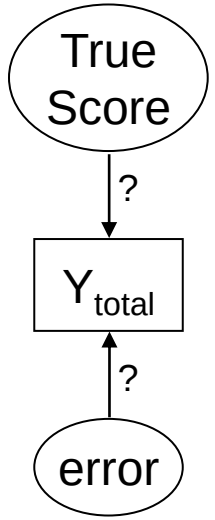
A Brief History of Test Theory…

- Motivated by problems in education and psychology
  - Education → Assessment of academic abilities
  - Psychology → Understand structure of intelligence or personality
  - Piecemeal approach; also barriers from technical presentation
  - Theories developed before availability of computing power, so approximations were developed that could actually be used (with remnants that unfortunately still get used, like alpha and EFA)

- 1904: Charles Spearman published two seminal papers
  - One showed how to estimate amount of error in test scores
    - Led to classical true score theory (aka, classical test theory)
  - Other showed how to recognize from test data that the tests measure just one psychological attribute in common ("G")
    - Led to common factor theory (aka, confirmatory factor analysis)

Classical Test Theory (CTT)

- In CTT, the **TEST** is the unit of analysis: $\boldsymbol{Y_{total} = T + e}$
  - **True score $\boldsymbol{T}$:**
    - Best estimate of latent trait: Mean over infinite replications
  - **Error $\boldsymbol{e}$:**
    - Expected value (mean) of 0; uncorrelated with T
    - Errors are supposed to wash out over repeated observations
  - **So the expected value of $\boldsymbol{T}$ is $\boldsymbol{Y_{total}}$**
  - 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 item responses, **items must be assumed exchangeable** (and more items is better)

Classical Test Theory, continued

- CTT unit of analysis is the TEST or SCALE (sum/mean of items)
  - Want to quantify how much of observed test score variance is due to "true score" variance versus "error" variance
  - "Error" is a unitary construct in CTT (and error is always 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 distinguish kinds of error
  - e.g., item variance, person variance, rater variance, occasion variance
  - Modern analog: mixed-effects (multilevel) models with crossed random effects for each (random) sampling dimension and their interactions

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
    - And no, a good alpha doesn't mean anything—stay tuned for why!
  - 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 indicators should '"go together" to measure the same thing
  - **CFA = Linear regression model** predicting each continuous observed outcomes ("indicators") from **latent** trait (unobserved) predictor(s)

- Differs from exploratory factor analysis (that is NOT a model):
  - In CFA *you* impose the number and content of factors
  - In CFA alternative models are COMPARABLE and TESTABLE

- Uses of confirmatory factor analysis models:
  - Analyze relationships among indicators that have normal, continuous distributions (or "incorrectly" to analyze ordinal response indicators)
  - Provide separation of persons, items, and occasions (as in any LTMM)

Confirmatory Factor Analysis (CFA)

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

  $$y_{is} = \mu_i + \lambda_i F_s + e_{is}$$ → **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** → part of the model (usually 1 latent trait per item)
  - Local Independence → e residuals are independent after controlling for factor(s)
  - The factor is the reason why item responses were correlated in the first place!
  - If not, you can **augment the model** to address unintended multidimensionality

- **Linear model** → a one-unit change in latent trait/factor F_s creates same increase in the expected response y_{is} along all points of y_{is}
  - Won't work well for binary or ordinal data… thus, we need another LTMM

- Items can now differ from each other in how much they relate to the latent trait, *but a "good item" is assumed equally good for everybody!*

A Brief History of Common Factor Theory

- 1900's: Spearman's "G" single-factor models
  - Development of techniques designed to find a common factor
  - Led to development of other IQ tests (Stanford-Binet, Wechsler)

- 1930's and 1940's: Thurstone elaborated Spearman's "G" unidimensional 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

A Brief History of Common Factor Theory

- 1940s: 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 in1970's)

- But this linear model pry should not be applied to binary, ordinal, or other not-continuous responses…
  - Predicted response will go past possible response options
  - 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 another set of latent trait measurement models)

Item Response Theory (IRT) Models

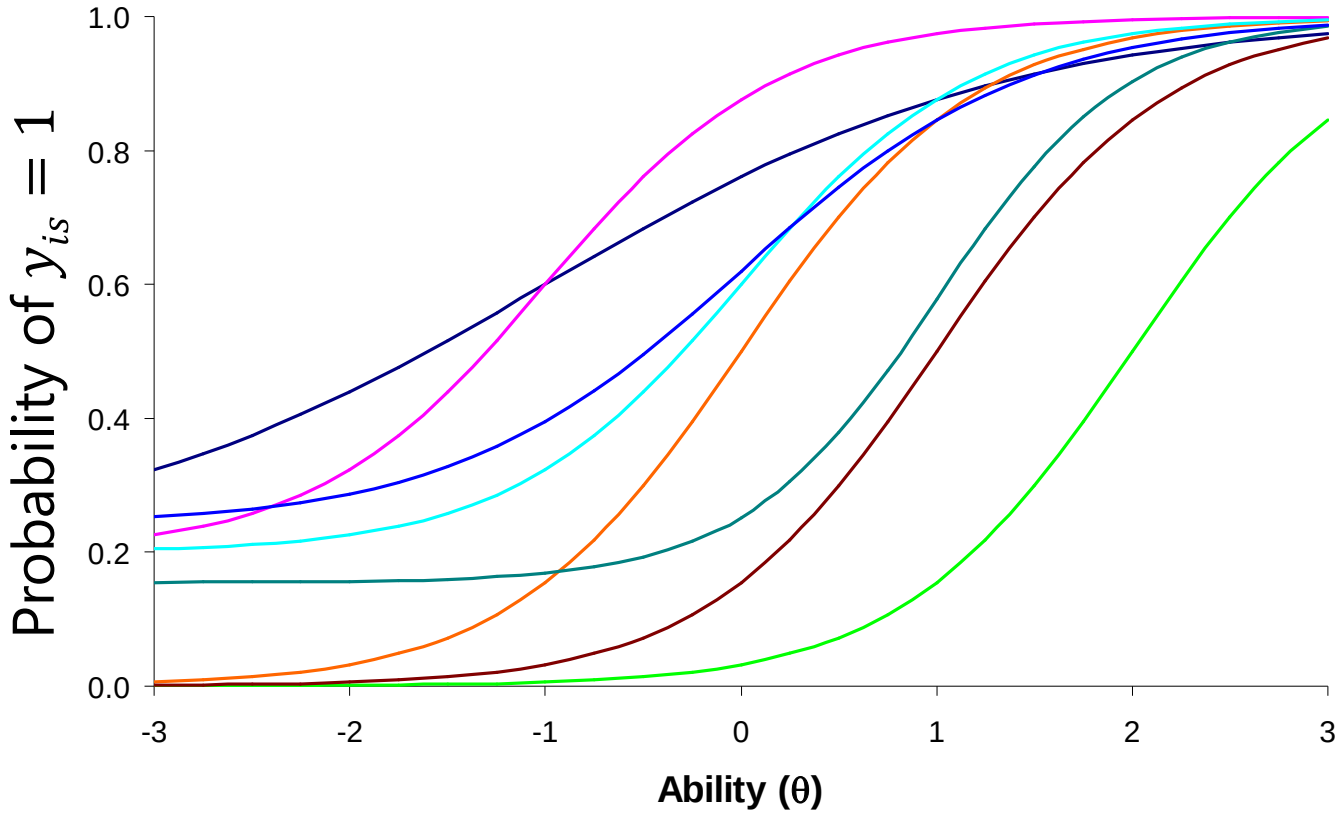
- Linear regression is to confirmatory factor models as to:
  - Logistic regression is to binary IRT models
  - Ordinal/nominal regression is to "polytomous" IRT models
  - IRT = generalized linear model predicting each categorical observed outcome indicator from latent predictors using link functions
  - Term "IRT" usually goes with full-information estimation (use all data)

- A "Rasch model" is a restricted version of an IRT model (but don't let any Rasch people hear you saying that)

- Uses of IRT models:
  - *Correctly* analyze categorical indicators (binary, ordinal, or nominal)
  - Examine sensitivity of measurement across range of latent trait
  - Provide separation of persons, items, and occasions (as in any LTMM)

Item "Characteristic" Curves

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



Note: Theta ability has a nonlinear relationship to probability of item response, but a linear relationship to its link-transformed mean…

Item Response Theory, continued

- The **IRT** unit of analysis is the individual **ITEM** (as in any LLTM) $\mathbf{Link}(\boldsymbol{y_{is}}) = \boldsymbol{a_i}(\boldsymbol{\theta_s} - \boldsymbol{b_i})$ → both items AND subjects matter
  - Response probability is predicted via a link (transformation) function (usually logit or probit, in which probit is called "ogive" in IRT)
  - 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 ($\boldsymbol{b_i}$), weighted by its discrimination ($\boldsymbol{a_i}$, how related the item is to the latent trait)
  - **"Item factor analysis" (IFA)** re-arranges IRT 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 trait level, and depends specifically on how well the items' difficulty matches subjects' traits

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 easier logit link instead (logit ≈ probit*1.7)
  - Elaborated in 1960s by Birnbaum (and others)

- 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 desirable properties if it fits…
  - … and a very different philosophical viewpoint (as "the" model)

A 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 in software by Jöreskog
- Item response theory (and Rasch) models are common factor models used for binary or ordinal responses…
  - Developed by Lord, Birnbaum, Rasch, and their students
- Confirmatory factor analysis are common factor models for continuous responses…
  - Approximation for ordinal data with varying degrees of success
- Latent traits can also be indicated by other kinds of non-normal responses (count, zero-inflated, two-part/hurdle)….
  - But they don't have special names (I'd call it "generalized SEM")
  - Other response data (e.g., eye fixation, RT) can be used, too!

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

- Explicit, testable models of dimensionality
- Concrete guidelines for selecting items to build scales
- Assess measurement sensitivity across range of latent trait (i.e., know where the "holes" of imprecision 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)
    - REML is not available for smaller samples (as it is in MLM software)
- 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 (within SEM) can do…
  - Bad items are still bad items, no matter what model is used
  - No, SEM is still not "causal" modeling

Practical Problems in Measurement

- To demonstrate the types of issues we will discuss related to instrument development and evaluation, consider the following two examples:
  - 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 sums 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 Test 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?
- Could different items have measured the same ability?
  - Equally well? Can multiple tests be made (with different items) so that the scores are interchangeable?
  - Could a computer be used to give the test adaptively?
- Are students with extreme scores (low or high) measured as accurately as students scoring in the middle?
  - Test floor? Test ceiling? Are floors and ceilings always 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 feel sad."
  - "I cry."
  - "I think about ending my life."

- 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 Instrument Development…

- Dimensionality: How many traits do these items measure?
  - Pro tip: if the trait name has a slash or an "and", it's not a single trait!
- Overall test accuracy vs. efficiency?
  - Do you need to add or remove items? What kind of items?
  - Add or remove response options?
- 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?
- Generalizability: Do your items 'work' for different kinds of people than were originally used to develop the instrument?
  - Sufficiently unbiased (i.e., only measures trait of interest)?
  - Sufficiently sensitive for different ability levels?

Defining Latent 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 (I use term "instrument")
- 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 Latent Constructs, continued

- The models we'll utilize each assume the construct to be a unidimensional and continuous latent variable
  - 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
  - Extreme traits can be unipolar or bipolar (see Tay & Jebb, 2018)

- 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 (Wilson, 2005)

- 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 must be orderable!

- Some examples of construct maps…

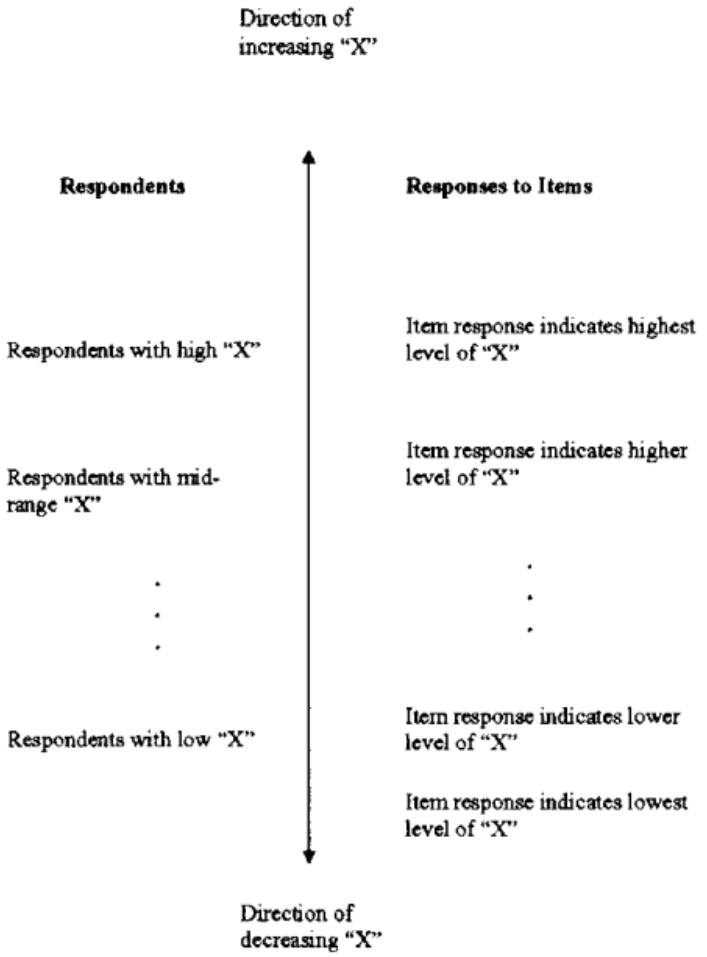


FIG. 2.1 A generic construct map in construct "X."

Template for a Construct Map

Left = PERSONS
- qualities
- characteristics

Right = ITEMS
- responses
- behaviors

From Wilson (2005)

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

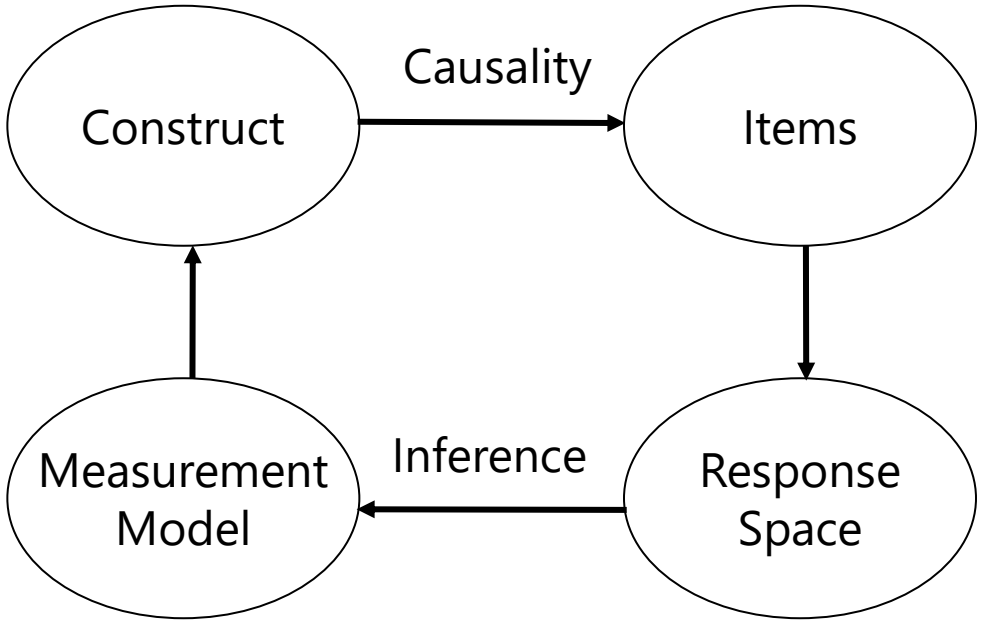
Instrument Construction

- 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 specific 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?
    - For existing instruments, has the dimensionality ever been tested???
  - 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 with a clearly defined referent
- Items should be easy to process (short, simple wording)
  - 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

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

Response formats DO NOT all have to be the same across items if you are using an LTMM to describe individual differences.

You can and should customize them to be most informative for the topic at hand.

The item above is an example of a useful "expanded format" (stay tuned!)

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)
    - Logit (logistic) or probit (ogive) 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 (much more data needed to estimate)
  - No as-easy measurement models 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
  - Latent trait = true score, factor score, latent factor, latent variable
  - Item construction is part art, part science
  - Seek as much info as possible before and after giving your items
- Response options should be carefully considered:
  - May be helpful to start with open-ended responses
  - Decide on optimal but fixed response categories eventually
- Measurement models provide basis for inference back to a person's position on the latent construct:
  - Specific LTMM is 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