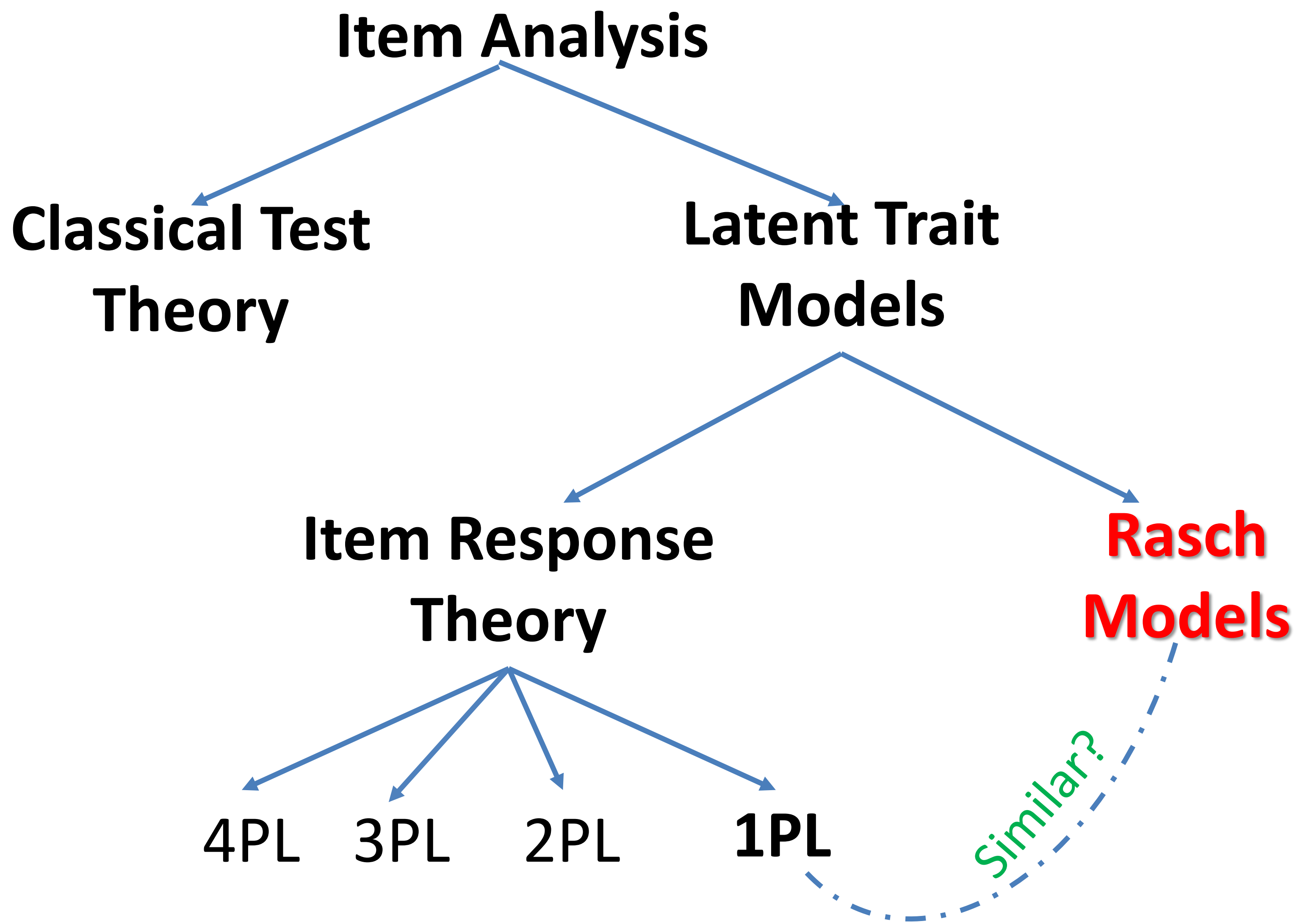
Workshop on Dichotomous Rasch Modeling with Ministep/Winsteps

Johnny Amora
Research Associate, De La Salle-College of Saint Benilde
Email: johnny.amora@gmail.com || Mobile: 0908-344-7184

PEMEA – Continuing Professional Development Program
DLS-CSB Auditorium
November 7, 2015

Contents


Item Analysis
Classical Test Theory
Latent Trait Models
Item Response Theory
Rasch Models
4PL
3PL
2PL
1PL
Similar?

Shortcomings of CTT

1. Examinee's ability is **exam-dependent**

For a fixed length test, examinee's ability is high if the test is easy; and examinee's ability is low if the test is difficult. Therefore, examinee's ability is exam-dependent.

2. Item/Test Difficulty is **group-dependent**

Item is easy if higher abilities take the test; and item is difficult if low abilities take the test. Therefore, item/test difficulty is group-dependent.

3. CTT is test-oriented.

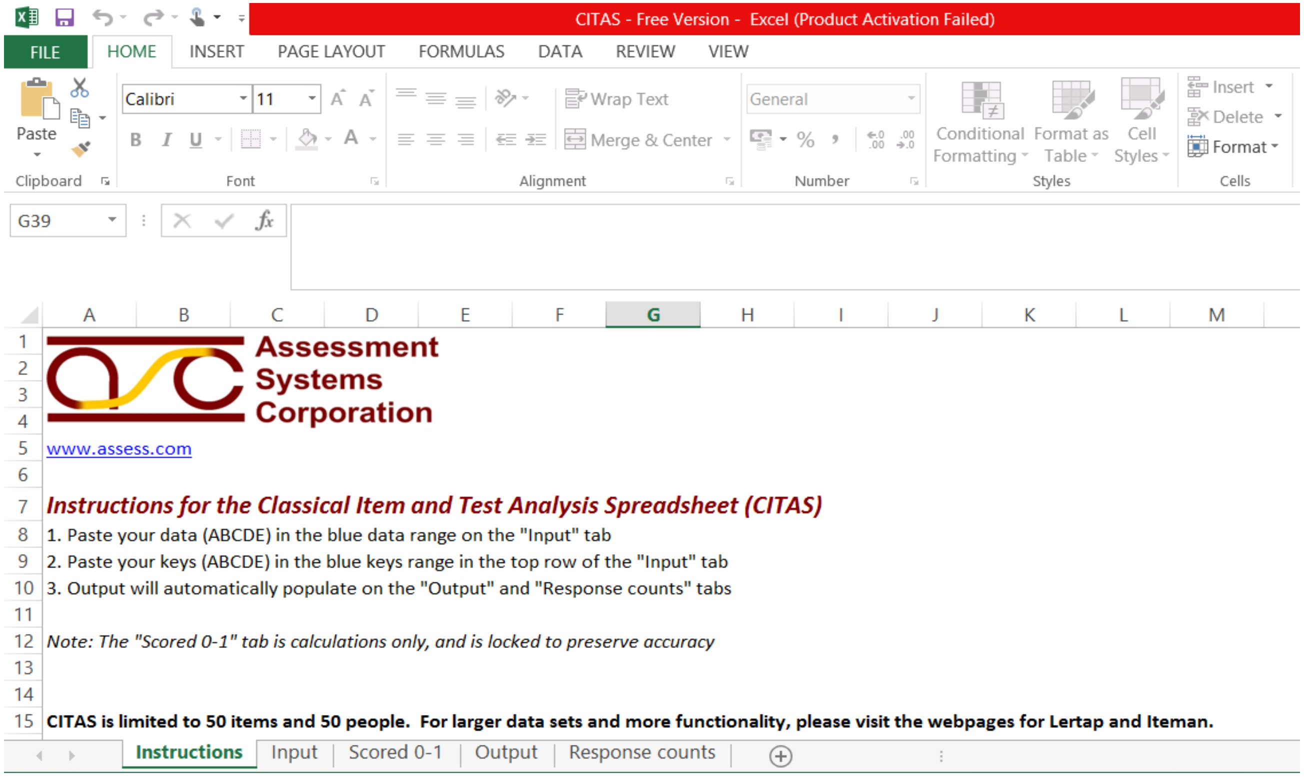
Score is given at the test level, but there is no basis in determining how well an examinee perform a particular item.

Limitations of CTT

1. Cannot predict (probabilistically) an examinee's response to an item.
2. Cannot predict individuals performance on certain items unless items have been administered to similar (comparable) individuals.
3. In adaptive testing, no mechanism exists in determining which item (*from an item pool*) is most appropriate to administer next.
4. Cannot determine how effective an item is at each level of ability.
5. Cannot estimate an examinee's ability from any given set of items

CTT Approach for Item Analysis using CITAS

CITAS is FREE!



Rasch Model

$$P(x_{vi} = 1 | \beta_v, D_i) = P_{vi} = \frac{e^{(\beta_v - D_i)}}{1 + e^{(\beta_v - D_i)}}$$

$$P(x_{vi} = 0 | \beta_v, D_i) = 1 - P_{vi}$$

P_{vi} = probability that person v gets a correct answer on item i, given his/her ability β and item difficulty D.

β_v = ability of examinee
D_i = difficulty of item i
e = 2.718 (*Euler's constant*)
X_{vi} = 1, 0 (*1 if correct answer; 0 if wrong*)
$-\infty < \beta_v < +\infty$ and $-\infty < D_i < +\infty$.

For example:

Suppose the ability of person v is β=3, while the difficulty level of item i is D=1. Then, the probability of getting a correct answer is .8808.

$$P_{vi} = \frac{e^{(3-1)}}{1+e^{(3-1)}} = \frac{e^{(2)}}{1+e^{(2)}} = .8808$$

Suppose the same person will attempt to answer an item with D=2. Then, the probability of getting a correct answer is .7311.

$$P_{vi} = \frac{e^{(3-2)}}{1+e^{(3-2)}} = \frac{e^{(1)}}{1+e^{(1)}} = .7311$$

Probabilities (P_{vi}) can be easily computed in Excel using the formula:

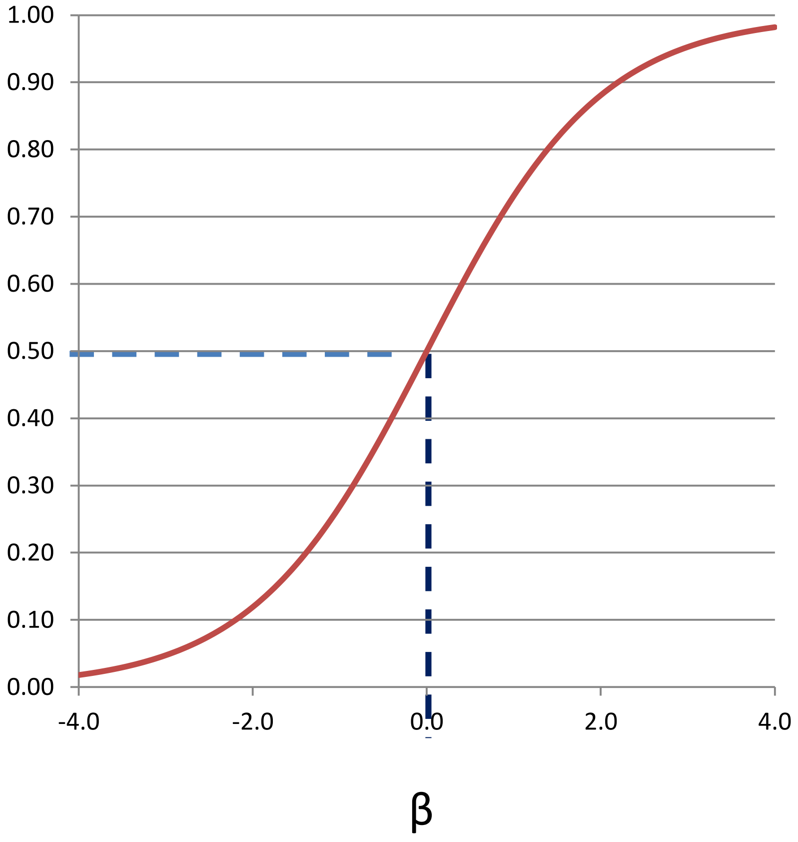
=Exp(B-D)/(1+EXP(B-D))

Excel outputs below:

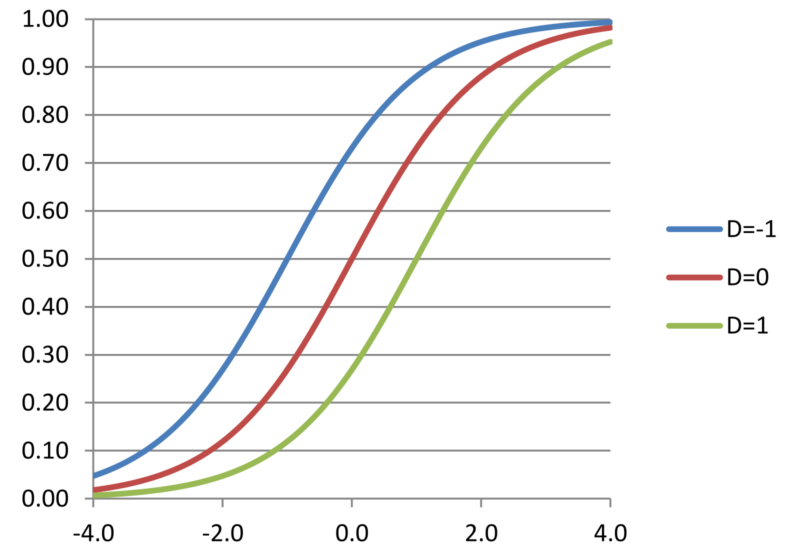
	Item1	Item2	Item3
B	**D=-1**	**D=0**	**D=1**
4.0	0.9933	0.9820	0.9526
3.9	0.9926	0.9802	0.9478
3.8	0.9918	0.9781	0.9427
3.7	0.9910	0.9759	0.9370
3.6	0.9900	0.9734	0.9309
3.5	0.9890	0.9707	0.9241
3.4	0.9879	0.9677	0.9168
3.3	0.9866	0.9644	0.9089
3.2	0.9852	0.9608	0.9002
3.1	0.9837	0.9569	0.8909
3.0	0.9820	0.9526	0.8808
2.9	0.9802	0.9478	0.8699
2.8	0.9781	0.9427	0.8581

Excelfile.xls

Item Characteristic Curves (ICCs)



Items with different difficulty levels



Notes:

For β=D, $P_{vi} = \frac{e^{(\beta_v \text{-} D_i)}}{1+e^{(\beta_v \text{-} D_i)}} = \frac{e^0}{1+e^0} = .5$

The probability of success is .5 if **person ability** matches the **item difficulty**.

For β>D, $P_{vi} > .5$

The probability of success is greater than .5 if **person ability** is higher than the **item difficulty** .

For β<D, $P_{vi} < .5$

The probability of success is lesser than .5 if the **person ability** is lesser than the **item difficulty**.

Odds for Success

Consider Person v:

Person v's odds of correctly getting a correct answer on item i, given his ability β_v and item difficulty D_I is defined as:

$$\frac{p_{vi}}{1-P_{vi}} = \frac{\frac{e^{(\beta_v - D_i)}}{1+e^{(\beta_v - D_i)}}}{1-\frac{e^{(\beta_v - D_i)}}{1+e^{(\beta_v - D_i)}}} = \frac{\frac{e^{(\beta_v - D_i)}}{1+e^{(\beta_v - D_i)}}}{\frac{1+e^{(\beta_v - D_i)} - e^{(\beta_v - D_i)}}{1+e^{(\beta_v - D_i)}}} = e^{(\beta_v - D_i)}$$

Taking Logarithm:

$$Log_e\left(\frac{P_{vi}}{1-P_{vi}}\right) = Log_e\left(e^{(\beta_v - D_i)}\right)$$

$$Log_e\left(\frac{P_{vi}}{1-P_{vi}}\right) = \beta_v - D_i$$

Person v's log odds of correctly getting a correct answer on item i, given his ability β_v and item difficulty D_I is defined as:

$$Log_e\left(\frac{P_{vi}}{1 - P_{vi}}\right) = \beta_v - D_i$$

<u>Consider another Person m</u>:

Similarly, Person m's log odds of getting a correct answer on the same item, given his ability B_m and item difficulty D_i:

$$Log_e\left(\frac{P_{mi}}{1 - P_{mi}}\right) = \beta_m - D_i$$

Comparing the abilities of Persons v and Person m:

To compare, we subtract the logarithm of odds:

$$(\beta_v - D_i) - (\beta_m - D_i) = Log_e\left(\frac{P_{vi}}{1-P_{vi}}\right) - Log_e\left(\frac{P_{mi}}{1-P_{mi}}\right)$$

$$\beta_v - \beta_m = Log_e\left(\frac{P_{vi}}{1-P_{vi}}\right) - Log_e\left(\frac{P_{mi}}{1-P_{mi}}\right)$$

Notice that the difference in the abilities of Person v and Person m does not involve D_i at all. This means that comparison on the person abilities does not depend on which particular item is used and so comparison is **"ITEM-FREE"**.

An analogous argument leads to "**PERSON-FREE**" comparisons of item difficulties.

Estimation of Rasch Parameters [1]

Winsteps implements te following methods of estimating Rasch parameters:

- JMLE (*Joint Maximum Likelihood Estimation by Wright and Panchapakesan*),
- PROX (*Normal Approximation Algorithm devised by Cohen (1979*)).

Estimation of Rasch Parameters [2]

Rasch measures are obtained by iterating through the data.

- **STEP 1**: Initially all unanchored parameter estimates (measures) are set to zero.
- **STEP 2:** Then the PROX method is employed to obtain rough estimates. Each iteration through the data improves the PROX estimates until they are usefully good.
- **STEP 3:** Then those PROX estimates are the initial estimates for JMLE which fine-tunes them, again by iterating through the data, in order to obtain the final JMLE estimates. The iterative process ceases when the convergence criteria are met.

Convergence Criteria

- Largest Logit change
  (default of Winsteps: LCONV=.0001 logits)

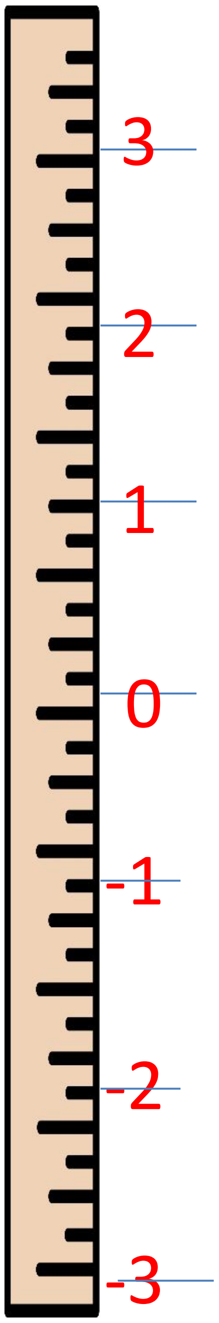
- Largest Score Residual
  (default of Winsteps: RCONV=.01 score points)

- MJMLE= 0 ; unlimited JMLE iterations

Rasch "Yardstick" [1]

- Rasch model creates a "yardstick" that can be used to measure both **Person Ability** and **Item Difficulty.**

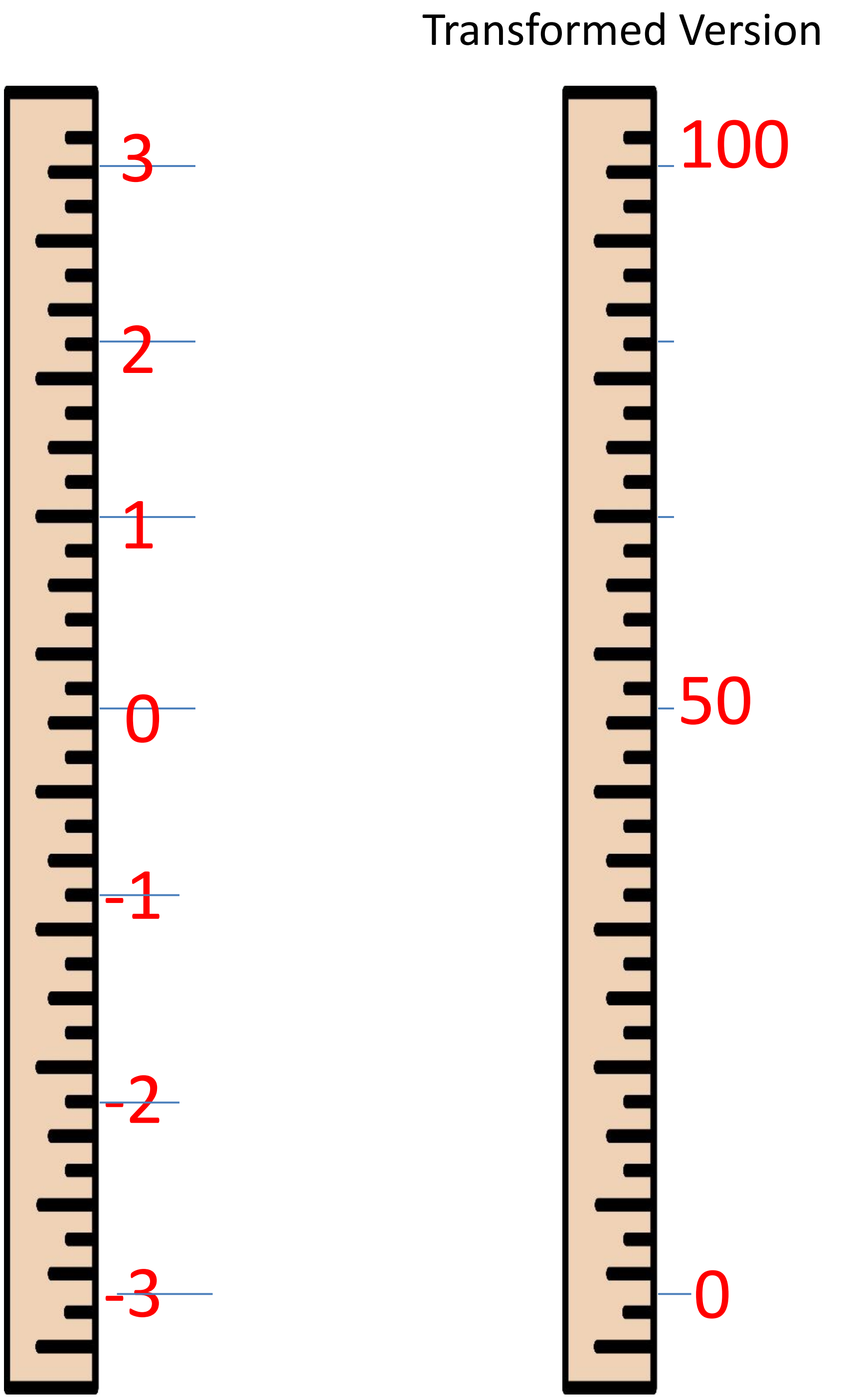
- **The values in the yardstick are logits, which range between -∞ and + ∞. But for application purposes, they range between -4 and +4 or between -3 and +3.**



Rasch "Yardstick" [2]

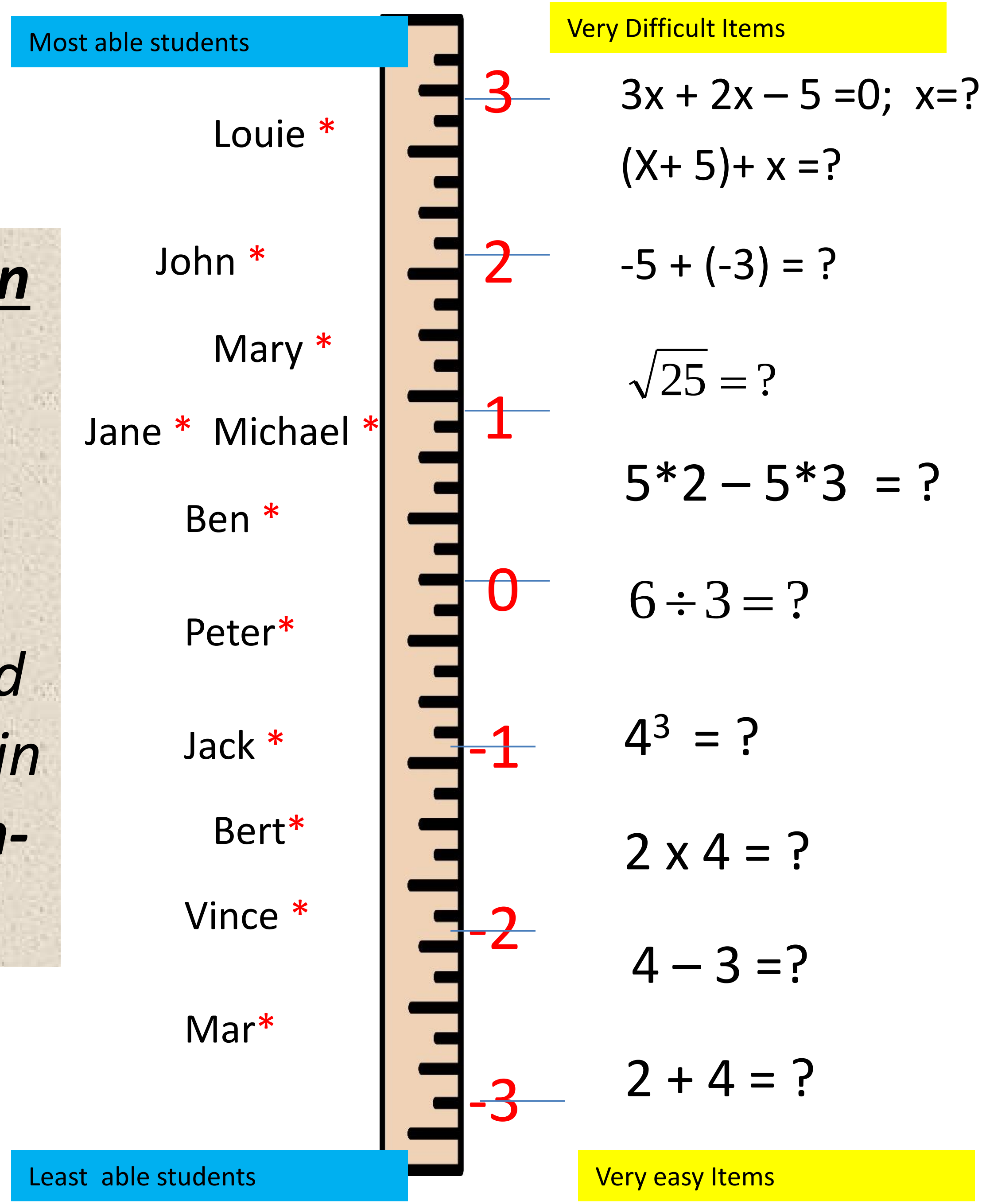
- The logits can be transformed so that the range would be understandable by non-technical users.

- For example, the logits can be transformed so that the values would fall between 0 and 100.



Person-Item Map [1]

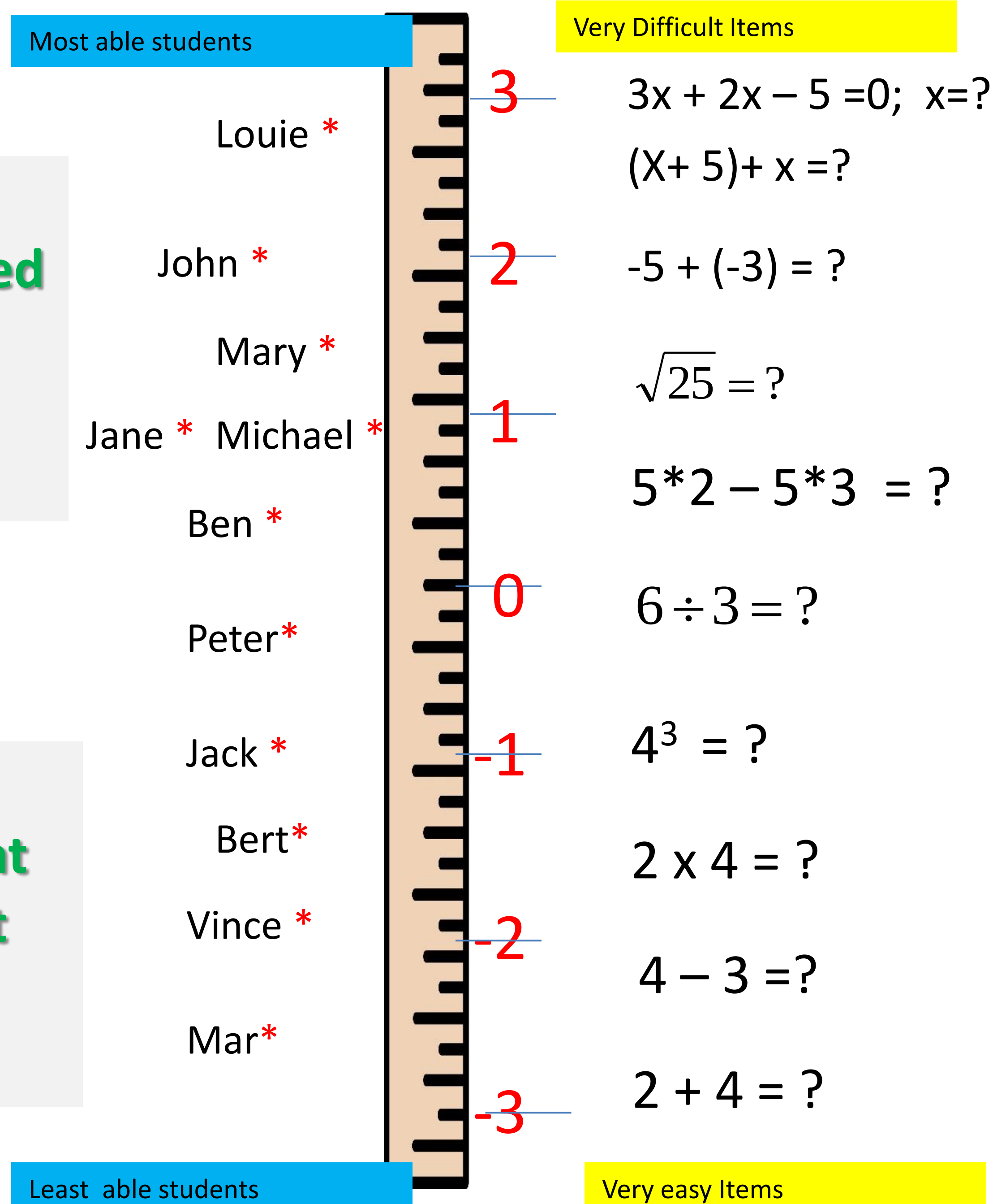
Because ***Person ability*** *and* ***item difficulty*** *are measured using a common "yardstick, then both person and items can be placed in a map, called* ***Person-Item map.***



Person-Item Map [2]

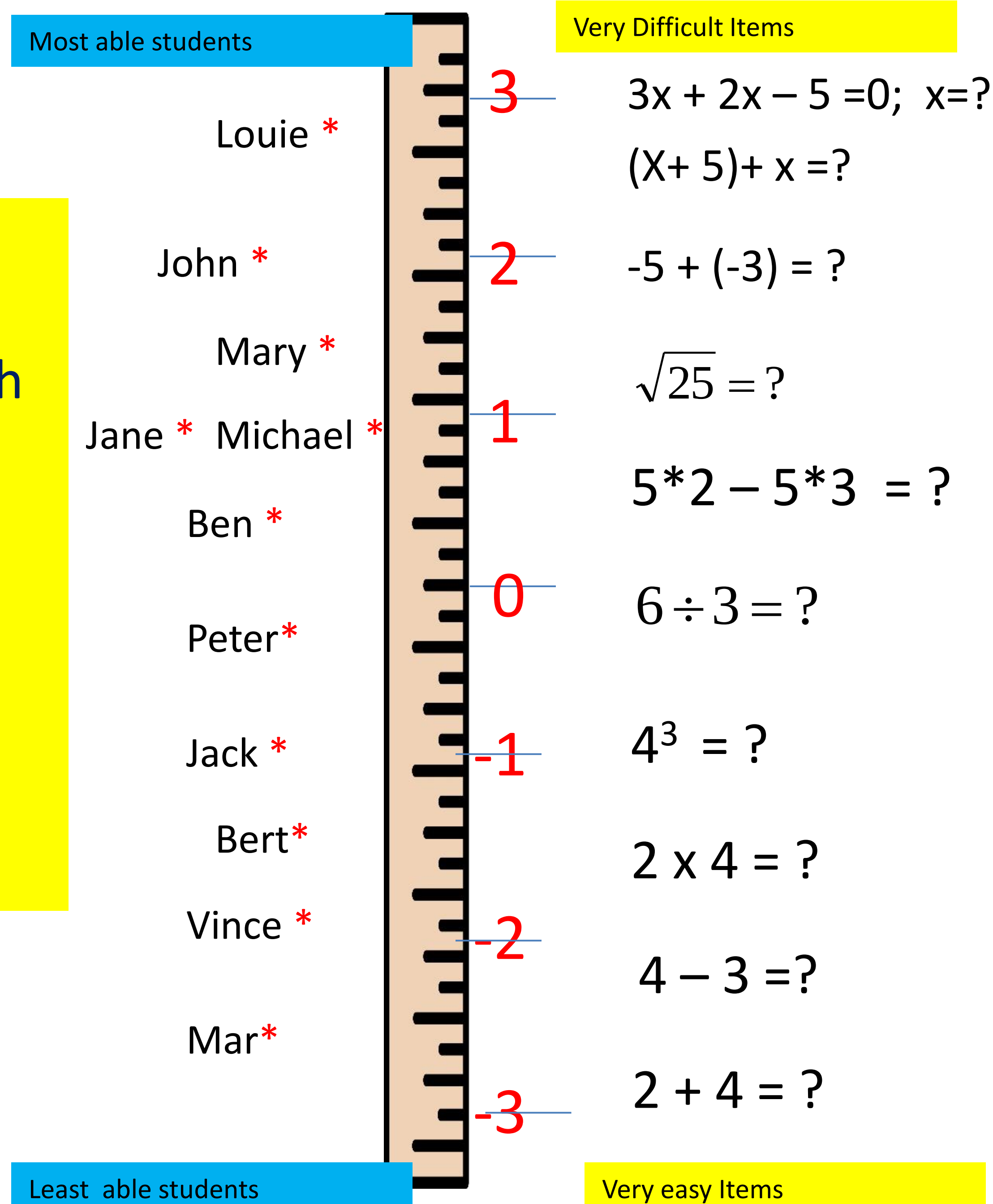
- **Items are hierarchically arranged from very easy (bottom) up to very difficult (top).**

- **Also, least able students are placed at the bottom and most able students at the top.**



Person-Item Map [3]

The nice with Rasch modeling is that we can determine which students are able to answer correctly which items. Or, we can determine which items can be answered correctly by which students.

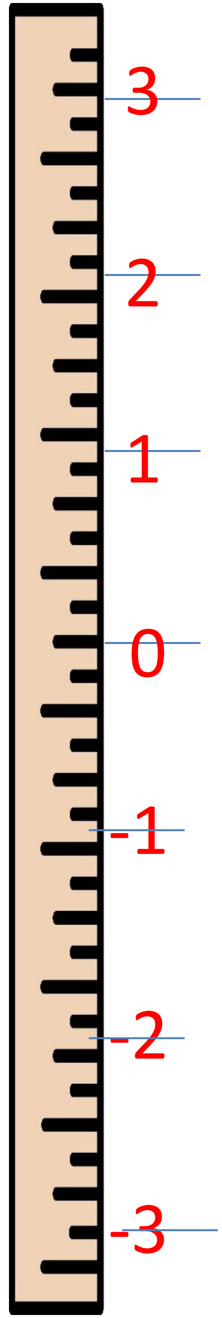


Things to consider in constructing a test using Rasch Model

- Validity: Construct Validity, Fit validity
- Reliability

Construct Validity:

Does the item difficulty hierarchy make sense?



$3x + 2x - 5 = 0;\ \ x = ?$

$(X + 5) + x = ?$

$-5 + (-3) = ?$

$\sqrt{25} = ?$

$5*2 - 5*3 = ?$

$6 \div 3 = ?$

$4^3 = ?$

$2 \times 4 = ?$

$4 - 3 = ?$

$2 + 4 = ?$

Louie *

John *

Mary *

Jane * Michael *

Ben *

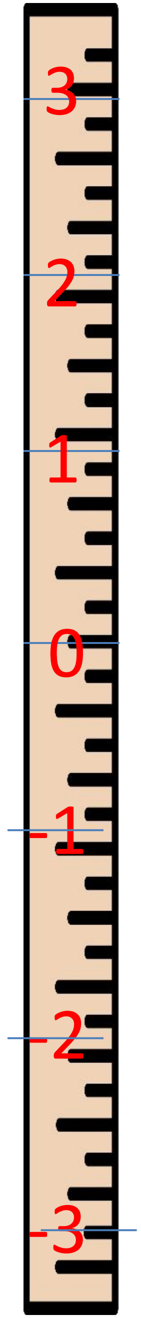
Peter*

Jack *

Bert*

Vince *

Mar*



$3x + 2x - 5 = 0$; $x = ?$

$(X + 5) + x = ?$

$-5 + (-3) = ?$

$\sqrt{25} = ?$

$5*2 - 5*3 = ?$

$6 \div 3 = ?$

$4^3 = ?$

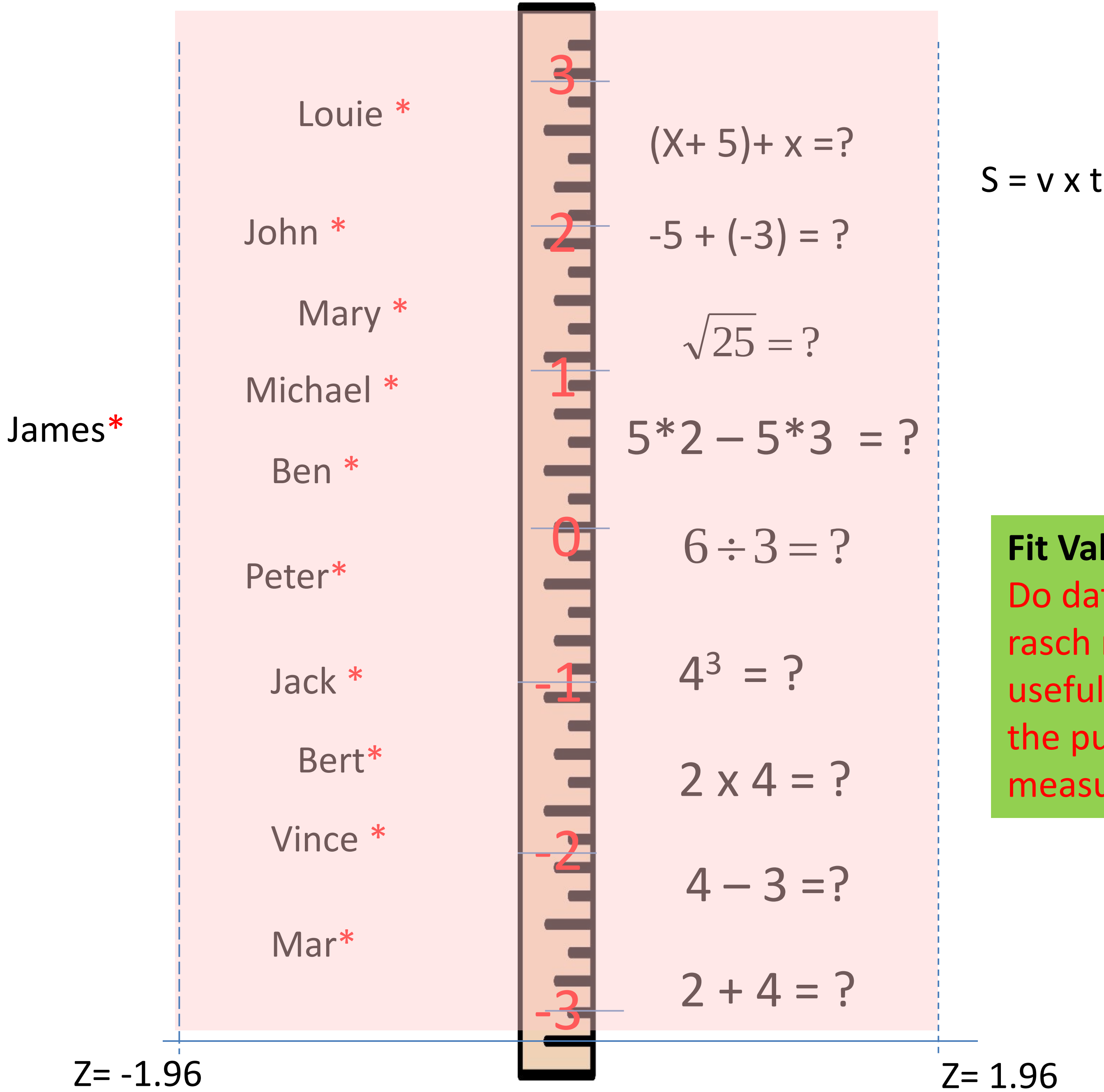
$2 \times 4 = ?$

$4 - 3 = ?$

$2 + 4 = ?$

Predictive Validity:

Does the person ability hierarchy make sense?



Fit Validity:
Do data fit the rasch model usefully well for the purpose of measurement?

Concept of Model and Data Fit

- Major concern of Rasch Modeling is its need for **unidimensionality.**
- Investigation of fit statistics determines whether the data are unidimensional in nature.
- Both **infit and outfit statistics** are evaluated to determine how **data-to-model** fit occurs for each item and person fit.

Infit and Outfit Statistics

- Infit statistics are sensitive to the <u>inlier</u> pattern of observations.
- Outfit statistics are sensitive to <u>outlier</u> observations.

Idealized Guttman Scale (Gutman, 1944)

Easiest items → **Hardest Items**

Most Able Student

	1	2	3	4	5	6
A	1	1	1	1	1	0
B	1	1	1	1	0	0
C	1	1	1	0	0	0
D	1	1	0	0	0	0
E	1	0	0	0	0	0

Least Able Student

Data with large infit statistics

Easy items **Difficult items**

1 1 1 0 0 1 1 0 0 0

Larger infit statistics because the 1's occurring in the middle-right section of the continuum and the 0's appearing in the middle-left section of the continuum are unexpected.

Data with large outfit statistics

0 0 1 1 1 0 0 0 1 1

Larger outfit statistics because observations at the extreme ends of the continuum are unexpected.

Fit Statistics as Indicator of Validity

Data adequately fitting the model is a key indicator of validity.

- Removal of misfitting persons and items that grossly misfit the model's expectation is acceptable.

[Removal of misfitting persons and items improves the precision of the measures produced.]

Formulas: *Fit Statistics*

- Outfit Mean Square: outlier-sensitive fit statistic. This is based on the conventional chi-square statistic.

Outfit Mean Square = average [(standardized residuals^2)] = chi-square/d.f.

- Infit Mean Square: inlier-pattern-sensitive fit statistic. This is based on the chi-square statistic with each observation weighted by its statistical information (model variance).

Infit Mean Square = average [(standardized residuals^2)* information)]

- **Z-Standardized:** statistical significance (probability) of the chi-square (mean-square) statistics occurring by chance when the data fit the Rasch model. The values reported are unit-normal deviates.

ZSTD probabilities: two-sided unit-normal deviates	
1.00	p= .317
1.96	.050
2.00	.045
3.00	.0027
4.00	.00006
5.00	.0000006

Infit Mean Square and Outfit Mean Square: Rule of Thumbs

> 2.0	**Distorts or degrades the measurement system.**
1.5 - 2.0	**Unproductive for construction of measurement, but not degrading.**
0.5 - 1.5	**Productive for measurement.**
<0.5	**Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations.**

Note: **1 = Expected Value (perfect fit)**

Reliability

- Reliability means reproducible of relative measure location.
- “High item reliability" means that there is a high probability that items estimated with high measures actually do have higher measures than items estimated with low measures.
- “High person reliability" means that there is a high probability that persons estimated with high measures actually do have higher measures than persons estimated with low measures.

Person Reliability

Person reliability depends chiefly on:

- **Sample ability variance. Wider ability range = higher person reliability.**
- **Length of test. Longer test = higher person reliability**
- **Sample-item targeting. Better targeting = higher person reliability**

How to increase person reliability?

– test persons with more extreme abilities (high and low)

– lengthen the test.

Person Reliability is independent of sample size. It is largely uninfluenced by model fit.

Item Reliability

Item reliability depends chiefly on

- **Item difficulty variance**. **Wide difficulty range = high item reliability**
- **Person sample size**. **Large sample = high item reliability**

How to increase item reliability?

– test more people.

Item Reliability is independent of test length. It is largely uninfluenced by model fit.

Sample Size Requirements

- Rasch is the same as any other statistical analysis with a small sample:
  - Less precise estimates (*bigger standard errors*)
  - Less powerful fit analysis
  - Less robust estimates

- Very small sample (*say, n=2 or 3 examinees*) provides a very unstable results, while very large sample (say, n=2000 or 3000) provides a very precise results. However, large sample is too expensive and time-consuming. So, how big a sample is necessary?

- Linacre (1994) provides the following sample size guidelines:

Item Calibrations stable within	Confidence	Minimum sample size range (best to poor targeting)	Size for most purposes
± 1 logit	95%	16 † -- 36	30 (minimum for dichotomies)
± 1 logit	99%	27 † -- 61	50 (minimum for polytomies)
± ½ logit	95%	64 -- 144	100
± ½ logit	99%	108 -- 243	150
Definitive or High Stakes	99%+ (Items)	250 -- 20*test length	250
Adverse Circumstances	Robust	450 upwards	500

Reference:
Linacre JM. (1994). Sample Size and Item Calibration Stability. Rasch Measurement Transactions, 7:4 p.328

Some Features of Rasch Model

1. Examinee performance on an unadministered item can be predicted.
2. Item and ability parameters can be estimated
3. Item parameter estimates are independent of the group of examinees who took the test
4. Examinee ability estimates are independent of the group of test items administered
5. Precision of ability estimates is known

Rasch versus 1PL IRT

Aspect	Rasch	IRT: One-parameter Logistic Model
Symbol	Rasch	1PL IRT, also 1PL
For Practical purposes	When each individual in the person sample is parameterized for item estimation, it is Rasch.	When the person sample is parameterized by a mean and standard deviation for item estimation, it is 1PL IRT.
Motivation	**Prescriptive:** Distribution-free person ability estimates and distribution-free item difficulty estimates on an additive latent variable	**Descriptive:** Computationally simpler approximation to the Normal Ogive Model of L.L. Thurstone, D.N. Lawley, F.M. Lord
Formulation: Exponential Form	$P_{ni} = \frac{e^{B_n - D_i}}{1 + e^{B_n - D_i}}$	$P_i(\theta) = \frac{e^{1.7(\theta - b_i)}}{1 + e^{1.7(\theta - b_i)}}$
Formulation: Logit-linear form	$\log_e\left(\frac{P_{ni}}{1 - P_{ni}}\right) = B_n - D_i$	$\log_e\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right) = 1.7(\theta - b_i)$

Aspect	Rasch	IRT: One-parameter Logistic Model
Students/persons	Person *n* of ability B_n in logits	Normally-distributed person sample of ability distribution θ, conceptualized as N(0,1), in probits; persons are incidental parameters
Items, multiple-choice questions, etc.; items are structural parameters	Item *i* of difficulty D_i in logits	Item *i* of difficulty b_i (the "one parameter") in probits
Nature of binary data	1 = "correct" 0 = "wrong"	1 = "correct" 0 = "wrong"
Probability of binary data	P_{ni} = probability that person *n* correctly answered item *i*	$P_i(\theta)$ = overall probability of "correct" by person distribution θ on item *i*

Aspect	Rasch	IRT: One-parameter Logistic Model
Local origin of scale: zero of parameter estimates	Average item difficulty, or difficulty of specified item. (Criterion-referenced)	Average person ability. (Norm-referenced)
Item discrimination	Item characteristic curves (ICCs) modeled to be parallel with a slope of 1 (the natural logistic ogive)	ICCs modeled to be parallel with a slope of 1.7 (approximating the slope of the cumulative normal ogive)
Fit evaluation	*Fit of the data to the model* Local, one parameter at a time	*Fit of the model to the data* Global, accept or reject the model
Data-model mismatch	Defective data do not support parameter separability in an additive framework. Consider editing the data.	Defective model does not adequately describe the data. Consider adding discrimination (2-PL), lower asymptote (guessability, 3-PL) parameters.

Aspect	Rasch	IRT: One-parameter Logistic Model
Minimum reasonable sample size	**30** **Linacre (1994)**	**200** **(Downing 2003)**
First conspicuous appearance	**Rasch, Georg. (1960)** Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.	**Birnbaum, Allan. (1968).** Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
First conspicuous advocate	**Benjamin D. Wright,** University of Chicago	**Frederic M. Lord,** Educational Testing Service
Widely-authoritative currently-active proponent	**David Andrich,** Univ. of Western Australia, Perth, Australia	**Ronald Hambleton**, University of Massachusetts

Source (of the comparison):
Linacre J.M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. Rasch Measurement Transactions, 19:3, 1032

Demonstration
&
Computer Hands-on

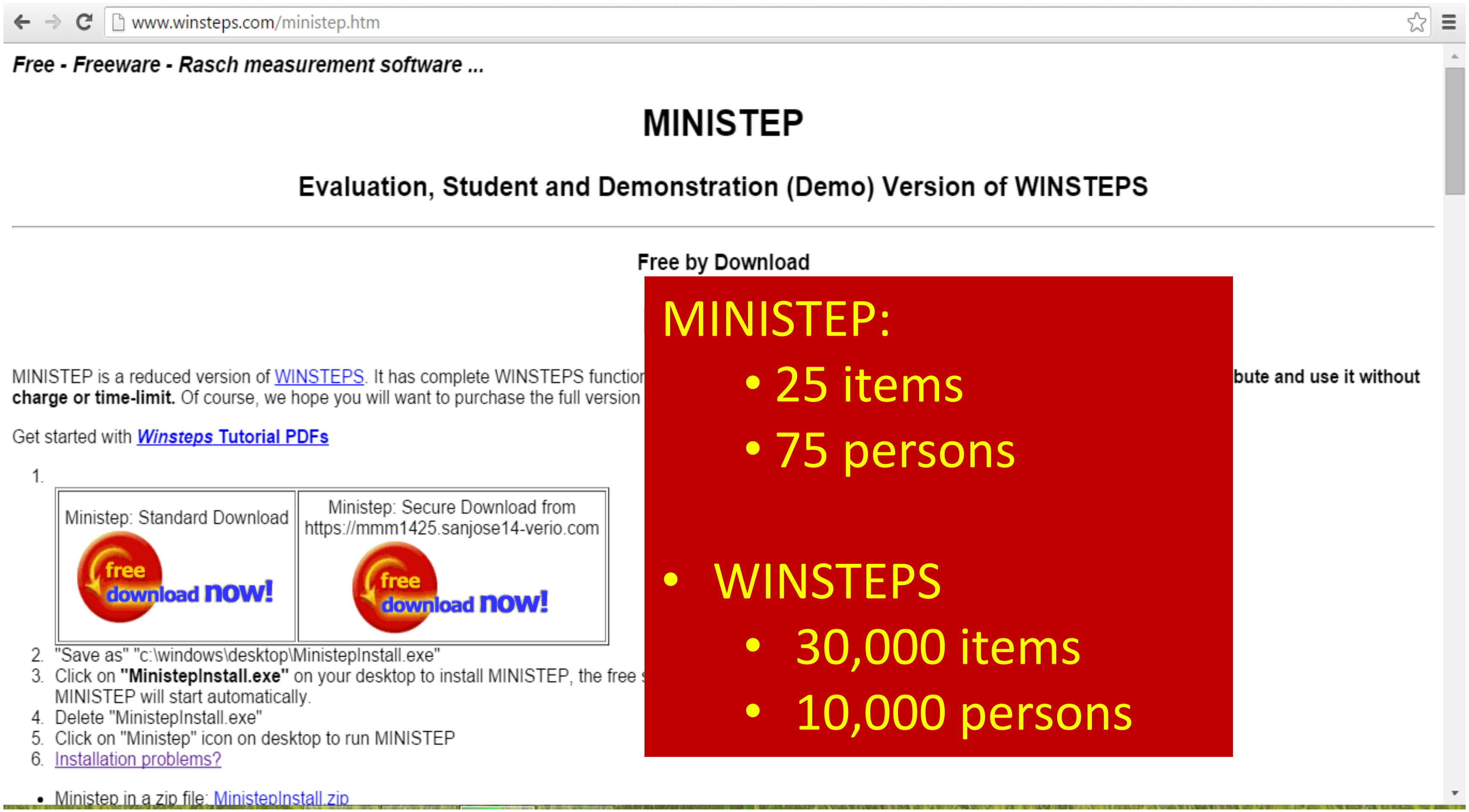
Rasch Analysis using Winsteps/Ministep

Download data and installers from this link:

https://www.dropbox.com/sh/vkxfosv6c630ftm/AAB-A2NqO2QG44J_Wa8CuvAsa?dl=0

Download Free Ministep

www.winsteps.com/ministep.htm



MINISTEP:
- 25 items
- 75 persons

- WINSTEPS
  - 30,000 items
  - 10,000 persons

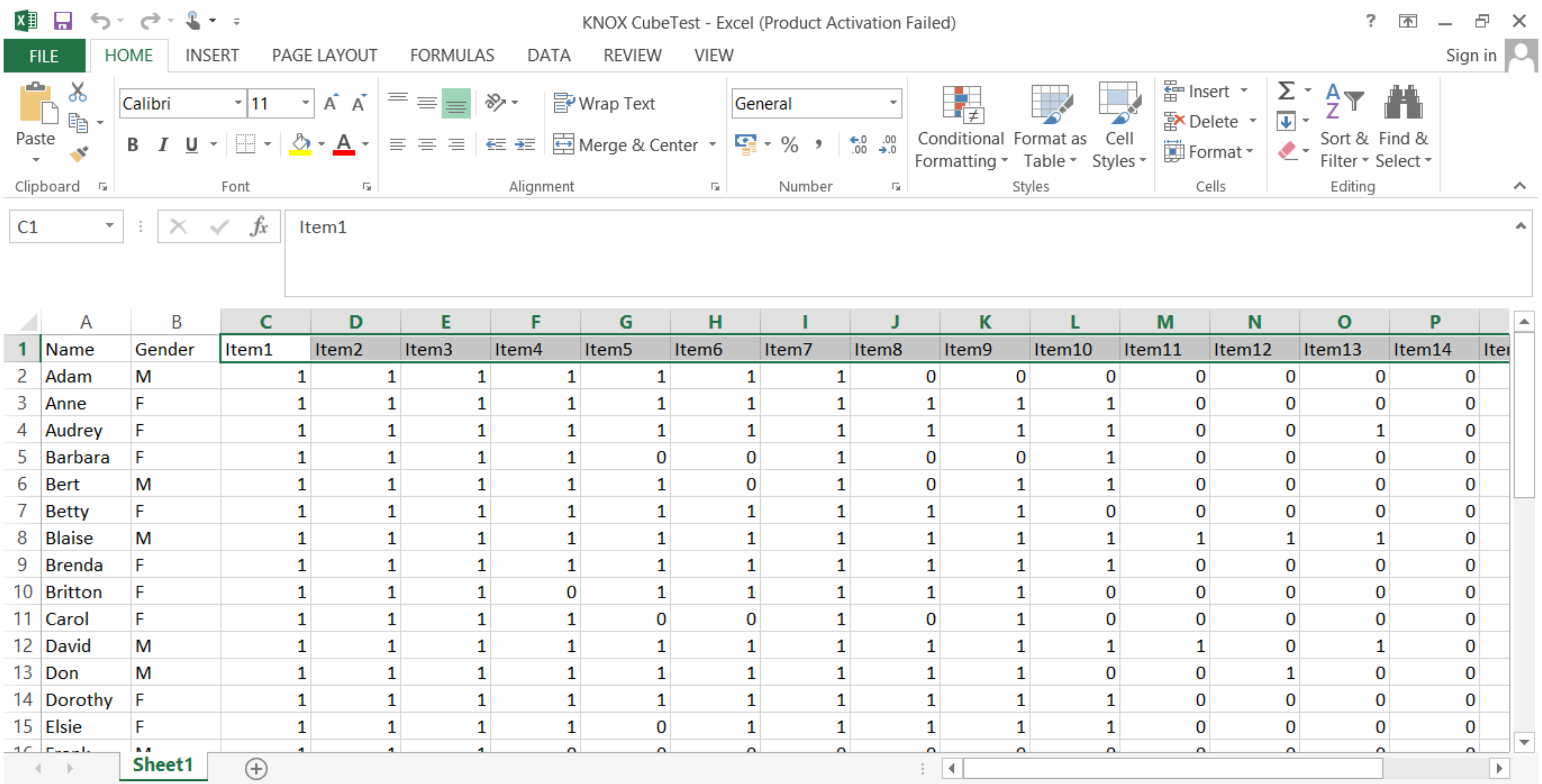
Example 1- Math Curriculum Test

	Items												
Persons	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	Raw Score
A	✔	✔	✔	✗	✗	✗	✗	✔	✔	✗	✗	✔	6
B	✔	✗	✔	✔	✗	✗	✗	✗	✔	✗	✗	✗	4
C	✔	✔	✔	✗	✔	✗	✗	✔	✔	✔	✔	✔	9
D	✔	✗	✔	✔	✗	✗	✗	✗	✔	✗	✗	✔	5
E	✗	✔	✔	✗	✗	✔	✗	✔	✔	✔	✔	✔	8
F	✔	✔	✔	✔	✔	✗	✗	✔	✔	✗	✗	✔	8
G	✔	✗	✔	✗	✗	✔	✗	✗	✔	✗	✔	✔	6
H	✔	✗	✔	✗	✗	✗	✗	✗	✗	✗	✗	✔	3
I	✔	✔	✔	✔	✗	✗	✗	✔	✔	✗	✗	✔	7
J	✔	✔	✔	✔	✔	✔	✔	✔	✔	✗	✔	✗	10
K	✔	✗	✔	✗	✗	✔	✗	✗	✔	✗	✔	✔	6
L	✗	✔	✔	✗	✗	✔	✗	✔	✔	✔	✔	✔	8
M	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0
N	✔	✔	✔	✔	✗	✔	✔	✔	✔	✔	✔	✔	11

12 items
14 persons

Example 2 – KNOX CUBE Test

18 items, 35 persons



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Name	Gender	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12	Item13	Item14
2	Adam	M	1	1	1	1	1	1	1	0	0	0	0	0	0	0
3	Anne	F	1	1	1	1	1	1	1	1	1	1	0	0	0	0
4	Audrey	F	1	1	1	1	1	1	1	1	1	1	0	0	1	0
5	Barbara	F	1	1	1	1	0	0	1	0	0	1	0	0	0	0
6	Bert	M	1	1	1	1	1	0	1	0	1	1	0	0	0	0
7	Betty	F	1	1	1	1	1	1	1	1	1	0	0	0	0	0
8	Blaise	M	1	1	1	1	1	1	1	1	1	1	1	1	1	0
9	Brenda	F	1	1	1	1	1	1	1	1	1	1	0	0	0	0
10	Britton	F	1	1	1	0	1	1	1	1	1	0	0	0	0	0
11	Carol	F	1	1	1	1	0	0	1	0	1	0	0	0	0	0
12	David	M	1	1	1	1	1	1	1	1	1	1	1	0	1	0
13	Don	M	1	1	1	1	1	1	1	1	1	0	0	1	0	0
14	Dorothy	F	1	1	1	1	1	1	1	1	1	1	0	0	0	0
15	Elsie	F	1	1	1	1	0	1	1	1	1	1	0	0	0	0

Q & A

References

Linacre J.M. (2005). *Rasch dichotomous model vs. One-parameter Logistic Model*. Rasch Measurement Transactions, 19:3, 1032

Linacre JM. (1994). Sample Size and Item Calibration Stability. Rasch Measurement Transactions, 7:4 p.328

Bond, T.G. & Fox, C. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd Ed).

Thank you!

For further questions, contact:

Johnny Amora
De La Salle-College of Saint Benilde

Email: johnny.amora@gmail.com **|| Mobile: 0908-344-7184**