EPSY 5221 Item Response Theory

IRT is a way of thinking about measurement: a probabilistic model

We give an item to a person and obtain an “item-person” interaction.

This results in a score with a probability.

What affects item-person interactions? Prior experience, skill, knowledge, energy level, the environment. Some of these characteristics are construct relevant (part of the true score) and others are measurement error, including systematic sources of measurement error (e.g., test-wiseness which also affects true scores) and random sources of measurement error (energy level).

CTT

Based on the idea of parallel forms: X = T + E, where *E*(X) = T

 because 

for individuals across two parallel forms, 

so, SD of X is SEM for an individual – an individual’s true score doesn’t vary.

All CTT item and test statistics are sample specific.

Observed scores and true scores are test and sample dependent – examinees will have lower true scores on difficult tests and higher true scores on easier tests.

IRT

Ability scores are more fundamental because they are test independent. Examinees come to a test administration with ability levels in relation to the construct being measured – not necessarily in relation to the test being administered.

Each examinee has an ability score that is defined in relation to the construct at the time of an assessment, and this remains invariant over samples of assessment tasks that may be used.

To predict the statistical and psychometric properties of any test administered to any group, IRT is more appropriate.

Most models assume that the dimensionality of the latent space is unidimensional – we measure one thing at a time.

The second assumption is local independence – performance on one item does not affect performance on another item. If persons are all of the same ability on the underlying dimension, there will be zero correlation among item responses.

We can define an item characteristic curve (ICC). This is a mathematical function that relates the probability of success on an item to the ability measured by the item or test that contains it. This relation is testable. We typically use a logistic function to describe this relation.

The 3-Parameter Logistic Model



Three parameters describe the relation of the probability of individuals at a certain level of ability responding to item ***i*** correctly.

b is the point of inflection, the location of the ICC on the ability continuum (difficulty)

when c=0, b is at θ where P(X = 1 | θ) = .5, otherwise where 1– [(1 – c) / 2]

a is a function of the slope of the curve at b

slope = 1.7 (0.25) a (1-c)

c is the lower asymptote

P(X = 1) for lowest scoring individual

e is the natural logarithm equal to 2.718

We can plot this curve for individuals of different abilities, P(θ). The curve is monotonically increasing.

IRT analysis involves obtaining the maximum likelihood estimates of the parameters based on large numbers of subjects. Maximum likelihood estimates are the values of the parameter that are most likely responsible for a specified response string.

The 2-Parameter Logistic Model



The 1-Parameter Logistic Model



And the Rasch model (slightly different parameterization), where *a* = 1



which is the same as the log-odds:

The metric of IRT is the logit, the natural logarithmic scale of the odds ratio: **Logit = Log(Odds)**

Here we can see that the probability of a correct response is a function of the difference between a person’s ability (θ) and the item location (b).

As an illustrative example, consider a person’s ability being equal to the item’s location:

When ability = item location: 

When ability > item location: 

When ability < item location: 

Benefits of IRT Modeling

The models link item responses to ability and item statistics are on the same scale as ability. This helps us to know where on the ability continuum the item makes a contribution to the measurement of a construct.

Information

Information tells us about the precision of estimation of person’s ability. The **item information function** shows us the contribution of a particular item to the measurement of ability. This is a function of (a) the slope of the item characteristic curve and (b) the variance of the test score at each ability level.

**Test information** is the sum of item information curves: I(θ)

Based on the inverse of the error variance, or the score variance at θ

 is the standard error of a test at θ, the precision of ability estimation.

🡪 Notice here that the standard error of measurement changes as a function of ability.

Invariance

Given θ, P(X = 1) is constant regardless of who is taking the test and what other items are on the test.

Estimation

Many different programs use different estimation procedures, mostly involving some form of maximum likelihood estimation.

An iterative procedure is used.

* An estimate of ability is obtained, usually from the raw score (proportion correct).
* Item difficulties are estimated
* Ability is re-estimated based on item difficulties
* Item parameters are then estimated based on ability estimates (re: # parameters)
* This procedure is continued until there is convergence in subsequent iterations

Test Construction

1. decide on test’s target information curve
2. select items with item information curves to fill the hard-to-fill areas under the target curve
3. cumulatively add the item information curves
4. continue filling areas until you reached the target.

Test “Bias” – still requires human review and judgement regarding bias

1. if the ICCs are identical for subgroups of examinees, the item is fair
2. if the ICC for one subgroup is above that for another at all ability levels, we have a case where one item is uniformly easier than the other for all ability levels
3. if the ICCs cross, then bias is more clearly indicated