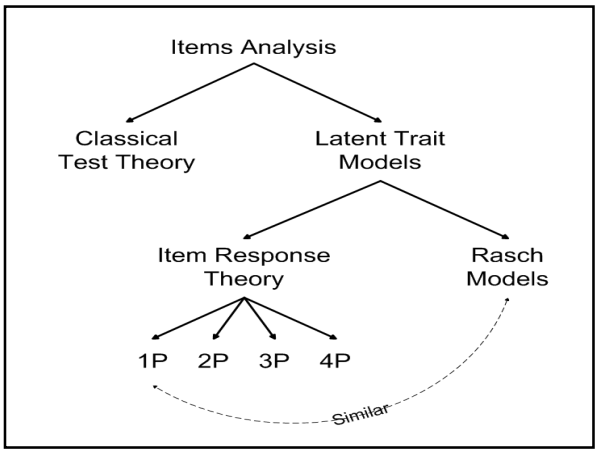
INTRODUCTION TO ITEM RESPONSE THEORY

Psy 427
Cal State Northridge
Andrew Ainsworth, PhD

Contents

What is item analysis in general?

- Item analysis provides a way of measuring the quality of questions - seeing how appropriate they were for the respondents and how well they measured their ability/trait.
- It also provides a way of re-using items over and over again in different tests with prior knowledge of how they are going to perform; creating a population of questions with known properties (e.g. test bank)



Classical Test Theory - Review

Classical Test Theory

- Classical Test Theory (CTT) - analyses are the easiest and most widely used form of analyses. The statistics can be computed by readily available statistical packages (or even by hand)
- Classical Analyses are performed on the test as a whole rather than on the item and although item statistics can be generated, they apply only to *that* group of students on *that* collection of items

Classical Test Theory

- CTT is based on the true score model
- In CTT we assume that the error :
  - Is normally distributed
  - Uncorrelated with true score
  - Has a mean of Zero

Classical Test Theory Statistics

- Difficulty (item level statistic)
- Discrimination (item level statistic)
- Reliability (test level statistic)

Classical Test Theory vs. Latent Trait Models

- Classical analysis has the test (not the item) as its basis. Although the statistics generated are often generalised to similar students taking a similar test; they only really apply to *those* students taking *that* test
- Latent trait models aim to look beyond that at the underlying traits which are producing the test performance. They are measured at item level and provide sample-free measurement

Latent Trait Models

- Latent trait models have been around since the 1940s, but were not widely used until the 1960s. Although theoretically possible, it is practically unfeasible to use these without specialized software.
- They aim to measure the underlying ability (or trait) which is producing the test performance rather than measuring performance per se.
- This leads to them being **sample-free.** As the statistics are not dependant on the test situation which generated them, they can be used more flexibly

Item Response Theory

Item Response Theory

- Item Response Theory (IRT) – refers to a family of latent trait models used to establish psychometric properties of items and scales
- Sometimes referred to as *modern psychometrics* because in large-scale education assessment, testing programs and professional testing firms IRT has almost completely replaced CTT as method of choice
- IRT has many advantages over CTT that have brought IRT into more frequent use

Three Basics Components of IRT

- Item Response Function (IRF) – Mathematical function that relates the latent trait to the probability of endorsing an item
- Item Information Function – an indication of item quality; an item's ability to differentiate among respondents
- Invariance – position on the latent trait can be estimated by any items with know IRFs and item characteristics are population independent within a linear transformation

IRT: Item Response Functions

IRT - Item Response Function

- Item Response Function (IRF) - characterizes the relation between a latent variable (i.e., individual differences on a construct) and the probability of endorsing an item.
- The IRF models the relationship between examinee trait level, item properties and the probability of endorsing the item.
- Examinee trait level is signified by the greek letter *theta* (θ) and typically has mean = 0 and a standard deviation = 1

IRT - Item Characteristic Curves

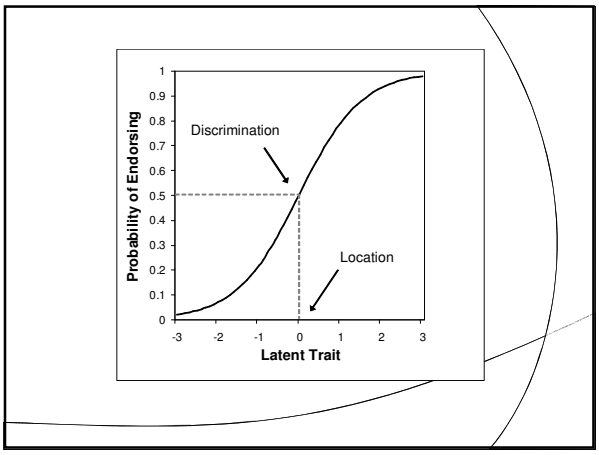
- IRFs can then be converted into Item Characteristic Curves (ICC) which are graphical functions that represents the respondents ability as a function of the probability of endorsing the item

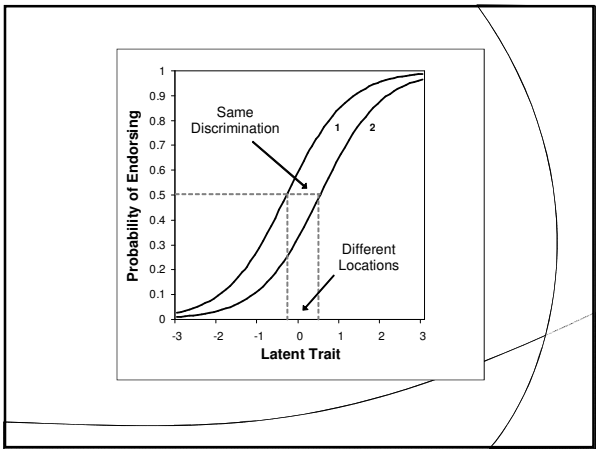
IRF – Item Parameters Location (b)

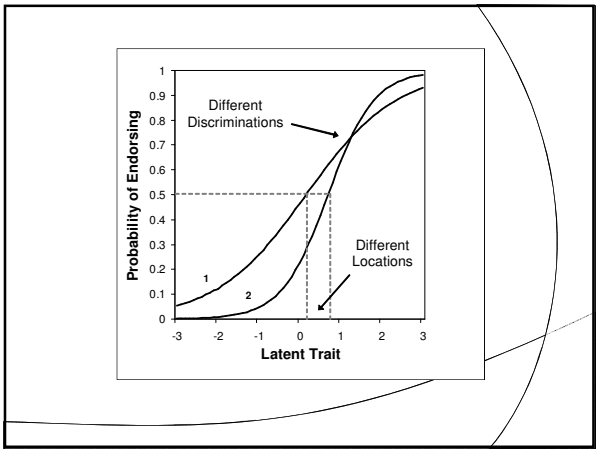
- An item's ***location*** is defined as the amount of the latent trait needed to have a .5 probability of endorsing the item.
- The higher the "b" parameter the higher on the trait level a respondent needs to be in order to endorse the item
- Analogous to difficulty in CTT
- Like Z scores, the values of *b* typically range from -3 to +3

IRF – Item Parameters Discrimination (a)

- Indicates the steepness of the IRF at the items location
- An items discrimination indicates how strongly related the item is to the latent trait like loadings in a factor analysis
- Items with high discriminations are better at differentiating respondents around the location point; small changes in the latent trait lead to large changes in probability
- Vice versa for items with low discriminations

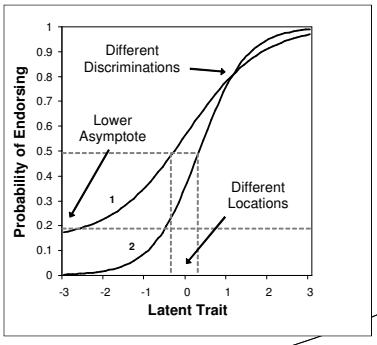

Probability of Endorsing
Discrimination
Location
Latent Trait


Probability of Endorsing
Same
Discrimination
1
2
Different
Locations
Latent Trait


Probability of Endorsing
Different
Discriminations
Different
Locations
1
2
Latent Trait

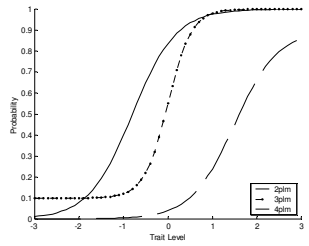
IRF – Item Parameters Guessing (c)

- The inclusion of a "c" parameter suggests that respondents very low on the trait may still choose the correct answer.
- In other words respondents with low trait levels may still have a small probability of endorsing an item
- This is mostly used with multiple choice testing…and the value should not vary excessively from the reciprocal of the number of choices.



IRF – Item Parameters Upper asymptote (d)

- The inclusion of a "d" parameter suggests that respondents very high on the latent trait are not guaranteed (i.e. have less than 1 probability) to endorse the item
- Often an item that is difficult to endorse (e.g. suicide ideation as an indicator of depression)



IRT - Item Response Function

- The 4-parameter logistic model

- Where
  - θ represents examinee trait level
  - *b* is the item difficulty that determines the location of the IRF
  - *a* is the item's discrimination that determines the steepness of the IRF
  - *c* is a lower asymptote parameter for the IRF
  - *d* is an upper asymptote parameter for the IRF

IRT - Item Response Function

- The 3-parameter logistic model

- If the upper asymptote parameter is set to 1.0, then the model is termed a 3PL.
- In this model, individuals at low trait levels have a non-zero probability of endorsing the item.

IRT - Item Response Function

- The 2-parameter logistic model

- If in addition the lower asymptote parameter is constrained to zero, then the model is termed a 2PL.
- In the 2PLM, IRFs vary both in their discrimination and difficulty (i.e., location) parameters.

IRT - Item Response Function

- The 1-parameter logistic model

- If the item discrimination is set to 1.0 (or any constant) the result is a 1PL
- A 1PL assumes that all scale items relate to the latent trait equally and items vary only in difficulty (equivalent to having equal factor loadings across items).

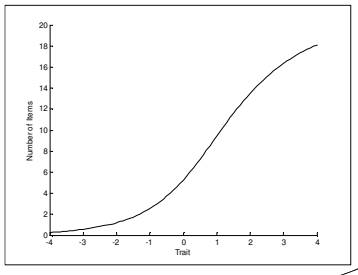
Quick Detour: Rasch Models vs. Item Response Theory Models

- Mathematically, Rasch models are identical to the most basic IRT model (1PL), however there are some (important) differences
- In Rasch the model is superior. Data which does not fit the model is discarded
- Rasch does not permit abilities to be estimated for extreme items and persons
- And other differences

IRT - Test Response Curve

- Test Response Curves (TRC) - Item response functions are additive so that items can be combined to create a TRC.
- A TRC is the latent trait relative to the number of items

IRT - Test Response Curve



IRT: Item Information Functions

IRT – Item Information Function

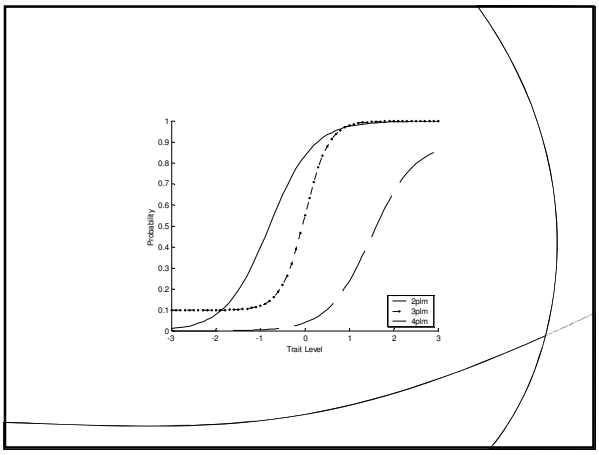
- Item Information Function (IIF) – Item reliability is replaced by item information in IRT.
- Each IRF can be transformed into an item information function (IIF); the precision an item provides at all levels of the latent trait.
- The information is an index representing the item's ability to differentiate among individuals.

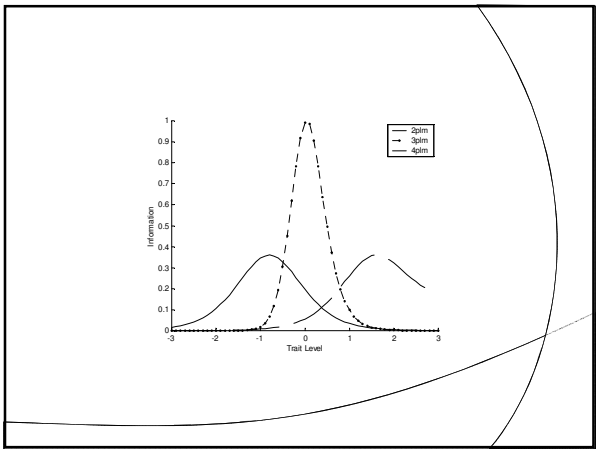
IRT – Item Information Function

- The standard error of measurement (which is the variance of the latent trait level) is the reciprocal of information, and thus, more information means less error.
- Measurement error is expressed on the same metric as the latent trait level, so it can be used to build confidence intervals.

IRT – Item Information Function

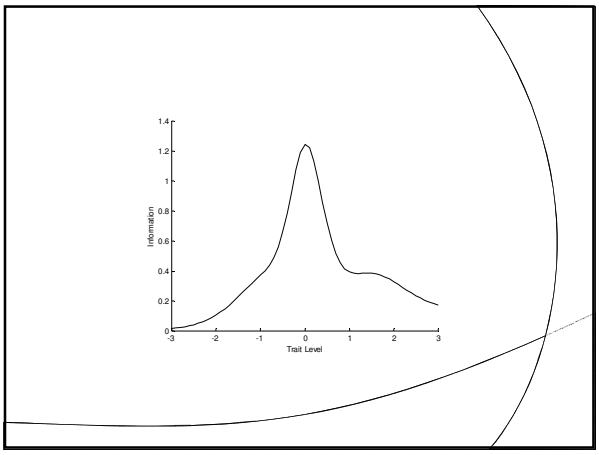
- Difficulty parameter - the location of the highest information point
- Discrimination - height of the information.
- Large discriminations - tall and narrow IIFs; high precision/narrow range
- Low discrimination - short and wide IIFs; low precision/broad range.

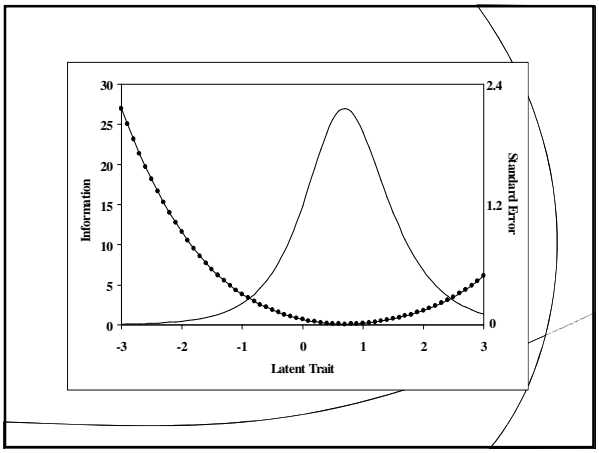




IRT – Test Information Function

- Test Information Function (TIF) – The IIFs are also additive so that we can judge the test as a whole and see at which part of the trait range it is working the best.





Item Response Theory Example

- The same 24 items from the MMPI-2 that assess Social Discomfort
- Dichotomous Items; 1 represents an endorsement of the item in the direction of discomfort
- Assess a 2pl IRT model of the data to look at the difficulty, discrimination and information for each item

IRT: Invariance

IRT - Invariance

- Invariance - IRT model parameters have an invariance property
  - Examinee trait level estimates do not depend on which items are administered, and in turn, item parameters do not depend on a particular sample of examinees (within a linear transformation).
- Invariance allows researchers to: 1) efficiently "link" different scales that measure the same construct, 2) compare examinees even if they responded to different items, and 3) implement computerized adaptive testing.

IRT: Assumptions

IRT - Assumptions

- Monotonicity - logistic IRT models assume a monotonically increasing functions (as trait level increases, so does the probability of endorsing an item).
- If this is violated, then it makes no sense to apply logistic models to characterize item response data.

IRT - Assumptions

- Unidimensionality – In the IRT models described above, individual differences are characterized by a single parameter, theta.
  - Multidimensional IRT models exist but are not as commonly applied
  - Commonly applied IRT models assume that a single common factor (i.e., the latent trait) accounts for the item covariance.
  - Often assessed using specialized Factor Analytic models for dichotomous items

IRT - Assumptions

- Local independence - The Local independence (LI) assumption requires that item responses are uncorrelated after controlling for the latent trait.
  - When LI is violated, this is called local dependence (LD).
  - LI and unidimensionality are related
  - Highly univocal scales can still have violations of local independence (e.g. item content, etc.).

IRT - Assumptions

- Local dependence:
  1. distorts item parameter estimates (i.e., can cause item slopes to be larger than they should be),
  2. causes scales to look more precise than they really are, and
  3. when LD exists, a large correlation between two or more items can essentially define or dominate the latent trait, thus causing the scale to lack construct validity.

IRT - Assumptions

- Once LD is identified, the next step is to address it:
  - Form testlets (Wainer & Kiely, 1987) by combining locally dependent items
  - Delete one or more of the LD items from the scale so local independence is achieved.

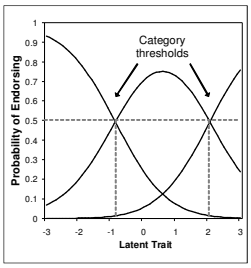
IRT - Assumptions

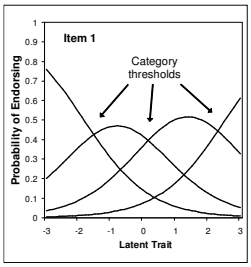
- Qualitatively homogeneous population - IRT models assume that the same IRF applies to all members of the population
  - Differential item functioning (DIF) is a violation of this and means that there is a violation of the invariance property
  - DIF occurs when an item has a different IRF for two or more groups; therefore examinees that are equal on the latent trait have different probabilities (expected scores) of endorsing the item.
  - No single IRF can be applied to the population

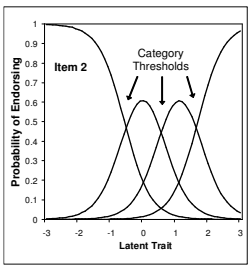
IRT: Applications

Applications

- Ordered Polytomous Items
  - IRT models exist for data that are not dichotomously scored
  - With dichotomous items there is a single difficulty (location) that indicates the threshold at which the probability switches from favoring one choice to favoring the other
  - With polytomous items a separate difficulty exists as thresholds between each sets of ordered categories

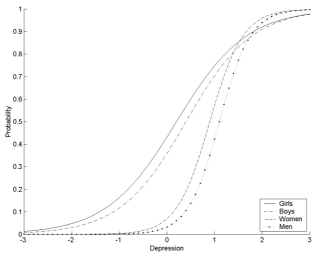

Category
thresholds
Probability of Endorsing
Latent Trait


Item 1
Category
thresholds
Probability of Endorsing
Latent Trait


Item 2
Category
Thresholds
Probability of Endorsing
Latent Trait

Applications

- Differential Item Functioning
  - How can age groups, genders, cultures, ethnic groups, and socioecomonic backgrounds be meaningfully compared?
  - Can be a research goal as opposed to just a test of an assumption
  - Test equivalency of test items translated into multiple languages
  - Test items influenced by cultural differences
  - Test for intelligence items that gender biased
  - Test for age differences in response to personality items



"Don't care about life"

Applications

- Scaling individuals for further analysis
  - We often collect data in multifaceted forms (e.g. multi-items surveys) and then collapse them into a single raw score
  - IRT based scores represent an optimal scaling of individuals on the trait
  - Most sophisticated analyses require at-least interval level measurement and IRT scores are closer to interval level than raw scores
  - Using scaled scores as opposed to raw scores has been shown to reduce spurious results

Applications

- Scale Construction and Modification
  - The focus is changing from creating fixed length, paper/pencil tests to creating a "universe" of items with known IRF's that can be used interchangeably
  - Scales are being designed based around IRT properties
  - Pre-existing scales that were developed using CTT are being "revamped" using IRT

Applications

- Computer Adaptive Testing (CAT)
  - As an extension of the previous slide, once a "universe" (i.e. test bank) of items with known IRFs is created they can be used to measure traits in a computer adaptive form
  - An item is given to the participant (usually easy to moderate difficulty) and their answer allows their trait score to be estimated, so that the next item is chosen to target that trait level
  - After the second item is answered their trait score is re-estimated, etc.

Applications

- Computer Adaptive Testing (CAT)
  - CA tests are at least twice as efficient as their paper and pencil counterparts with no loss of precision
  - Primary testing approach used by ETS
  - Adaptive form of the Headache Impact Survey outperformed the P and P counterpart in reducing patient burden, tracking change and in reliability and validity (Ware et al., 2003)

Applications

- Test Equating
  - Participants that have taken different tests measuring the same construct (e.g. Beck depression vs. CESD), but both have items with known IRFS, can be placed on the same scale and compared or scored equivalently
  - Equating across grades on math ability
  - Equating across years for placement or admissions tests