Comparação dos modelos de Rasch e de três parâmetros nas calibrações dos parâmetros dos itens do pré-teste da Prova Nacional para o Ingresso na Carreira Docente

Ricardo Primi (Laboratório de Avaliação Psicológica e Educacional – LabAPE, Universidade São Francisco, Itatiba)
&
Alexandre José de Souza Peres (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP)

Batalha ...

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \quad vs \quad P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

- Modelagem (tradição estatística) vs criação de medidas substanciais (tradição psicológica)

Debate entre entre Modelo de Rasch *vs* TRI de três parâmetros

Rasch

- Construir testes a partir de um modelo medida
  - Parâmetro *a* pode indicar DIF requerendo um modelo mais complexo
  - Por não possuir o parâmetro c não se ajustaria a testes de múltipla escolha
- Intervalar
  - Consequência: justifica uso de estatísticas paramétricas (todo o resto do mundo estaria errado!)
- "Perde" mais itens
  - Qual o impacto de variações de *a* e *c* afetam os parâmetros do modelo de Rasch ?

3-parâmetros

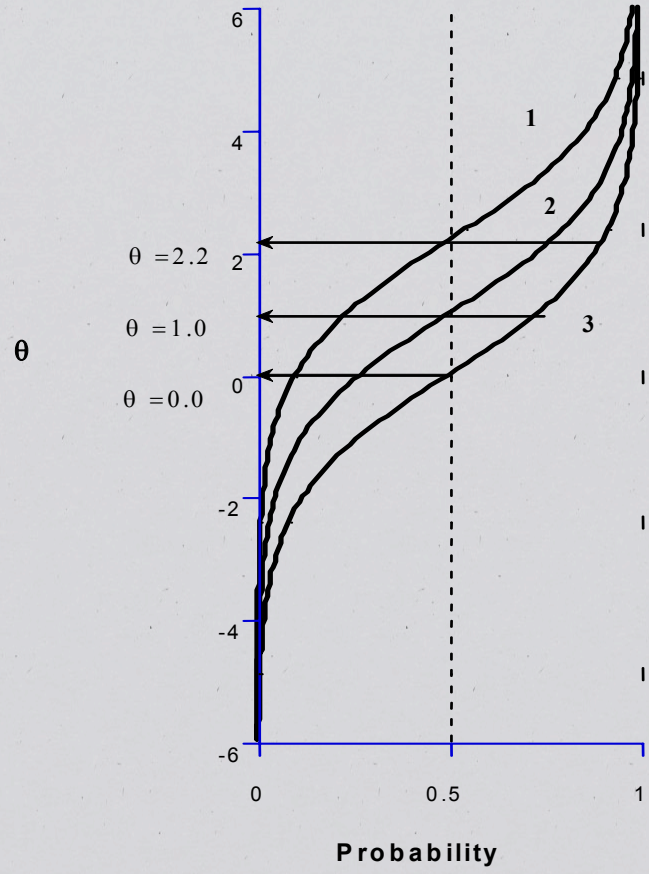
- Descrever os dados
  - O modelo de Rasch sempre seria pior do que o modelo de 3-parâmetros pois não se ajusta tão bem aos dados
  - Mais adequado para testes de múltipla escolha
- Não satisfaz as condições de medida intervalar
  - Qual o sentido de uma unidade ? (métrica arbitrária)
- "Salva" mais itens
  - Até que ponto itens com baixo *a* e alto *c* deveriam ser usados?

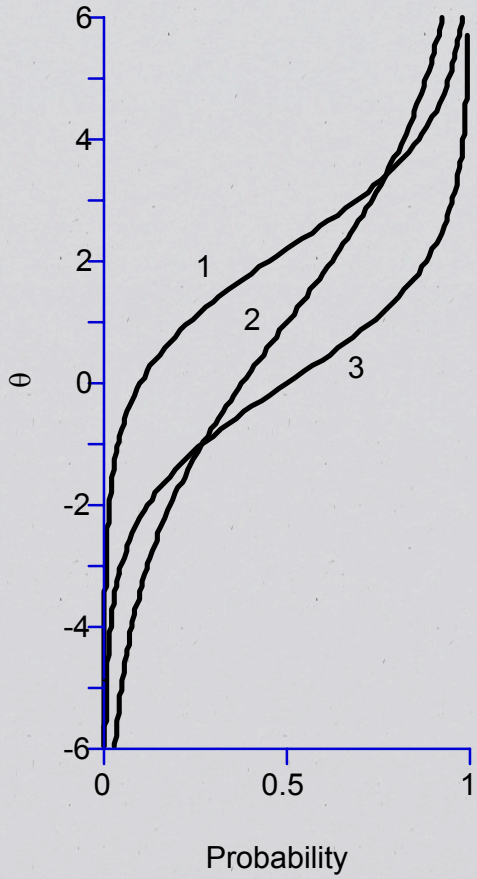
- *www.rasch.org/rmt/rmt61a.htm*

Birnbaum Model: 3-PL **For 2-PL, set c_i=0** **For 1-PL, set a_i=1.7, c_i=0**	**Rasch Model**
Allan Birnbaum 1957$ / 1968	Georg Rasch 1952$ / 1960
imitates data	**defines** measures
contrived to fit observed MCQ ICC's	**derived** to construct scientific measurement
$\log\left[\frac{P_{\theta i}-c_i}{1-P_{\theta i}}\right] = a_i(\theta - b_i)$	$\log\left[\frac{P_{ni}}{1-P_{ni}}\right] = (B_n - D_i)$
$\sum_i a_i X_{\theta i} = \sum_i a_i P_{\theta i} \rightarrow \theta$ $\sum_\theta \theta X_{\theta i} = \sum_\theta \theta P_{\theta i} \rightarrow a_i$ Shared $X_{\theta i}$ causes θ <–> a_i feedback: divergence	$\sum_i X_{ni} = \sum_i P_{ni} \rightarrow B_n$ $\sum_n X_{ni} = \sum_n P_{ni} \rightarrow D_i$ inevitable convergence
MCQ dichotomies **only** [1992: Eiji Muraki's Generalized Partial Credit Model]	**any ordered observation** dichotomy, rating, ranking, counting
guessing accepted reliable item asset	**guessing rejected** unreliable person liability
discrimination variation **welcomed** as a useful item scoring weight	**discrimination** variation **rejected** as a misleading item bias interaction
crossed ICC's accepted natural and unavoidable	**crossed ICC's rejected** prevents construct validity

Figure 1. Comparison of Rasch and Birnbaum Models.
($ first written report)

- Wilson, M. (2003). On Choosing a Model for Measuring. *Methods of Psychological Research* Online, 8(3), 1-22.





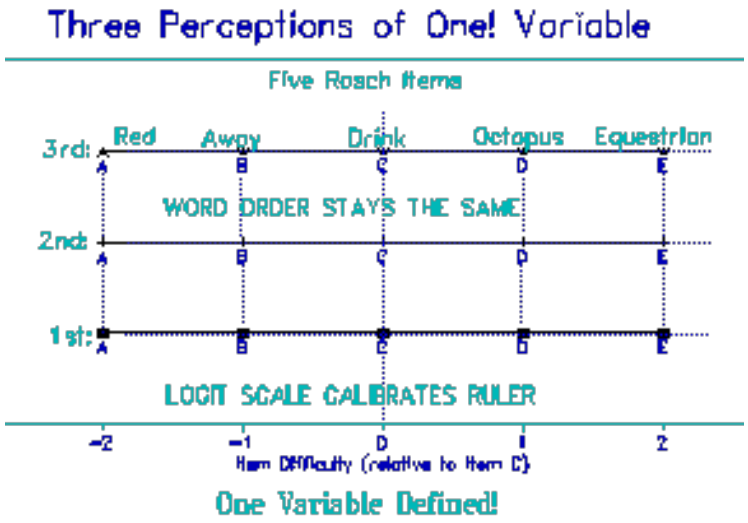


Figure 3. Five Rasch Items and Three Ability Levels
1st = Low ability; 2nd = Medium ability; 3rd = High Ability
Notice the 3 identical item-difficulty hierarchies (advancing from left to right)

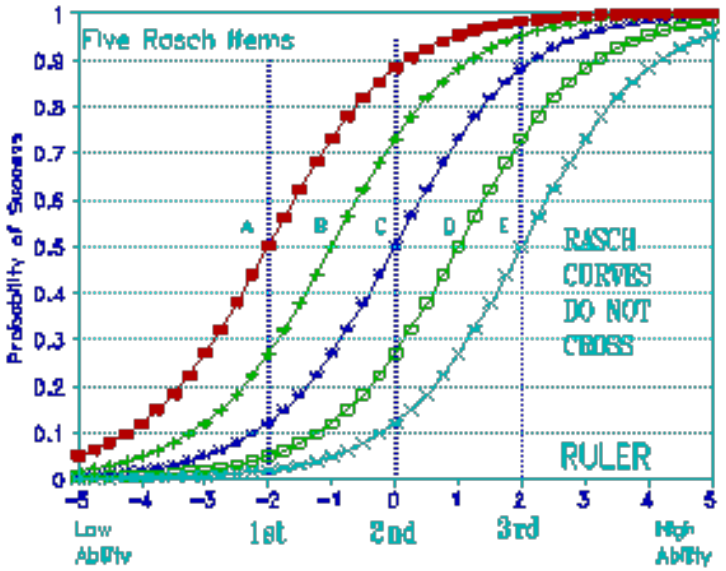


Figure 4. Five Rasch Curves and Three Ability Levels
1st = Low ability; 2nd = Medium ability; 3rd = High Ability

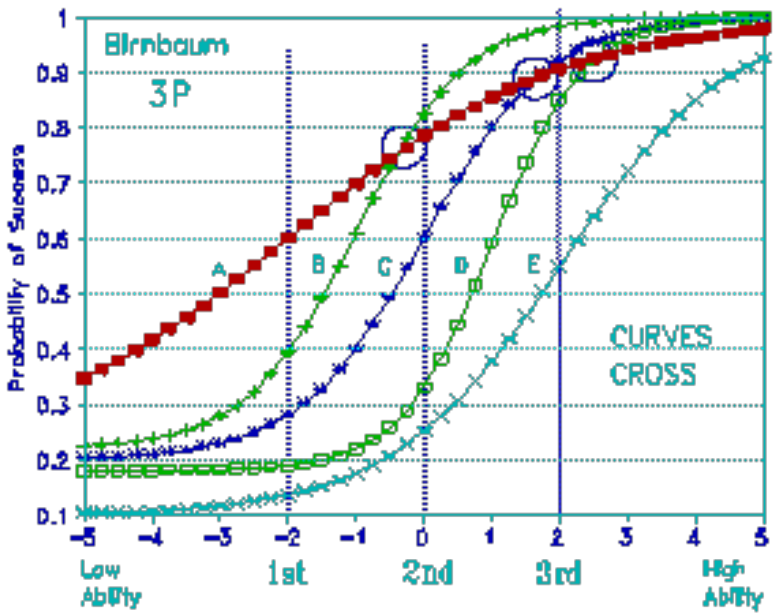


Figure 5. Five Birnbaum Curves and Three Ability Levels
1st = Low ability; 2nd = Medium ability; 3rd = High Ability

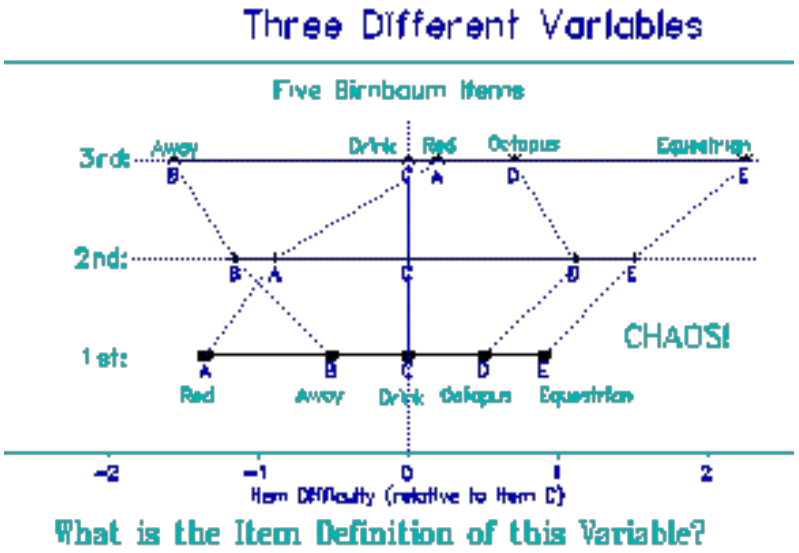


Figure 6. Five Birnbaum Items and Three Ability Levels
1st = Low ability; 2nd = Medium ability; 3rd = High Ability
Notice the 3 different item-difficulty hierarchies (advancing from left to right)

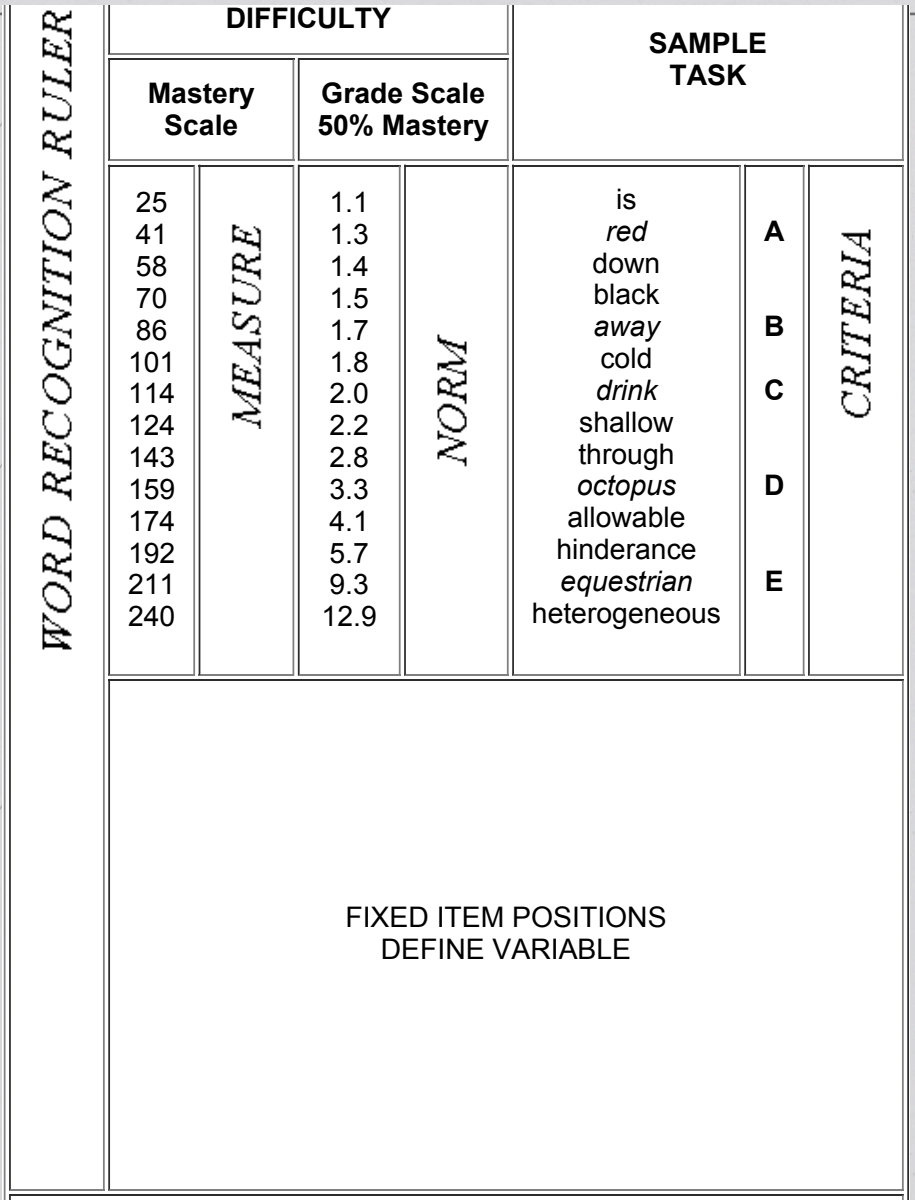


Figure 2. A useful, linear, invariant measuring instrument.

Impacto do *c* nas calibrações do modelo de Rasch

Uso do modelo de Rasch em testes de múltipla escolha

- Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics*, *37*(3), 417-442.

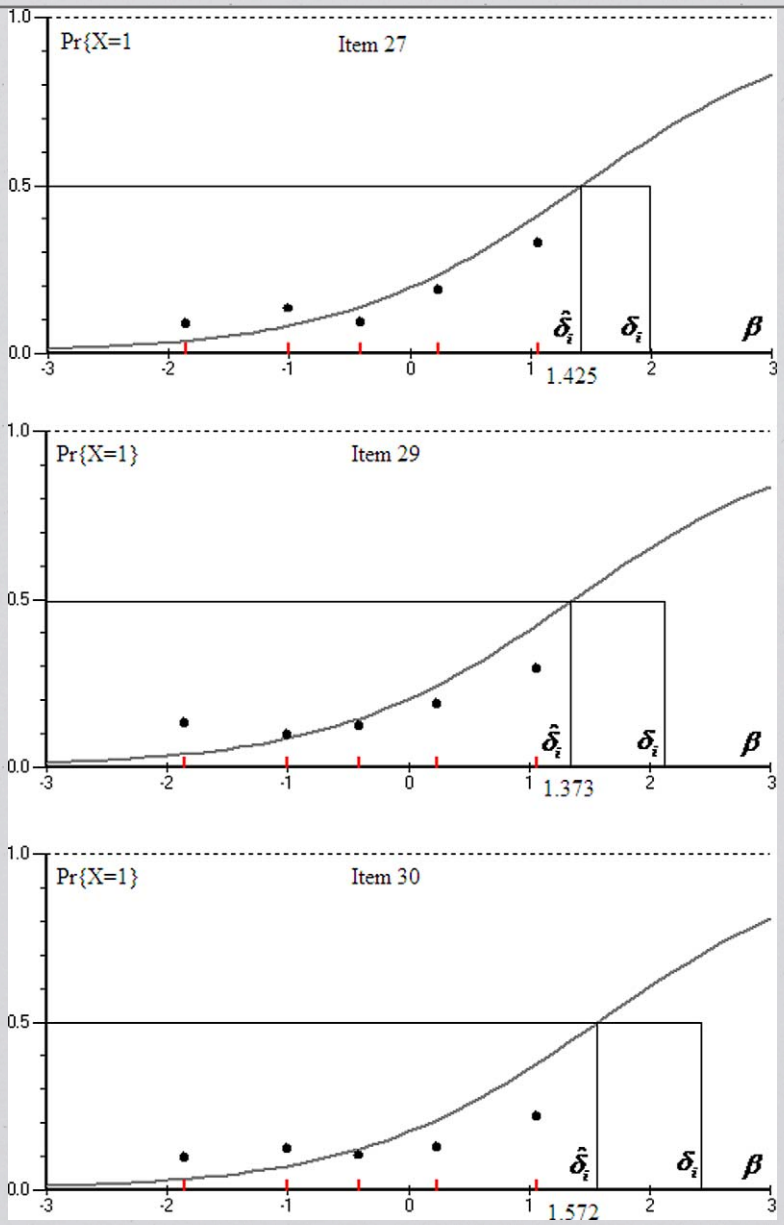


FIGURE 3. *ICCs of 3 items from the first analysis of the simulated example which contains guessed responses showing the bias of* $\hat{\delta}_i < \delta_i$.

Parâmetro *c* como moderador !

$$\Pr\{X_{ni} = 1\} = c_i + (1 - c_i)P = c_i + P - c_iP = P + c_i(1 - P).$$

$$\Pr\{X_{ni} = 1\} = P + c_i(1 - P)^{y},$$

y redutor do efeito do *c*
Em pessoas com alta habilidade
relativa ao item

Probabilidade de acerto
segundo modelo de Rasch
(sem acerto ao acaso)

é .. moderada pelo *c* multiplicado
pela distância entre a habilidade
da pessoa *vs* dificuldade do item Q = 1-P

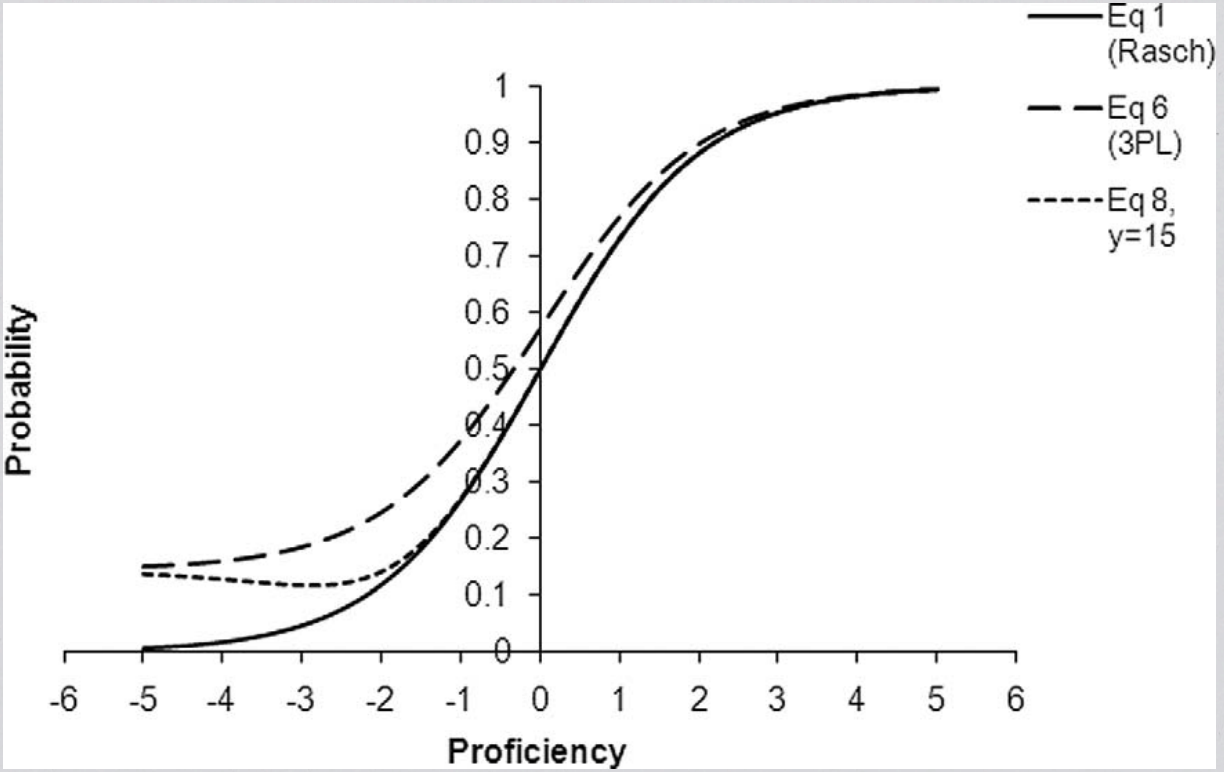


FIGURE 2. *Item characteristic curves for the Rasch model, the 3PL and the generalized guessing model with* $y = 15$.

Caminho do meio ...

(porque não o melhor dos dois mundos?)

Questões basais

- Quanto, de fato, os resultados dos sujeitos e dos itens diferem ao utilizarmos um ou outro modelo ?

Prova Docente

- “A Prova Nacional de Concurso para Ingresso na Carreira Docente tem o objetivo principal de subsidiar os Estados, o Distrito Federal e os Municípios na realização de concursos públicos para a contratação de docentes para a educação básica. Trata-se de uma prova anual, a ser aplicada de forma descentralizada em todo o país para os candidatos ao ingresso na carreira docente das redes de educação básica.”

Matriz de Referência

Processos (competências):

P1. A articulação de conhecimentos para compreensão de aspectos culturais, ambientais, políticos, econômicos, científicos e tecnológicos da sociedade contemporânea.

P2. A promoção de ações de inclusão, de valorização da diversidade e singularidade dos alunos e de respeito aos direitos educativos no contexto da comunidade escolar.

P3. O planejamento do trabalho pedagógico para orientar os processos de construção de conhecimento.

P4. O desenvolvimento de metodologias e recursos pertinentes para alcançar os objetivos do trabalho pedagógico.

P5. A organização de procedimentos avaliativos que permitam reorientar a prática educacional.

P6. A comunicação com coerência e coesão por meio de textos escritos.

Objetos de conhecimento:

- Políticas Educacionais (POL)
- Organização e Gestão do Trabalho Pedagógico (OGTP)
- Desenvolvimento e Aprendizagem (DES)
- Língua Portuguesa e seu Ensino (LP)
- Matemática e seu Ensino (MAT)
- História e seu Ensino (HIS)
- Geografia e seu Ensino (GEO)
- Ciências da Natureza e seu Ensino (CIEN)
- Arte e seu Ensino (ART)
- Educação Física e seu Ensino (EF)

Mapa de itens do pré-teste

Objetos (áreas) de conhecimento	Processos (competências)					
	P1	P2	P3	P4	P5	Total
Políticas Educacionais (POL)		36				36
Organização e Gestão do Trabalho Pedagógico (OGTP)			12	12	12	36
Desenvolvimento e Aprendizagem (DES)			12	12	12	36
Língua Portuguesa (LP)	6		8	8	8	30
Matemática (MAT)	6		8	8	8	30
História (HIS)	6		8	8	8	30
Geografia (GEO)	6		8	8	8	30
Ciências da Natureza (CN)	6		8	8	8	30
Arte (ART)	6		8	8	8	30
Educação Física (EF)			8	8	8	24
Total	36	36	80	80	80	312

Dados do pré-teste da Prova Docente

- Uso de Blocos Balanceados Incompletos (BIB) e distribuição de cadernos em espiral.
  - Equalização e link por: desenho de grupo equivalente e itens comuns
- 312 itens divididos em dois subconjuntos de 156 itens (B1 e B2)
- Formaram-se: 26 blocos de 12 itens -> 52 cadernos de 36 itens
  - Itens comuns entre cadernos (12)
  - Amostras randomicamente equivalentes aos cadernos
  - Amostras representativas para estimação de correlações entre qualquer par de item (para se calcular a correlação entre eles).
- N = 10.588 pessoas (professores e estudantes): 5.759 B1 e 4.829 B2
- AFE e AFC testando-se a unidimensionalidade

Objetivos e Método

- Comparar as estimativas dos parâmetros pelo modelo de Rasch e 3-parâmetros.

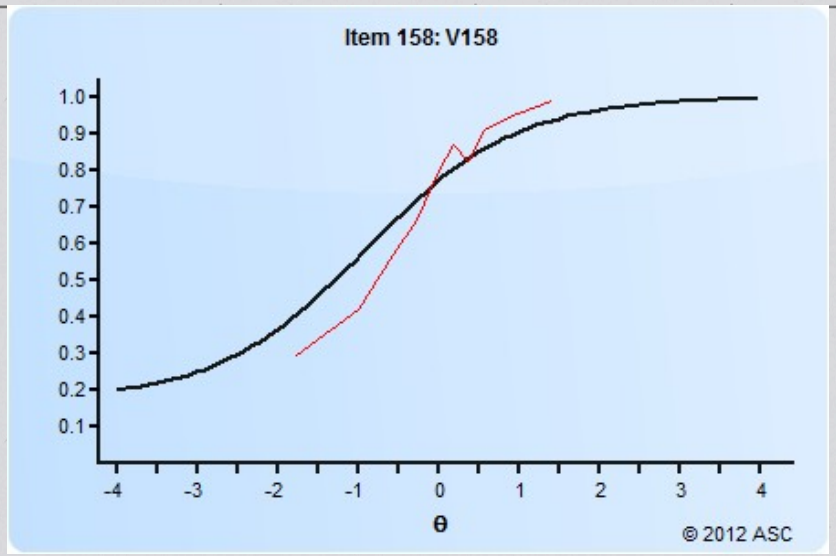
- Calibração dos parâmetros:
  - 3-parâmetros: XCALIBRE 4.1 (Guyer, & Thompson, 2012) que implementa o método de estimação de máxima verossimilhança marginal (MML) e EAP para os Thetas
  - 1-parâmetro (modelo de Rasch): WINSTEPS (Linacre, 2009) que implementa o método de máxima verosimilhança conjunta (JMLE)
  - Zero na escala de habilidade para identificar a métrica
- Análise correlacional entre os parâmetros
- Calibrados itens com cargas aceitáveis no fator geral
- Estimativas dos sujeitos não otimizadas !!

Índices de ajuste no modelo de Rasch

$$Outfit_i = \frac{\sum_{n=1}^{N} \frac{r_{ni}^2}{V_{ni}}}{N}$$

$$Infit_i = \frac{\sum_{n=1}^{N} r_{ni}^2}{\sum_{n=1}^{N} V_{ni}}$$

	Exemplos de padrões de resposta à 12 itens em 3 níveis de dificuldade			Indices de Ajuste	
Padrões	*Fácil*	*Médio*	*Difícil*	*INFIT*	*OUTFIT*
1. Padrão ajustado	1110	1011	1000	<1,3	<1,3
2. Descuido/Desatenção	**0**111	1111	0000	<1,3	>>1,3
3. "Chute" com sorte	1110	1110	000**1**	<1,3	>>1,3
4. Conhecimento específico	1111	000**1**	**11**00	>>1,3	<1,3
5. Padrão assistematico	**0000**	**0111**	**1111**	>>1,3	>>1,3



Item information

Seq.	ID	Model	Key	Scored	Num Options	Domain	Flags
158	V158	3PL	B	Yes	4	1	

Classical statistics

N	P	S-Rpbis	T-Rpbis	Alpha w/o
1100	0.724	0.336	0.506	0.738

IRT parameters

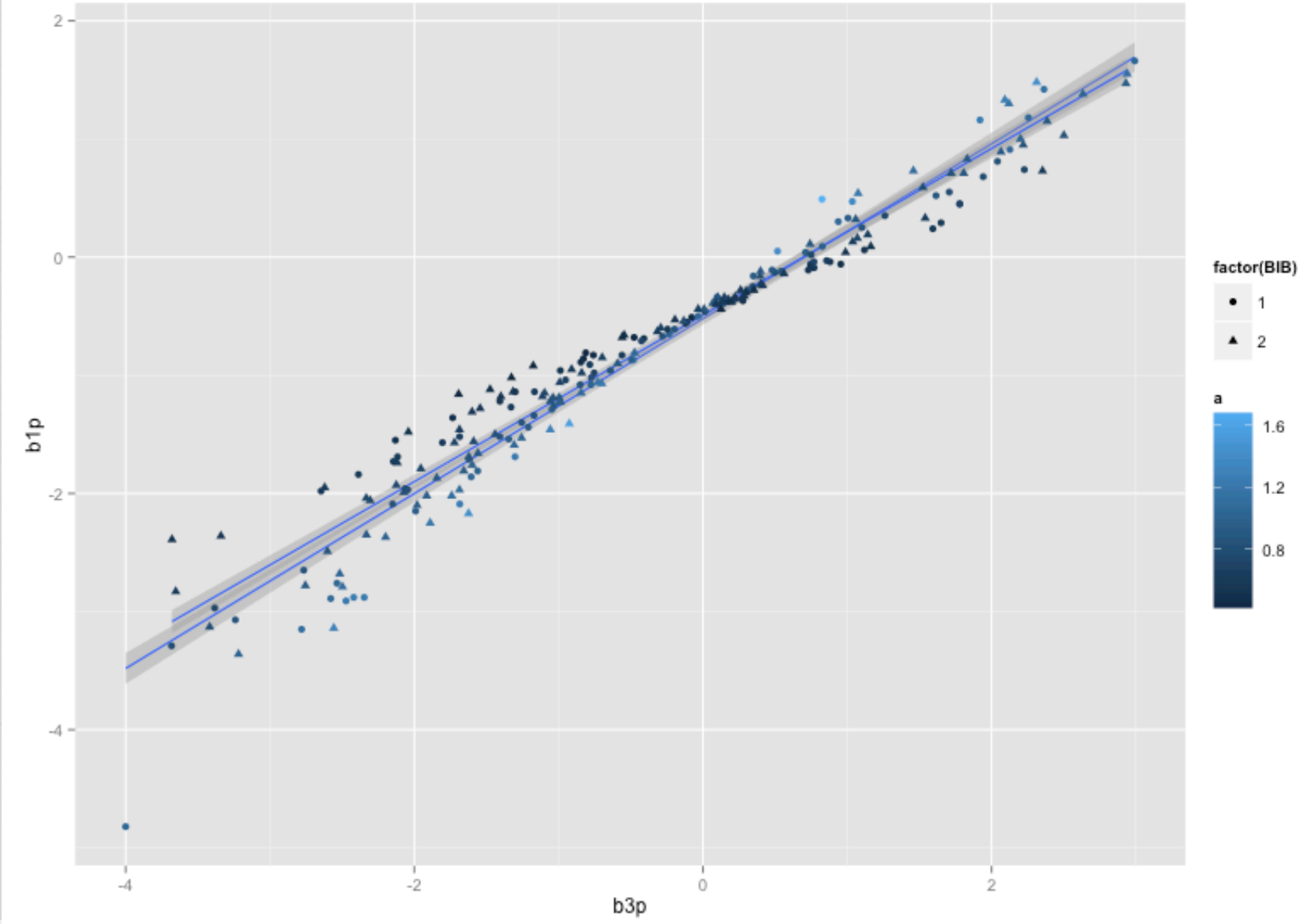
a	b	c	a SE	b SE	c SE	Chi-sq	df	p	z Resid	p
1.060	-0.844	0.172	0.036	0.037	0.019	37.561	15	0.001	0.413	0.679

Option statistics

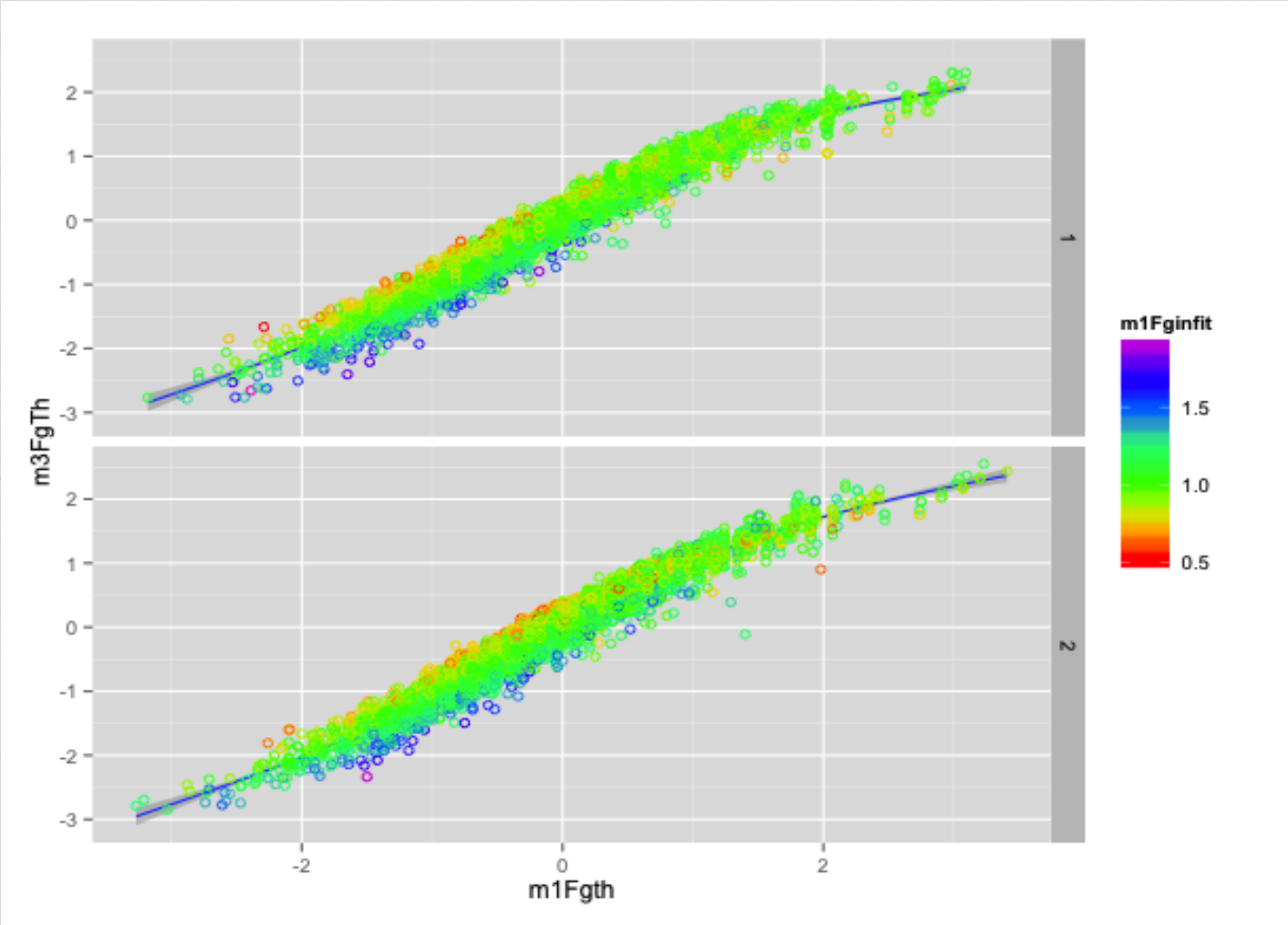
Option	N	Prop.	S-Rpbis	T-Rpbis	Mean	SD	
A	40	0.036	-0.112	-0.163	-0.774	0.944	
B	796	0.724	0.336	0.506	0.273	0.761	**KEY**
C	133	0.121	-0.141	-0.242	-0.604	0.722	
D	125	0.114	-0.233	-0.339	-0.874	0.885	
Omit	6	0.005	-0.126	-0.130	-1.608	1.176	
Not Admin	9488				0.003	0.557	

Resultados

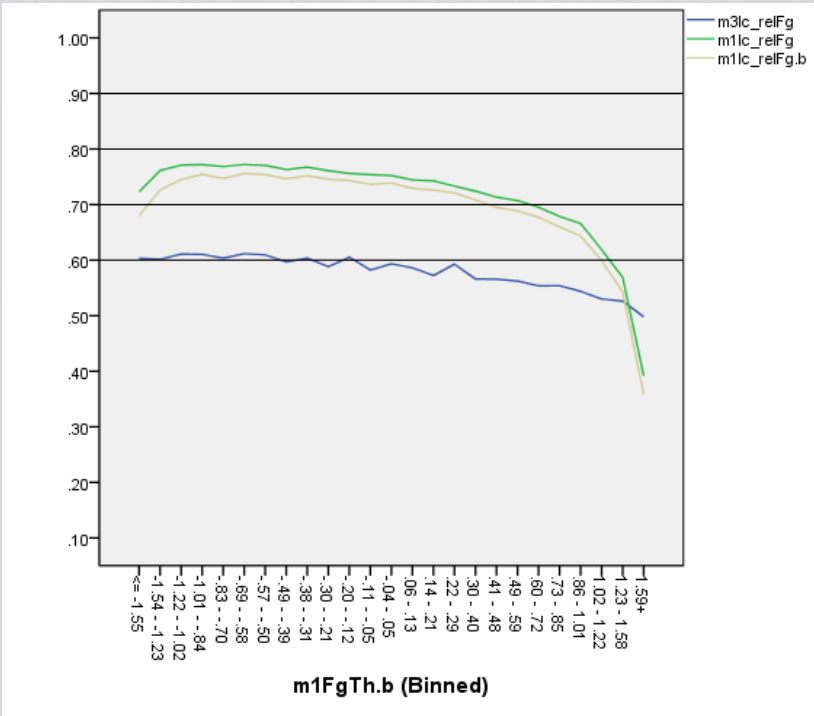
	M	DP	N	*1*	*2*	*3*	*4*	*5*
BIB 1								
1. b 3p	-0,405	1,490	111	1				
2. b 1p	-0,821	1.134	111	**0,973****	1			
3. a	0,824	0,247	111	0,007	-0,061	1		
4. c	0,192	0,011	111	-0,403**	-0,417**	-0,574**	1	
5. Infit	0,996	0,054	111	0,340**	0,364**	-0,848**	0,332**	1
6. Outfit	0,979	0,117	111	0,496**	0,586**	-0,747**	0,145	0,867**
BIB 2								
1. b 3p	-0,487	1,579	118	1				
2. b 1p	-0,834	1,137	118	**0,978****	1			
3. a	0,812	0,246	118	0,050	-0,048	1		
4. c	0,177	0,009	118	-0,544**	-0,554**	-0,549**	1	
5. Infit	0,996	0,057	118	0,367**	0,442**	-0,814**	0,207*	1
6. Outfit	0,983	0,129	118	0,459**	0,576**	-0,658**	-0,034	0,876**

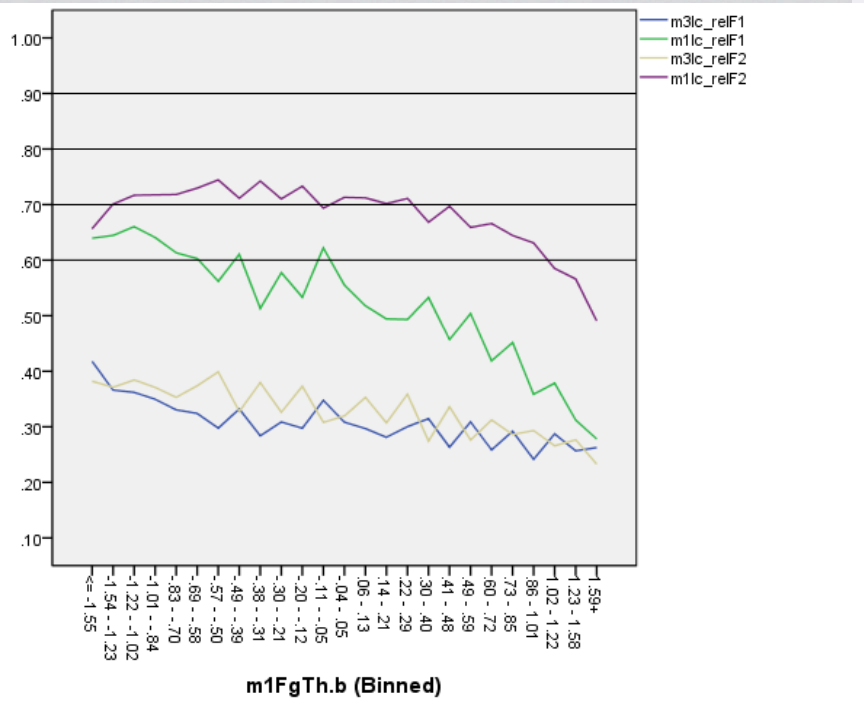

2
0
-2
-4
b1p
-4
-2
0
2
b3p
factor(BIB)
1
2
a
1.6
1.2
0.8

Escores Theta	M	DP	N	m1Fgth	m1F1th	m1F2th
BIB 1						
m3FgTh	0,013	0,879	5463	**0,974****	0,832**	0,781**
m3F1Th	-0,004	0,802	5463	0,849**	**0,947****	0,575**
m3F2Th	0,021	0,737	5463	0,826**	0,588**	**0,936****
m1Fgth	0,005	0,884	5463			
m1F1th	0,109	1,085	5463			
m1F2th	0,003	1,112	5463			
BIB 2						
m3FgTh	0,013	0,883	4769	**0,975****	0,704**	0,917**
m3F1Th	-0,027	0,688	4769	0,716**	**0,941****	0,590**
m3F2Th	0,011	0,860	4769	0,927**	0,575**	**0,967****
m1Fgth	0,004	0,870	4769			
m1F1th	0,356	1.235	4769			
m1F2th	0,011	0,947	4769			


m3FgTh
m1Fgth
m1Fginfit
1.5
1.0
0.5
1
2
2
1
0
-1
-2
-3
-2
0
2

Precisão Local (Curva de informação na escala precisão 0-1)

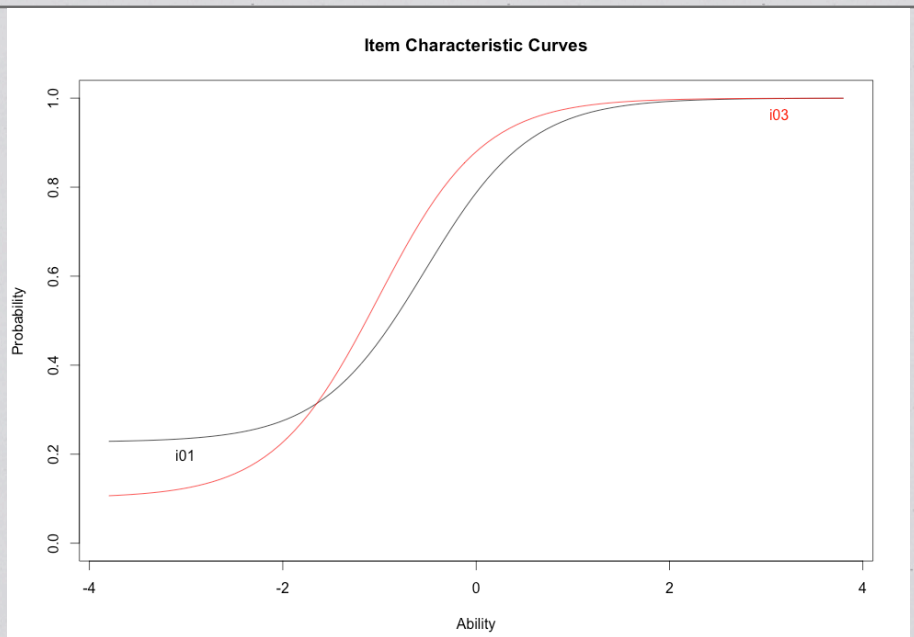


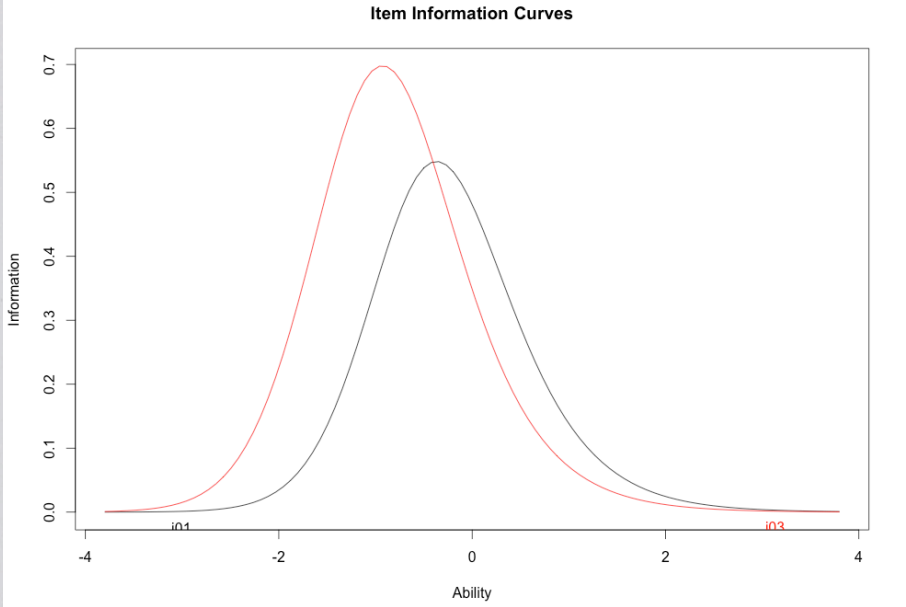


- Lembrança: prova com 36 itens – não é o formato final da Prova Docente!

Conclusões

- Correlação entre as pontuações theta, segundo os dois modelos, que estão altamente correlacionadas de 0,94 a 0,97
- Independente do modelo tem-se pontuações bastante similares para os sujeitos.
  - As variações que se observam nas Figuras 3, 4 e 5 dos thetas calculados segundo o modelo de três parâmetros para um mesmo theta (e consequentemente o mesmo escore total) no modelo de Rasch terá a ver com os parâmetros de discriminação dos itens em causa já que no modelo de Rasch "o escore total é uma estatística suficiente. Assim, uma mesma estimativa de escore theta é recebida independentemente de qual itens o sujeito acertou ou errou. Para o modelo de dois parâmetros, o qual contém itens com diferentes discriminações, a estimativa do nível de escore theta depende de exatamente quais itens se acertou e errou. Acertando-se itens relativamente mais discriminativos leva a estimativas de theta mais altas" (Embretson & Reise, 2006, p.60).
- A variação da estimativa no modelo de três parâmetros está relacionada ao índice infit dos sujeitos. Quanto mais ajustado o padrão de resposta do sujeito mais ele será atraído para a média.
- A precisão pelo modelo de Rasch é superestimada já que não modela o "c"
  - Linacre (2013): "In general, the information in a response that fits the dichotomous Rasch model contains more statistical information than an equivalent response that fits the 3-PL model. This is because of the information that is lost due to the lower asymptote, c parameter. For example, suppose that the c parameter is 0.99, then there is almost no information in the response. If the item-sample targeting is at 70% success and c=.2, we expect a 3-PL response to contain around 67% of the information in a Rasch response"
- Solução: estimar b e theta no Rasch adaptativamente (usando o comando CUTLOW no Winsteps) eliminando itens muito difíceis para estimar as habilidades de pessoas com baixa habilidade (Andrich e cols. 2012).

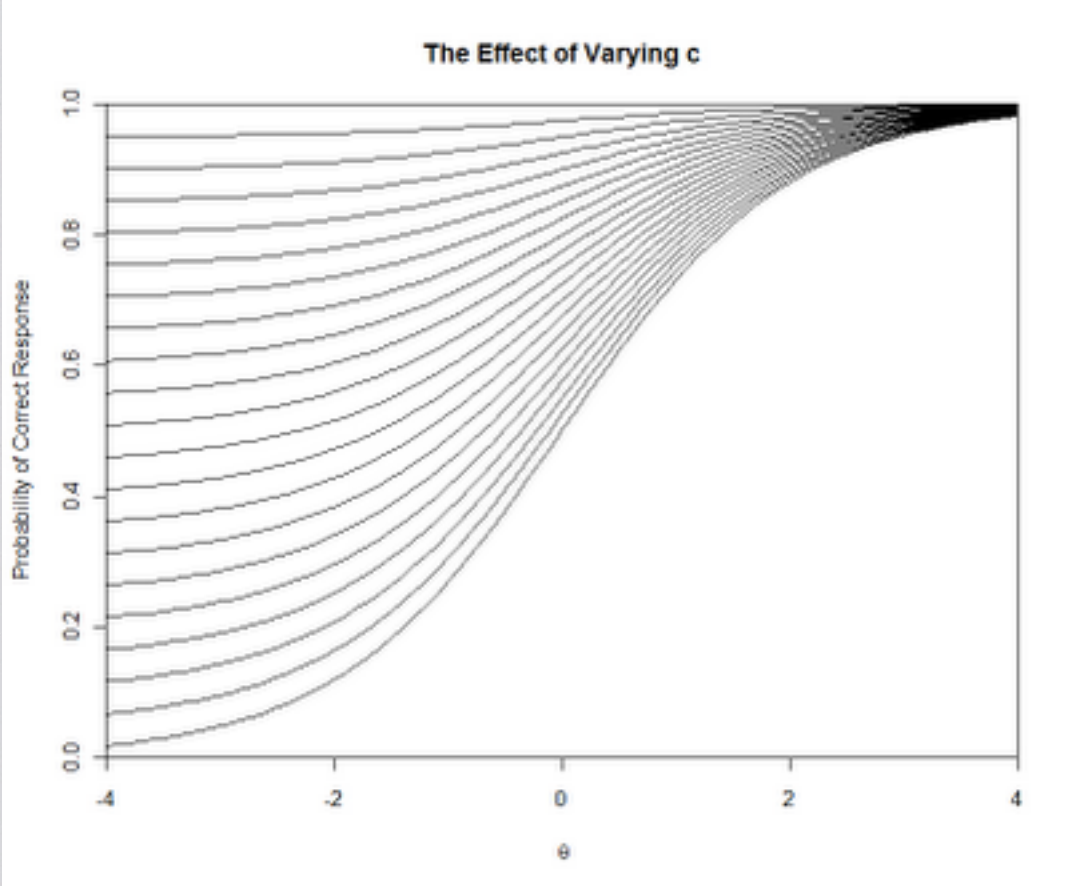




Coefficients:

	Gussng	Dffclt	Dscrmn
i01	0.227	-0.526	1.841
i03	0.101	-1.013	1.843

http://www.econometricsbysimulation.com/2012/09/playing-around-with-irt-graphs.html



Referências

- Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. Em J. Leighton, & M. Gierl, (Eds.). *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85-115). Cambridge University Press.
- Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5(8). Retrieved [date] from http://www.jtla.org.
- Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, 15(1), 136-153.
- Hambleton, H. K., & Swaminatham, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer.