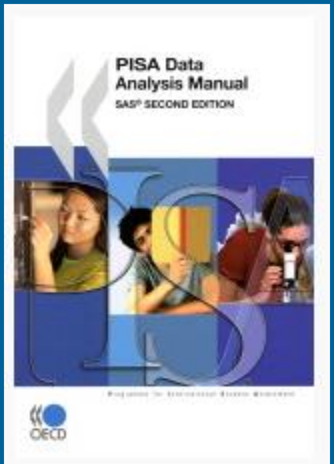
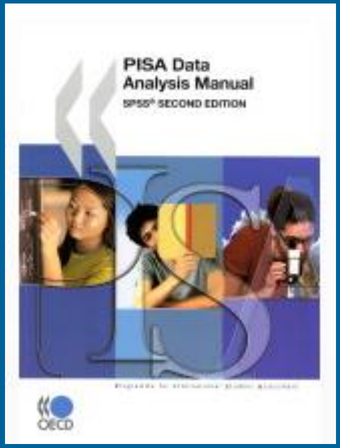


Scaling of the Cognitive Data and Use of Student Performance Estimates



Guide to the PISA Data Analysis Manual

OECD
BETTER POLICIES FOR BETTER LIVES

Classical Test Theory *versus* Item Response Theory

- Questions
  - What is the *Rasch* model?
  - Why do International Surveys use the *Rasch* model or another IRT model?
  - Which kind of information do we get from the *Rasch* model?
- Test design
  - Experts groups request lots of items
  - School principals want to limit testing time
- ⇒ Incomplete design

Classical Test Theory *versus* Item Response Theory

- Incomplete Block Design

Booklet	*Part 1*	*Part 2*	*Part 3*	*Part 4*
1	M1	M2	M4	S1
2	M2	M3	M5	S2
3	M3	M4	M6	P1
4	M4	M5	M7	P2
5	M5	M6	R1	M1
6	M6	M7	R2	M2
7	M7	R1	S1	M3
8	R1	R2	S2	M4
9	R2	S1	P1	M5
10	S1	S2	P2	M6
11	S2	P1	M1	M7
12	P1	P2	M2	R1
13	P2	M1	M3	R2
UH	M-R-S-P	M-R-S-P		

Classical Test Theory *versus* Item Response Theory

- Impossible comparisons
  - How can we compare the difficulty of two items from two different test booklets?
  - How can we compare the performance of two students who have answered two different test booklets?
- Long time ago
  - Test booklets have exactly the same difficulty and therefore the differences in score reflects differences in abilities
    - OR / AND
  - The randomization of test allocation guarantees the comparability of the sub-populations and therefore differences in item parameters reflect differences in item difficulties

Item Response Theory

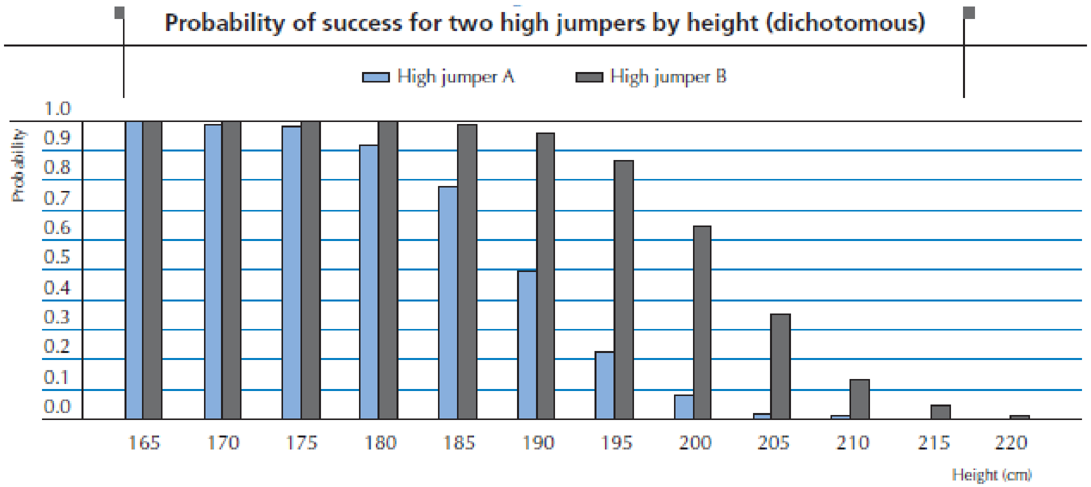
- IRT models solve this problem:
  - None of the above assumptions has to be made
  - As far as there is a link between the test booklets
    - Item difficulties can be compared
    - Student performances can be compared

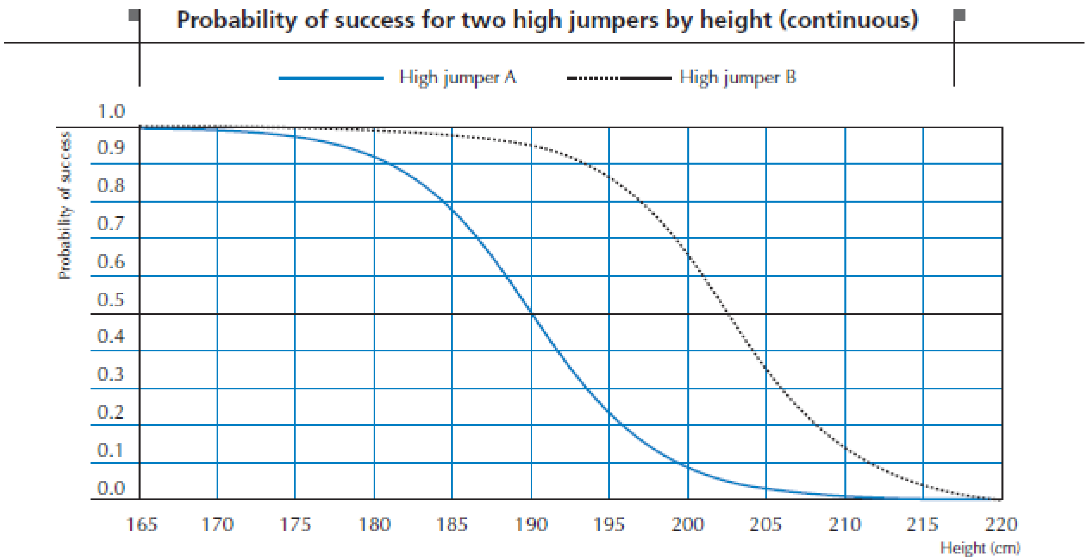
Item Response Theory

- What is the performance of the jumper? Is it…
  - Individual record?
  - Individual record during an official event?
  - The mean of performance during a particular period of time?
  - The most frequent performance during a particular period of time?

- Use of the logistic regression

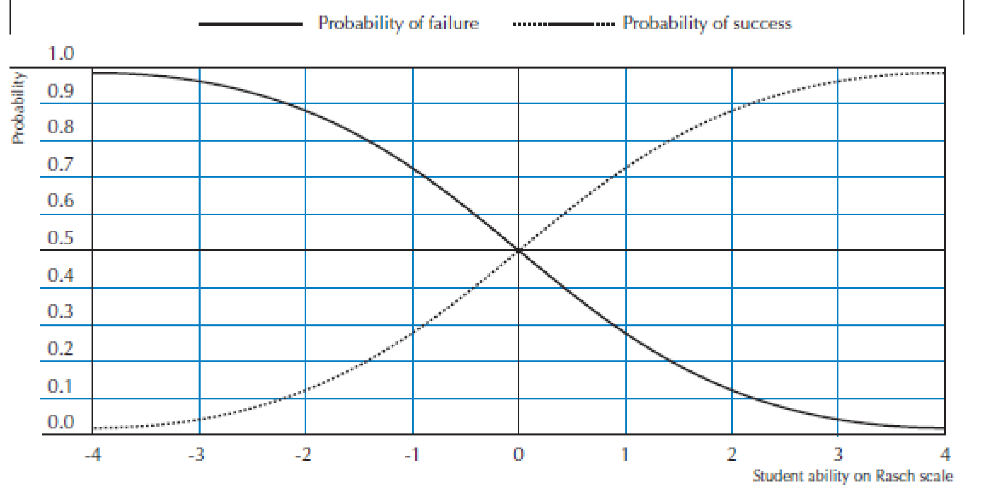
Rasch Item Response Theory

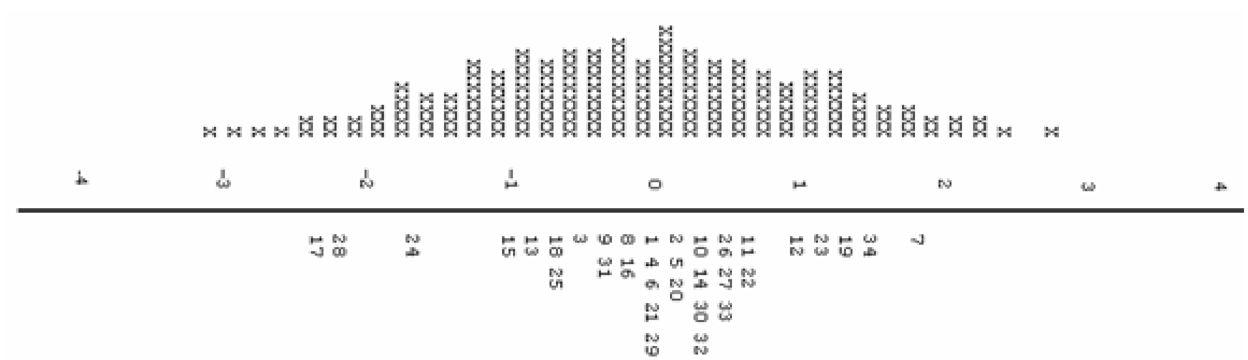




Rasch Item Response Theory

Probability of success to an item of difficulty zero as a function of student ability





Rasch Item Response Theory

- How can we predict a probability of success or failure?
  - Linear regression

$$Y_i = \beta_0 + \beta_1 X_i$$

  - The dependent variable can vary from -∞ to + ∞

$$e^{(\beta_0 + \beta_1 X_i)} \in \left]0,+\infty\right[$$

$$\frac{e^{(\beta_0 + \beta_1 X_i)}}{1 + e^{(\beta_0 + \beta_1 X_i)}} \in \left]0,1\right[$$

Rasch Item Response Theory

- Rasch IRT Model
  - One-parameter logistic model

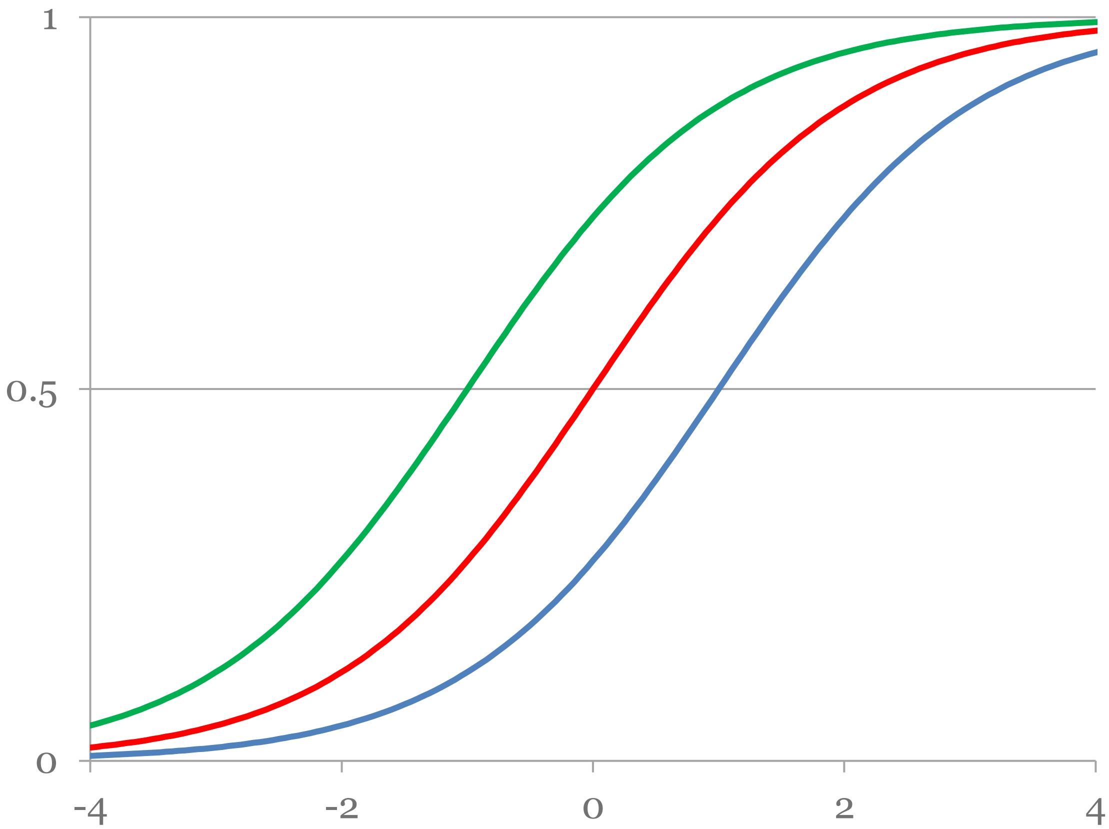
$$P\left[X_{ij}=1 \middle| \beta_i,\delta_j\right]=\frac{\exp(\beta_i-\delta_j)}{1+\exp(\beta_i-\delta_j)}=\frac{e^{(\beta_i-\delta_j)}}{1+e^{(\beta_i-\delta_j)}}$$

$$P\left[X_{ij}=0 \middle| \beta_i,\delta_j\right]=\frac{1}{1+\exp(\beta_i-\delta_j)}=\frac{1}{1+e^{(\beta_i-\delta_j)}}$$

$$\Rightarrow \quad \frac{\exp(\beta_i-\delta_j)}{1+\exp(\beta_i-\delta_j)}+\frac{1}{1+\exp(\beta_i-\delta_j)}=\frac{1+\exp(\beta_i-\delta_j)}{1+\exp(\beta_i-\delta_j)}=1$$

Rasch Item Response Theory

- Item Characteristics Curves (ICC)

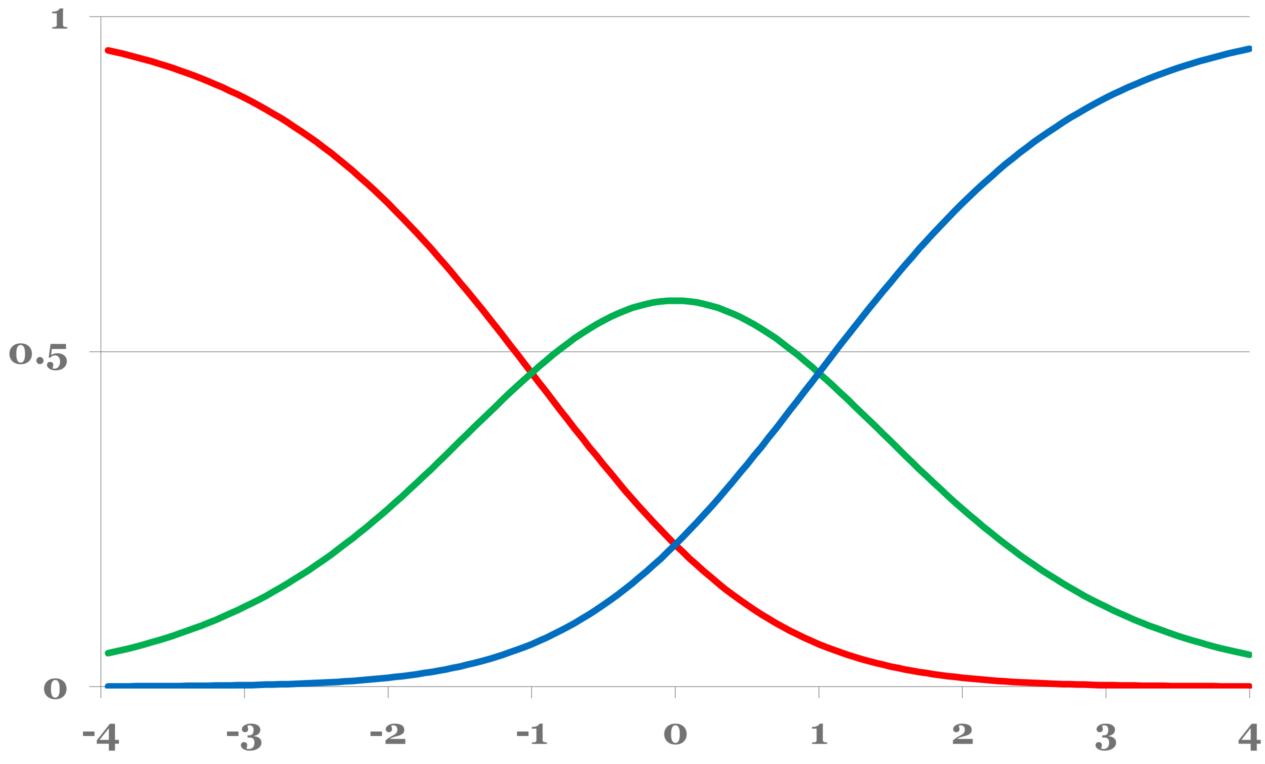


Rasch Item Response Theory

Student performance	Item Difficulty	Probability	Student performance	Item Difficulty	Probability
-2	-2	0.50			
-1	-1	0.50			
0	0	0.50			
1	1	0.50			
2	2	0.50			
-2	-3	0.73	-2	-1	0.27
-1	-2	0.73	-1	0	0.27
0	-1	0.73	0	1	0.27
1	0	0.73	1	2	0.27
2	1	0.73	2	3	0.27
-2	-4	0.88	-2	0	0.12
-1	-3	0.88	-1	1	0.12
0	-2	0.88	0	2	0.12
1	-1	0.88	1	3	0.12
2	0	0.88	2	4	0.12

Rasch Item Response Theory

- Item Characteristics Curves (ICC) Partial Credit Item



Rasch Item Response Theory

- Partial Credit Item

$$P(X_{ni}=2)=\frac{\exp(2\beta_n-2\delta_j-t_{i1}-t_{i2})}{1+\exp(\beta_n-\delta_j-t_{i1})+\exp(2\beta_n-2\delta_j-t_{i1}-t_{i2})}$$

$$P(X_{ni}=1)=\frac{\exp(\beta_n-\delta_j-t_{i1})}{1+\exp(\beta_n-\delta_j-t_{i1})+\exp(2\beta_n-2\delta_j-t_{i1}-t_{i2})}$$

$$P(X_{ni}=0)=\frac{1}{1+\exp(\beta_n-\delta_j-t_{i1})+\exp(2\beta_n-2\delta_j-t_{i1}-t_{i2})}$$

Rasch Item Response Theory

High Achievers ***Difficult items***

```
                         |
                         |
                        X|
                         |
                        X|
  2                    XX|
                       XX|7
                       XX|
                      XXX|34
                      XXX|19
                     XXXX|23
  1                XXXXXX|12
                   XXXXXX|
                    XXXXX|11 22
                   XXXXXX|26 27 33
                  XXXXXXX|10 14 30 32
                  XXXXXXX|2 5 20
  0              XXXXXXXX|1 4 6 21 29
               XXXXXXXXXX|8 16
                  XXXXXXX|9 31
                XXXXXXXXX|3
                 XXXXXXXX|18 25
                 XXXXXXXX|13
 -1               XXXXXXX|15
                 XXXXXXXX|
                   XXXXXX|
                  XXXXXXX|
                     XXXX|24
                     XXXX|
 -2                 XXXXX|
                      XXX|28
                       XX|17
                       XX|
                       XX|
                        X|
 -3                     X|
                        X|
                        X|
                         |
```

Low Achievers ***Easy items***

Rasch Item Response Theory

- Step 1: item calibration
  - Different methods
    - 1) Joint Maximum Likelihood (JML)
    - 2) Conditional Maximum Likelihood (CML)
    - 3) Marginal Maximum Likelihood (MML)
    - 4) Bayesian modal estimation
    - 5) Markov Chain Monte Carlo (MCMC).
  - Relative scale (*Celsius* scale)

Rasch Item Response Theory

- Step 2: Student proficiency estimates
  - Test of 4 items

Raw score	*Response patterns*
0	(0,0,0,0)
1	(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)
2	(1,1,0,0), (1,0,1,0), (1,0,0,1), (0,1,1,0), (0,1,0,1), (0,0,1,1)
3	(1,1,1,0),(1,1,0,1), (1,0,1,1), (0,1,1,1)
4	(1,1,1,1)

Rasch Item Response Theory

- Step 2: Student proficiency estimates
  - Probability to observe a response pattern (1100)

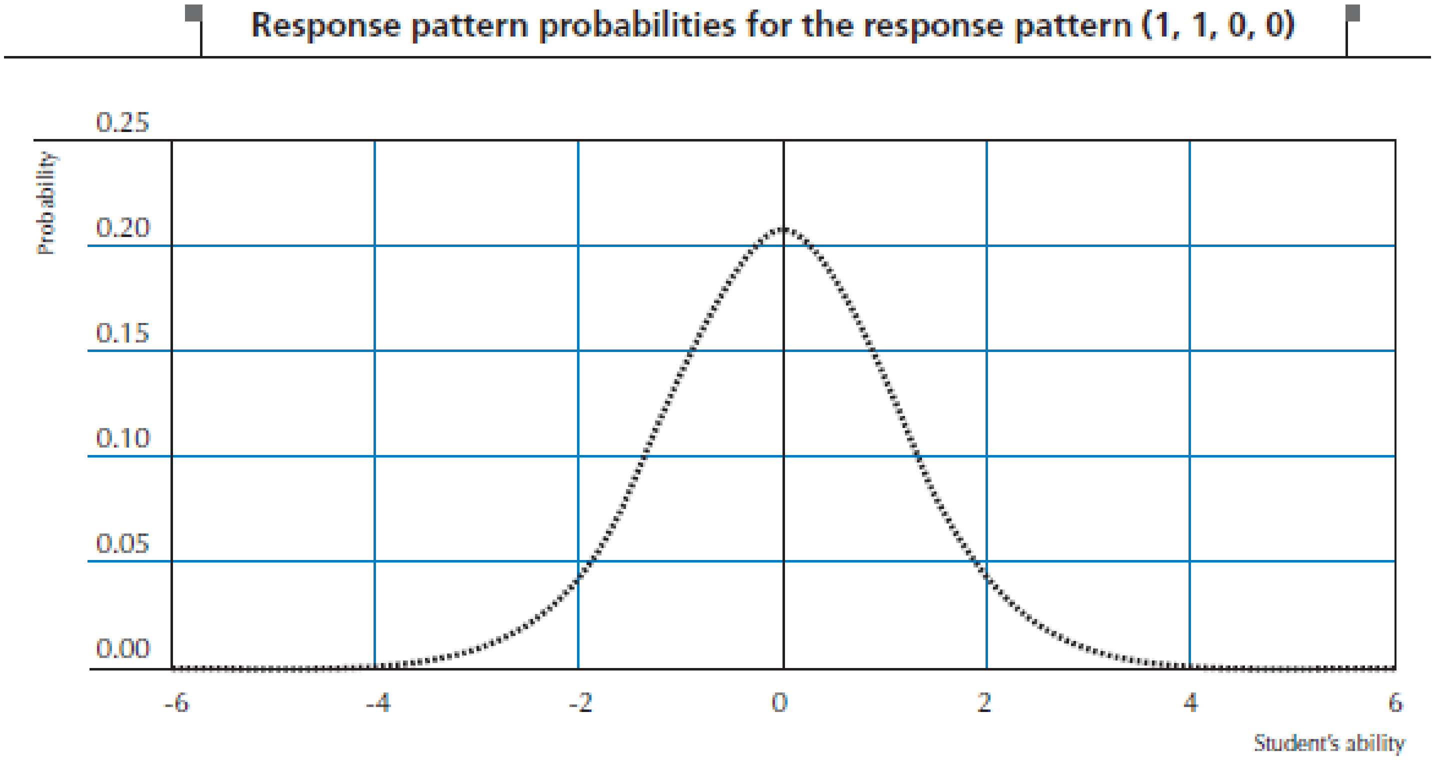
			B=-1	*B*=0	*B*=1
Item 1	*D*=-1	Response=1	0.50	0.73	0.88
Item 2	*D*=-0.5	Response=1	0.38	0.62	0.82
Item 3	*D*=0.5	Response=0	0.82	0.62	0.38
Item 4	*D*=1	Response=0	0.88	0.73	0.50
Global P			0.14	0.21	0.14

$$P\left[X_{ij}=1\middle|\beta_i,\delta_j\right]=\frac{e^{(\beta_i-\delta_j)}}{1+e^{(\beta_i-\delta_j)}}$$

$$P\left[X_{ij}=0\middle|\beta_i,\delta_j\right]=\frac{1}{1+e^{(\beta_i-\delta_j)}}$$

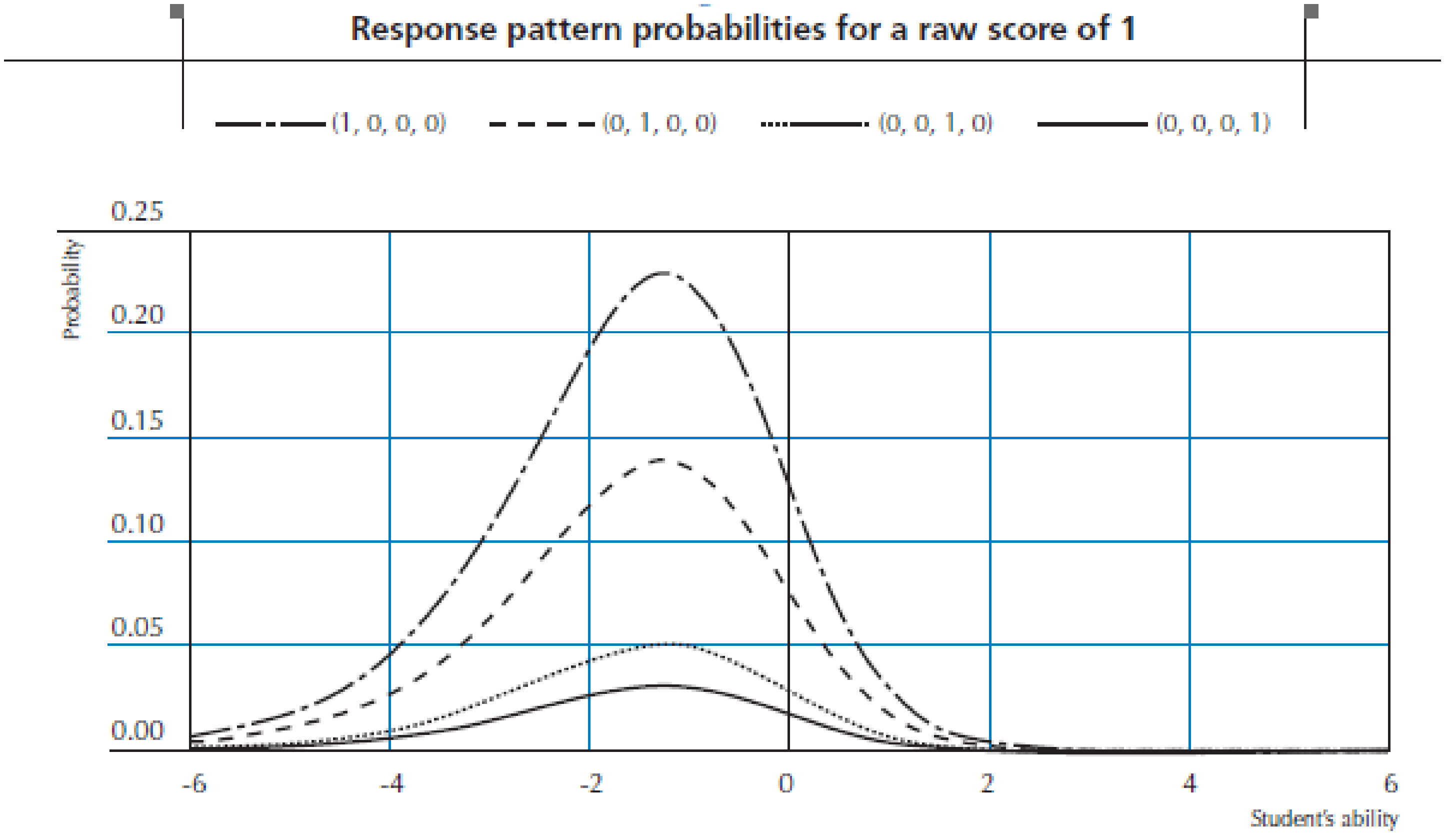
Rasch Item Response Theory

- Step 2: Student proficiency estimates
  - Likelihood function for a response pattern (1, 1, 0, 0)



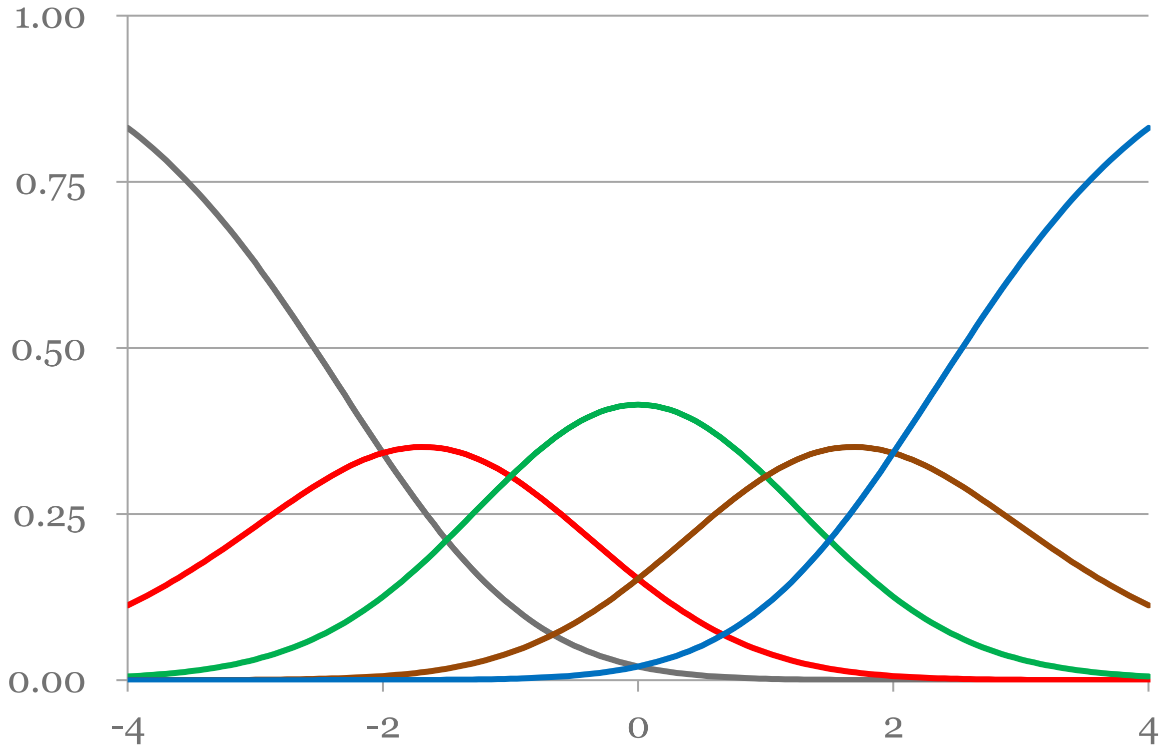
Rasch Item Response Theory

- Step 2: Student proficiency estimates
  - Likelihood functions for a score of 1



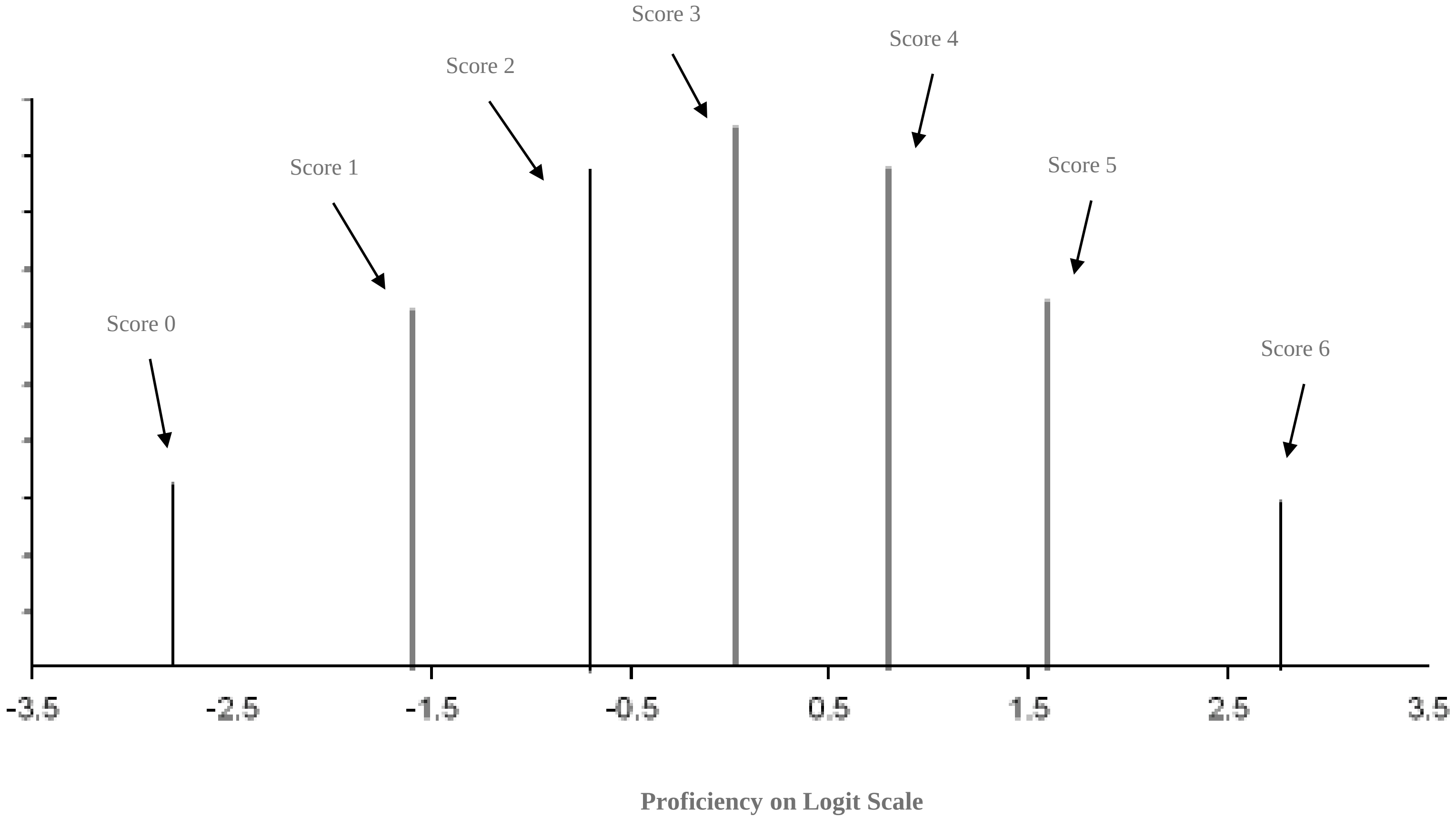
Rasch Item Response Theory

- Step 2: Student proficiency estimates
  - Likelihood functions for a score of [0,0,0,0], [1,0,0,0], [1,1,0,0], [1,1,1,0], [1,1,1,1]



Rasch Item Response Theory

Distribution of MLE : test of 6 items



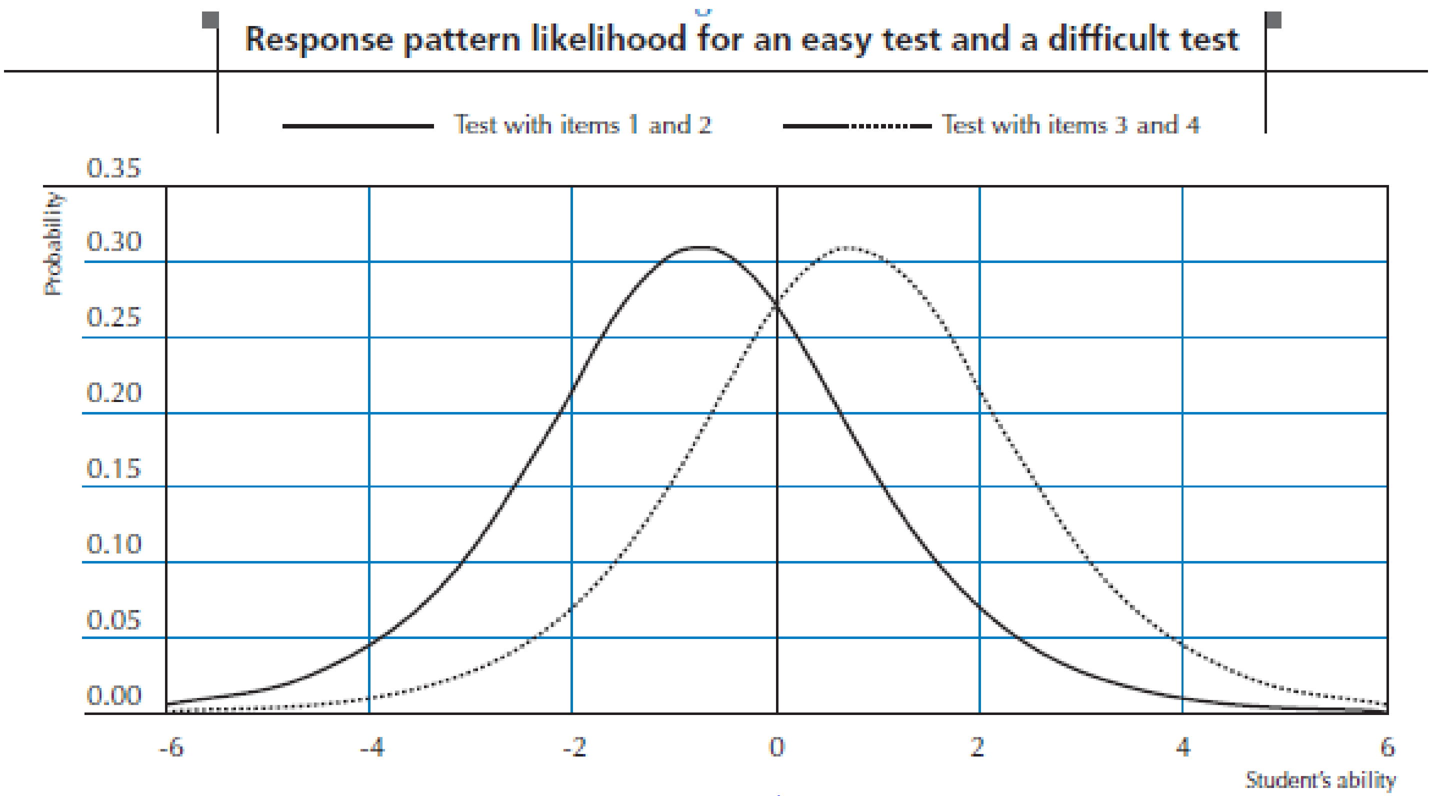
Rasch Item Response Theory

- Easy test administered to low achievers, difficult test administered to high achievers
- Likelihood functions for response pattern [1,0]

			B=-1	*B*=1
Item 1	*D*=-1	Response=1	0.50	
Item 2	*D*=-0.5	Response=0	0.62	
Item 3	*D*=0.5	Response=1		0.62
Item 4	*D*=1	Response=0		0.50
Global P			0.31	0.31

Rasch Item Response Theory

- Easy test administered to low achievers, difficult test administered to high achievers
- Likelihood functions for response pattern [1,0]



Other IRT models

- Models with 1, 2 or 3 parameters (1-, 2- or 3-Parameter Logistic Models)
  - 1 parameter:
    - Item difficulty
  - 2 parameters :
    - Item difficulty
    - Item discrimination
  - 3 parameters :
    - Item difficulty
    - Item discrimination
    - Guessing

Other IRT models

- 1, 2 and 3 PL IRT models

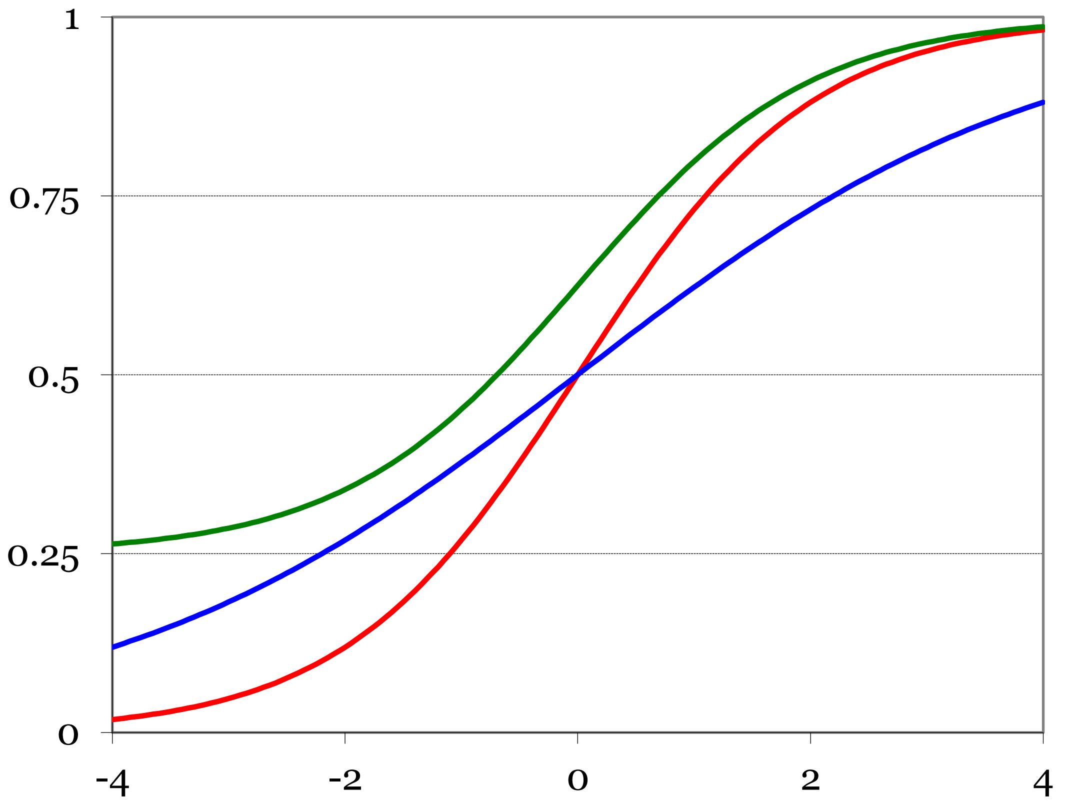
$$p(x_{ij}=1|\theta_j,b_i)=\frac{\exp^{(\theta_j-b_i)}}{1+\exp^{(\theta_j-b_i)}}=\frac{1}{1+\exp^{-(\theta_j-b_i)}}$$

$$p(x_{ij}=1|\theta_j,b_i,a_i)=\frac{\exp^{a_i(\theta_j-b_i)}}{1+\exp^{a_i(\theta_j-b_i)}}=\frac{1}{1+\exp^{-a_i(\theta_j-b_i)}}$$

$$p(x_{ij}=1|\theta_j,b_i,a_i,c_i)=c_i+(1-c_i)\frac{\exp^{a_i(\theta_j-b_i)}}{1+\exp^{a_i(\theta_j-b_i)}}=c_i+\frac{1-c_i}{1+\exp^{-a_i(\theta_j-b_i)}}$$

Other IRT models

- 1, 2 and 3 PL IRT models



$\delta = 0$

$\delta = 0$
$a = 0.5$

$\delta = 0$
$a = 0.5$
$c = 0.25$

Other IRT models

- Generalized Partial Credit Model

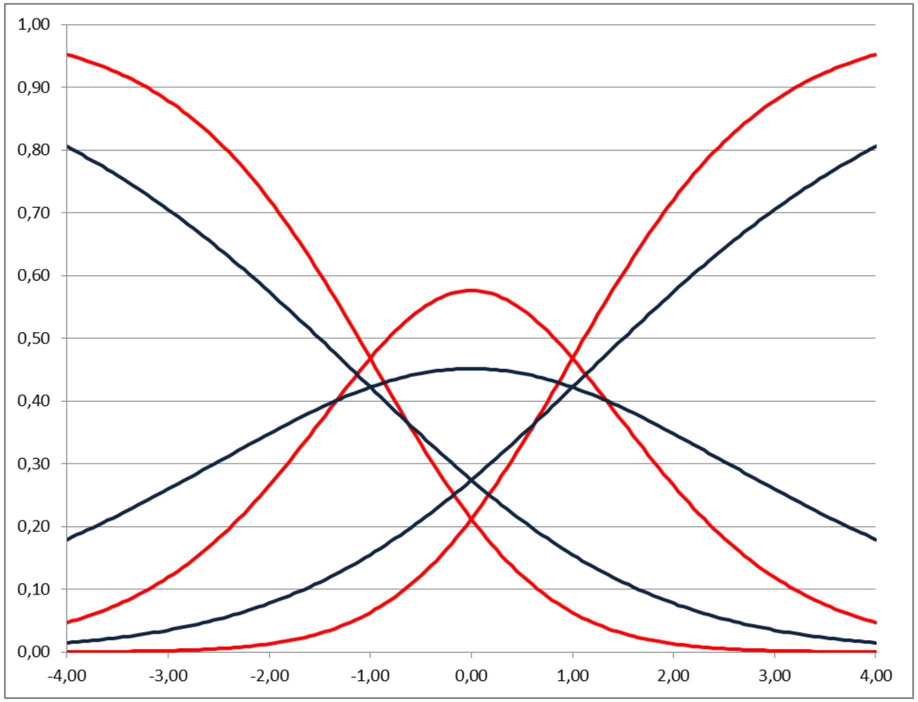
$$P(X_{ni}=0)=\frac{1}{1+\exp(a(\beta_n-\delta_j-t_{i1}))+\exp(a(2\beta_n-2\delta_j-t_{i1}-t_{i2}))}$$

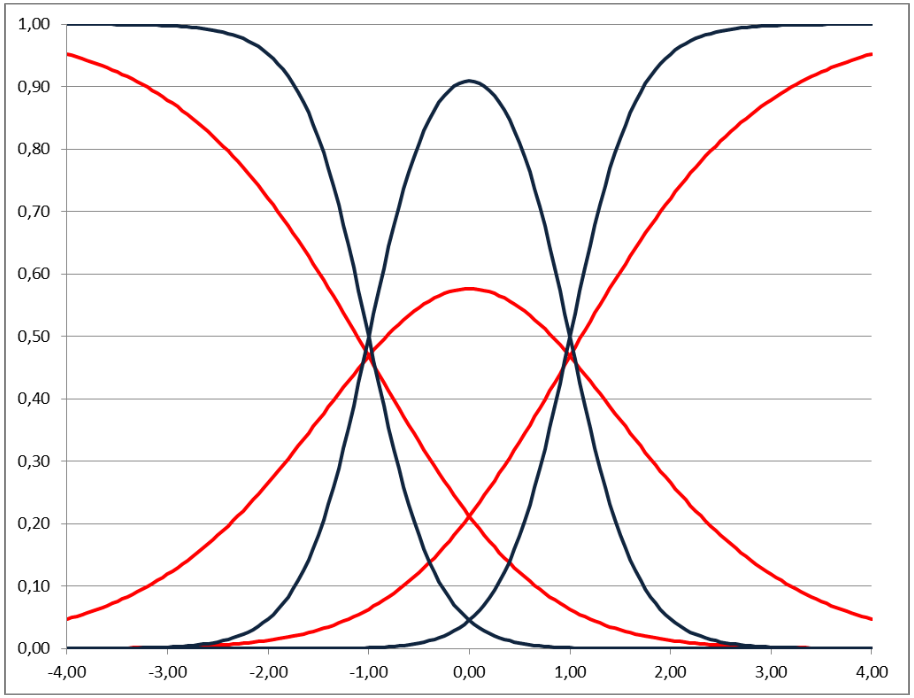
$$P(X_{ni}=1)=\frac{\exp(a(\beta_n-\delta_j-t_{i1}))}{1+\exp(a(\beta_n-\delta_j-t_{i1}+\exp(a(2\beta_n-2\delta_j-t_{i1}-t_{i2}))}$$

$$P(X_{ni}=2)=\frac{\exp(a(2\beta_n-2\delta_j-t_{i1}-t_{i2}))}{1+\exp(a(\beta_n-\delta_j-t_{i1}))+\exp(a(2\beta_n-2\delta_j-t_{i1}-t_{i2}))}$$

Other IRT models

- Generalized Partial Credit Model





a=0.5

a=3

Item Response Theory

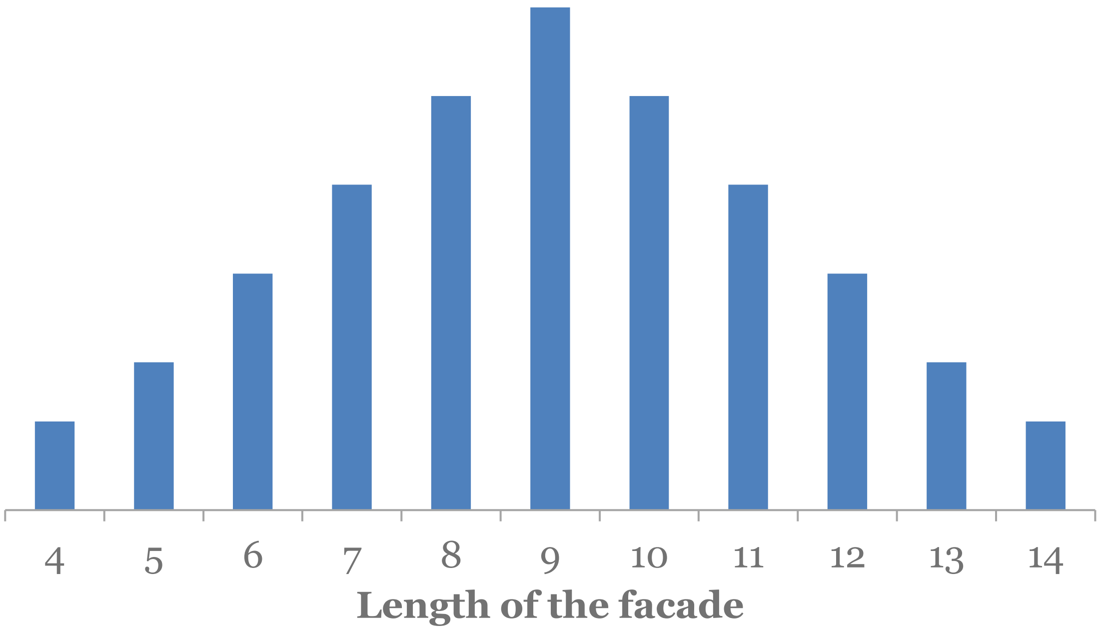
- Student IRT estimates in PISA
  - Non Cognitive scales : Weighted Likelihood Estimate (WLE)
    - Student Contextual Questionnaire data
      - *Student Reading Enjoyment, Sense of Belonging, Self Concept in Mathematics, Self Efficacy in Mathematics*
    - School Contextual Questionnaire data
      - *Shortage of teachers, Teacher morale*
  - Cognitive scales : Plausible Values
    - What are plausible values?
    - Why do we use them?
    - How to analyze plausible values?

Plausible Values (PVs)

- Purpose of Educational Assessments
  - **Estimating the proficiency of particular students**
    (minimize measurement error of individual estimates)
  - **Estimating the population proficiency (mean, STD…)**
    (minimize error when generalizing to the population)

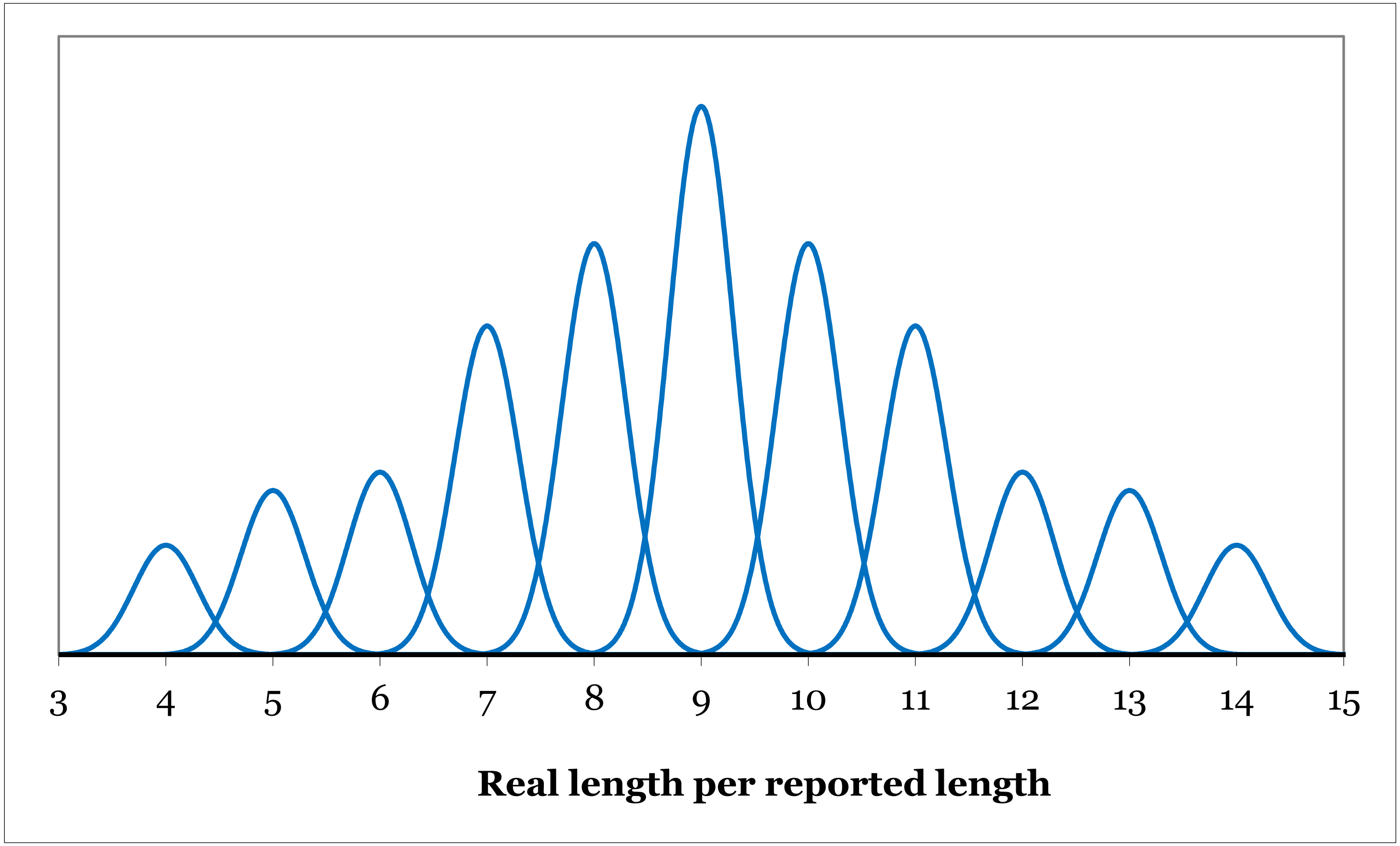
Plausible Values

- New tax on the façade length of the building



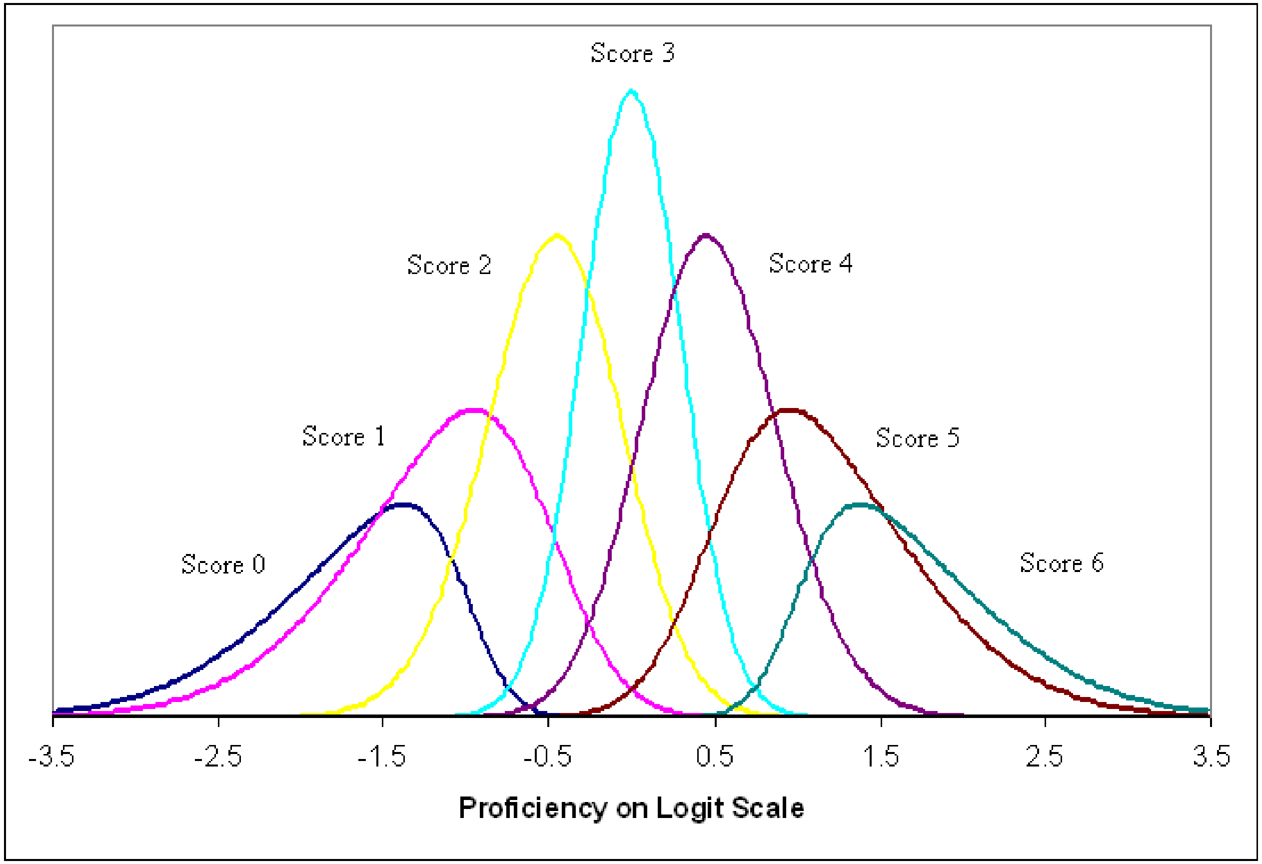
Plausible Values

- Real length



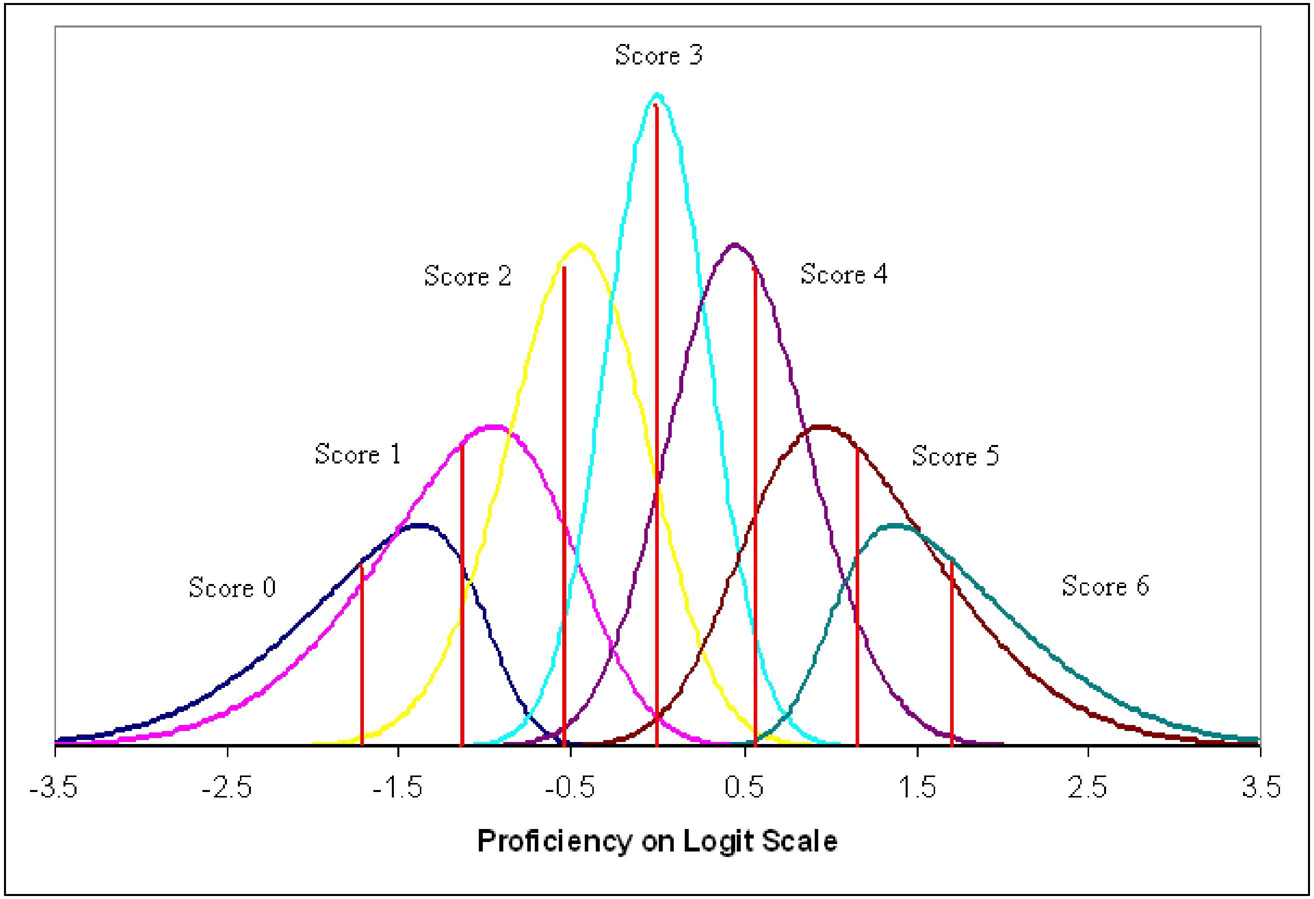
Plausible Values

- Posterior distributions for test scores on 6 dichotomous items



Plausible Values

- EAP – Expected A Posteriori Estimator



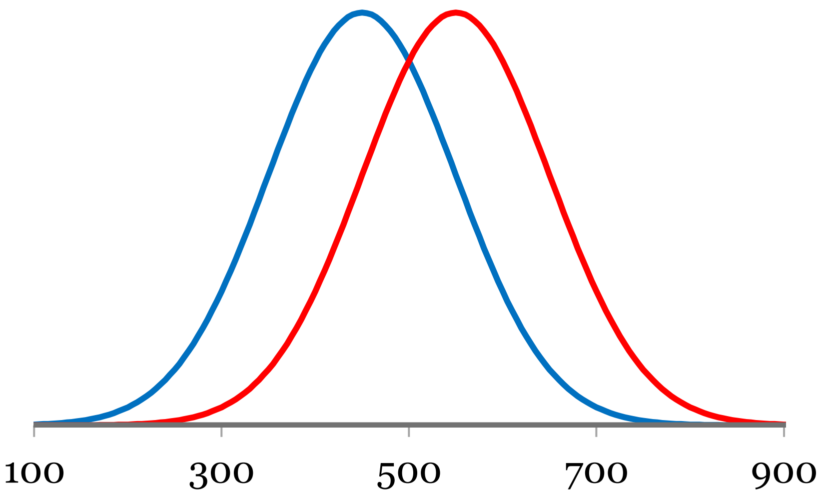
Plausible Values

- Methodology of PVs
  - Aim is building a continuum from a discontinuous variable to prevent biased inferences
  - Mathematically computing posterior distributions around test scores
  - Drawing 5 random values for each assessed individual from the posterior distribution for that individual
- Individual estimates from PVs
  - Expected A Posteriori estimate (EAP), i.e. the mean of posterior distribution
  - Not a one to one relationship with raw score, unlike WLE

Plausible Values

- Assuming normal distribution: $N(\mu, \sigma^2)$

- Model sub-populations: $N(\mu + \alpha X, \sigma^2)$
  - X=0 for boy
  - X=1 for girl



- Generalization $N(\mu + \alpha X + \beta Y + \gamma Z + ..., \sigma^2)$

Plausible Values

- Simulating data for assessing biases in WLE, EAP, PVs estimates
  - Generating item responses
  - True abilities are known
  - True relationships between abilities and background variables are known
  - Within- and between-school variances are known
  - Simulated item responses are used to estimate WLEs, EAPs and PVs with ConQuest

Structure of the simulated data

School ID	Student ID	Sex	HISEI	Item 1	Item 2	...	Item 14	Item 15
001	01	1	32	1	1		0	0
001	02	0	45	1	0		1	0
...	...							
150	5 249	0	62	0	0		1	1
150	5 250	1	50	0	1		1	1

Plausible Values

- Data file
  - 150 schools with 35 students each
  - TRUE ability scores
  - Background variables:
    - HISEI
    - Gender (dichotomous), Gender (continuous)
  - School mean
  - Raw Score, WLE,
    EAP and PVs *without* conditioning,
    EAP and PVs *with* conditioning

Plausible Values

Means and variances for the latent variables and the different student ability estimators

	Mean	Variance
Latent variable	0.00	1.00
WLE	0.00	1.40
EAP	0.00	0.75
PV1	0.01	0.99
PV2	0.00	0.99
PV3	0.00	1.01
PV4	0.00	1.01
PV5	-0.01	0.00
Average of the 5 PV statistics	0.00	1.00

Plausible Values

Percentiles for the latent variables and the different student ability estimators

	P5	P10	P25	P50	P75	P90	P95
Latent variable	-1.61	-1.26	-0.66	0.01	0.65	1.26	1.59
WLE	-2.15	-1.65	-0.82	-0.1	0.61	1.38	1.81
EAP	-1.48	-1.14	-0.62	-0.02	0.55	1.08	1.37
PV1	-1.68	-1.29	-0.71	-0.03	0.64	1.22	1.59
PV2	-1.67	-1.31	-0.69	-0.03	0.62	1.22	1.58
PV3	-1.67	-1.32	-0.70	-0.02	0.64	1.21	1.56
PV4	-1.69	-1.32	-0.69	-0.03	0.63	1.23	1.55
PV5	-1.65	-1.3	-0.71	-0.02	0.62	1.2	1.55
Average of the 5 PV statistics	-1.67	-1.31	-0.70	-0.03	0.63	1.22	1.57

Plausible Values

Correlation between HISEI, gender and the latent variable, the different student ability estimators

	HISEI	GENDER
Latent variable	0.40	0.16
WLE	0.33	0.13
EAP	0.46	0.17
PV1	0.41	0.15
PV2	0.42	0.15
PV3	0.42	0.13
PV4	0.40	0.15
PV5	0.40	0.14
Average of the 5 PV statistics	0.41	0.14

Plausible Values

Between- and within-school variances

	Between-school variance	Within-school variance
Latent variable	0.33	0.62
WLE	0.34	1.02
EAP	0.35	0.38
PV1	0.35	0.61
PV2	0.36	0.60
PV3	0.36	0.61
PV4	0.35	0.61
PV5	0.35	0.61
Average of the 5 PV statistics	0.35	0.61

Plausible Values

- Note on conditioning
  - When analyzing relationships between ability and background variables, only PVs derived from a conditional model that includes the background variables as regressors give reliable population estimates.

Plausible Values

- How to analyze Plausible Values?

Weight	PV1	PV2	PV3	PV4	PV5
Final	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
Replicate 1	$\hat{\mu}_{1\_1}$	$\hat{\mu}_{2\_1}$	$\hat{\mu}_{3\_1}$	$\hat{\mu}_{4\_1}$	$\hat{\mu}_{5\_1}$
Replicate 2	$\hat{\mu}_{1\_2}$	$\hat{\mu}_{2\_2}$	$\hat{\mu}_{3\_2}$	$\hat{\mu}_{4\_2}$	$\hat{\mu}_{5\_2}$
Replicate 3	$\hat{\mu}_{1\_3}$	$\hat{\mu}_{2\_3}$	$\hat{\mu}_{3\_3}$	$\hat{\mu}_{4\_3}$	$\hat{\mu}_{5\_3}$
....	....	....	....	....	....
....	....	....	....	....	....
Replicate 80	$\hat{\mu}_{1\_80}$	$\hat{\mu}_{2\_80}$	$\hat{\mu}_{3\_80}$	$\hat{\mu}_{4\_80}$	$\hat{\mu}_{5\_80}$
Sampling variance	$\sigma^2_{(\hat{\mu}_1)}$	$\sigma^2_{(\hat{\mu}_2)}$	$\sigma^2_{(\hat{\mu}_3)}$	$\sigma^2_{(\hat{\mu}_4)}$	$\sigma^2_{(\hat{\mu}_5)}$

Plausible Values

- Estimated mean is the AVERAGE of the mean for each PV

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M}\hat{\mu}_i$$

- <u>*Sampling variance*</u> is the AVERAGE of the sampling variance for each PV

$$\hat{\sigma}^2_{(\hat{\mu})} = \frac{1}{M}\sum_{i=1}^{M}\hat{\sigma}^2_{(\hat{\mu}_i)}$$

- Where $\hat{\sigma}^2_{(\hat{\mu}_i)} = \frac{1}{20}\sum_{j=1}^{80}\left(\hat{\mu}_{ij} - \hat{\mu}_i\right)^2$

Plausible Values

- *Measurement variance* computed as:

$$\hat{\sigma}^2_{(PV)} = \frac{1}{M-1}\sum_{i=1}^{5}\left(\hat{\mu}_i - \hat{\mu}\right)^2$$

- Total *Standard Error* computed from measurement and Sampling Variance as:

$$\hat{\sigma}_{(\hat{\mu}_{PV})} = \sqrt{\hat{\sigma}^2_{(\hat{\mu})} + (1 + \frac{1}{M})\hat{\sigma}^2_{(PV)}}$$

Plausible Values

- How to analyze Plausible Values?

- μ can be replaced by any statistic,
  - SD
  - Percentile
  - Correlation coefficient
  - Regression coefficient
  - R-square
  - etc.

Plausible Values

- How to analyze Plausible Values?

Weight	PV1	PV2	PV3	PV4	PV5
Final	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Replicate 1	$\hat{\beta}_{1\_1}$	$\hat{\beta}_{2\_1}$	$\hat{\beta}_{3\_1}$	$\hat{\beta}_{4\_1}$	$\hat{\beta}_{5\_1}$
Replicate 2	$\hat{\beta}_{1\_2}$	$\hat{\beta}_{2\_2}$	$\hat{\beta}_{3\_2}$	$\hat{\beta}_{4\_2}$	$\hat{\beta}_{5\_2}$
Replicate 3	$\hat{\beta}_{1\_3}$	$\hat{\beta}_{2\_3}$	$\hat{\beta}_{3\_3}$	$\hat{\beta}_{4\_3}$	$\hat{\beta}_{5\_3}$
....	....	....	....	....	....
....	....	....	....	....	....
Replicate 80	$\hat{\beta}_{1\_80}$	$\hat{\beta}_{2\_80}$	$\hat{\beta}_{3\_80}$	$\hat{\beta}_{4\_80}$	$\hat{\beta}_{5\_80}$
Sampling variance	$\sigma^2_{(\hat{\beta}_1)}$	$\sigma^2_{(\hat{\beta}_2)}$	$\sigma^2_{(\hat{\beta}_3)}$	$\sigma^2_{(\hat{\beta}_4)}$	$\sigma^2_{(\hat{\beta}_5)}$

Plausible Values

- **Five steps** for analyzing Plausible Values

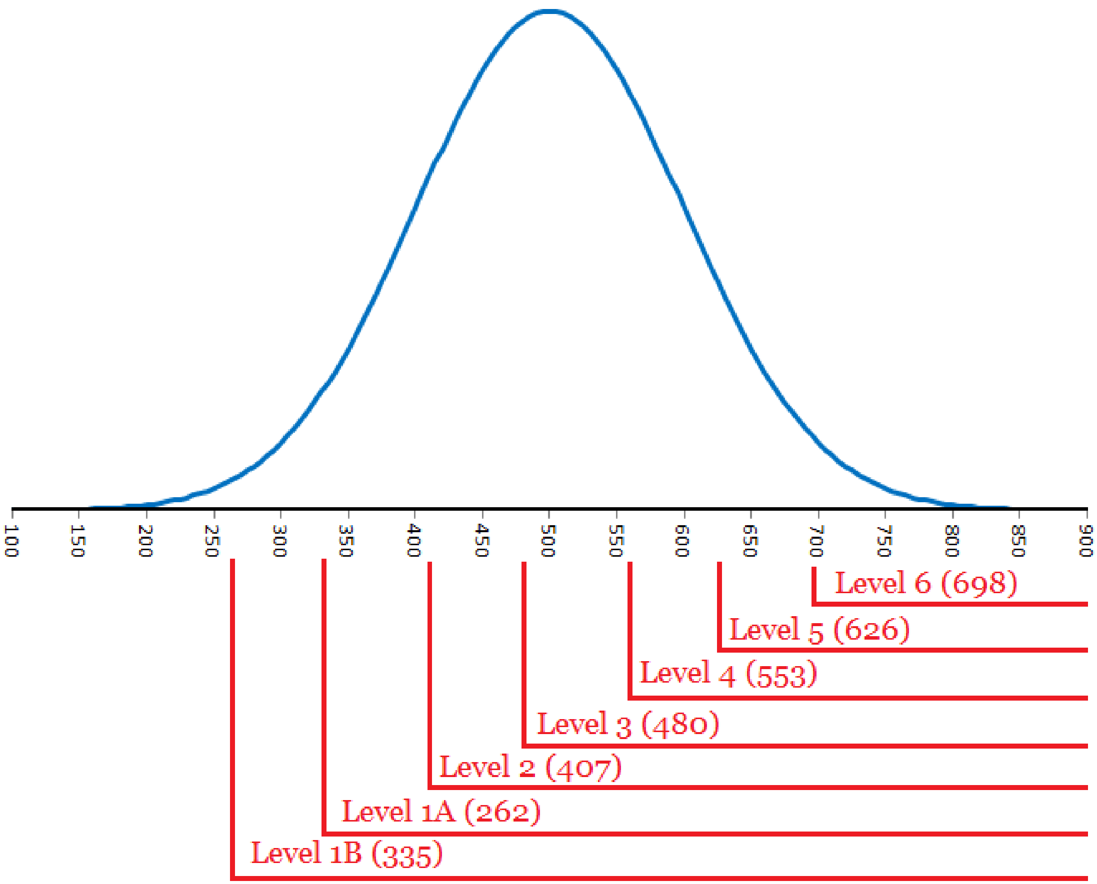
1. Compute estimate for each PV
2. Compute *final estimate* by averaging 5 estimates from (1)
3. Compute *sampling variance* (average of sampling variance estimates for each PV)
4. Compute *imputation variance* (measurement error variance)
5. Compute *final standard error* by combining (3) and (4)

Remaining issues with PVs

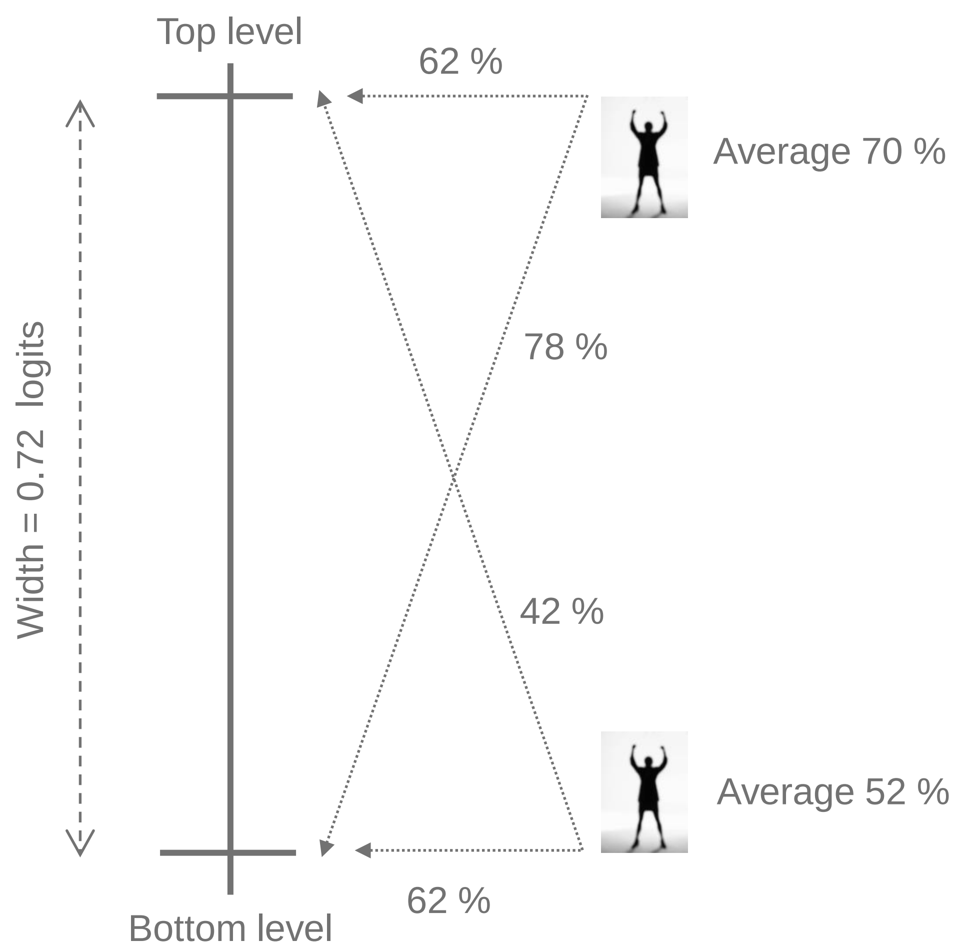
- Use of Proficiency levels
- Biased / unbiased shortcuts
- Correlations between PVs
- Computation of trend indicators

Proficiency levels

- Proficiency level in reading (PISA 2009)



Proficiency levels



One proficiency level (RP62)

Proficiency levels

Level	Lower score limit	Percentage of students able to perform tasks at each level or above (OECD average)	Characteristics of tasks
6	698	0.8% of students across the OECD can perform tasks at Level 6 on the reading scale	Tasks at this level typically require the reader to make multiple inferences, comparisons and contrasts that are both detailed and precise. They require demonstration of a full and detailed understanding of one or more texts and may involve integrating information from more than one text. Tasks may require the reader to deal with unfamiliar ideas, in the presence of prominent competing information, and to generate abstract categories for interpretations. *Reflect and evaluate* tasks may require the reader to hypothesise about or critically evaluate a complex text on an unfamiliar topic, taking into account multiple criteria or perspectives, and applying sophisticated understandings from beyond the text. A salient condition for *access and retrieve* tasks at this level is precision of analysis and fine attention to detail that is inconspicuous in the texts.

1b	262	98.9% of students across the OECD can perform tasks at least at Level 1b on the reading scale	Tasks at this level require the reader to locate a single piece of explicitly stated information in a prominent position in a short, syntactically simple text with a familiar context and text type, such as a narrative or a simple list. The text typically provides support to the reader, such as repetition of information, pictures or familiar symbols. There is minimal competing information. In tasks requiring interpretation the reader may need to make simple connections between adjacent pieces of information.

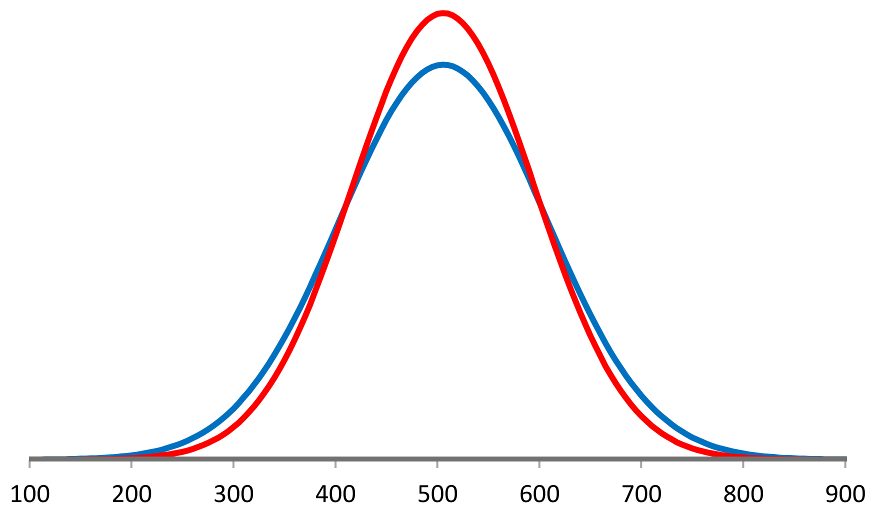
Proficiency levels

- How to analyze the proficiency levels:

				L4	L3	L3	L3	L3	
	Country code 3-character	School ID 5-digit	Student ID 5-digit	Plausible value in reading	Plausible value in reading	Plausible value in reading	Plausible value in reading	Plausible value in reading	FINAL STUDENT WEIGHT
1	BEL	00001	00001	684.84	625.52	604.67	625.52	623.91	1.00
2	BEL	00001	00002	548.04	584.58	562.34	570.28	585.37	1.13
3	BEL	00001	00003	557.77	495.24	571.40	524.10	480.81	1.00
4	BEL	00001	00004	536.37	553.20	532.36	531.56	523.54	1.00
5	BEL	00001	00005	624.94	648.77	620.17	616.99	608.25	1.09
6	BEL	00001	00006	660.13	656.95	643.45	589.43	628.35	1.09

Proficiency levels

	PVs	≈EAP
Mean	505.9	505.9
STD	101.8	99.16
Level 6	1.12	0.97
Level 5	9.95	4.76
Level 4	25.03	11.61
Level 3	25.97	20.17
Level 2	20.31	26.06
Level 1A	11.71	26.34
Level 1B	4.76	9.56
Below	1.15	0.52



Proficiency levels

- Recoding of 5 PVs into 5 Plausible Levels

The 405 percentage estimates for a particular proficiency level

Weight	PV1	PV2	PV3	PV4	PV5
Final	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_5$
Replicate 1	$\hat{\pi}_{1\_1}$	$\hat{\pi}_{2\_1}$	$\hat{\pi}_{3\_1}$	$\hat{\pi}_{4\_1}$	$\hat{\pi}_{5\_1}$
Replicate 2	$\hat{\pi}_{1\_2}$	$\hat{\pi}_{2\_2}$	$\hat{\pi}_{3\_2}$	$\hat{\pi}_{4\_2}$	$\hat{\pi}_{5\_2}$
Replicate 3	$\hat{\pi}_{1\_3}$	$\hat{\pi}_{2\_3}$	$\hat{\pi}_{3\_3}$	$\hat{\pi}_{4\_3}$	$\hat{\pi}_{5\_3}$
....	....	....	....	....	....
....	....	....	....	....	....
Replicate 80	$\hat{\pi}_{1\_80}$	$\hat{\pi}_{2\_80}$	$\hat{\pi}_{3\_80}$	$\hat{\pi}_{4\_80}$	$\hat{\pi}_{5\_80}$
Sampling variance	$\sigma^2_{(\hat{\pi}_1)}$	$\sigma^2_{(\hat{\pi}_2)}$	$\sigma^2_{(\hat{\pi}_3)}$	$\sigma^2_{(\hat{\pi}_4)}$	$\sigma^2_{(\hat{\pi}_5)}$

Proficiency levels

- Estimated percentage is the AVERAGE of the percentage for each PV

$$\hat{\pi} = \frac{1}{M}\sum_{i=1}^{M}\hat{\pi}_i$$

- *Sampling variance* is the AVERAGE of the sampling variance for each PV

$$\hat{\sigma}^2_{(\hat{\pi})} = \frac{1}{M}\sum_{i=1}^{M}\hat{\sigma}^2_{(\hat{\pi}_i)}$$

- Where $\hat{\sigma}^2_{(\hat{\pi}_i)} = \frac{1}{20}\sum_{j=1}^{80}\left(\hat{\pi}_{ij} - \hat{\pi}_i\right)^2$

Proficiency levels

- *Measurement variance* computed as:

$$\hat{\sigma}^2_{(PV)} = \frac{1}{M-1}\sum_{i=1}^{5}\left(\hat{\pi}_i - \hat{\pi}\right)^2$$

- Total *standard error* computed from measurement and sampling variance as:

$$\hat{\sigma}_{(\hat{\pi}_{PV})} = \sqrt{\hat{\sigma}^2_{(\hat{\pi})} + \left(1 + \frac{1}{M}\right)\hat{\sigma}^2_{(PV)}}$$

Biased / unbiased shortcut

- Plausible values should **never** be aggregated at the student level:
  - Underestimation of the STD
  - Underestimation of the % of students at the lowest and highest proficiency levels and overestimation of median proficiency levels
  - Overestimation of lowest percentiles and underestimation of highest percentiles
  - Overestimation of correlation coefficients
  - …

Biased / unbiased shortcut

- Mean estimates are not biased if PVs aggregated at the student level but Standard Errors
  - Will be underestimated
  - Will not incorporate measurement errors

	BEL N=8501	BEL 05611 N=1723	BEL 05602 N=208
PV1	2.30	6.00	8.24
PV2	2.31	6.07	7.24
PV3	2.31	6.07	5.59
PV4	2.24	5.87	7.60
PV5	2.32	6.07	8.67
5PV	2.35	6.06	7.86
≈EAP	2.29	5.99	7.40

Biased / unbiased shortcut

- Computing 405 estimates sometimes is too time consuming
- Using one PV :
  - gives unbiased population estimates
  - gives unbiased sampling variance
  - does not allow the computation of the imputation variance
- Therefore, with one PV only, SE does only reflect sampling variance, not measurement / imputation variance

Biased / unbiased shortcut

Therefore, an unbiased shortcut consists of :

- Computing one sampling variance (i.e. PV1)
- Computing 5 population estimates using full student weight (one on each PVs)
- Averaging 5 estimates from (2) to obtain *final population estimate*
- Computing imputation variance
- Combining (1) and (4) to obtain *final standard error*

Biased / unbiased shortcut

- In summary

Weight	PV1	PV2	PV3	PV4	PV5
Final	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
Replicate 1	μ_{1_1}				
Replicate 2	μ_{1_2}				
Replicate 3	μ_{1_3}				
....	....				
....	....				
Replicate 80	μ_{1_80}				
Sampling variance	$\sigma^2_{(\hat{\mu}_1)}$				

- It saves 4 times 80 replicates, i.e. 320 estimates → This unbiased shortcut requires 85 estimates instead of 405

Biased / unbiased shortcut

- A comparison between the full computation and the shortcut computation

	Mean estimate in science	
	Full computation	Shortcut computation
	S.E.	S.E.
AUS	2.26	2.21
AUT	3.92	3.98
BEL	2.48	2.49
CAN	2.03	2.00
CHE	3.16	3.25
CZE	3.48	3.40
DEU	3.80	3.80
DNK	3.11	3.06
ESP	2.57	2.60
FIN	2.02	2.05
FRA	3.36	3.34
GBR	2.29	2.23
GRC	3.23	3.29
HUN	2.68	2.63
IRL	3.19	3.18
ISL	1.64	1.59
ITA	2.02	2.02
JPN	3.37	3.45
KOR	3.36	3.41
LUX	1.05	1.14
MEX	2.71	2.64
NLD	2.74	2.77
NOR	3.11	3.07
NZL	2.69	2.67
POL	2.34	2.37
PRT	3.02	3.02
SVK	2.59	2.57
SWE	2.37	2.28
TUR	3.84	3.82
USA	4.22	4.20

Correlation / regression between PVs

- What are the correlations between student proficiency estimates in mathematics and in science?
  - Should we compute 5 by 5 correlation coefficients?
- What is the relationship between mathematic proficiency estimates and student social background, under control of student proficiency estimates in reading?
  - Partial correlation
    - Should we compute 5 by 5 correlation coefficients?
  - Regression coefficient
    - Should we compute 5 by 5 regression coefficients?

Correlation / regression between PVs

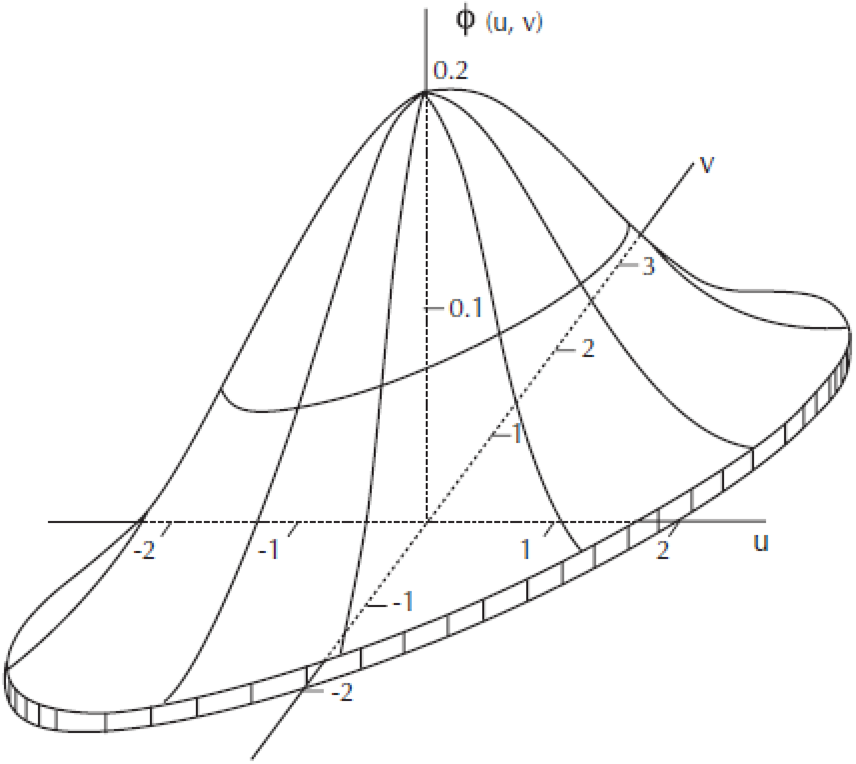
- Correlation coefficients between PVs in Reading and:
  - Mathematic subdomain Space and Shape (SS1-SS5)
  - Combined mathematics (Math1-Math5)
    - PISA 2003, Belgium

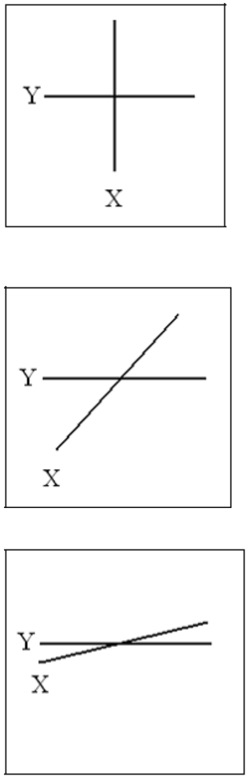
	SS1	SS2	SS3	SS4	SS5	Math1	Math2	Math3	Math4	Math5
Read1	0.722	0.720	0.717	0.715	0.715	0.820	0.805	0.801	0.801	0.801
Read2	0.716	0.718	0.718	0.716	0.716	0.802	0.821	0.802	0.801	0.801
Read3	0.717	0.720	0.716	0.712	0.712	0.807	0.809	0.822	0.802	0.802
Read4	0.724	0.721	0.723	0.720	0.720	0.803	0.807	0.803	0.821	0.821
Read5	0.718	0.724	0.721	0.719	0.719	0.803	0.804	0.801	0.802	0.802

Correlation / regression between PVs

Two-dimensional distribution

A two-dimensional distribution





Correlation / regression between PVs

- PISA 2000
  - 3 D scaling M/R/S
  - 5 D scaling R1/R2/R3/M/S
- PISA 2003
  - 4 D scaling M/R/S/PS
  - 7 D scaling M1/M2/M3/M4/R/S/P
- PISA 2006
  - 5 D scaling M/R/S & 2 attitudinal dimensions
  - 5D scaling M/R/S1/S2/S3
- PISA 2009
  - 3 D scaling M/R/S
  - 5 D scaling R1/R2/R3/M/S
  - 4 D scaling R4/R5/M/S

Correlation / regression between PVs

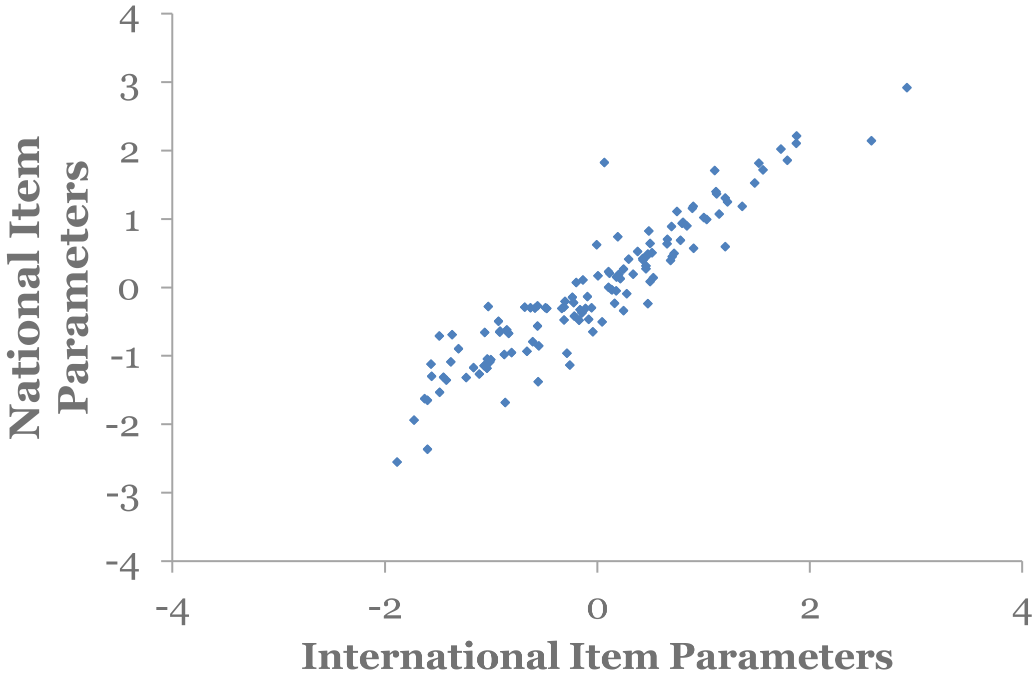
- In summary, this means that:
  - The correlation between the major combined scale and minor domain scales can be computed;
  - The correlation between minor domain scales can be computed;
  - The correlation between subdomain scales can be computed (except in 2009: correlation can be computed between processes or between type of texts, but not between a type of text and a reading process);
  - The correlation between a subdomain scale and the combined major domain scale <u>should not</u> be computed;
  - The correlation between any subdomain scale and any minor domain <u>should not be</u> computed.

Trend estimates

- A subset of items from major domains are selected as link items for subsequent PISA cycles.
  - PISA 2000 reading material: 37 units and 129 items
  - PISA 2003 reading material: 8 units and 28 items
- Item by Country interactions
  - In a country, some items might be easier/harder than expected due to:
    - Mistranslation
    - More curriculum emphasis
    - Cultural bias
    - …

Trend estimates

- PISA 2000 DEU & international IRT item parameters in Reading



Trend estimates

- Easier/harder at the item level and at the unit level

	Changes in the 2003 Reading mean estimates without…							
Pays	***Unit 1***	***Unit 2***	***Unit 3***	***Unit 4***	***Unit 5***	***Unit 6***	***Unit 7***	***Unit 8***
AUS	1.82	-3.32	-1.09	4.07	-3.60	-6.17	4.30	2.77
AUT	3.34	-2.66	2.48	1.58	0.40	-4.17	-4.79	2.13
BEL	6.79	-1.03	0.07	1.87	1.67	-0.53	-5.54	0.24
CAN	-0.06	-6.64	1.41	3.87	-6.64	-5.31	3.32	6.62
CHE	5.60	-3.40	3.68	3.21	1.68	-2.21	-6.66	-3.05
CZE	0.60	-3.31	-0.21	-0.93	0.37	2.82	-3.90	7.44
DEU	5.73	-0.79	1.07	4.39	1.77	-2.25	-6.24	-1.70
	Average differences at the unit level between International and DEU item parameters							
DEU	-0.384	0.083	-0.039	-0.286	0.093	0.135	0.214	0.041

Trend estimates

- Depending on the selected items for the linking, countries might be advantaged or disadvantaged.
- To reflect this additional uncertainty, PISA computes Linking Errors:
  - Based on the shift in the international item parameter estimates between two PISA cycles, i.e. *Item by Cycle Interactions*;
  - Take into account the hierarchical structure of the data (i.e. items embedded within units);
  - Do not take into account the *Item by Country Interaction*

Trend estimates

- Linking errors estimates

Domains	*PISA cycles*	*Linking errors*
Reading	2000-2003	5.32
	2000-2006	4.96
	2000-2009	5.39
	2003-2006	4.48
	2003-2009	4.09
	2006-2009	4.07
Math	2003-2006	1.35
	2003-2009	1.99
	2006-2009	1.33
Science Interim	2000_2003	3.11
Science	2006-2009	2.57

Trend estimates

- How to use the linking errors?
- Mean estimates in Reading for Poland:
  - PISA 2000 : 479 (4.5)
  - PISA 2003 : 497 (2.9)
  - Linking Error Reading (2000, 2003) : 5.307

$$SE_{(\hat{\mu}_{2003}-\hat{\mu}_{2000})} = \sqrt{\sigma^2_{(\hat{\mu}_{2000})} + \sigma^2_{(\hat{\mu}_{2003})} + \sigma^2_{Linking}}$$

$$SE_{(\hat{\mu}_{2003}-\hat{\mu}_{2000})} = \sqrt{4.5^2 + 2.9^2 + 5.309^2} = 7.68$$

$$z = \frac{(\hat{\mu}_{2003} - \hat{\mu}_{2000})}{SE_{(\hat{\mu}_{2003}-\hat{\mu}_{2000})}} = \frac{497 - 479}{7.68} = \frac{18}{7.69} = 2.34$$

Trend estimates

- When should we inflate the standard errors by the linking errors?
  - Linking errors need only to be considered when comparisons are being made between results from different data collections and then usually when group means are being compared;
  - Ex:
    - Differences in country mean estimates from 2 cycles
    - Differences in subgroup (boys or girls, natives …) mean estimates from 2 cycles
  - But not in the gender difference shift between 2 cycles