IRT
Chapter 1
Background

A typical scenario to consider the use of item response theory (IRT)

Where - Testing company that specializes in the development and analysis of achievement and aptitude test (awarding high school diplomas, promoting students from one grade to the next, evaluating the quality of education, identifying workers in need of training, and credentialing practitioners in a wide variety of professions)

Why - Because classical test theory (CTT) that have been used for years has laminations and IRT offers solutions to many of the limitations. Clients suggest, and sometimes require the use of IRT.

Purpose of this book

(a) introduce the basic concepts and most popular models of IRT
(b) address parameter estimation and available computer programs
(c) demonstrate approaches to assessing model-data fit
(d) describe the scales on which abilities and item characteristics are reported
(e) describe the application of IRT to test construction, detection of differential item functioning (DIF), equating, and adaptive testing.

Limitations of CTT ($X = T + E$)

1. Group dependent: A "hard" or "easy" test depends on the group of examinees who took the test. The difficulty of a test item is defined as "the proportion of examinees in a group of interest who answer the item correctly."
    Consequences - Group-dependent item indices are of limited use when constructing tests for examinee populations that are dissimilar to the population of examinees with which the item indices were obtained (e.g., field testing, item banks).
2. Test dependent: An examinee's ability is defined only in terms of a particular test. (i.e., High ability with easy test, lower ability with hard test).
    Consequences - It is difficult to compare examinees who take different tests.

3. Reliability and standard error of measurement (SEM): The assumption of equal errors of measurement for all examinees (in CTT) is implausible.

4. Test oriented rather than item oriented: CTT provides no information on how well a particular examinee might do when confronted with a test item, which makes DIF analyses and test construction of targeted ability difficult.

5. Others: the design of tests, the identification of biased items, adaptive testing, and the equating of test scores.

Advantages of IRT include

(a) item characteristics that are not group-dependent
(b) scores describing examinee proficiency that are not test-dependent
(c) a model that is expressed at the item level rather than at the test level
(d) a model that does not require strictly parallel tests for assessing reliability, and
(e) a model that provides a measure of precision for each ability score.

IRT
Chapter 2
Concepts, Models, and Features

Basic Ideas

Two postulates:

(a) The performance of an examinee on a test item (or the probability of an examinee answering the item correctly (P)) can be predicted by a set of factors called traits, latent traits, or abilities (theta(s) or θ(s)), and

(b) P and theta(s) can be described by a monotonically increasing function called an item characteristic function (ICF) or item characteristic curve (ICC).

☺ Draw some possible models.

Desirable features when a given IRT models fits the data,

1. Invariance of item and ability parameters
   - Ability estimates are not test-dependent
   - Item indices are not group- dependent
2. Standard errors for individual ability estimates

Assumptions

A. Unidimensionality (for unidimensional IRT models only)

B. Local Independence
   p. 10

☺ List practical testing situations when you think A may not tenable.

☺ Explain that A implies B, but B does not imply A.

Popular Models in IRT

One-Parameter Logistic Model (1PL) p.12
Two-Parameter Logistic Model (2PL) p.15
Three-Parameter Logistic Model (3PL) p. 17

The Property of Invariance

The property of invariacne of item and ability parameters is the cornerstone of IRT and its major distinction from CTT, and makes possible such important applications as equating, item banking, DIF, and CAT.

The property is a well-known feature of the linear regression model.

When the regression model holds, the slope and intercept of the line will be the same in any subpopulation on X.

☺ Explain above, and also show that correlation is NOT invariant among subpopulations.

p. 20, p. 21

	Using CTT		Using IRT	
	p-value	discrimination	a	b
Total group	.5	.65	.8	-.1
Low ability group	.2	.56	.8	-.1
High ability group	.8	.47	.8	-.1

The property of invariance holds only when

(a) the model fit the data, and

(b) describing the population not the sample.

The above table refers to the invariance of item parameters. The invariance of ability parameters can be explained similarly.

Other Promising Models

pp. 26-28

IRT
Chapter 3
Ability and Item Parameter Estimation

How do you get item parameter estimates (a, b,, and c) and ability estimates (θ)?

Similar to obtaining coefficients in the regression analysis, but different in the following points:
1. Non-linear models
2. unobservable variable θ

When the item parameters are known:

Maximum likelihood estimate (MLE)

$$L(u_i,u_2,\ldots u_n|\theta) = \prod P_j^{u_j} Q_j^{1-u_j}$$

Example. Table 3.1 (p. 35)

For $\theta = 0$

	1	2	3	4	5
P	.16	.33	.51	.78	.91
Q	.84	.67	.49	.22	.09

☺ How do you get P = .16 for Item 1?

Examinee 3 had (0 0 0 1 1).
At $\theta = 0$
L = (.84)(.67)(.49)(.78)(.91) = 0.195

Now you try other θ values. See p.52 Table 3.4. The θ ranges from -1.0 to 0. You see L is maximum at -0.5. Figure 3.1 is a graphic display. The y-axis is "log" likelihood (ln L). The logarithm of L is used because maximum of L is the same as maximum of ln L and ln L is easier to handle mathematically than the plain L. In practice, the Newton-Raphson procedure is used to obtain the maximum.

Problems of MLE.
1. The likelihood function may not have a maximum value. See Figure 3.2. Examinee 1 had a response pattern (11100) and (10100). ☺ Explain why Examinees 1 and 2's responses are considered to be "aberrant". See Table 3.1.
2. No solutions to all 0 or all 1 response patterns.

Overcoming the problems of MLE -- Bayesian estimation (pp. 38-39). (The basic idea is to modify the likelihood function to incorporate any prior information we may have about the ability parameters.) Bock and Mislevy (1982) - the Expected A Posteriori (EAP) estimate.

When θ's are known:

Use the similar process as when the item parameters are known except:

- Estimating (at most) 3 unknown parameters (a, b, and c) as opposed to just 1 (θ). - a multivariate version of the above process is employed (Newton-Raphson procedure in the multivariate form).
- Assuming examinee independence as opposed to local independence of items.

When items parameters AND θ's are both unknown:

Indeterminacy problem.

☺ Show $P(\theta) = P(\theta^*)$ given the transformation $\theta^* = \alpha\theta + \beta$, $b^* = \alpha b + \beta$, and $a^* = a/\alpha$.

What to do - Choose an arbitrary scale for θ or b (the mean of 0 and SD of 1).

Joint maximum likelihood estimation (JMLE)

Stage 1 - Initial values of θs are chosen. (The log ratio of # correct to # incorrect score for each examinee provides good starting values.) Then the ability parameters are standardized to eliminate the indeterminacy problem, and treating θ's as known, the item parameters are estimated.

Stage 2 - Treating the item parameters are known, θ's are estimated.

Stage 1 and Stage 2 are repeated until the values of the estimates do not change between two successive estimation stages.

Problems of JMLE

Improper estimates for certain response patterns.
Not "consistent" (asymptotically unbiased) estimates.

Alternative to JMLE

Marginal Maximum Likelihood (MML) estimation (Mislevy and Bock,1984, BILOG)

If we consider the examinees as having been selected randomly from a population, then, by specifying a distribution for the ability parameters, we can integrate them out of the likelihood function.

Still estimation may fail (primarily due to the estimation of the c-parameter). BILOG puts a prior distribution on c by default.

Standard Errors of Estimates

For θ's, $SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$, where $I(\theta)$ is the information function (see Chapter 6).

For a, b, and c, the variance-covariance matrix can be calculated.

IRT
Chapter 4
Assessment of Model-Data Fit

The advantage of item response models can be obtained only when the fit between the model and the test data of interest is satisfactory.

How do you check the model fit?

I. Checking Assumptions

1. Unidimensionality
- Linear factor analysis — Non-linear factor analysis (McDonald, etc.)
- Local independence — DIMTEST (Stout, etc.)

2. Equal Discrimination Indices (For 1PL)

3. Minimal Guessing

4. Nonspeeded Test Administration

II. Checking Expected Model Features

1. Invariance of Ability Parameter Estimates (Figures 4.3, 4.4)

2. Invariance of Item Parameter Estimates (Figures 2.6, 4.2)

III. Checking Model Predications

1. Residual Analysis - A residual is the difference between observed item performance for a subgroup of examinees and the subgroup's expected item performance. Yen's Q1 statistic (Equation 4.1). H_0: good fit. See Figures 4.5-4.10.

2. Computer Simulation Methods (4.11-4.13)

3. Item Misfit Statistics (4.14-4.16)

IRT
Chapter 5
The Ability Scale

CTT

Score (X)

$E(X) = \tau$

where X is the number-right score and τ is the true score.

Transformation

linear transformation - scaled scores
non-linear transformation - stanines, percentiles

Drawback

test dependent, examinees dependent

IRT

Score (θ)

$$\tau = \sum_{j=1}^{n} P_j(\theta)$$

where θ is defined in the interval $(-\infty, \infty)$.
Test characteristic curve (TCC) (☺ See Table 5.2. Can you draw a TCC?)

Transformation

linear transformation - e.g., the Woodcock-Johnson scale ($w_\theta = 9.1\theta+500$)
non-linear transformation - e.g., Logit scale (pp. 80-84) - 1PL only

domain score $\pi = \frac{1}{n}\sum_{j=1}^{n} P_j(\theta)$

Advantage

test independent, examinees independent
The true score can be computed on a set of items NOT administered to the examinee!

IRT
Chapter 6
Item and Test Information and Efficiency Functions

Item Information Function

$$I_i(\theta) = \frac{2.89a_i^2(1 - c_i)}{[c_i + e^{1.7a_i(\theta - b_i)}][c_i + e^{-1.7a_i(\theta - b_i)}]^2}$$

$I_i(\theta)$ is the "information" provided by item i at θ.

The role of b, a, and c parameters in the item information function:

(a) information is higher when b values is close to θ than when the b value is far from θ.
(b) information is generally higher when the a parameter is high, and
(c) information increases as the c parameter goes to zero.

When $c_i = 0$, $\theta_{max} = b_i$. When $c_i > 0$, an item provides the maximum information at an ability level slightly higher than its difficulty.

Examples: (pp. 92-94)

Test Information Function

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

Standard Error of Estimation

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

Relative Efficiency

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)}$$

where $I_A(\theta)$ and $I_B(\theta)$ are the information functions for Test A and B, respectively. If ,for example, $I_A(\theta) = 25$ and $I_B(\theta) = 20$, then $RE(\theta) = 1.25$, and it is said that at θ, Test A is functioning as if it were 25% longer than Test B.

IRT
Chapter 7
Test Construction

Advantages of IRT over CTT in Test Construction

1. Item parameters are invariant, overcoming the problems of classical item indices.
(i.e., field testing : Groups may differ in ability distributions. item banking : If the experimental items are numerous, obviously not all can be embedded in one test. Thus, those items are embedded in different tests taken by different examinees.)

2. Item difficulty and examinee ability are measured on the same sclae, making it possible to select items that are most useful in certain regions of the ability scale.

3. IRT permits the selection of items based on the amount of information the items contribute to the total amount of information needed in the test to meet the test specifications.

Basic Approach

A procedure using item information functions to build tests (Lord, 1977)

1. Decide the target information function.
2. Select items from the item back with item information functions that will fill up areas under the target information function.
3. Calculate the test information function.
4. Continue the process until 1 and 3 match to a satisfactory degree.

Issues (pp. 101-103)

How to determine the target information function (NRT vs. CRT)?
How to use the above approach to ensure content validity?
How can the above approach help the gain score analysis (pre-post tests)?
Easier items on the pre-test, harder items on the post-test
How can the above approach help the classification error rates for the test with a cut-off score?

Problems

Content validity
Inflation of the discrimination parameter

Examples

Broad ability test (Figure 7.1)
Criterion-referenced test (Figures 7.2. - 7.4))

IRT
Chapter 8
Identification of Potentially Biased Test Items

Background

Item Bias vs. Differential Item Functioning

Definition

1. "An item shows DIF if the majority and minority groups differ in their mean performance on the item." --- Ignores a real between-group difference (a.k.a. "Impact")

2. "An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right."

IRT Methods for Detecting DIF

#2 restated: "An item shows DIF if the item response functions across different subgroups are not identical"."

First step is to put item parameters from separate calibrations onto a common scale. (See Chapter 9)

1. Comparison of Item Parameters (Lord's chi-square method)

H_0: $b_1 = b_2$; $a_1 = a_2$; $c_1 = c_2$

$\chi^2 = (a_{diff}\ b_{diff}\ c_{diff})' \sum^{-1} (a_{diff}\ b_{diff}\ c_{diff})$

where Σ is the variance-covariance matrix of the differences between the parameter estimates. The test statistic is asymptotically distributed as chi-square with p degrees of freedom where p is the number of parameters compared.

Note:
(a) Exclude the c parameter due to its large standard error.
(b) Criticism #1: Significant chi-square when $ICC_1 \approx ICC_2$
(c) Criticism #2: The distribution of the test statistic is known only asymptotically (how large the sample size must be?); furthermore, the asymptotic distribution is applicable only when item parameters are estimated in the presence of known ability parameters.
(d) Criticism #3: The chi-square statistic has a higher than expected false-positive rate.

2. Area Between Item Characteristic Curves
    (a) Numerical procedures (Rudner et al., 1980)
    (b) Area Measures (Raju, 1988)
    Note: The c is assumed to be the same for both groups. If not, the area is infinite.

3. More IRT-Based DIF Indices After 1991
    Closed interval area measures (Kim and Cohen, 1991)
    Likelihood Ratio Test (Thissen, Steiberg, & Wainer, 1988, 1993)
    DFIT (Raju, van der Linden, & Fleer, 1995)

Non-IRT Methods

Mantel-Haenszel Technique (Holland & Thayer, 1988)
Logistic Regression Procedures (Swaminathan & Rogers, 1990)
SIBTEST (Shealy & Stout, 1993)

Example

Table 8.1
Figure 8.1 (uniform DIF)
Figure 8.2 (non-uniform DIF)

IRT
Chapter 9
Test Score Equating

Background

Comparability of test scores across different tests (X and Y) measuring the same ability is an issue of considerable importance to test developers. Suppose the score on test X is converted to the metric of test Y:

X		Y	
x	➜ y*	y	y* = f(x)

Classical Methods of Equating

Basic Design

Design A - Two randomly assigned groups take on of the forms (X or Y)
(randomized two groups)- need large N

Design B - X and Y are both administered (XY order or YX order)
(randomized two groups) - need long testing time

Design C - X plus anchor test, Y plus anchor test
(randomized or **non-randomized two groups**)

Equipercentile equating (See C & A p.461)
Linear equating (See C & A p. 458)

Equipercentile or linear equating is generally adequate with Designs A, B, and C (randomized two groups). IRT equating is most useful with Design C with non-randomized two groups.

Problems inherent in classical methods:

Equity (p. 125)
Symmetry
Independence

In an IRT framework, "scaling" rather than "equating" in necessary.

Scaling in IRT

- When item parameters are known, ability estimates are on the same scale (no scaling necessary).
- When they are unknown, we need to solve the problem of scale indeterminancy.

$$\theta^* = \alpha\,\theta + \beta,\ b^* = \alpha\, b + \beta,\ \alpha^* = \frac{a}{\alpha}$$

☺ Recall P = P* when the above transformation is applied.

☺ If different groups of examinees take different tests, is it possible to equate the two tests using IRT?

Common item design (DIF, anchor equating)
Common examinee design

Determination of the Scaling Constants (Anchor Test)

1. Regression Method (not symmetry)
2. Mean and Sigma Method
3. Robust Mean and Sigma Method
4. Test Characteristic Curve (TCC) Method

Other methods

5. Divgi (1985)'s squared difference of item parameters (minimum chi-square method)
6. Haebara's (1980) Item Characteristic Curve (ICC) Method

Which one to use?

1-3 use only b-parameter and ignore the a-parameter
Research findings:
Way and Tang (1991) - Among (2,4,5,6), 4,5,6 better than 2.
Kim and Cohen (1991) - Among (3,4,5) , when used with small n, 4 is the best.
when n is large, 3, 4, 5 are about the same.

Guidelines for anchor items:

- # of anchor items (rule of thumb = 20-25%)
- acceptable range of b's (not too easy or too hard in one group)

Examples

Example 1 (p. 136)
Example 2 (p. 137)

IRT
Chapter 10
Computerized Adaptive Testing

Background

Binet (1908) - Intelligence Testing

Lord (1960's)

- Fixed-length tests were inefficient for most examinees, especially for low and high ability examinees.
- Tests can be shortened.

Computers - store test info, produce, administer, and score.

Promise of IRT

IRT is suitable for CAT ☺ Why?

Unidimensionality assumption

3PL model most appropriate

Basic Approach

Advantage (See p. 147)

☺ What about disadvantages?

Research Areas

1. Choice of IRT models
2. Item bank
3. Starting point
4. Selection of subsequent test items
5. Scoring/ability estimation
6. Stopping rule

Example (p. 149)

IRT
Chapter 11
Future Directions of Item Response Theory

Polytomous unidimensional IRT

Dichotomous and polytomous multidimensional IRT