Item Response Theory (IRT) models: Reducing bias in small samples

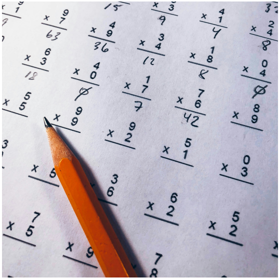
FOS Seminar | Brunei R User Group

Haziq Jamil
Assistant Professor in Statistics
`https://haziqj.ml`

4 September 2024

Joint work with Ioannis Kosmidis (Warwick)

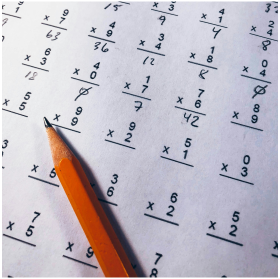
Introduction



Context

In *educational assessments*, data Y are composed of several test items from students. Each item is marked **correct** ($Y = 1$) or **wrong** ($Y = 0$).

Introduction



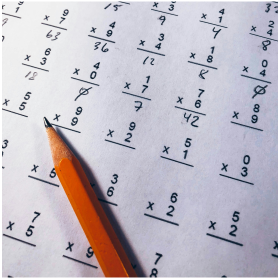
Context

In *educational assessments*, data Y are composed of several test items from students. Each item is marked **correct** ($Y = 1$) or **wrong** ($Y = 0$).

It is common to enquire, from this set of data, the **reliability** and **validity** of the assessment, including:

1. How **difficult** is each test item?

Introduction



Context

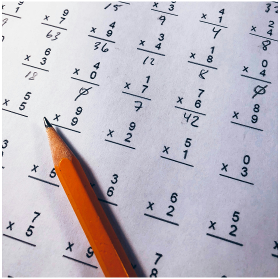
In *educational assessments*, data Y are composed of several test items from students. Each item is marked **correct** ($Y = 1$) or **wrong** ($Y = 0$).

It is common to enquire, from this set of data, the **reliability** and **validity** of the assessment, including:

1. How **difficult** is each test item?
2. How well does each item **discriminate** between students of different ability levels?

Universiti Brunei Darussalam

Introduction



Context

In *educational assessments*, data Y are composed of several test items from students. Each item is marked **correct** ($Y = 1$) or **wrong** ($Y = 0$).

It is common to enquire, from this set of data, the **reliability** and **validity** of the assessment, including:

1. How **difficult** is each test item?
2. How well does each item **discriminate** between students of different ability levels?
3. Can I accurately estimate students' **abilities**?

The IRT family of models provides a statistical framework for addressing these sorts of questions.

Example

A typical data set

Student	Item1	Item2	Item3	Item4	Item5
1	1	1	1	1	1
2	0	1	1	1	1
3	1	1	0	1	1
4	1	1	1	1	0
5	1	1	1	1	0
6	0	0	1	1	0
7	1	0	0	0	0
8	0	0	0	1	0
9	1	0	0	0	0
10	0	0	0	0	0

Example (cont.)

Simple scores and item difficulties

Student	Item1	Item2	Item3	Item4	Item5	**Score**
1	1	1	1	1	1	**5**
2	0	1	1	1	1	**4**
3	1	1	0	1	1	**4**
4	1	1	1	1	0	**4**
5	1	1	1	1	0	**4**
6	0	0	1	1	0	**2**
7	1	0	0	0	0	**1**
8	0	0	0	1	0	**1**
9	1	0	0	0	0	**1**
10	0	0	0	0	0	**0**
Difficulty	**4**	**5**	**5**	**3**	**7**	

Example (cont.)

Item discrimination

Student	Item1	Item2	Item3	Item4	Item5	**Score**
1	1	1	1	1	1	**5**
2	0	1	1	1	1	**4**
3	1	1	0	1	1	**4**
4	1	1	1	1	0	**4**
5	1	1	1	1	0	**4**
Difficulty	**1**	**0**	**1**	**0**	**2**	
6	0	0	1	1	0	**2**
7	1	0	0	0	0	**1**
8	0	0	0	1	0	**1**
9	1	0	0	0	0	**1**
10	0	0	0	0	0	**0**
Difficulty	**3**	**5**	**4**	**3**	**5**	

The Item Response Theory (IRT) model

- Let $Y_{si} \in \{0, 1\}$ represent the binary response of a subject $s \in \{1, \ldots, n\}$ to a set of test items indexed by $i = 1, \ldots, p$.
- Assume independent Bernoulli responses, i.e.

$$Y_{si} = \begin{cases} 1 \text{ (correct)} & \text{w.p. } \pi_{si} \\ 0 \text{ (wrong)} & \text{w.p. } 1 - \pi_{si} \end{cases}$$

The Item Response Theory (IRT) model

- Let $Y_{si} \in \{0, 1\}$ represent the binary response of a subject $s \in \{1, \dots, n\}$ to a set of test items indexed by $i = 1, \dots, p$.
- Assume independent Bernoulli responses, i.e.

$$Y_{si} = \begin{cases} 1 \text{ (correct)} & \text{w.p. } \pi_{si} \\ 0 \text{ (wrong)} & \text{w.p. } 1 - \pi_{si} \end{cases}$$

- We can model the probability of success using the two-parameter logistic model (2PL) defined by:

$$\pi_{si}(\boldsymbol{z}, \boldsymbol{\theta}) := \Pr(Y_{si} = 1 \mid \boldsymbol{z}, \boldsymbol{\theta}) = \frac{e^{a_i(z_s - b_i)}}{1 + e^{a_i(z_s - b_i)}},$$

  where
  - $\boldsymbol{z} = (z_1, \dots, z_n)^\top$ are the **latent traits** of the subjects; and
  - $\boldsymbol{\theta} = (a_i, b_i)_{i=1}^p$ are the **model parameters**, including
    - item **difficulty** parameters b_i (location) and
    - item **discrimination** parameters a_i (scale).

The Item Response Theory (IRT) model

- Let $Y_{si} \in \{0, 1\}$ represent the binary response of a subject $s \in \{1, \ldots, n\}$ to a set of test items indexed by $i = 1, \ldots, p$.
- Assume independent Bernoulli responses, i.e.

$$Y_{si} = \begin{cases} 1 \text{ (correct)} & \text{w.p. } \pi_{si} \\ 0 \text{ (wrong)} & \text{w.p. } 1 - \pi_{si} \end{cases}$$

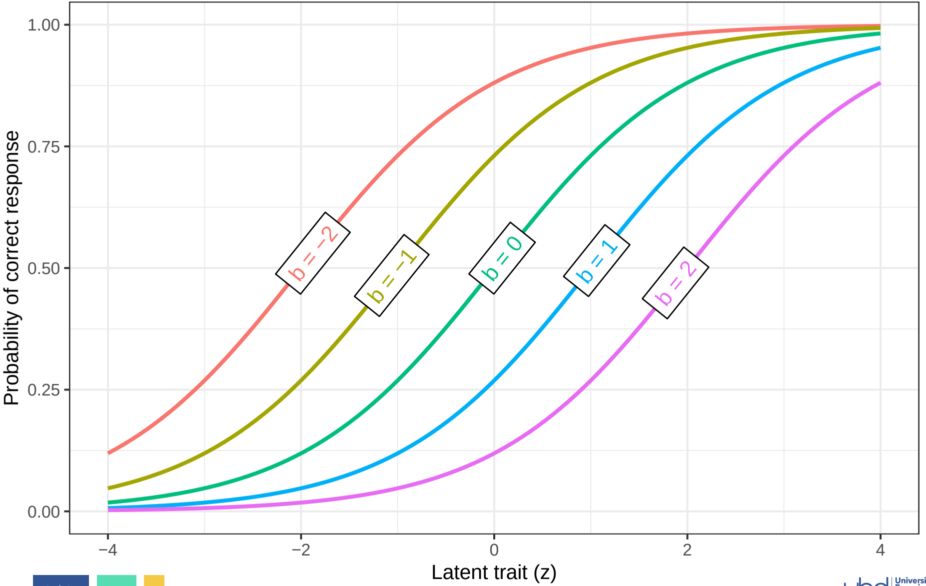
- We can model the probability of success using the two-parameter logistic model (2PL) defined by:

$$\operatorname{logit} \Pr(Y_{si} = 1 \mid \boldsymbol{z}, \boldsymbol{\theta}) = \log \frac{\pi_{si}(\boldsymbol{z}, \boldsymbol{\theta})}{1 - \pi_{si}(\boldsymbol{z}, \boldsymbol{\theta})} = a_i(z_s - b_i),$$

  where
  - $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ are the **latent traits** of the subjects; and
  - $\boldsymbol{\theta} = (a_i, b_i)_{i=1}^{p}$ are the **model parameters**, including
    - item **difficulty** parameters b_i (location) and
    - item **discrimination** parameters a_i (scale).

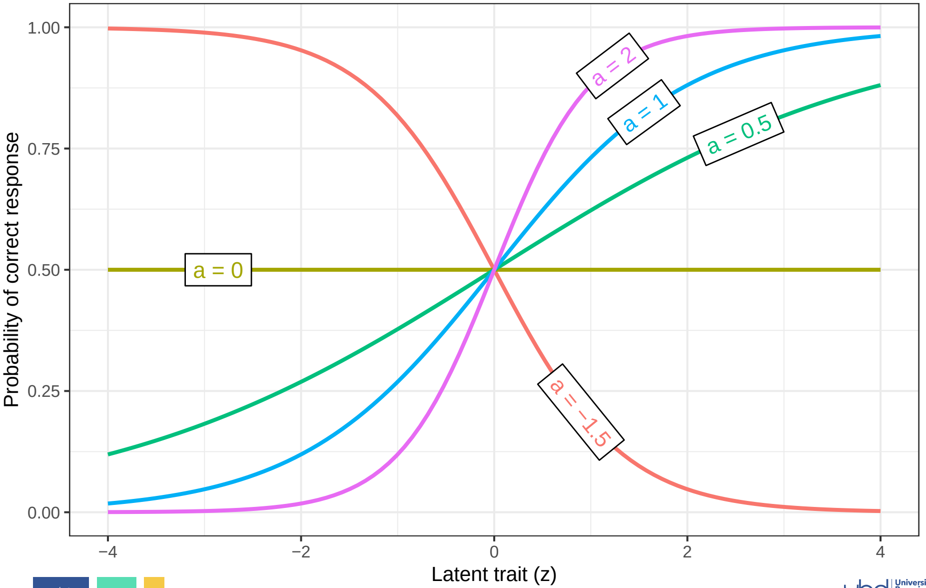
Interpretation

Effect of item difficulties on response probabilities

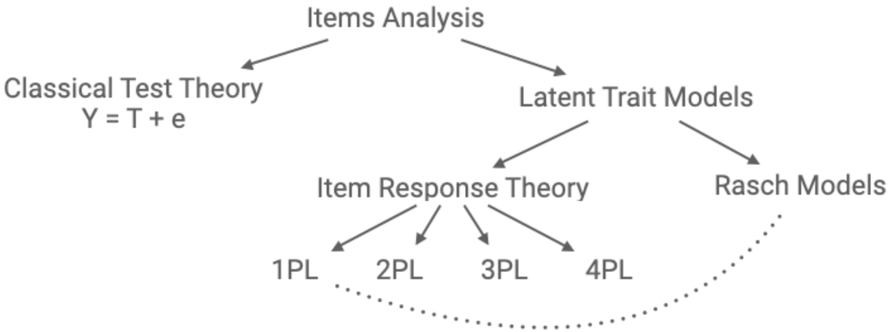


Interpretation

Effect of item discriminations on response probabilities



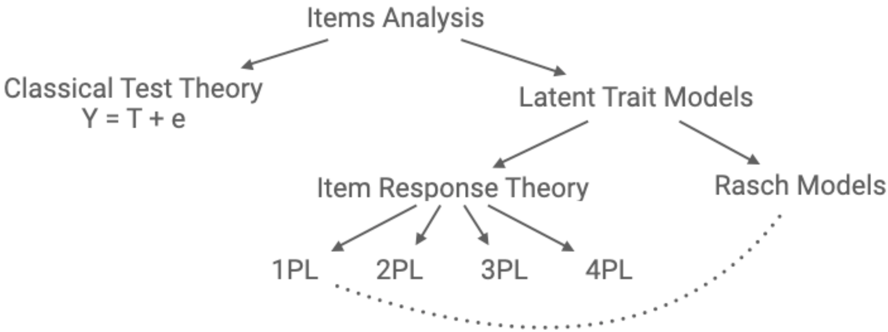
Family of IRT models



- The 2PL IRT model is a special case of the wider class of IRT models

$$\pi_{si}(\boldsymbol{z}, \boldsymbol{\theta}) := \Pr(Y_{si} = 1 \mid \boldsymbol{z}, \boldsymbol{\theta}) = c_i + (1 - c_i)\frac{e^{a_i(z_s - b_i)}}{1 + e^{a_i(z_s - b_i)}}.$$

Family of IRT models

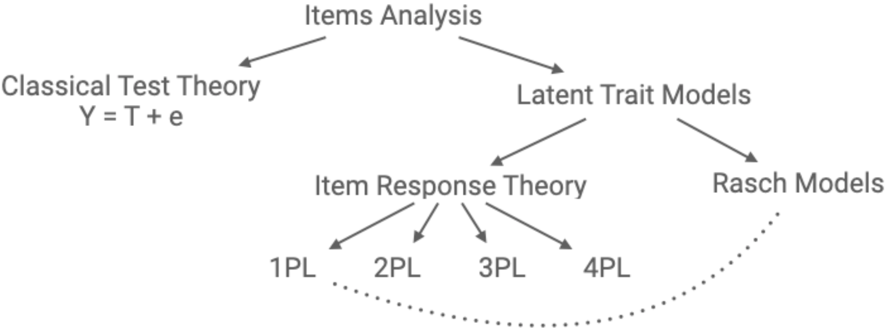


- The 2PL IRT model is a special case of the wider class of IRT models

$$\pi_{si}(\boldsymbol{z}, \boldsymbol{\theta}) := \Pr(Y_{si} = 1 \mid \boldsymbol{z}, \boldsymbol{\theta}) = c_i + (1 - c_i)\frac{e^{a_i(z_s - b_i)}}{1 + e^{a_i(z_s - b_i)}}.$$

- The above is the **3PL IRT** model, where c_i is the *guessing* parameter.

Family of IRT models

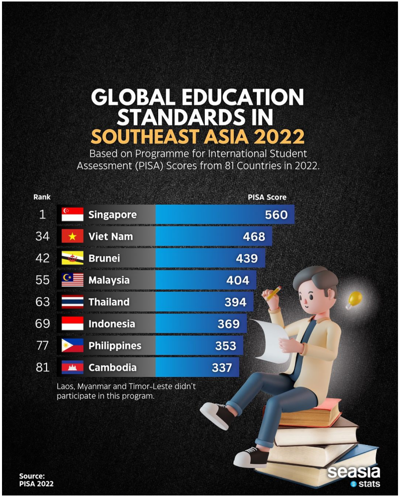


- The 2PL IRT model is a special case of the wider class of IRT models

$$\pi_{si}(\boldsymbol{z}, \boldsymbol{\theta}) := \Pr(Y_{si} = 1 \mid \boldsymbol{z}, \boldsymbol{\theta}) = c_i + (1 - c_i)\frac{e^{a_i(z_s - b_i)}}{1 + e^{a_i(z_s - b_i)}}.$$

- The above is the **3PL IRT** model, where c_i is the *guessing* parameter.
- When $c_i = 0$ and $a_i = 0$ for all $i = 1, \dots, p$, then we have the **1PL IRT** model, commonly known as the *Rasch model*.

Program for International Student Assessment (PISA)

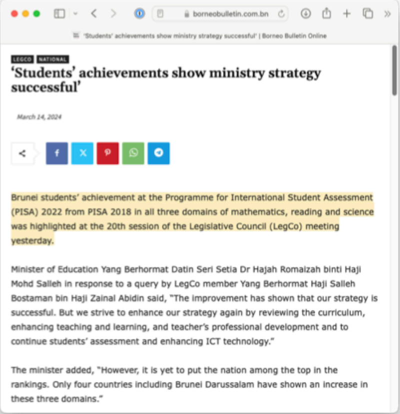


Credit: https://seasia.co/

- An international assessment that measures 15 year-old students' reading, mathematics, and science literacy (primarily among OECD nations).
- PISA primarily makes use of the Rasch (1PL) model for
  - **Scoring students**: Estimate students' abilities (latent traits).
  - **Item calibration**: Ensure items are appropriately challenging and can effectively differentiate students.
  - **Reporting outcomes**: Country and trends analyses.
  - **Diagnostic information**: Identify strengths and weakness in specific areas.

In Brunei

secara *stratified* sahaja. Tetapi *yes* 'kitani' sudah dapat, أَلْحَمْدُ لِلّٰه, 'kitani' dapat mencari juga beberapa *predictive tools* yang dapat digunakan dan *the team*, أَلْحَمْدُ لِلّٰه, dapat menggunakan *predictive tools* seperti *rush model analysis RN Conquest* untuk mengira *linear* dan *multiple regression* dan *other predictive analysis*. Jawapannya ada.



borneobulletin.com.bn

'Students' achievements show ministry strategy successful' | Borneo Bulletin Online

LEGCO NATIONAL

'Students' achievements show ministry strategy successful'

March 14, 2024

Brunei students' achievement at the Programme for International Student Assessment (PISA) 2022 from PISA 2018 in all three domains of mathematics, reading and science was highlighted at the 20th session of the Legislative Council (LegCo) meeting yesterday.

Minister of Education Yang Berhormat Datin Seri Setia Dr Hajah Romaizah binti Haji Mohd Salleh in response to a query by LegCo member Yang Berhormat Haji Salleh Bostaman bin Haji Zainal Abidin said, "The improvement has shown that our strategy is successful. But we strive to enhance our strategy again by reviewing the curriculum, enhancing teaching and learning, and teacher's professional development and to continue students' assessment and enhancing ICT technology."

The minister added, "However, it is yet to put the nation among the top in the rankings. Only four countries including Brunei Darussalam have shown an increase in these three domains."

Source: MMN Hansard 13/3/24 (am) & Borneo Bulletin 14/3/24

Software

Many software packages available, ranging from expensive commercial software (flexMIRT™, IRTPRO™, PARSCALE[a]) to free and open-source (e.g. in R: `{mirt}`, `{ltm}`, `{lavaan}`[b]).

The software mentioned in the MMN Hansard is Acer's ConQuest.

[a] Annual licence fee of $10,600!
[b] 21,000+ citations.

$659.00

ACER ConQuest 5 Multiple Standard Licence – Windows

✔ In stock

Estimation via MML

- Maximum marginal likelihood (MML) estimation [c.f. joint maximum likelihood (JML)] requires an additional assumption: $z_s \overset{\text{iid}}{\sim} N(0,1)$.

Estimation via MML

- Maximum marginal likelihood (MML) estimation [c.f. joint maximum likelihood (JML)] requires an additional assumption: $z_s \overset{\text{iid}}{\sim} N(0,1)$.
- Then the MML involves maximisation of the likelihood

$$L(\theta) = \prod_{s=1}^{n} \int \prod_{i=1}^{p} \pi_{si}(z,\boldsymbol{\theta})^{y_{si}} \big(1 - \pi_{si}(\boldsymbol{z},\boldsymbol{\theta})\big)^{1-y_{si}} \phi(z_s)\, \mathrm{d} z_s$$

  where $\phi(\theta_s)$ is the standard normal density function.
- This intractable integral is usually overcome using quadrature rules.

Estimation via MML

- Maximum marginal likelihood (MML) estimation [c.f. joint maximum likelihood (JML)] requires an additional assumption: $z_s \overset{\text{iid}}{\sim} N(0,1)$.
- Then the MML involves maximisation of the likelihood

$$L(\theta) = \prod_{s=1}^{n} \int \prod_{i=1}^{p} \pi_{si}(z, \boldsymbol{\theta})^{y_{si}} \big(1 - \pi_{si}(\boldsymbol{z}, \boldsymbol{\theta})\big)^{1-y_{si}} \phi(z_s) \, \mathrm{d} z_s$$

  where $\phi(\theta_s)$ is the standard normal density function.
- This intractable integral is usually overcome using quadrature rules.
- Some remarks:
  - It can be shown that bias is of $O(n^{-1})$, so in finite samples the bias is typically non-zero (Lord, 1986), though generally less biased than JML.
  - Parameters are consistent only when model is correctly specified (Bock & Aitkin, 1981).
  - MML is more robust to sample size variations and provides more stable item parameter estimates (Engelen, 1987).

{ltm} R package

The {ltm} package is available on CRAN. Example using Law School Admission Test (LSAT) from the US.

```
# install.packages("ltm")
library(ltm)
head(LSAT)  # contained within {ltm}
```

```
  Item 1 Item 2 Item 3 Item 4 Item 5
1      0      0      0      0      0
2      0      0      0      0      0
3      0      0      0      0      0
4      0      0      0      0      1
5      0      0      0      0      1
6      0      0      0      0      1
```

{ltm} R package (cont.)

Fit a 2PL model

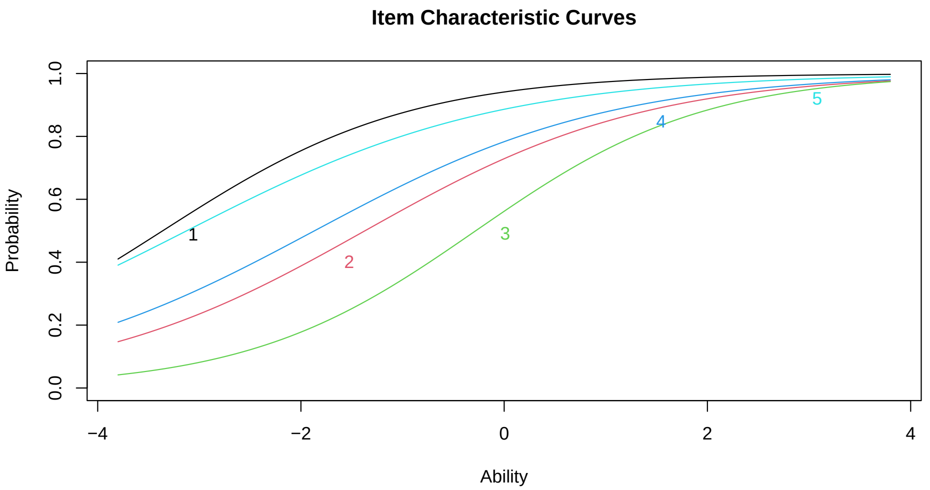
```
(fit <- ltm(LSAT ~ z1, IRT.param = TRUE))
```

```
Call:
ltm(formula = LSAT ~ z1, IRT.param = TRUE)

Coefficients:
        Dffclt  Dscrmn
Item 1  -3.360   0.825
Item 2  -1.370   0.723
Item 3  -0.280   0.890
Item 4  -1.866   0.689
Item 5  -3.124   0.657

Log.Lik: -2466.653
```

{ltm} R package (cont.)

Plot Item Characteristic Curves (ICC)

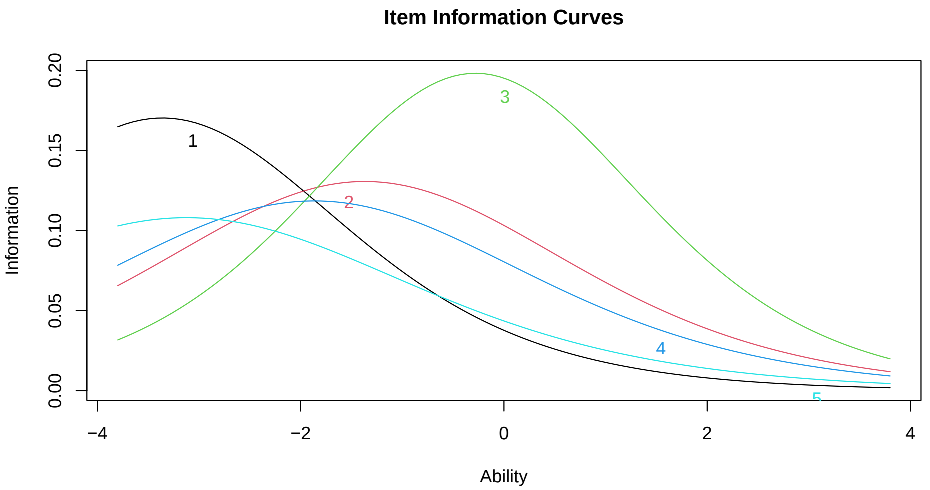
```
plot(fit)
```



{ltm} R package (cont.)

Plot Item **Information** Curves (IIC)

```
plot(fit, type = "IIC")
```



{ltm} R package (cont.)

Fit Rasch models

```r
# constr sets the (common) discrimination parameter to 1
(fit <- rasch(LSAT, constr = cbind(length(LSAT) + 1, 1)))
```

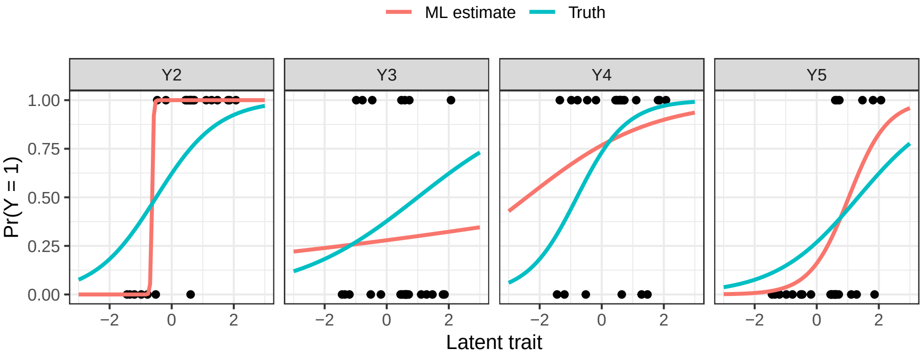
```
Call:
rasch(data = LSAT, constraint = cbind(length(LSAT) + 1, 1))

Coefficients:
Dffclt.Item 1  Dffclt.Item 2  Dffclt.Item 3
       -2.872         -1.063         -0.258
Dffclt.Item 4  Dffclt.Item 5         Dscrmn
       -1.388         -2.219          1.000

Log.Lik: -2473.054
```

Bias problem

In practice, sample size can be limited.

- Small-scale educational assessments, or
- Pilot studies (before deploying the test proper).

Standard IRT model estimates can be biased due to small sample sizes. We explore an empirical bias adjustment method to mitigate bias issues.



Sources of (parameter) bias

Besides small sample sizes...

- Departure from normality, e.g. [can be treated using robust ML]
  - skewed latent traits (Wall et al., 2012); or
  - zero-inflated distributions (Wall et al., 2015).
- Model misspecification
  - Incorrect functional form (e.g. 2PL instead of 3PL)
  - Dimensionality (assuming unidimensional model for multidimensional data), model incorrectly assumes all items measure a single common trait when there are multiple underlying abilities
- Differences in response styles. E.g. careless respondents (Hong & Cheng, 2019) or tendency to use extreme categories
- Etc.

Bias correction

estimator → $\hat\theta$; improved estimator → $\tilde\theta$; bias function → $B_G(\theta_0)$; possibly intractable → E_G; unknown true value → θ_0

$$\hat\theta - \tilde\theta = B_G(\theta_0) := \mathrm{E}_G(\hat\theta - \theta_0)$$

					Requirements		
	Method	Model	$B_G(\theta_0)$	Type	$\mathrm{E}(\cdot)$	$\partial\cdot$	$\hat\theta$
1	Asymptotic bias correction	full	analytical	explicit	✓	✓	✓
2	Adjusted score functions	full	analytical	implicit	✓	✓	✗
3	Bootstrap	partial	simulation	explicit	✗	✗	✓
4	Jackknife	partial	simulation	explicit	✗	✗	✓
5	Indirect inference	full	simulation	implicit	✗	✗	✓
6	Explicit RBM	partial	analytical	explicit	✗	✓	✓
7	Implicit RBM	partial	analytical	implicit	✗	✓	✗

1–Efron (1975), Cordeiro and McCullagh (1991); 2–Firth (1993), Kosmidis and Firth (2009); 3–Efron and Tibshirani (1994), Hall and Martin (1988); 4–Quenouille (1956), Efron (1982); 5–Gourieroux et al. (1993), MacKinnon and Smith Jr (1998)

Empirical bias reducing adjustments

- Kosmidis and Lunardon (2024) introduces a novel general framework for reducing the bias in M-estimation, which is derived from asymptotically unbiased estimating functions.

Empirical bias reducing adjustments

- Kosmidis and Lunardon (2024) introduces a novel general framework for reducing the bias in M-estimation, which is derived from asymptotically unbiased estimating functions.
- Briefly, $\hat{\theta}$ is an M-estimator if $\hat{\theta} = \arg\min_\theta \sum_{s=1}^{n} \rho_s(\theta)$, or results from the solution (van der Vaart, 1998) of

$$\sum_{s=1}^{n} \psi_s(\theta) = 0.$$

[E.g. maximum likelihood: $\psi_s(\theta) = \nabla \log L_s(\theta)$.]

Empirical bias reducing adjustments

- Kosmidis and Lunardon (2024) introduces a novel general framework for reducing the bias in M-estimation, which is derived from asymptotically unbiased estimating functions.
- Briefly, $\hat{\theta}$ is an M-estimator if $\hat{\theta} = \arg\min_\theta \sum_{s=1}^n \rho_s(\theta)$, or results from the solution (van der Vaart, 1998) of

$$\sum_{s=1}^{n} \psi_s(\theta) = 0.$$

  [E.g. maximum likelihood: $\psi_s(\theta) = \nabla \log L_s(\theta)$.]
- For M-estimators, it is possible to write down the bias function as

$$\mathrm{E}_G(\hat{\theta} - \theta_0) = b(\theta_0) + O(n^{-3/2}),$$

  where $b(\theta_0)$ may be approximated empirically by a function of derivatives of $\psi_s(\theta)$.
- Then, a reduced-bias estimator is simply $\hat{\theta} - b(\hat{\theta})$.

Implicit reduced bias M-estimators (iRBM)

- The estimator $\tilde{\theta}^{(\mathrm{iRBM})}$ is obtained from

$$\tilde{\theta}^{(\mathrm{iRBM})} = \arg\max_{\theta} \left\{ \log L(\theta) - \frac{1}{2} \mathrm{tr}\left[j(\theta)^{-1} e(\theta) \right] \right\}, \quad \text{where}$$

  - $j(\theta) = -\sum_{s=1}^{n} \nabla^2 \log L_s(\theta)$ is the observed information matrix,
  - $e(\theta) = \sum_{s=1}^{n} \nabla \log L_s(\theta) \nabla \log L_s(\theta)^\top$ is the cross-products of the scores.

Implicit reduced bias M-estimators (iRBM)

- The estimator $\tilde{\theta}^{(\mathrm{iRBM})}$ is obtained from

$$\tilde{\theta}^{(\mathrm{iRBM})} = \arg\max_{\theta} \left\{ \log L(\theta) - \frac{1}{2} \mathrm{tr}\left[j(\theta)^{-1} e(\theta) \right] \right\}, \quad \text{where}$$

  - $j(\theta) = -\sum_{s=1}^{n} \nabla^2 \log L_s(\theta)$ is the observed information matrix,
  - $e(\theta) = \sum_{s=1}^{n} \nabla \log L_s(\theta) \nabla \log L_s(\theta)^\top$ is the cross-products of the scores.
- The iRBM estimator is consistent and asymptotically normal, i.e.

$$\sqrt{n}(\tilde{\theta}^{(\mathrm{iRBM})} - \theta_0) \xrightarrow{\mathrm{D}} \mathrm{N}\left(0, j(\theta_0)^{-1} e(\theta_0) j(\theta_0)^{-\top}\right).$$

  and has smaller bias than the M-estimator $\hat{\theta}$.

Implicit reduced bias M-estimators (iRBM)

- The estimator $\tilde{\theta}^{(\text{iRBM})}$ is obtained from

$$\tilde{\theta}^{(\text{iRBM})} = \arg\max_{\theta} \left\{ \log L(\theta) - \frac{1}{2} \operatorname{tr}\left[j(\theta)^{-1} e(\theta) \right] \right\}, \quad \text{where}$$

  - $j(\theta) = -\sum_{s=1}^{n} \nabla^2 \log L_s(\theta)$ is the observed information matrix,
  - $e(\theta) = \sum_{s=1}^{n} \nabla \log L_s(\theta) \nabla \log L_s(\theta)^\top$ is the cross-products of the scores.
- The iRBM estimator is consistent and asymptotically normal, i.e.

$$\sqrt{n}(\tilde{\theta}^{(\text{iRBM})} - \theta_0) \xrightarrow{\text{D}} \text{N}\left(0, j(\theta_0)^{-1} e(\theta_0) j(\theta_0)^{-\top}\right).$$

  and has smaller bias than the M-estimator $\hat{\theta}$.
- Components of the estimated θ may "blow up" under certain data configurations (e.g. perfect separation). To mitigate this, a shrinkage factor can be applied to obtain a *penalised* iRBM estimator from

$$\tilde{\theta}^{(\text{iRBMp})} = \arg\max_{\theta} \left\{ \log L(\theta) - \frac{1}{2} \operatorname{tr}\left[j(\theta)^{-1} e(\theta) \right] - \frac{1}{n} \|\theta\|^2 \right\}.$$

Explicit reduced bias M-estimators (eRBM)

- Another estimator with the same bias properties as the iRBM is the eRBM, which is obtained via

$$\tilde{\theta}^{(\text{eRBM})} = \hat{\theta} + j(\hat{\theta})^{-1} A(\hat{\theta}),$$

where

$$A(\hat{\theta}) = -\frac{1}{2} \nabla \operatorname{tr} \left\{ j(\theta)^{-1} e(\theta) \right\} \Big|_{\theta=\hat{\theta}}.$$

Explicit reduced bias M-estimators (eRBM)

- Another estimator with the same bias properties as the iRBM is the eRBM, which is obtained via

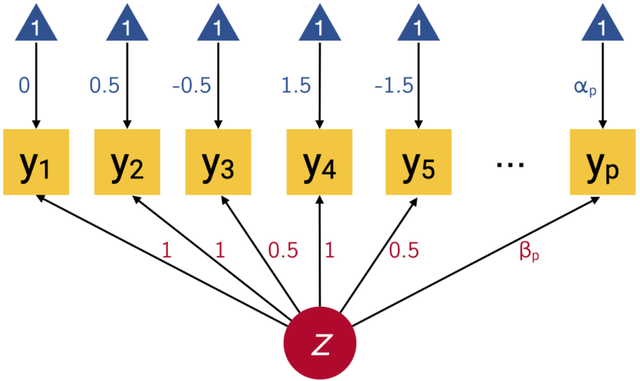
$$\tilde{\theta}^{(\text{eRBM})} = \hat{\theta} + j(\hat{\theta})^{-1} A(\hat{\theta}),$$

  where

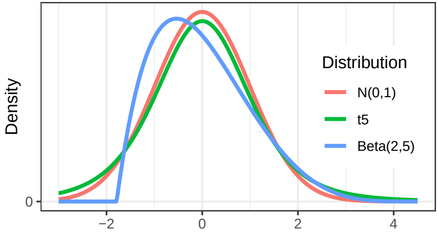
$$A(\hat{\theta}) = -\frac{1}{2} \nabla \operatorname{tr} \left\{ j(\theta)^{-1} e(\theta) \right\} \Big|_{\theta=\hat{\theta}}.$$

- Operationally the eRBM is simpler, though requiring accurate computation of the bias term to be effective (usually involving numerical routines).
- One downside: No saving infinite estimates.

Simulation setup

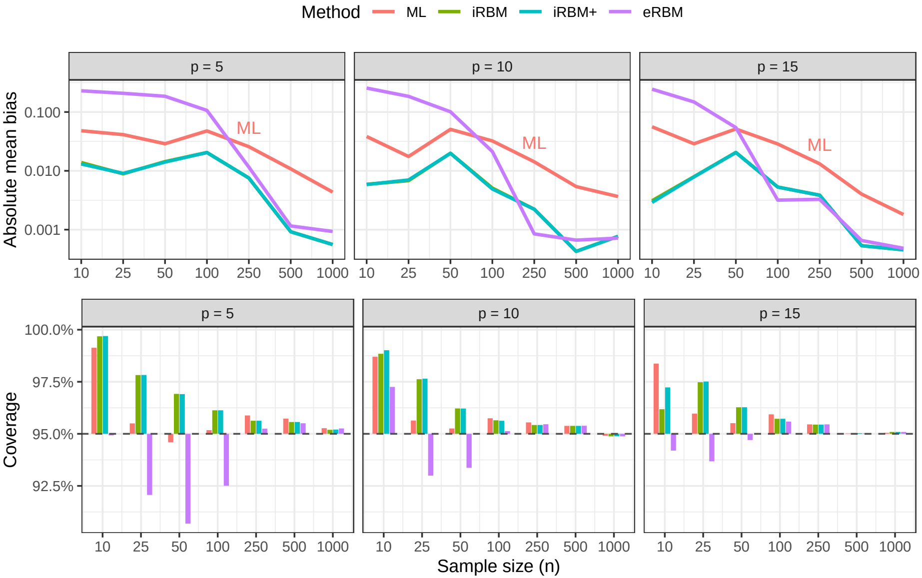


1. $n \in \{10, 25, 50, 100, 250, 500, 1000\}$
2. $p \in \{5, 10, 15\}$
3. Departure from normality:
   - $z \sim \mathrm{N}(0, 1)$
   - $z \sim t_5$
   - $z \sim \mathrm{Beta}(2, 5)$ (centred and scaled)



Simulation results

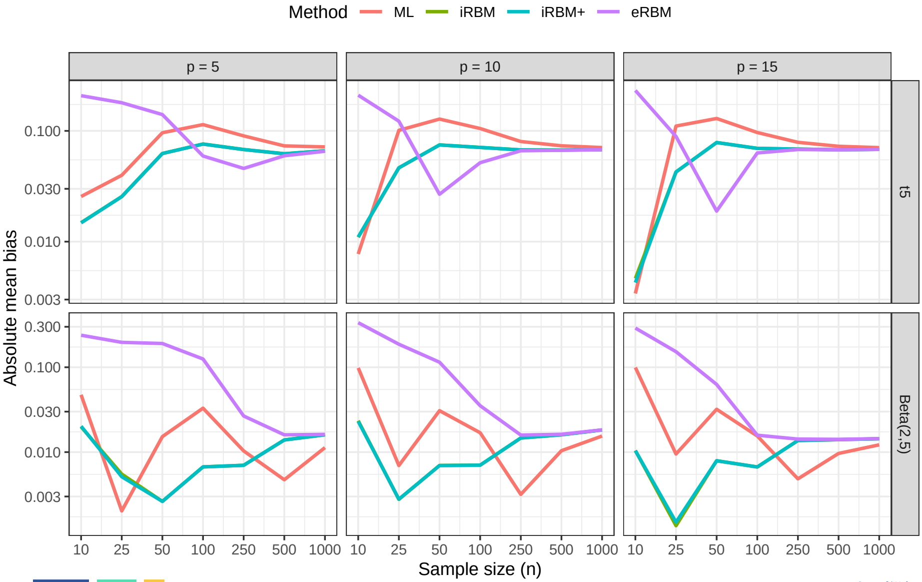
Effects of sample size and no. of items



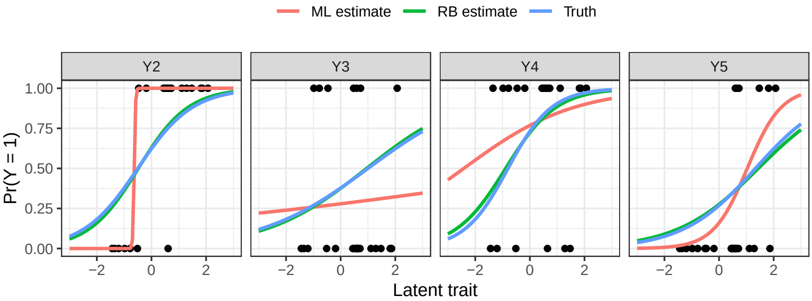
Universiti Brunei Darussalam

Simulation results

Effects of sample size and departure from normality



Conclusions



- Small sample size and departure from normality can lead to bias in the parameter estimates of IRT models.
- The iRBM and eRBM estimators are effective in reducing bias for the 2PL IRT model in small samples when the normality assumption holds.
- Way forward:
  - Comparison to other bias reduction methods (e.g. bootstrap, jackknife, etc.).
  - Investigate the performance of the iRBM and eRBM estimators in more complex IRT models (3PL and multidimensional IRT models).
  - Refine simulations to include more complex departures from normality.

Universiti Brunei Darussalam

End

Thank you!

References

Bock, R. D., & Aitkin, M. (1981).Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Cordeiro, G. M., & McCullagh, P. (1991).Bias correction in generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *53*(3), 629–643.

Efron, B. (1975).Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 1189–1242.

Efron, B. (1982, January). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970319

Efron, B., & Tibshirani, R. J. (1994, May). *An Introduction to the Bootstrap*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Engelen, R. J. H. (1987).A review of different estimation procedures in the Rasch model.

Firth, D. (1993).Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.1093/biomet/80.1.27

Gourieroux, C., Monfort, A., & Renault, E. (1993).Indirect inference. *Journal of Applied Econometrics*, *8*(S1), S85–S118. https://doi.org/10.1002/jae.3950080507

Hall, P., & Martin, M. A. (1988).On bootstrap resampling and iteration. *Biometrika*, *75*(4), 661–671.

Hong, M. R., & Cheng, Y. (2019).Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods*, *51*(2), 573–588. https://doi.org/10.3758/s13428-018-1150-4

References

Kosmidis, I., & Firth, D. (2009).Bias reduction in exponential family nonlinear models. *Biometrika*, *96*(4), 793–804.

Kosmidis, I., & Lunardon, N. (2024).Empirical bias-reducing adjustments to estimating functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *86*(1), 62–89. https://doi.org/10.1093/jrsssb/qkad083

Lord, F. M. (1986).Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. *Journal of Educational Measurement*, *23*(2), 157–162.

MacKinnon, J. G., & Smith Jr, A. A. (1998).Approximate bias correction in econometrics. *Journal of Econometrics*, *85*(2), 205–230.

Quenouille, M. H. (1956).Notes on bias in estimation. *Biometrika*, *43*(3/4), 353–360.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802256

Wall, M. M., Guo, J., & Amemiya, Y. (2012).Mixture Factor Analysis for Approximating a Nonnormally Distributed Continuous Latent Factor With Continuous and Dichotomous Observed Variables. *Multivariate Behavioral Research*, *47*(2), 276–313. https://doi.org/10.1080/00273171.2012.658339

Wall, M. M., Park, J. Y., & Moustaki, I. (2015).IRT modeling in the presence of zero-inflation with application to psychiatric disorderseverity. *Applied Psychological Measurement*, *39*(8), 583–597.