A (very) Brief Overview of Item Response Theory

PSY 600K
Dr. Alyssa Gibbons
May 1, 2018

Agenda

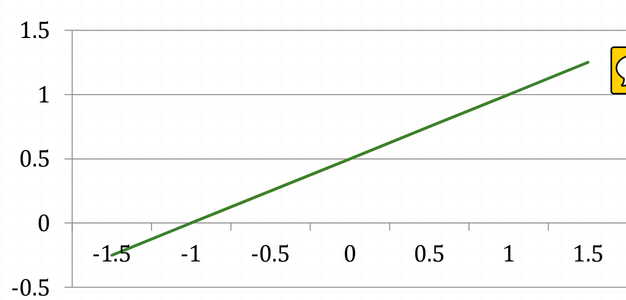
- Intro to IRT:
  - Limitations of CTT
  - Key assumptions
  - Item characteristic curves
  - The 3PL model
  - Local independence

Classical Test Theory is Great, But…

- CTT parameters are *population dependent*.
  - To use CTT to predict how an individual will respond to an item, you have to have pretested that item with people similar to the one you are interested in.
  - For example, math item difficulties from a population of HS students don't tell us anything about item difficulties for college students.
  - You can't use CTT item parameters to estimate the true score for a person from a new population.
- CTT assumes we are measuring equally well (or poorly) across the entire ability distribution.
  - SEM is the same across all possible true scores.

CTT is a Rough Approximation

- The CTT model posits a linear relationship between item response and the underlying latent construct.
  - Limited set of response options…
  - … but a theoretically large possible range of true scores.
- Can predict off-the-scale responses at extreme true score values.



IRT Improves on CTT By…

- Estimating individuals' true score on the construct, **not** using total test score as a proxy.
  - And we can determine how precise these estimates are.
- This allows us to predict performance on an item for an individual with a given true score/ability, **across examinee populations**.
  - IRT item parameters are **not** population dependent.
- Resolving the improper estimates problem by more accurately modeling the relationship between the latent trait and the item response.

IRT Applications

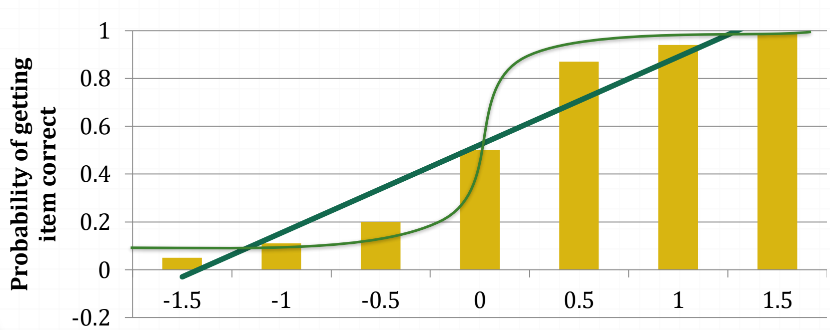
- Maintain test security by being able to create parallel forms from a large test bank.
- Obtain precise estimates of the latent trait (e.g., in high-stakes applications) and know just how good your estimate is.
- Identify mismeasured individuals (those whose scores do not reflect their ability); detect cheating.
- Evaluate the equivalence of translated versions of the same test.
- Detect bias at the test or item level.
- Etc.

Key IRT Assumptions and Conventions

- We are measuring an underlying latent (not observable) characteristic of individuals.
  - In IRT notation, an individual's standing on the latent trait (i.e., true score) is written as θ.
  - We often write $P(\theta)$ – this is **not** the probability **of** θ, but the probability of a **correct/positive response given** $\boldsymbol{\theta}$.
- We assume that θ is distributed **continuously** (and can be measured on an interval scale).
  - Do *not* need to assume that it is distributed normally.
- We scale θ to have a mean of 0 and SD of 1.
  - It's unobservable – we can scale it however we want!
- Early models were developed for **binary** data.
  - There are models for more complex data, but we won't get to them today.

Item Characteristic Curves

- If we plot the actual probability of getting the item right against θ we often get something like this:
  - Equal changes in θ do **not** mean equal changes in probability.
- The linear CTT model doesn't fit so well.
  - But a curvilinear function works much better!



The Ogive Function

- This S-shaped pattern looks like the **normal ogive** function.
  - Normal ogive – cumulative version of the normal probability distribution.
  - Area under the standardized normal curve from a *z*-score of -∞ to $a_i(\theta - b_i)$.
  - Naturally bounded by 0 and 1
- Height of this function at θ = probability of a correct response for an examinee with ability θ.
- We could integrate (calculus) to estimate a_i and b_i for a particular item.
  - Or we can avoid the calculus by using the **logistic ogive** function instead.
  - The logistic ogive is nearly identical to the normal one if we add a scaling constant.

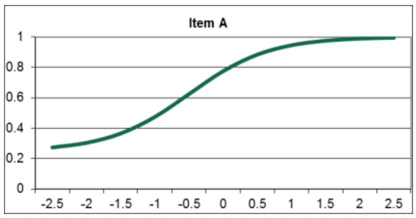
The 3PL Model

- The most general IRT model for binary data is the **3-parameter logistic model**, or 3PL:

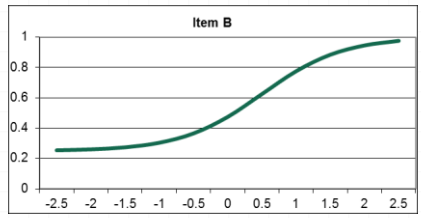
$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{[1 + \exp\{-Da(\theta - b_i)\}]}$$

- Gives us the height of the ICC at any given value of θ.
  - a = steepness of the curve at its inflection point (the point where the curve changes direction) – indicator of *discrimination*
  - b = value of θ at the inflection point – indicator of *difficulty*
  - c = lower asymptote of the ICC – *guessing parameter* – probability of a correct response even given very low θ.

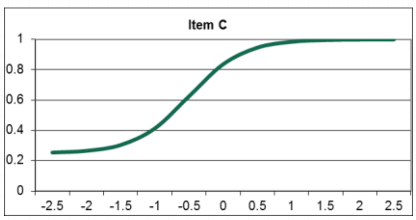
A Few 3PL ICCs



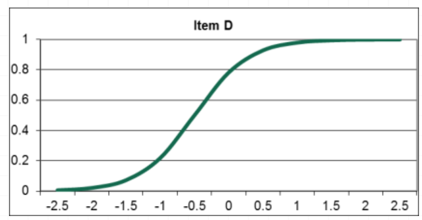
a = 1 *b* = -0.5 *c* = .25



a = 1 *b* = 0.5 *c* = .25



a = 1.5 *b* = -0.5 *c* = .25



a = 1.5 *b* = -0.5 *c* = .0

Interpreting the Parameters: *a* = item discrimination

- Increasing *a* means that our item makes a sharper distinction between examinees.
  - Why is this a good thing?
- When *a* is large, the curve is steep around *b* and flat elsewhere.
  - In other words, high *a* gives us good discriminating power *around b*, but not at other points on the continuum.
  - This is very different from classical test theory, where we are used to thinking of item parameters as applying at all levels of θ.
- Items with low *a* values have some (though probably not much) discriminating power through a broader range of values.
  - These items may sometimes be useful if we don't know the range of θ.
- *a* must be a positive real number.
  - Usually ranges from about .30 to 2.0

Interpreting the Parameters: b = item difficulty

- At what point in the theta continuum does this item measure the best?
  - No longer about the number of people who got it right!
- Can be any real number; usually between -3.0 and 3.0
- Now a low difficulty parameter = an easier item.
  - The farther to the right the ICC is shifted, the harder the item.

Interpreting the Parameters: c = guessing

- For very low values of θ, the last term of our item response function becomes essentially zero, so our best estimate of $P_i(\theta) = c_i$.
- c must be between 0 and 1.
  - c often comes close to $1/m$ for multiple choice items where m is the number of response options.
  - However, in practice, c_i is often less than the probability of guessing randomly. Why?
- Including a c_i parameter means we will have smaller discrimination (a_i) parameters.
  - Restricts our ICC to fall between c_i and 1 (instead of 0 and 1).

Simpler Models

- 2PL: all $c = 0$
  - Appropriate when "guessing" is not expected to be an issue.
  - In this model, b_i is the value of θ for which the probability of getting a correct response is 0.5.
    - This is *not* the case in the 3PL model.
- 1PL or Rasch: all $c = 0$ and all $a = 1$.
  - In other words, all items are assumed to be equally discriminating – all that varies is difficulty.
  - This model is quite popular in some circles and unpopular in others – why do you think that is?

Questions?

For next time: More .

Read: R & M 11.1 -11.7

Reading Response: Explain, in simple language, what local independence is. Is it a good thing or a bad thing for our IRT models? Why?