Minicurso

MC-12 - AVALIAÇÃO EDUCACIONAL: ENTENDENDO A TEORIA DA RESPOSTA AO ITEM (ABE)

Ministrantes: Jorge Luis Bazán (USP) e Mariana Curi (USP)
jlbazan@icmc.usp.br, mcuri@icmc.usp.br

Público alvo: Professores do ensino básico

Sala: *AT 04 - Sala 82*
De 14/7/2015 à 17/7/2015 - das 08h00 às 10h00

Ementa:

Avaliações são importantes para a análise da qualidade educacional e do impacto de políticas públicas nacionais e estaduais nos diferentes níveis educacionais. O propósito deste minicurso é apresentar as características das avaliações, o processo de elaboração de instrumentos e de questões e introduzir os conceitos dos modelos adotados, enfatizando os modelos da Teoria da Resposta ao Item. Adicionalmente, a interpretação dos modelos da TRI será mostrada em outros contextos além do âmbito educacional. Combinando metodologias expositivas e oficina, o minicurso tem com resultado a formação de indivíduos com capacidade crítica para o entendimento dos resultados que se divulgam usando TRI.

AULA 1. AVALIAÇÃO, MEDIÇÃO E PSICOMETRIA (Expositiva)

Conteúdo: Avaliação educacional e em outras áreas. As avaliações educacionais em larga escala. Medidas de avaliação. O que é Psicometria? Avaliação de normas ou de critérios. Principais métodos de medição: Teoria Clássica de Testes (TCT) e Teoria de Resposta ao Item (TRI).

Ministrante: Jorge Luís Bazán

AULA 2. ELABORAÇÃO DE QUESTÕES OU ITENS E SUA ANÁLISE USANDO METODOLOGIA TRADICIONAL (Oficina)

Conteúdo. *Principais medidas e critérios para a interpretação dos resultados.* Matrizes de referência. Tipos de itens. Recomendações para elaboração de itens. Exemplos. Oficina de redação de itens. Análise clássica de itens (qualitativa e quantitativa) baseada na TCT. Respostas correta, não resposta e valores perdidos. Identificação de viés. Principais medidas e critérios para a interpretação dos resultados.

Ministrante: Jorge Luís Bazán

AULA 3. INTRODUÇÃO À TEORIA DA RESPOSTA AO ITEM (Expositiva)

Conteúdo. Limitações da TCT. A TRI e seus modelos. Vantagens e desvantagens. Interpretação do principal modelo da TRI. Suposições e obtenção dos resultados. Exemplos de uso da TRI nas avaliações (educacionais e outras). Futuro da avaliação: testes adaptativos informatizados.

Ministrante: Mariana Curi

AULA 4. ANALISANDO ITENS USANDO TEORIA DA RESPOSTA AO ITEM (Oficina)

Conteúdo. Interpretação dos parâmetros de item no modelo da TRI. A Escala do traço latente medido. Comparação dos resultados da TRI e análise clássica. Análise de resultados de avaliações nacionais e internacionais usando TRI: SAEB, PISA, ENEM, TERCE.

Material complementar

Jorge Luis Bazán (USP)
http://www.icmc.usp.br/pessoas/jlbazan/

Mariana Curi (USP)
http://www.icmc.usp.br/pessoas/mcuri/

Minicurso

MC-12 - AVALIAÇÃO EDUCACIONAL: ENTENDENDO A TEORIA DA RESPOSTA AO ITEM (ABE)

Ministrantes: Jorge Luis Bazán (USP) e Mariana Curi (USP)
jlbazan@icmc.usp.br, mcuri@icmc.usp.br

Público alvo: Professores do ensino básico

Sala: *AT 04 - Sala 82*
De 14/7/2015 à 17/7/2015 - das 08h00 às 10h00

AULA 1. AVALIAÇÃO, MEDIÇÃO E PSICOMETRIA (Expositiva)

Conteúdo:
Avaliação educacional e em outras áreas. As avaliações educacionais em larga escala. Medidas de avaliação. O que é Psicometria?. Principais métodos de medição: Teoria Clássica de Testes (TCT) e Teoria de Resposta ao Item (TRI).

Ministrante: Jorge Luís Bazán

TÓPICOS

1. AVALIAÇÃO
2. PSICOMETRIA
3. MODELOS DE MEDICAO
4. MODELOS DE TESTES CLASSICOS O TEORIA CLASSICA DOS TESTES
5. MODELOS DE RESPOSTA AO ITEM
6. TCT VS TRI

1. AVALIAÇÃO

Avaliação é importante para

- Melhora dos processos dos governos federais, estaduais e municipais
- Processos da administração e governabilidade
- Melhora da eficiência e eficácia (serviços, setores produtivos, mercados, competitividade)
- Melhora na qualidade do sistema educacional
- Melhores critérios de seleção e avaliação de estudantes e profissionais

A avaliação tem a ver com:

- Desenvolvimento de instrumentos (escalas, questionários e suas propriedades)

- Melhor definição dos propósitos de pesquisa (objetivos e resultados)

- Melhores modelos matemáticos e estatísticos

- Melhores sistemas de computo para bases de dados, analises e aplicações de provas

- Desenvolvimento de critérios para propor políticas usando os resultados da avaliação

Isto induz ao desenvolvimento metodológico da avaliação e a um debate ao respeito de sua aplicabilidade.

No Brasil, poucos conhecem os aspectos técnicos da avaliação.

O que é avaliação educacional?

O processo de avaliação educacional está relacionado à produção de informações sobre o aprendizado. Isto é algo que está bastante presente no cotidiano escolar e na educação superior: usualmente, os professores aferem o aprendizado dos seus alunos através de diversos instrumentos (observações, questionários, escalas, listas, registros, provas etc.) e indicam, a partir daí, o que precisa ser feito para que seus alunos possam avançar no sistema escolar.

O que é avaliação em larga escala?

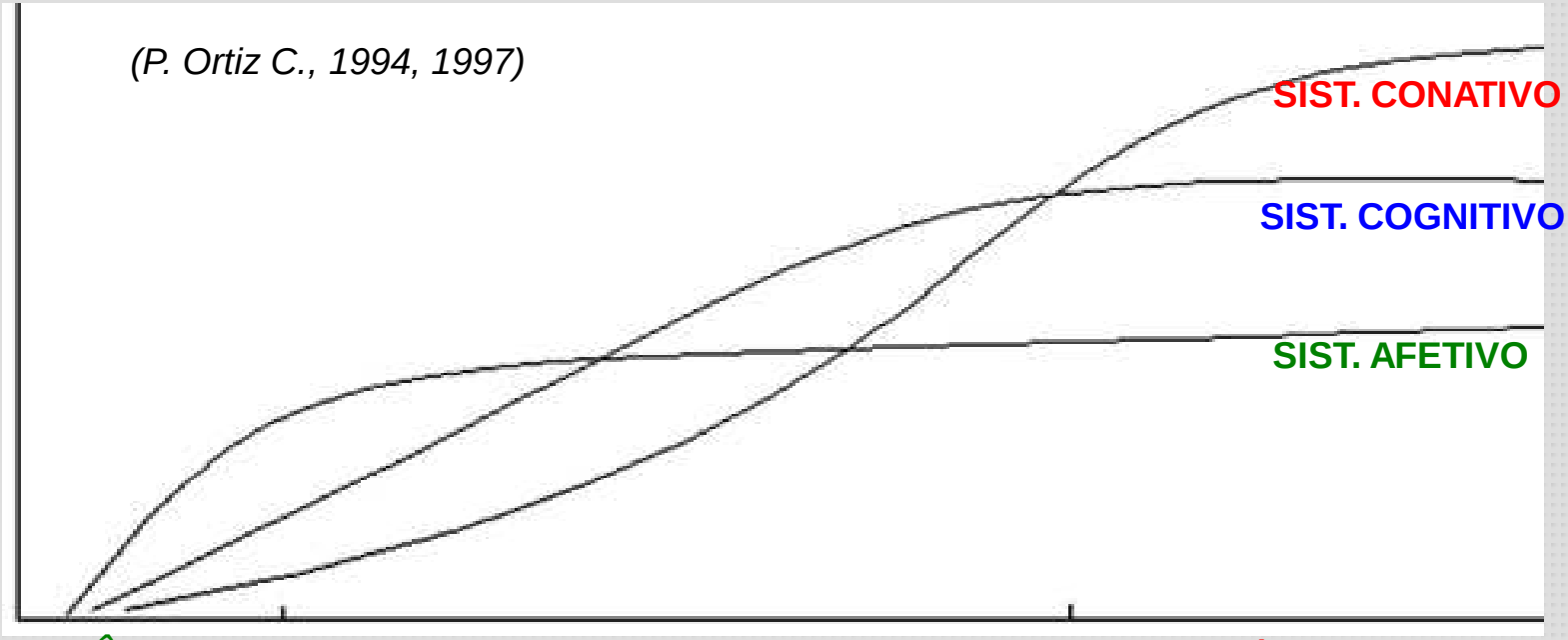
Nas últimas décadas, junto com às avaliações tradicionais na salas de aula, outro tipo de avaliação educacional tem ganhado espaço: são as avaliações externas, geralmente em larga escala, isto é, aplicada simultaneamente a grandes amostras ou censos em forma padronizada incluindo as vezes alunos professores, diretores e coordenadores. Exemplos ENEM, SAEM, PISA.

Estas avaliações têm objetivos e procedimentos diferenciados das avaliações tradicionais de salas de aula. Por exemplo para certificação, o credenciamento, o diagnóstico e a rendição de contas.

O que se avalia?

Em geral as avaliações individuais privilegiam o sistema de avaliação cognitiva não em tanto outros sistemas podem ser avaliados embora eles sejam menos discutidos.

O DESENVOLVIMENTO DA CONSCIÊNCIA E A PERONALIDADE REFLETE A HISTORIA DA



INFÂNCIA 1:

OS SENTIMENTOS REFLETEM A ESTRUTURA TRADICIONAL

INFÂNCIA 2:

OS CONHECIMENTOS REFLETEM A ESTRUTURA CULTURAL

ADOLESCÊNCIA:

AS MOTIVAÇÕES REFLETEM A ESTRUTURA ECONÓMICA

2. PSICOMETRIA

Contexto

	PSICOLOGIA	
ESTATÍSTICA	Profissão A	Ciência B
Metodologias para I	Enfoque Quantitativo	Paradigma quantitativo
Profissão II	Consultoria Estatística	Psicometria
Ciência	Novos paradigmas quantitativos	Psicologia Matemática

Aplicações da Estatística em Psicologia

O que é Psicometria?

- É o campo de estudo relacionado com a teoria e técnica da medição psicológica, incluindo a medição de conhecimentos, habilidades, atitudes e traços de personalidade e a medição educacional.

- A área está principalmente associada com a construção e validação de instrumentos de medição, como questionários, provas, escalas, inventários e testes, entre outros.

- A psicometría tem dois tarefas de pesquisa principais:

(i) A construção de instrumentos e procedimentos de medição, e

(Ii) o desenvolvimento e aperfeiçoamento de abordagens teóricas para a medição.

- Profissionais da psicologia, psicometristas são científicos envolvidos no planejamento do teste para tentar medir diferentes características humanas.

- A área sofreu um rápido crescimento desde a sua criação. Os testes psicométricos são utilizados em escolas, organizações, empresas, governos, forças armadas, e, claro, em ambientes hospitalares e clínicos.

- Todos os expertos em psicometría devem ter pelo menos um Maestría, e a maioria tem Doutorado.

- Por causa de que a Psicometría é considerada uma área da psicologia, uma licenciatura em Psicologia não é incomum como formação previa.

- Os graduados em Psicometria costumam trabalhar nos departamentos de Psicologia, mas não e infrequente encontrar muitos especialistas tem uma graduação em Estatística.

- De acordo com um recente artigo no Journal Washington Monthly, psicometristas (muitas vezes chamado de "test makers") estão em grande demanda.

- Cada vez são mais requeridos testes, e não há especialistas suficientes em psicometría para atender a demanda.

- Em uma sociedade com uma cultura de medição, qualquer psicólogo especialista em psicometria não deve ter dificuldade em encontrar emprego.

Como é a construção de um instrumento psicométrico?

"Brincando com a altura [*]

(*) Prof CAW Glas - University of Twente - Holanda

ABE - SINAPE 2006.

Questionário para medir altura: alguns itens

- 1. Na cama, eu frequentemente sinto frio nos pés.
- 2. Eu frequentemente desço as escadas de dois em dois degraus.
- 3. Eu acho que me daria bem em um time de basquete.
- 4. Como policial, eu impressionaria muito.
- 5. Na maioria dos carros eu me sinto desconfortável.
- 6. Eu literalmente olho para meus colegas de cima para baixo.
- 7. Você é capaz de pegar um objeto no alto de um armário, sem usar escada?
- 8. Você abaixa quando vai passar por uma porta?

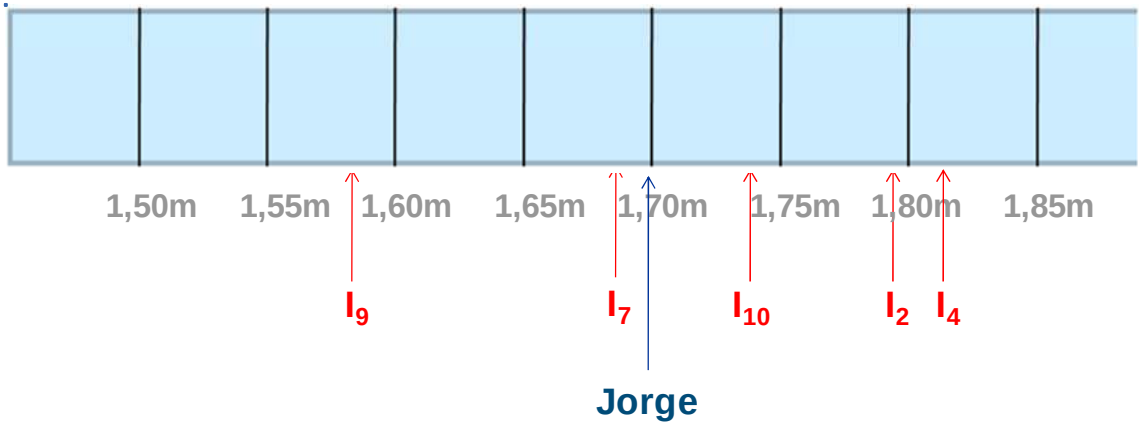
- 9. Você consegue guardar a bagagem no porta-malas do avião?
- 10. Você regulava o banco do carro para trás?
- 11. Normalmente quando você está andando de carona lhe oferecem o banco

  da frente?
- 12. Quando você e várias pessoas vão tirar fotos, formando-se três fileiras, onde ninguém ficará agachado, você costuma ficar atrás?
- 13. Você tem dificuldade para se acomodar no ônibus?
- 14. Em uma fila, por ordem de tamanho, você é sempre colocado atrás?

Formatos itens:

Dicotômica: Sim - Não, verdadeiro ou falso, certo ou errado.
Politômicos: nunca, raramente, a metade do tempo, muitas vezes, sempre.

Posição de examinados e itens na mesma escala



O que mede um teste?

- Um teste ou medida pode ser visto com um conjunto de questões de auto-relato (também chamado de "itens"), cujas respostas são pontuadas e de alguma forma agregadas para obter uma pontuação composta.

- As características essenciais são:
  - Uma série de perguntas as quais os indivíduos respondem
  - Um escore composto que surge a partir da pontuação das respostas para as perguntas.

- O conjunto resultante de perguntas é referido como uma "escala", "teste" ou "medida". Em geral, um instrumento psicométrico.

Dois tipos de resultados estão disponíveis a partir dos itens, mas precisasse notar que o importante não é tanto o formato da pergunta se não o formato da resposta ou pontuação.

Pontuações binárias (resposta dicotômicas), (a) os itens que estão qualificados como resposta *correta* ou *incorreta* em teste de rendimento(por exemplo, no caso de múltipla escolha), ou (b) itens que são classificados dicotomicamente de acordo com um tipo de pontuação, ou escala de personalidade (ie, *verdadeiro - falso, de acordo com - em desacordo).*

Respostas a Item ordinais (respostas graduadas, Likert, tipo Likert, ou item politomos); envolvendo mais de duas opções de pontuação tais, como uma escala de 5 pontos, *em total acordo* até *em total discordo* ou em uma escala de personalidade ou medida atitude.

Como é determinada a qualidade dos instrumentos psicométricos?

- As considerações de validade e confiabilidade dos instrumentos psicométricos pelo geral são vistos como elementos essenciais para determinar a qualidade de qualquer teste.

- Associações Profissionais e usuários muitas vezes têm estas preocupações dentro de contextos mais amplos no desenvolvimento de critérios para avaliar a qualidade de qualquer teste num determinado contexto.

- ***The Standards for Educational and Psychological Testing*** **(1999)** é um conjunto de criterios de avaliação desenvolvidos pela American Educational Research Association (AERA), American Psychological Association (APA), e o **National Council on Measurement in Education (NCME).**.

Parte I: Construção de Testes, Avaliação e Documentação

1. Validade
2. Erros de medida e confiabilidade
3. Desenvolvimento de teste e revisão.
4. Escalas, Normas e comparabilidade dos escores
5. Administração de teste, Qualificação e Relatórios
6. Documentação de apoio para o testes

Parte II: Equidade dos Testes

7. Teste de equidade e uso do teste
8. Os direitos e as responsabilidades dos examinadores.
9. Testes individuais de pessoas de diversa procedência linguística
10. Teste individuais para pessoas deficientes

Parte III: Aplicações de teste

11. As responsabilidades de usuários de teste
12. Avaliação e Medição Psicológica
13. Avaliação e Medição Educacional
14. Avaliação e Certificação do trabalho
15. Teste de Avaliação de Programas e Políticas Públicas

- No pais utilizasse as normas de 1954 e não as de 1999.

- Há uma necessidade de reformulação de disciplinas em Estatística e Psicologia, por exemplo de Teoria da Resposta ao Item, Variáveis latentes, Estatísticas na Psicologia, Medição Psicológica, Construção de Testes, Psicometria, etc

3. MODELOS DE MEDIÇÃO

Quais são os princípios de medição?

- Se você quiser medir o quanto de habilidade uma pessoa, tem, você deve ter uma escala de medição, ou seja, uma regra com uma métrica.

- Esta regra deve ser utilizada para determinar que capacidade uma determinada pessoa tem.

- A aproximação habitual é definir uma medida da capacidade e desenvolver um teste que consiste num determinado numero de itens sob a definição (perguntas).

- Cada um desses itens mede alguma faceta de uma particular habilidade de interesse.

- Assumisse que cada examinado que responde a um item de um teste tem certa quantidade da capacidade subjacente.

- Assim, podemos considerar que cada examinando tem um valor numérico, denotado por θ que toma o lugar da sua posição na escala de habilidade.

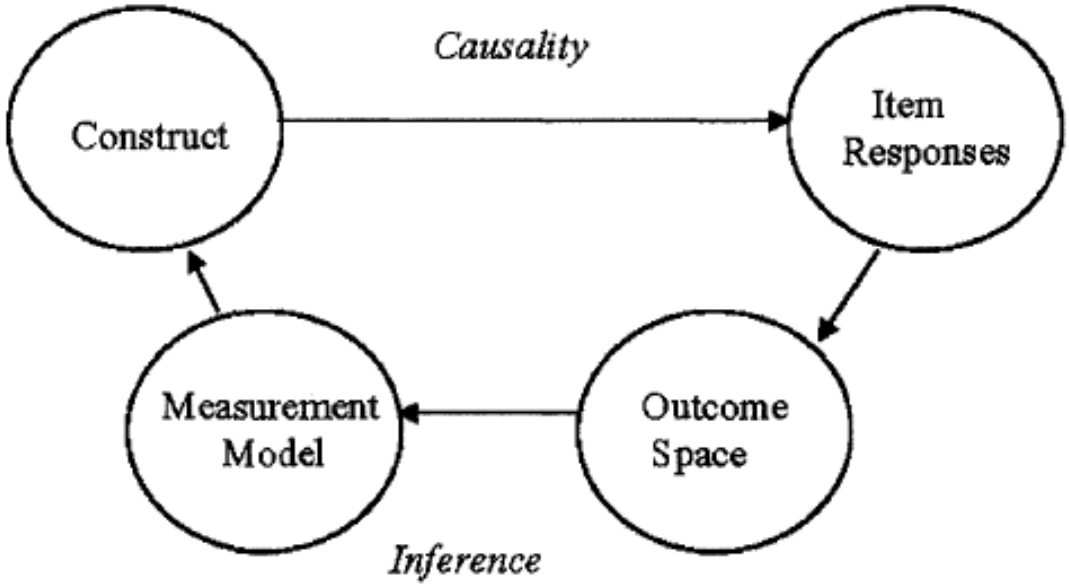


FIG. 1.8 The "four building blocks" showing the directions of causality and inference.

Wilson, Mark (2005). Constructing Measures: An Item Response Modeling Approach. Lawerence Erlbaum Associates, Inc Publisher.

http://www-gse.berkeley.edu/faculty/MWilson/MWilson.html

- *Modelamento do constructo* é uma estratégia para o desenvolvimento de um instrumento usando cada um dos quatro tijolos da construção:
    - Mapa do constructo
    - Plano para o desenvolvimento dos itens
    - Espaço dos resultados
    - Modelo de medição
- Há uma necessidade de reformulação de disciplinas em Estatística em Medicina ou Psicologia por exemplo incluindo tópicos de Teoria da Resposta ao Item, Variáveis latentes, Estatísticas na Psicologia, Medição Psicológica, Construção de Testes, Psicometria, etc

O que é o modelo de medição?

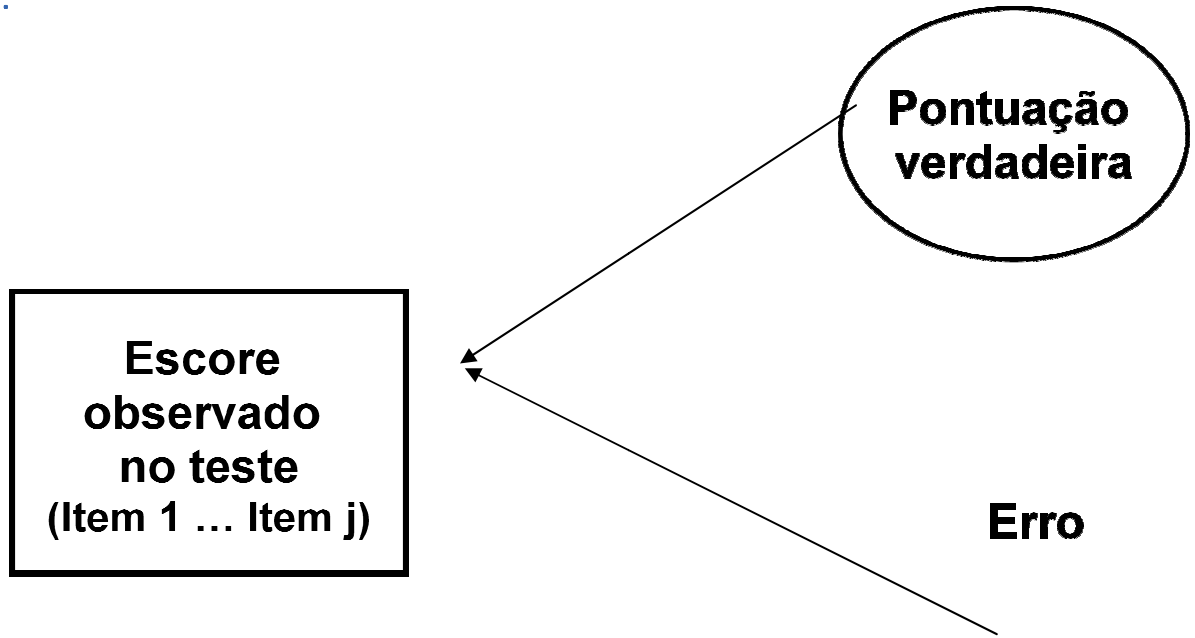
Utilizado para relacionar as variáveis observadas, registradas e medidas (respostas aos itens) com as variáveis latentes (habilidade).

- Modelo de Teoria Clássica
- Modelo de Resposta ao Item

Eles não são as únicas, mas eles são as mais consolidadas

4. MODELOS DE TESTES CLASICOS OU TEORICA CLASSICA DOS TESTES

Ela expressa uma relação linear entre o verdadeiro valor de habilidade e o escore de habilidade observado.



Escore observado = pontuação verdadeira + erro

- O resultado do teste o escore de linha é a soma das pontuações recebidas sobre os itens do teste.
- Tradicionalmente, a teoria da medição foi estabelecida baseado num analise de escala- ou de nível do teste baseado em métodos de correlação.

- Os resultados são, é claro, não segmentados (ou seja, você não tem ideia de como uma pessoa com determinado valor no teste executa a um nível particular de habilidade), há uma única e simples medida geral do desempenho.

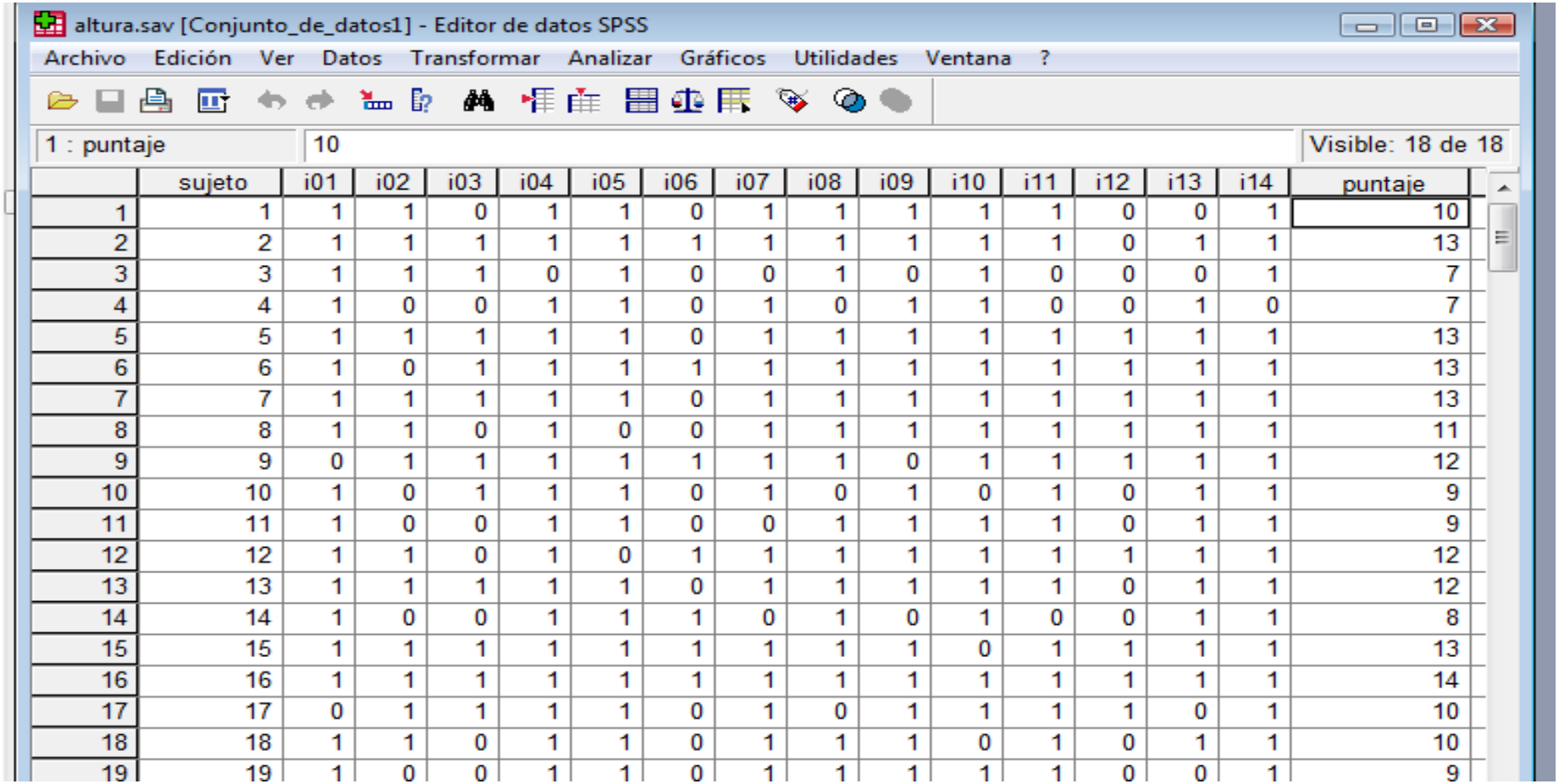
- A principal ferramenta estatística é o ANOVA dos efeitos aleatórios, ou análise de componentes de variância, cujo principal objetivo é medir a quantidade de erro na medida.

- Um conjunto de índices que fazem parte dos analises de itens como proporção de acerto, porcentagem de omissão, discriminação, correlação pergunta-prova, alfa de cronbanch se o item é desconsiderado entre outros incluindo média e variância são comunmente estudados.

- Uma medida general de consistencia interna da prova baseado no Alfa de Cronbach é visto como uma medida apropriada neste contexto.

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\left(\sum_{i=1}^{k} \sigma_i \rho_{i\,X}\right)^2}\right]$$

- Esta teoria é valida para qualquer formato de pontuação dos itens. É aplicado tanto para itens dicotômicos quanto para itens politômicos o qualquer subtipo.


altura.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

1 : puntaje 10 Visible: 18 de 18

	sujeto	i01	i02	i03	i04	i05	i06	i07	i08	i09	i10	i11	i12	i13	i14	puntaje
1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	1	10
2	2	1	1	1	1	1	1	1	1	1	1	1	0	1	1	13
3	3	1	1	1	0	1	0	0	1	0	1	0	0	0	1	7
4	4	1	0	0	1	1	0	1	0	1	1	0	0	1	0	7
5	5	1	1	1	1	1	0	1	1	1	1	1	1	1	1	13
6	6	1	0	1	1	1	1	1	1	1	1	1	1	1	1	13
7	7	1	1	1	1	1	0	1	1	1	1	1	1	1	1	13
8	8	1	1	0	1	0	0	1	1	1	1	1	1	1	1	11
9	9	0	1	1	1	1	1	1	1	0	1	1	1	1	1	12
10	10	1	0	1	1	1	0	1	0	1	0	1	0	1	1	9
11	11	1	0	0	1	1	0	0	1	1	1	1	0	1	1	9
12	12	1	1	0	1	0	1	1	1	1	1	1	1	1	1	12
13	13	1	1	1	1	1	0	1	1	1	1	1	0	1	1	12
14	14	1	0	0	1	1	1	0	1	0	1	0	0	1	1	8
15	15	1	1	1	1	1	1	1	1	1	0	1	1	1	1	13
16	16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
17	17	0	1	1	1	1	0	1	0	1	1	1	1	0	1	10
18	18	1	1	0	1	1	0	1	1	1	0	1	0	1	1	10
19	19	1	0	0	1	1	0	1	1	1	1	1	0	0	1	9

Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se eleimina el elemento
i01	10.05	3.090	.143	.471
i02	9.98	3.261	.057	.490
i03	10.32	2.804	.240	.442
i04	9.91	3.161	.259	.450
i05	9.97	3.168	.147	.468
i06	10.47	2.990	.141	.475
i07	9.92	3.170	.230	.454
i08	9.96	2.975	.331	.426
i09	10.05	3.236	.036	.500
i10	9.98	3.130	.171	.463
i11	9.91	3.053	.384	.428
i12	10.49	2.929	.183	.461
i13	10.02	3.092	.164	.464
i14	9.89	3.358	.060	.483

Estadísticos de fiabilidad

Alfa de Cronbach	N de elementos
.481	14

Recursos - TCT

Software

- Pacotes estatisticos (Excel, SPSS, SAS)
- ITEMAN (disponível a partir de http://www.assess.com/xcart/home.php?cat=18)

Leitura

Matlock-Hetzel (1997) *Basic Concepts in Item and Test Analysis* available at www.ericae.net/ft/tamu/Espy.htm

5. MODELOS DE RESPUESTA AL ITEM

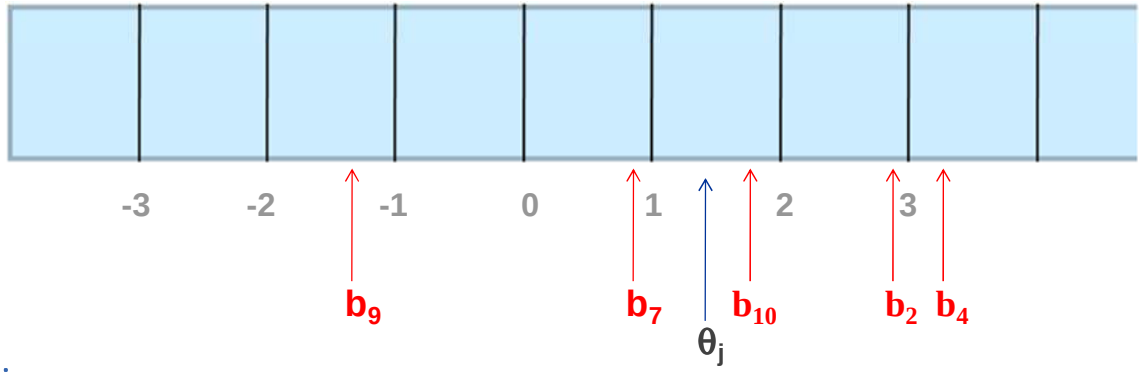
De acordo com a chamada Teoria de Resposta ao Item (TRI), o interesse primário esta em saber se o examinando tem um determinado item correto ou não, ao invés de saber a pontuação total

- Especifica como o traço latente e as características do item estão relacionados com as respostas das pessoas aos itens.

- Modelo mas simples: modelo Rasch

P (resposta correta item) = função { nível de habilidade, dificuldade do item}

Posição de examinados e itens numa mesma escala



θ_j: traço latente do examinando (parâmetro da pessoa: "habilidade")

b_i: dificuldade do item "traço latente" (parâmetro do item)

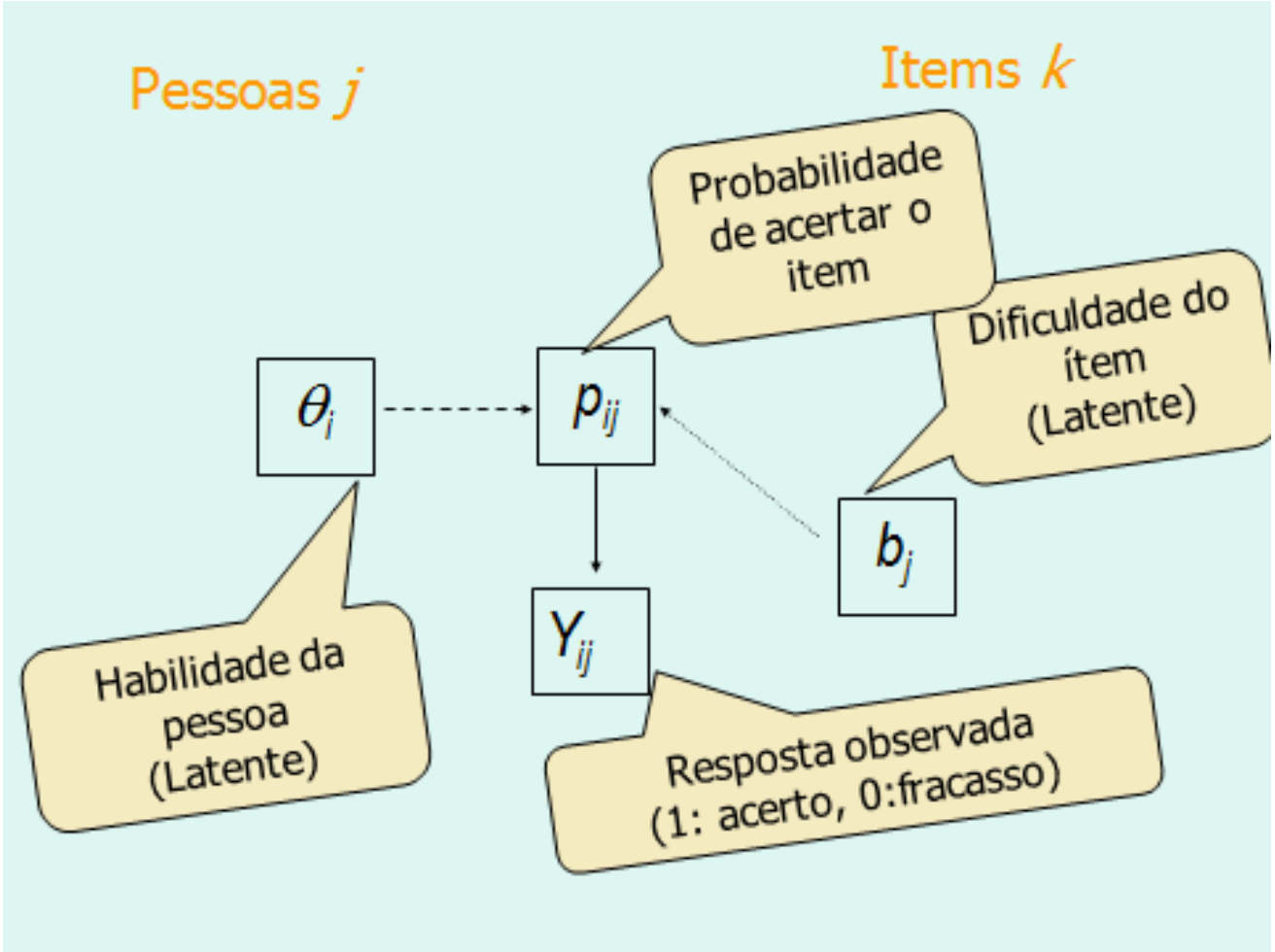
> 0 examinado está "acima" do item

$(\theta_j - b_i) \approx 0$ é considerado" próximo "do item

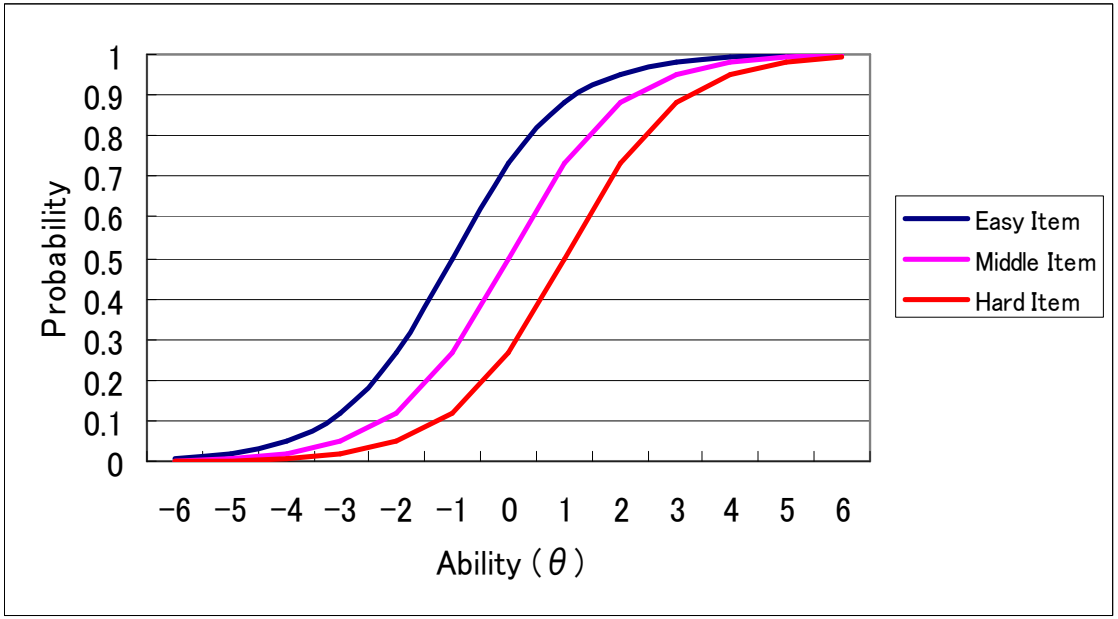
<0 é considerado "abaixo" do item

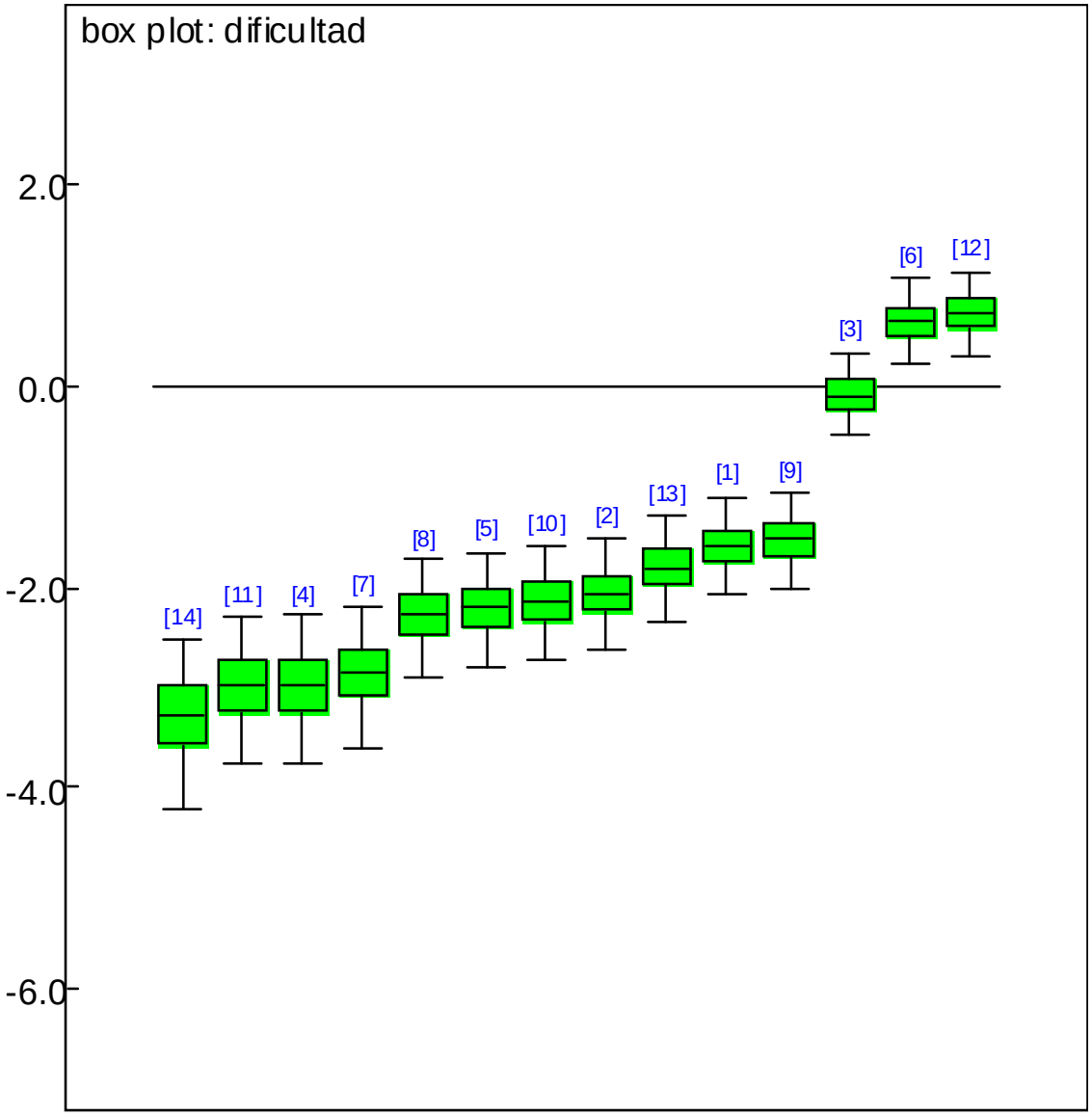
- Baseado nas respostas dos itens de um teste desejasse estimar:
  - parâmetros dos Itens (calibração)
  - Traços latentes dos examinados
  - Parâmetros da população (distribuição dos traços latentes): média, desvio padrão, etc

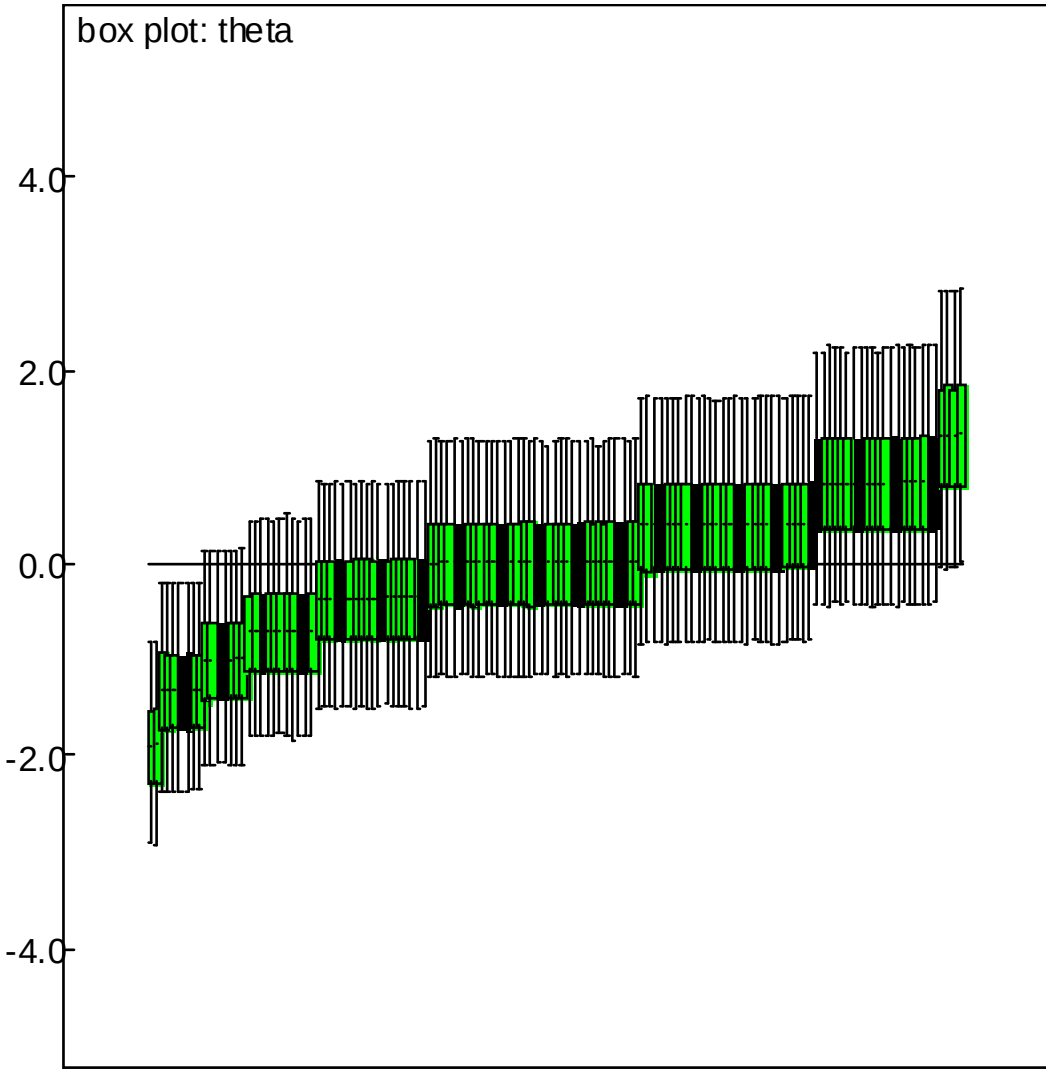
- A probabilidade de uma resposta "correta" para um item é modelada como função da habilidade do examinando e os parâmetros do item.


Pessoas j
Items k
Probabilidade de acertar o item
Dificuldade do ítem (Latente)
θ_i
p_{ij}
b_j
Y_{ij}
Habilidade da pessoa (Latente)
Resposta observada (1: acerto, 0:fracasso)

Curvas características dos itens




box plot: dificultad
2.0
0.0
-2.0
-4.0
-6.0
[14]
[11]
[4]
[7]
[8]
[5]
[10]
[2]
[13]
[1]
[9]
[3]
[6]
[12]


box plot: theta
4.0
2.0
0.0
-2.0
-4.0

Software Psicometrico

- IRTPRO
- Winstep
- Rascal
- Bilog
- Conquest
- Quest
- Winmira
- RUMM2020
- Param3PL
- Logimo
- MSP
- LPCM-WIN
- RSP

- T-Rasch
- ICL-WIN
- LEM
- Multilog
- Xcalibret

Estatísticos

- SAS
- R
- Stata
- WinBUGS
- Systat
- OpenStat

Pacotes no R

http://cran.r-project.org/web/views/Psychometrics.html

Classical Test Theory (CTT):

- The CTT package can be used to perform a variety of tasks and analyses associated with classical test theory: score multiple-choice responses, perform reliability analyses, conduct item analyses, and transform scores onto different scales.
- Functions for correlation theory, meta-analysis (validity generalization), reliability, item analysis, inter-rater reliability, and classical utility are contained in the psychometric package.
- The cocron package provides functions to statistically compare two or more alpha coefficients based on either dependent or independent groups of individuals.
- The CMC package calculates and plots the step-by-step Cronbach-Mesbach curve, that is a method, based on the Cronbach alpha coefficient of reliability, for checking the unidimensionality of a measurement scale.
- Cronbach alpha, kappa coefficients, and intra-class correlation coefficients (ICC) can be found in the psy package. Functions for ICC computation can be also found in the packages psych, psychometric and ICC.
- A number of routines for scale construction and reliability analysis useful for personality and experimental psychology are contained in the package psych.
- QME (not on CRAN) computes measures from generalizability theory.

Item Response Theory (IRT):

- The eRm package fits extended Rasch models, i.e. the ordinary Rasch model for dichotomous data (RM), the linear logistic test model (LLTM), the rating scale model (RSM) and its linear extension (LRSM), the partial credit model (PCM) and its linear extension (LPCM) using conditional ML estimation. Missing values are allowed.
- The package ltm also fits the simple RM. Additionally, functions for estimating Birnbaum's 2- and 3-parameter models based on a marginal ML approach are implemented as well as the graded response model for polytomous data, and the linear multidimensional logistic model.
- TAM fits unidimensional and multidimensional item response models and also includes multifaceted models, latent regression models and options for drawing plausible values.
- The mirt allows for the analysis of dichotomous and polytomous response data using unidimensional and multidimensional latent trait models under the IRT paradigm. Exploratory and confirmatory models can be estimated with quadrature (EM) or stochastic (MHRM) methods. Confirmatory bi-factor and two-tier analyses are available for modeling item testlets. Multiple group analysis and mixed effects designs also are available for detecting differential item functioning and modelling item and person covariates.
- IRTShiny provides an interactive shiny application for IRT analysis.
- The mcIRT package provides functions to estimate the Nominal Response Model and the Nested Logit Model. Both are models to examine multiple-choice items and other polytomous response formats. Some additional uni- and multidimensional item response models (especially for locally dependent item responses) and some exploratory methods (DETECT, LSDM, model-based reliability) are included in sirt.

- The pcIRT estimates the multidimensional polytomous Rasch model and the Mueller's continuous rating scale model.
- Thurstonian IRT models can be fitted with the kcirt package.
- MultiLCIRT estimates IRT models under (1) multidimensionality assumption, (2) discreteness of latent traits, (3) binary and ordinal polytomous items.
- Conditional maximum likelihood estimation via the EM algorithm and information-criterion-based model selection in binary mixed Rasch models are implemented in the mRm package and the psychomix package. The mixRasch package estimates mixture Rasch models, including the dichotomous Rasch model, the rating scale model, and the partial credit model.
- The PP package includes estimation of (MLE, WLE, MAP, EAP, ROBUST) person parameters for the 1,2,3,4-PL model and the GPCM (generalized partial credit model). The parameters are estimated under the assumption that the item parameters are known and fixed. The package is useful e.g. in the case that items from an item pool/item bank with known item parameters are administered to a new population of test-takers and an ability estimation for every test-taker is needed.
- The equateIRT package computes direct, chain and average (bisector) equating coefficients with standard errors using Item Response Theory (IRT) methods for dichotomous items.
- kequate implements the kernel method of test equating using the CB, EG, SG, NEAT CE/PSE and NEC designs, supporting gaussian, logistic and uniform kernels and unsmoothed and pre-smoothed input data.
- SNSequate provides several methods for test equating. Besides of traditional approaches (mean-mean, mean-sigma, Haebara and Stocking-Lord IRT, etc.) it supports methods such that local equating, kernel equating (using Gaussian, logistic and uniform kernels), and IRT parameter linking methods based on asymmetric item characteristic functions including functions for obtaining standard errors.

- The EstCRM package calibrates the parameters for Samejima's Continuous IRT Model via EM algorithm and Maximum Likelihood. It allows to compute item fit residual statistics, to draw empirical 3D item category response curves, to draw theoretical 3D item category response curves, and to generate data under the CRM for simulation studies.
- The difR package contains several traditional methods to detect DIF in dichotomously scored items. Both uniform and non-uniform DIF effects can be detected, with methods relying upon item response models or not. Some methods deal with more than one focal group.
- The package lordif provides a logistic regression framework for detecting various types of differential item functioning (DIF).
- DIFlasso implements a penalty approach to Differential Item Functioning in Rasch Models. It can handle settings with multiple (metric) covariates.
- A set of functions to perform Raju, van der Linden and Fleer's (1995) Differential Item and Item Functioning analyses is implemented in the DFIT package. It includes functions to use the Monte Carlo Item Parameter Replication (IPR) approach for obtaining the associated statistical significance tests cut-off points.
- The catR package allows for computarized adaptive testing using IRT methods.
- The mirtCAT package provides tools to generate an HTML interface for creating adaptive and non-adaptive educational and psychological tests using the shiny package. Suitable for applying unidimensional and multidimensional computerized adaptive tests using IRT methodology and for creating simple questionnaires forms to collect response data directly in R.
- The package plRasch computes maximum likelihood estimates and pseudo-likelihood estimates of parameters of Rasch models for polytomous (or dichotomous) items and multiple (or single) latent traits. Robust standard errors for the pseudo-likelihood estimates are also computed.

- Explicit calculation (not estimation) of Rasch item parameters (dichotomous and polytomous) by means of a pairwise comparison approach can be done using the pairwise package.
- A multilevel Rasch model can be estimated using the package lme4, nlme, and MCMCglmm with functions for mixed-effects models with crossed or partially crossed random effects. The ordinal package implements this approach for polytomous models. An infrastructure for estimating tree-structured item response models of the GLMM family using lme4 is provided in irtrees.
- Nonparametric IRT analysis can be computed by means if the mokken package. It includes an automated item selection algorithm, and various checks of model assumptions. In relation to that, fwdmsa performs the Forward Search for Mokken scale analysis. It detects outliers, it produces several types of diagnostic plots.
- This KernSmoothIRT package fits nonparametric item and option characteristic curves using kernel smoothing. It allows for optimal selection of the smoothing bandwidth using cross-validation and a variety of exploratory plotting tools.
- The RaschSampler allows the construction of exact Rasch model tests by generating random zero-one matrices with given marginals.
- The irtProb package is designed to estimate multidimensional subject parameters (MLE and MAP) such as personnal pseudo-guessing, personal fluctuation, personal inattention. These supplemental parameters can be used to assess person fit, to identify misfit type, to generate misfitting response patterns, or to make correction while estimating the proficiency level considering potential misfit at the same time.
- cacIRT computes classification accuracy and consistency under Item Response Theory. Currently, only works for 3PL IRT models (or 2PL or 1PL) and only for independent cut scores.

- The package irtoys provides a simple common interface to the estimation of item parameters in IRT models for binary responses with three different programs (ICL, BILOG-MG, and ltm, and a variety of functions useful with IRT models.
- The CDM estimates several cognitive diagnosis models (DINA, DINO, GDINA, RRUM, LCDM, pGDINA, mcDINA), the general diagnostic model (GDM) and structured latent class analysis (SLCA).
- Gaussian ordination, related to logistic IRT and also approximated as maximum likelihood estimation through canonical correspondence analysis is implemented in various forms in the package VGAM.
- Two additional IRT packages (for Microsoft Windows only) are available and documented on the JSS site. The package mlirt computes multilevel IRT models, and cirt uses a joint hierarchically built up likelihood for estimating a two-parameter normal ogive model for responses and a log-normal model for response times.
- Bayesian approaches for estimating item and person parameters by means of Gibbs-Sampling are included in MCMCpack. In addition, the pscl package allows for Bayesian IRT and roll call analysis.
- The latdiag package produces commands to drive the dot program from graphviz to produce a graph useful in deciding whether a set of binary items might have a latent scale with non-crossing ICCs.
- The purpose of the rpf package is to factor out logic and math common to IRT fitting, diagnostics, and analysis. It is envisioned as core support code suitable for more specialized IRT packages to build upon.
- The classify package can be used to examine classification accuracy and consistency under IRT models.
- WrightMap provides graphical tools for plotting item-person maps.

Livros recomendados

- Baker, Frank (2001). The Basics of Item Response Theory available at http://ericae.net/irt/baker/
- Baker, F. and Kim, S. (2004). Item Response Theory: Parameter Estimation Techniques . Marcel Dekker Inc. New York
- Bond, T.G and Fox, C.M (2001). Applying the Rasch Model: Fundamental Measurement in the Human Sciences Lawrence Erlbaum Associates
- De Boeck, P., & Wilson, M. (Eds.) (2004). Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach. New York: Springer
- Embretson, S. and Reise, S. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

- Fox, J.-P. (2010). Bayesian Item Response Modeling: Theory and Applications New York: Springer.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park: Sage.
- Lord (1980) Applications of Item Response Theory to Practical Testing Problems
- McDonald, R. P. (1999). Test theory: A unified approach. Mahwah, NJ: Lawrence Erlbaum.
- Thissen, D., & Wainer, H. (Eds.). (2001). Test Scoring. Mahwah, NJ: Lawrence Erlbaum.

- Van der Linden, W.J. & Hambleton, R.K. (Eds.) (1997). Handbook of modern item response theory. New York: Springer.

Sites

- http://edres.org/irt/
- http://work.psych.uiuc.edu/irt/tutorial.asp
- http://psychcentral.com/psypsych/Item_response_theory

Software e Livros

- http://www.ssicentral.com
- http://www.assess.com

6. TCT vs TRI

Modelo dos testes clássico	Modelo de resposta ao item
O modelo é expresso a nível de teste	O modelo é expresso a nível do item
As características do Item são dependentes da amostra	As características do item são independentes da amostra (Invariância de Item)
Estimados da habilidade dependem dos itens	Estimativas da habilidade independente dos itens (Invariância de pessoas)
O mesmo erro de medição para todos examinados	O erro de medição é para cada nível de habilidade
Teste mais longos são mais confiáveis do que os testes mais curtos	Pequeno testes podem ser mais confiáveis do que testes longos

- Embora as bases teóricas da TRI aconteceram entre 1950 e 1960, o métodos não foram amplamente utilizados até os 70, devido à complexidade na estimativa.
- TCT é usado ainda mas a aproximação TRI está cada vez mas predominante
- A TRI trata novos problemas como multidimensionalidade, diferenciabilidade, testes de adaptação, teste de velocidades (speedeness), testlet, testes longitudinais, o equalizações
- Há uma extensa bibliografia cada vez mais na TRI, como software livre e comercial.

7. AVALIACOES EM DIFERENTES CONTEXTOS

1. Escala Global de Atitudes frente a Estatística obtida de
Aparicio, A. (2015). AVALIAÇÃO DAS ATITUDES NO CURSO DE ESTATÍSTICA: CONTEXTOS UNIVERSITÁRIOS LATINO-AMERICANOS. Teses de doutorado. FEA USP.

Veja também

http://www.ime.unicamp.br/sinape/sites/default/files/sumissao%20de%20trabalho%20AparicioEstradaBazan%2019Sinape.pdf

2. Prova de conhecimentos em Matemática 6ta serie
Bazán, J. L, Branco, M. D. , Bolfarine, H. (2006). A skew item response model. Bayesian Analysis, 1 (2006), pp. 861–892.

Prova de conhecimentos para 6ta Série

Leia com atenção cada questão e responda marcando com uma X sua resposta.

1. Em que alternativa os seguintes números estão ordenados do maior ao menor?

A) 567; 756; 765	B) 756; 765; 567	C) 765; 567; 756	D) 765, 756; 567

2. Indica a desigualdade correta.

A) 1/2 > 3/4	B) 7/6 <1/2	C) 3/4 > 7/6	D) 1/2<11/4

3. Um metro de pano custa S.I. 65. Quanto será pago por 0,5 metro?

A) S/. 302,50	B) S/. 121,00	C) S/. 30,25	D) S/. 30,05

4. Pepe dividiu um número entre 17, obtendo-se um quociente de 9 e um residuo de 2. Qual é o número?

A) 155 B) 171 C) 187 D) 306

5. Ao fazer a divisão: 960 ÷ 87 o quociente e o residuo obtido é?

A) quociente: 12; residuo: 16
B) quociente: 11; residuo: 13
C) quociente: 11; residuo: 3
D) quociente: 3; residuo: 11

6. O preço de uma blusa é S /. 30. Se Ana comprou com 20% de desconto, quanto pagou pela blusa?

A) S/. 50 B) S/. 24 C) S/. 20 D) S/. 6

7. Faça a seguinte operação de frações: 2/3 + 3/4

A) 5/7 B) 17/7 C) 6/12 D) 17/12

8. Pela compra de 100 litros de vinho paga-se S /. 1200. Quanto será pago por 200 litros?

A) S/.1 200	B) S/.1 400	C) S/. 1 500	D) S/. 2 400

9. Resolva as seguintes operações com decimais: **0,75 - 0,2 + 1,2 - 0,30**

A) 2,45	B) 2,05	C) 1,45	D) 0,45

10. Se o lado de um quadrado é de 3 cm, qual é seu perímetro?

A) 3 cm	B) 6 cm	C) 9 cm	D) 12 cm

11. Luisa, Dora e Maria compraram pano. Luisa comprou médio metro, Dora comprou 75 cm y Maria comprou 50 cm. Quais delas compraram a mesma quantidade de pano?

A) Luisa e Dora	B) Dora e María	C) Luisa e María	D) Ninguem

12. Um tanque recebe 4,5 litros de água por minuto. Quantos litros de água vai ter o tanque em uma hora e meia?

A) 4050 litros
B) 405 litros
C) 7,2 litros
D) 6,75 litros

13. Qual das seguintes figuras têm retas paralelas?

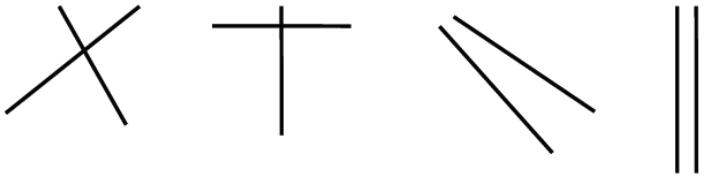
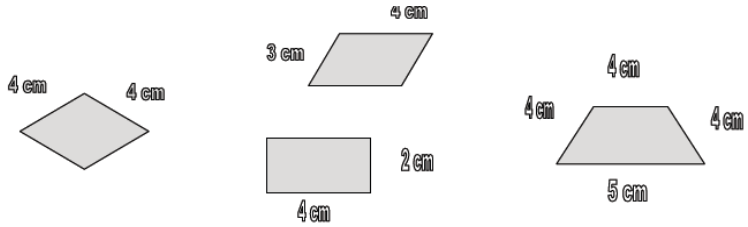


Figura 1 Figura 2 Figura 3 Figura 4

A) A figura 4
B) A figura 3
C) A figura 2
D) A figura 1

14. Observe as seguintes figuras.



Qual é a soma de todos os lados do losango?

A) 17 cm B) 16 cm C) 14 cm D) 11 cm

3. Sternberg, R. J. (1997). Construct validation of a triangular love scale. *European Journal of Social Psychology, 27*(3), 313-335.
http://vivanautics.com/pdf/Sternberg1997.pdf

Ver também a versão em português
http://www.redalyc.org/pdf/261/26118733005.pdf

Minicurso

MC-12 - AVALIAÇÃO EDUCACIONAL: ENTENDENDO A TEORIA DA RESPOSTA AO ITEM (ABE)

Ministrantes: Jorge Luis Bazán (USP) e Mariana Curi (USP)
jlbazan@icmc.usp.br, mcuri@icmc.usp.br

Público alvo: Professores do ensino básico

Sala: *AT 04 - Sala 82*
De 14/7/2015 à 17/7/2015 - das 08h00 às 10h00

AULA 2. ELABORAÇÃO DE QUESTÕES OU ITENS E SUA ANÁLISE USANDO METODOLOGIA TRADICIONAL (Oficina)

Conteúdo. *Principais medidas e critérios para a interpretação dos resultados.* Matrizes de referência. Tipos de itens. Recomendações para elaboração de itens. Exemplos. Oficina de redação de itens. Análise clássica de itens (qualitativa e quantitativa) baseada na TCT. Respostas correta, não resposta e valores perdidos. Identificação de viés. Principais medidas e critérios para a interpretação dos resultados.

Ministrante: Jorge Luís Bazán

TÓPICOS

1. MARCO METODOLÓGICO DA AVALIACAO
2. MATRIZES DE REFERENCIA OU MAPA DO CONSTRUTO
3. ELABORAÇÃO DE ITENS OU PLANEJAMENTO DA MEDIDA
4. ANALISE CLASSICA DE ITENS OU MODELO DE MEDIÇÃO CLASSICO
5. CRITERIOS PARA INTERPRETACAO DE RESULTADOS OU DO ESPAÇO DE RESULTADOS
6. ANALISIS DE ITENS USANDO SOFTWARE

1. MARCO METODOLÓGICO DA AVALIAÇÂO

Nos podemos considerar como marco metodológico uma versão adaptada da proposta de Duckor, Draney e Wilson (2009) também discutido em Wilson (2005).

Estes autores apresentam uma proposta para a construção de medidas com base em quatro etapas e princípios do sistema de avaliação, os quais são apresentados na Figura 1.

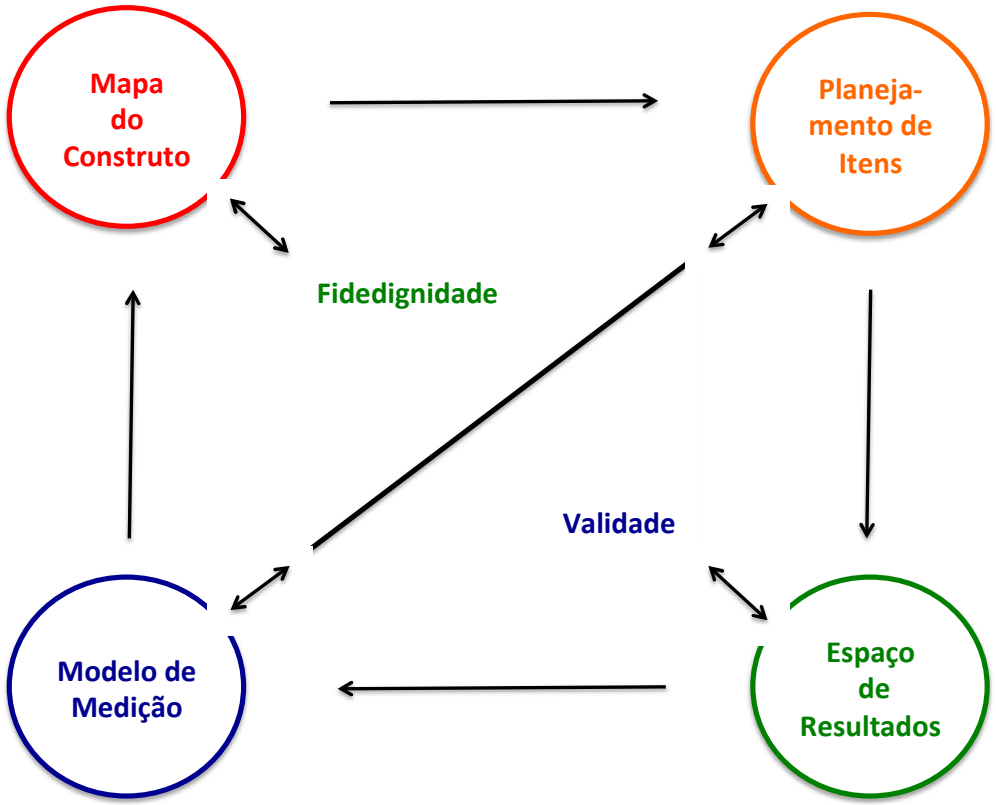


Figura 1. Relações entre os quatro blocos para a construção de medidas (tomado de Duckor, Draney e Wilson, 2009).

O processo da construção de uma medida inicia-se na 1) definição do mapa do construto, segue com o 2) planejamento de itens, a 3) definição do espaço de resultados, o qual define, e finalmente 4) o modelo de medição a ser considerado. Neste processo as Etapas 1-3 envolvem a fidedignidade[1] da medida enquanto que as Etapas 2-4 envolvem a validade[2] da mesma.

[1] A fidedignidade é uma característica da medida que faz referencia ao grau de consistência ou reprodutibilidade das medidas quando os procedimentos das avaliações são replicados sob as mesmas condições.

[2] A validade da medida faz referencia ao grau pelo qual a evidência e a teoria suportam as interpretações a partir dos valores das medidas.

Na pratica esse processo não necessariamente é explicito, isto é, os elaboradores ou construtores de medidas não necessariamente seguem esse processo no nível de detalhe discutido na proposta dos autores. Entretanto, quando se deve analisar uma medida é requerido avaliar cada uma dessas etapas. .

A analise de toda medida enfatiza tópicos que envolvem diferentes objetivos como: revisão, descrição, crítica ou proposta. Neste documento, propomos a classificação destes diferentes objetivos, dependendo do foco em que eles se centram. Por exemplo alguns trabalhos enfatizam sua revisão, descrição, crítica ou proposta na definição do mapa de construto, mas outros podem ser melhor classificados como centrando seus objetivos no modelo de medição.

Note em nossa proposta, trocamos a ordem das Etapas 3 e 4. Isto é, uma vez que as medidas já estão definidas, primeiro revisamos o modelo de medição e em seguida o espaço de resultados adotado. Neste caso, a sequencia fica semelhante a um planejamento estatístico usual.

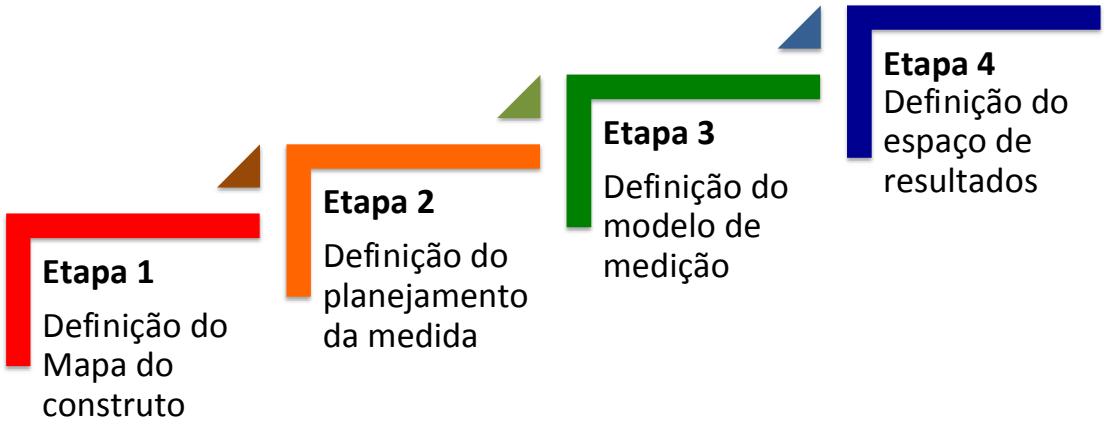


Figura 2. Etapas para a avaliação de construção de medidas (adaptado de Duckor, Daney e Wilson, 2009).

Quadro 1. Etapas de avaliação do processo de construção de medidas.

Etapa	**Nome da Etapa**	**Definição**	**Pergunta**	**Planejamento estatístico**
I	A definição do Mapa do construto	definição de aquilo que está sendo medido.	que vai ser medido?	Definição da pesquisa ou interação entre o pesquisador e o analista de dados
II	A definição do Planejamento da medida	definição do formato de avaliação ou instrumento e as unidades de observação o fontes de informação (alunos, diretores, etc.) processo, amostragem ou instrumentos, bases de dados	Como vai ser medido?	Definição dos instrumentos, amostragem, processo de captura de dados, elaboração de base de dados

Etapa	Nome da Etapa	Definição	Pergunta	Planejamento estatístico
III	Modelo de medição	definição do modelo de medição (modelo estatístico) que é aplicado em II		Definição do modelo estatístico ou técnica de análise de dados a serem adotadas
IV	definição da apresentaçã o do espaço de resultados	definição da forma de apresentação dos resultados finais e sua interpretação e uso que é aplicado ao processo em III.		Definição do modelo de reporte de resultados

2. MATRICES DE REFERENCIA O MAPA DO CONSTRUTO

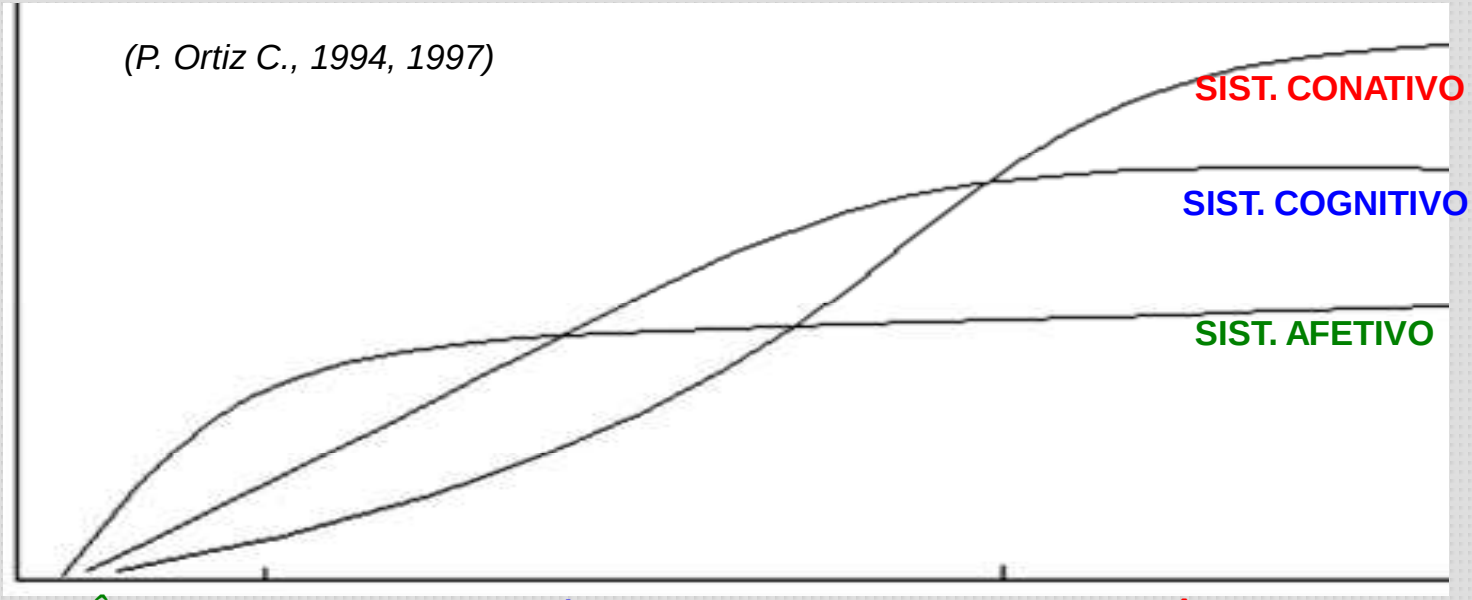
Mapa do construto

Um instrumento é sempre secundário. Há sempre uma finalidade para a qual é necessário um instrumento e do contexto em que este será utilizado (ou seja, envolvendo algum tipo de decisão).

Trata-se de uma idéia ou conceito que é o objeto teórico do nosso interesse em o avaliado conhecido comunmente como *construto* .

O construto pode ser parte de um modelo teórico de uma cognição pessoal - como a sua compreensão de determinado conjunto de conceitos ou atitude em relação a alguma coisa - ou pode ser alguma outra variável psicológica, ou o desempenho de determinado domínio educacional , etc. Ou pode estar relacionada com um grupo , em vez de um indivíduo ou podem ser um objecto inanimado complexo .

O DESENVOLVIMENTO DA CONSCIÊNCIA E A PERONALIDADE REFLETE A HISTORIA DA



INFÂNCIA 1:

OS SENTIMENTOS REFLETEM A ESTRUTURA TRADICIONAL

INFÂNCIA 2:

OS CONHECIMENTOS REFLETEM A ESTRUTURA CULTURAL

ADOLESCÊNCIA:

AS MOTIVAÇÕES REFLETEM A ESTRUTURA ECONÓMICA

Há uma infinidade de teorias - o importante aqui é ter uma estrutura para proporcionar motivação e estrutura para o construto a serem medido.

A idéia de construir um mapa é um conceito mais preciso do que falar de construto.

Supõe-se que o construto a ser medido tem uma forma particularmente simples, estendida de um extremo a outro, de alto ao um baixo valor, a partir de um pequeno a grande valor, de positiva para negativ, ou de forte para fraco.

Há alguma complexidade no que acontece entre os valores extremos mas estamos interessado principalmente na ubicao de um entrevistado é entre um extremo e outro .

Em particular, podem ser definidos níveis qualitativos entre os extremos - estes são importantes e úteis na interpretação -

Este ponto ainda é uma idéia latente antes que algo claro. Embora os níveis qualitativos são definíveis , presume-se que os entrevistados podem estar em qualquer lugar no continuum do construto subjacente

Em resumo um mapa do construto pode ser considerado uma variável latente unidimensional .

Muitos construtos são mais complexos do que isto, por exemplo, pode ser multidimensional, mas isso não é uma barreira para a modelagem que fazemos, porque cada uma dessas dimensões pode ser considerado unidimensional e, portanto, podemos ter um mapa de construto para cada um deles dimensões.

Matriz de referencia

É o consenso do que e quanto deve conhecer o docente de ensino médio ao respeito de sua especialidade.

Este consenso é representado numa tabela de especificações ou matriz de referencia a qual geralmente é de dupla entrada onde em geral temos nas linhas conteúdos e nas colunas níveis cognitivos para estabelecer os pesos dos mesmos

Quadro 2. Exemplo de Matriz de referencia para uma prova educacional cognitiva

	Níveis Cognitivos			
	Nivel 1	Nivel 2	Nivel 3	Total
Conteúdo 1				
Conteúdo 2				
Conteúdo 3				
Conteúdo 4				
Conteúdo 5				
Total				

Um exemplo: Avaliação docente

Uma parte importante na formação de professores é estabelecer seu nível de desempenho ou de domínio nos principais conteúdos de ensino do currículo de sua especialidade, identificando estes segundo níveis cognitivos.

Em relação aos aspectos cognitivos vários esquemas ou quadros de referências têm sido propostos para ter em conta na preparação de provas. A modo de exemplo consideramos três níveis gerais os quais são apropriadas para medir diretamente através de um exame escrito

Outros níveis podem ser medido de forma abrangente com formatos maiores a um exame escrito, como a entrevista, as questões de desenvolvimento e avaliação de registros e atividades.

Quadro 3. Complexidade das tarefas baseada em níveis cognitivos para avaliações de professores de Ensino Médio

Nivel	Nome	Descrição
I	Gestão de informação	É o conjunto de questões que avaliam a capacidade em matéria de gestão de conceitos, termos e símbolos relacionados com as competências desejáveis que cada professor deve desenvolver dentro de um determinado assunto. Isso corresponde a um nível primário de cognição ligada ao passivo e concreto . Ele inclui reconhecimentos , descrições , sistemas, interpretações literais
II	Gestão de processos	É o conjunto de questões que avaliam a capacidade em matéria de gestão e implementação de estratégias (relacionamentos através de conceitos, imagens e procedimentos) relacionados com as competências desejáveis que cada professor deve desenvolver dentro de um determinado assunto. Isso corresponde a um nível primário de cognição ligado ao operacional e concreto. Involve o nível I, mas implica a aplicação directa desse nível para situações familiares .
III	Reflexão	É o conjunto de questões que avaliam a capacidade relativa à resolução de situações-problema envolvendo reflexão relacionada com as competências desejáveis que cada professor deve desenvolver dentro de um determinado assunto. Este nível corresponde a um nível secundário da cognição , vinculada ao operacional e ao hipotético dedutivo. Ele inclui os níveis 1 e 2

Quadro 4. Lista de conteúdos e níveis cognitivos (NC) a serem avaliados em provas hipotéticas por áreas dos professores

Àreas	Lista de conteúdos avaliados	Conteúdos	NC avaliados	Número de NC
Matemática y Lógico matemática	Ecuaciones, Triángulos, Conjuntos, Divisibilidad, Cuatro operaciones, Productos Notables	6	I, II y III	3
Letras	Gramática, Normativa, Redacción, Literatura peruana y Razonamiento verbal	5	I, II y III	3
Ciencia y ambiente	Anatomía y fisiología humana, Botánica, Anatomía y fisiología animal, Reinos biológicos, Ecología, Materia	6	I, II y III	3
Biología	Citología, División celular, Anatomía y fisiología animal, Genética y fisiología celular, Anatomía y fisiología humana y Botánica	6	I, II y III	3
Química	Estructura atómica, Enlaces químicos, Nomenclatura inorgánica, Cálculos químicos y Química orgánica	5	I, II y III	3
Física	Cinemática, Trabajo y energía, Electrodinámica, Estática / dinámica y Electromagnetismo	5	I, II y III	3
Ciencias Sociales	Perú prehispánico, Antropogénesis, Feudalismo, Perú siglo XIX – XX, El universo y Elementos básicos de economía	6	I, II y III	3
Filosofía	Religión y filosofía oriental, Filosofía y religión en el esclavismo, Filosofía en el feudalismo y cristianismo, Renacimiento, Filosofía en el capitalismo y Disciplinas filosóficas	6	I, II y III	3
Psicología	Ramas de la psicología, Estados de la conciencia, Bases biológicas del psiquismo humano, Bases socioculturales del psiquismo humano, Personalidad y Desarrollo humano	6	I, II y III	3
Unidocencia (Inicial, I Ciclo y II Ciclo)	Teoría de conjuntos, Cuatro operaciones, Gramática, Comprensión lectora, El cuerpo humano y Las regiones del Perú / Actividades económicas	6	I, II y III	3
Inglés	Vocabulario, Lenguaje, Reading y Writing	4	I y II	2
Educación Física	Habilidades y destrezas, Psicomotricidad, Juego, Capacidades físicas, Aprendizaje motor y Deporte	6	I, II y III	3
Computación	Diseño Macromedia, Diseño gráfico y Ofimática	3	II	1
Arte	Conceptos básicos de Arte, Historia del Arte, Arte y Cultura y Función social del Arte	4	I, II y III	3

LENGUAJE Y LITERATURA

	CONTENIDOS	NIVELES COGNITIVOS I	II	III	PREGUNTAS POR CONTENIDO
1	Gramática	2	2	3	**7**
2	Normativa	1	2	3	**6**
3	Redacción	0	0	1	**1**
4	Literatura peruana	1	2	2	**5**
5	Razonamiento verbal	2	3	6	**11**
	PREGUNTAS POR NC	**6**	**9**	**15**	**30**
	PESO DE NC	**20%**	**30%**	**50%**	**100%**

Quando um mapa do construto é postulado pela primeira vez, é muitas vezes menos desenvolvidas do que aqui é apresentado. A melhora do mapa é obtido por meio de vários processos a medida que o instrumento é desenvolvido.

Esses processos incluem:
a) Explicar o construto a outras pessoas usando o mapa do construto;
b) Criar itens que você acredita que levam ao entrevistado a responder os níveis do mapa de construto;
c) Testar esses itens com uma amostra de respondentes e
d) Analisar os dados resultantes para verificar se os resultados são consistentes com as suas intenções expressas pela mapa do construto.

3. ELABORAÇÃO DE ITENS OU PLANEJAMENTO DA MEDIDA

Em seguida, logo após de ter o mapa do construto o medidor deve pensar em alguma forma como este construto teórico pode se manifestar em uma situação do mundo real.

No início, não será mais do que um palpite, um contexto em que se acredita que o construto deve estar envolvido, de fato, aquele em que o construto deve desempenhar um papel decisivo nesta situação.

Ainda este palpite se tornará mais cristalizada e se tornará em certos padrões.

A relação entre os itens e o construto não é necessariamente da forma como esta foi descrita. Muitas vezes, os itens podem ser pensados primeiro e o construto pode ser mais tarde elucidado.

Um item também pode assumir muitas formas, tais como múltipla escolha e Likert (tipos de itens de escolha forçada). Há muitas variações sobre isto. O entrevistado também pode produzir uma resposta livre de uma forma tal como um teste, entrevista ou desempenho (concertos, experimento científico, desenho).

Os itens variam em conteúdo e modo : perguntas de entrevista normalmente têm uma ampla gama de muitos aspectos de um tópico ; questões ou tarefas de um desempenho cognitivo podem ser apresentados dependendo das respostas a alguns itens iniciais; questoes em uma pesquisa podem usar diferentes conjuntos de opções e algumas resposta pode ser forçado e de resposta livre

Na situação mostrada na Figura 1,4 o medidor assume que o entrevistado " tem " uma certa quantidade da construto e que seu valor no construto é a causa das respostas dos itens no instrumento que o medidor usa.

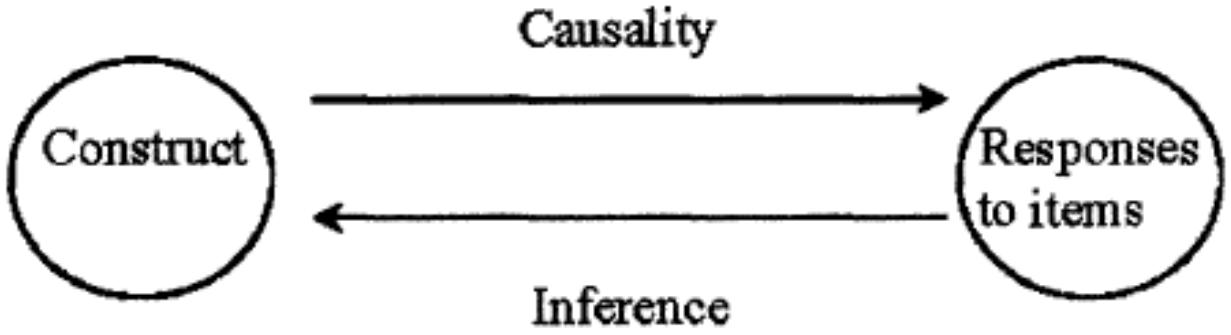


FIG. 1.4 A picture of the construct modeling idea of the relationship between degree of construct possessed and item responses.

No entanto, este agente causal é latente. O medidor não pode observar diretamente o construto. Em vez disso observa as respostas aos itens metros e então, inferir o construto subjacente a estas observações.

Note-se que a idéia de causalidade é uma suposição e a análise não fornece evidência dessa causalidade, na verdade, esta relação pode ser mais complexa.

Exemplo na Área de Língua

Nível I:

Qual é o substantivo relacionado à palavra abecedário?

a) Alfabeto b) Vogais c) Letras d) Maiúsculas e) Consonantes

Nível II:

Em qual dos itens abaixo estão escritos somente substantivos próprios?

a) Pedro, Augusto, Brasil, Bonito, Inteligente, América, Europa

b) Pedro, Augusto, Brasil, América, Europa

c) Pedro, Augusto, Brasil, Feio, Cachorro, América, Europa

d) Feio, Bonito, Cachorro, Inteligente

e) América, Europa, Bonito, Cachorro, Inteligente

Nivel III:
Associe os substantivos com seus respectivas classificações

I.	horário	a. primitivo
II.	caracol	b. derivado
III.	flotilha	c. composto
IV.	couve	d. coletivo

a) Ia, IIc, IIId, IVb b) Ib, IIa, IIId, IVc c) Ib, IIC, IIIa, IVd
d) Id, IIa, IIIb, IVc e) Ia, IIb, IIId, IVc

Atividades 1

Definir os conteúdos para avaliação de professores de uma determinada área de interesse (conteúdos mais relevantes usando programação anual)
5 minutos

Determinar a distribuição de questoes da matriz de referencia de sua área.
5 minutos.

Escrever um exemplo de item del área escolhida num nível cognitivo particular e conteúdo específico .
10 minutos

4. ANALISE CLASSICA DE ITENS OU MODELO DE MEDIÇÃO CLASSICO

Referências iniciais

- Gulliksen, H. (1950). Theory of Mental Tests. New York: John Wiley and Sons.
- Lord, F.M., Norvick, M.R. (1968). Statistical Theories of Mental Test Score. Reading: Addison-Wesley.
- Vianna, H.M. (1987). Testes em Educação. São Paulo: Ibrasa.

4.1. ANALISE ESTATÍSTICA DOS ITENS DE UM TESTE

- O objetivo geral em construção de testes é a obtenção de um teste de tamanho mínimo o qual deve produzir escores confiáveis e válidos para o uso que se desejar dar ao teste.
- Isto normalmente é alcançado testando um grande número de itens e selecionando aqueles que mais contribuem para a validade e confiabilidade do teste.
- Esses itens são identificados através do processo chamado *análise de itens*.
- Os parâmetros dos itens geralmente examinados são: função de variância e correlação com o critério.

- A análise estatística dos itens tem por finalidade estudar o *comportamento psicométrico* tanto de cada um de eles como de todo o conjunto.
- Usando vários métodos estatísticos e fazendo uso da interpretação da informação este analise nos permite garantir a validade e confiabilidade do instrumento que é construído.
- Os métodos que fazem parte da analise de itens dentro do enfoque da chamada Teoria Clássica dos Testes (TCT) se baseiam num conjunto de técnicas estadísticas de tipo descritivo as quais são interpretadas usando critérios empíricos e não supõem um modelo probabilístico.

Tamanho de amostra

Com respeito ao tamanho da amostra Lazarte (1995) diz que "não existe uma regra absoluta sobre o tamanho da amostra.

Certamente, uma análise de itens para testes nacionais envolve uma amostra grande e obtida cuidadosamente.

Para teses e outros trabalhos feitos por alunos recomenda-se amostras nos 200".

Uma regra empírica recomendada por Nunnally (1987) é usar entre 5 e 10 indivíduos para cada item no teste a ser analisado.

Devemos levar em conta que uma amostra adicional será necessária para estudar a validação cruzada da que falaremos logo.

Etapas da análise estatística dos itens

Segundo Ezcurra (1995), os passos a considerar numa análise estatística dos itens são:

a) Selecionar uma amostra representativa de indivíduos em que é aplicado o teste piloto, que deve ser pelo menos igual ou maior do que os 200 casos, se o teste é para uma pesquisa e 1000, se o teste é para uso comercial.

b) Qualificar os testes de acordo com a grelha de correção.

c) Preparar o banco de dados, utilizando o seguinte modelo:

Tabela 1 Banco de dados típico para análise estatística dos itens

Pessoas	Itens I1 I2 . . . Ij . . . Ik
P1	X_{11} X_{12} X_{1j} X_{1k}
P2	X_{21} X_{22} X_{2j} X_{2k}
.	. . . . .
.	. . . . .
.	. . . . .
Pi	X_{i1} X_{i2} X_{ij} X_{ik}
.	. . . . .
.Pn	X_{n1} X_{n2} X_{nj} X_{nk}

onde X_{ij} representa o valor ou escore obtido pelo indivíduo i no item j, que pode ser dicotômico ou policotômico.

Processo de Análise de Itens

Lazarte (op. cit) considera o seguinte processo numa análise de itens:

1. Decidir quais propriedades do escore total são importantes (ou seja, maximizar a variabilidade, maximizar a predição de critérios externos, etc.)
2. Identificar os parâmetros dos itens mais relevantes para estas propriedades do escore total.
3. Aplicar os itens para uma amostra de examinados que seja semelhante à população para a qual o teste está sendo construído.
4. Obter as estatísticas dos itens especificados no passo 2.
5. Estabelecer um plano para selecionar os itens ou identificar e revisar aqueles que estão com defeito.
6. Selecione um grupo final de itens.
7. Avaliar se o teste satisfaz o objetivo no passo 1, utilizando um estudo de validação cruzada.

Processamento dos dados psicométricos dos itens.

O processamento dos dados para obter as propriedades psicométricas dos itens, de acordo Ezcurra (op. cit.) envolve realizar os seguintes tipos de análise de forma obrigatória:

a) Obter a distribuição de frequência dos escores totais e de cada subteste (se o teste tem subtestes).
b) Representar graficamente (polígonos de frequência ou histogramas) as distribuições de frequência dos escores totais e de cada subteste.
c) Calcular a média, variância, desvio padrão, assimetria e curtose da distribuição dos escores totais e parciais de cada subteste.
d) Obter a dificuldade do item (as proporções da resposta correta para cada item), e corrigir para evitar o efeito do acaso, assim como a proporção da escolha de cada um dos distratores (outras alternativas de resposta propostas) incluídos.

e) Calcular a variância e desvio padrão de cada item, assim como a média e o desvio padrão do escore total e dos escores parciais dos indivíduos que escolheram a resposta correta.
f) Calcular a dificuldade de cada item.
g) Calcular o poder discriminativo de cada item.
h) Calcular o coeficiente de validade de cada item.

Opcional:

i) Calcular a matriz de correlação entre os sub-testes, e entre o escore total e cada sub-teste.
j) Calcular a análise de regressão múltipla dos sub-testes, e sob o escore total de modo que a partir da estimação dos coeficientes de regressão parcial possa-se fazer o peso para cada sub-teste.
k) Calcular a análise fatorial da matriz de intercorrelação dos itens para estabelecer a existência de fatores comuns.

4.2. TIPO DE ANALISES DOS ITENS

Os principais tipos de análise estatística utilizados hoje de preferência nos testes de desempenho, atuação ou aptidão (Nuria Cortada Kohan, 1968; Magnusson, 1990; Kline 1986, Nunnally, 1987), são:

A. Dificuldade do Item, média e variância

Itens dicotômicos.

São os mais comuns nos testes de aptidão. O item média corresponde à proporção de examinados que responderam o item "corretamente". Para o item i essa proporção, p_i, é chamada de dificuldade do item ou índice de dificuldade. Também pode ser apresentada como o percentual de pessoas que responderam corretamente o item através de:

$$Dif = \frac{Número\ de\ indivíduos\ que\ responderam\ corretamente\ o\ item}{Número\ de\ indivíduos\ availados} \times 100$$

Estas proporções podem ser ainda maiores se considerarmos que a resposta correta pode ser obtida se algumas alternativas obviamente erradas são eliminadas. Em muitos testes de aptidão usados nos EUA as dificuldades do item reportadas variam geralmente entre 0,6 e 0,8, em parte devido a esse fenômeno de adivinhar.

Portanto, para itens de múltipla escolha é aconselhável obter, para além da média e da variância do item, a distribuição de frequências para as alternativas que foram escolhidas pelos avaliados. As alternativas que não são a resposta correta são chamadas distratores. Esta distribuição pode indicar se existem distratores que não atraem nenhuma resposta, ou que atraem a maioria das respostas sem ser a correta, etc.

Por exemplo, na tabela adjacente o item 1 é difícil, porque um dos distratores atrai a maioria dos indivíduos. No item 2, dois distratores não funcionam em absoluto. No item 3, temos o caso clássico de um item com distratores aceitáveis.

	Alternativas (%)				
Item	A	B	C	D	p_i
1	24	4	56	16*	0,16
2	92*	0	8	0	0,92
3	20	20	8	52*	0,52

Para os itens dicotômicos, a variância da amostra do item deve ser descartada pois não fornece informação sobre as diferenças entre os avaliados.
Um item oferece a maior quantidade de informação sobre as diferenças entre os avaliados, quando $p_i = 0.5$ (Dif = 50%), e portanto a variância é maximizada.

Por isso, recomenda-se selecionar os itens em um intervalo de cerca de 0.5 (alguns autores sugerem entre 0,3 e 0,7).

Se o teste é para selecionar indivíduos, os itens mais difíceis são recomendados.

Tabela 2 Classificação do nível de dificuldade dos itens dicotômicos *

CLASSIFICAÇÃO	ÍNDICE DE DIFICULDADE
MUITO FÁCIL	DE 0.75 A 0.99
FÁCIL	DE 0.55 A 0.74
INTERMEDIÁRIO	DE 0.45 A 0.54
DIFÍCIL	DE 0.25 A 0.44
MUITO DIFÍCIL	DE 0.05 A 0.24

* Tomado de Ezcurra (op. cit)

Itens Politômicos

Os mais comuns nas escalas de Atitudes. Neste caso é requerido obter independentemente a média e a variância dos itens.

A media é equivalente de p_i nos itens dicotômicos, pero agora não tem interpretação de dificuldade.

A variância dos itens nos ajuda a escolher aqueles itens no sentido que procuramos aqueles com a maior variância possível

B. Discriminação do item.
Mede o grau em que o item é capaz de estabelecer diferenças entre os indivíduos com altos e baixos níveis de uma habilidade, aptidão ou conhecimento que está sendo avaliado.

O objetivo de qualquer teste é fornecer informação sobre as diferenças individuais no construto medido pelo teste, ou num critério externo, que o teste supostamente prediz. Portanto, estamos interessados em obter índices que mostram como efetivamente um item discrimina entre os avaliados que têm altos escores no critério e aqueles que têm baixos escores.

Na ausência de um critério externo, o escore total do mesmo teste é utilizado. Assim, o objetivo é identificar itens que os indivíduos que tem altos escores respondem corretamente com uma alta probabilidade, enquanto que os indivíduos com baixos escores respondem incorretamente.

Um item que é respondido igualmente de forma correta por indivíduos com escores altos e baixos, não discrimina bem entre esses dois grupos e não seria útil.

Um item que é respondido corretamente pelos indivíduos de escore baixo, e incorretamente pelos de alto escore, é um item com a discriminação negativa e não é desejável.

Índice de Discriminação

Este índice aplica-se só aos itens dicotômicos. Determina-se na distribuição dos escores do critério, um ou dois pontos de corte e classifica-se aos avaliados em grupos com escores abaixo e acima desses pontos de corte. Por exemplo, dividir em duas metades e classificar indivíduos na metade inferior e superior, dividir no terço superior e o terço inferior, etc.

Por exemplo, no seguinte:

- Grupo superior, que representa o 27% dos casos com escores totais maiores.
- Grupo intermediário, que representa o 46% dos casos com escores intermediários.
- Grupo baixo, que representa o 27% dos casos com escores totais menores.

Deles separam-se os grupos extremos

Uma vez que os dois grupos foram identificados, o índice de discriminação, D_i, do item I é obtido como:

$$D_i = p_{iS} - p_{il}$$

onde p_{iS} é a proporção de indivíduos no grupo superior que respondeu o item corretamente, e p_{il} é a proporção de corretas do grupo inferior.
De outra forma como regra geral, no grupo superior e no grupo inferior, são calculados separadamente para cada item a percentagem de indivíduos que responderam corretamente, ambos dados são subtraídos e o resultado final é a discriminação que têm o item, sua fórmula é:

Disc. = % de resposta correta no item i, do grupo Superior – % de resposta correta no item i do grupo Inferior

Disc. pode variar entre -1 e 1. Os valores positivos indicam que o item discrimina em favor do grupo superior, os negativos indicam que o item é discriminador ou que favorece ao grupo inferior.

Tabela 3 Classificação da discriminação dos itens dicotômicos *

CLASSIFICAÇÃO	DISCRIMINAÇÃO
MUITO BOA DISCRIMINAÇÃO	DE 0.40 A 0.99
DISCRIMINAÇÃO ACEITÁVEL	DE 0.30 A 0.39
DISCRIMINAÇÃO INTERMEDIÁRIA	DE 0.20 A 0.29
DISCRIMINAÇÃO INACEITÁVEL	DE 0.05 A 0.19

* Tomado de Ezcurra (op. cit.).

C. Validade do item.

Mede o grau no qual um item mede validamente aquela capacidade que deseja-se medir.

c1) Índices de correlação de validação do item

Todos esses índices correlacionam o escore no item com o escore obtido no critério externo, ou, na ausência de critérios externos, o escore total obtido no mesmo teste.

Em geral, todos esses índices são chamados *correlações item-total*. Quando o item é policotômico (como um item Likert), a correlação entre o item e o total é a correlação de Pearson entre outros casos receberam novos nomes, como veremos logo.

Ao usar o escore total do mesmo teste como critério, as correlações são modificados para eliminar a contribuição ao escore total do item estudado. Este tipo de correlação é chamado *correlação item-total com o item removido.*

Geralmente os coeficientes de correlação item-teste são utilizados para quantificar, os mais usados são:

a) Correlação r de Pearson:

É usada em situações em que as duas variáveis correlacionadas são contínuas. Utiliza-se a seguinte fórmula:

$$\rho_{i\,X} = \frac{\sigma_{i\,X}}{\sigma_i \sigma_X},$$

e para corrigir o resultado utiliza-se a seguinte fórmula:

$$\rho_{\,i\,(X-i\,)} = \frac{\rho_{iX}\sigma_X - \sigma_i}{\sqrt{\sigma_X^2 + \sigma_i^2 - 2\rho_{i\,X}\sigma_X\sigma_i}}$$

Onde:

$\rho_{\,i\,(X-1)}$ = Correlação corrigida item-teste.

$\rho_{\,i\,X}$ = Correlação item-teste.

σ_X = Desvio padrão dos escores totais dos indivíduos avaliados.
σ_i = Desvio padrão dos escores do item.
σ_{iX} = Covariância entre o item e o escore total.

Quanto mais próximo o coeficiente é de 1 é melhor, e aceita-se como critério empírico para aceitar o item que o resultado obtido deve ser, pelo menos, superior ou igual a 0.20.

b) Correlação bisserial:

É usada em situações em que uma variável que se correlaciona é contínua e a outra é dicotômica. É a correlação produto-momento de Pearson entre uma variável dicotômica (0 ou 1) e uma variável contínua. É o caso típico de itens dicotômicos. A fórmula para calcular essa correlação é dada por:

$$\rho_{\rho bis} = \frac{\mu_{i+} - \mu_X}{\sigma_X}\sqrt{p_i\ q_i}$$

Onde:

μ_{i+} = Média no critério (a média dos escores totais) dos indivíduos que respondem corretamente o item i.

μ_X = Média ou média dos escores totais de todos os indivíduos no teste.

σ_X = Desvio padrão dos escores totais dos indivíduos avaliados.

p_i = Proporção de indivíduos que respondem corretamente o item i. (Dificuldade do item i)

Quanto mais próximo o resultado é do valor 1, o coeficiente será melhor, e aceita-se como critério empírico que este deve ser, pelo menos, superior ou igual a 0.20 para ter em conta o item.

Versão corrigida

Na maioria dos casos para calcular o escore total e analisar um item, o resultado do mesmo está incluído no escore total, se o número de itens é grande (25 ou mais), isso não é um problema. Se não for o caso, é necessário corrigir esta situação pois introduze ao resultado final um aumento do mesmo por efeito da autocorrelação, em geral pode-se corrigir a correlação removendo o item do total utilizando a seguinte fórmula:

$$\rho_{\text{pbis c}} = \frac{\rho_{\text{pbis}}\sigma_X - \sigma_i}{\sqrt{\sigma_X^2 + \sigma_i^2 - 2\rho_{\text{pbis}}\sigma_X\sigma_i}}$$

Aqui $\mathbf{p}_{\text{pbis}}$ é a correlação bisserial original entre o item e o escore total do critério, s_x é o desvio padrão do item, e $\mathbf{p}_{\text{ibis c}}$ é a correlação bisserial corrigida quando o item i é removido do escore total. Note-se que esta equação pode ser aplicada a qualquer tipo de correlação original, e não apenas à ponto-bisserial.

c) <u>Coeficiente Phi:</u>

Quando os itens dicotômicos devem correlacionar-se com os critérios dicotômicos (ingressante vs. não ingressante, sucesso vs. falha, etc.), a extensão da correlação produto-momento de Pearson é chamada coeficiente Phi. Como a covariância entre os itens i e o critério dicotômico X, e suas respectivas variâncias são uma função da proporção de indivíduos que passam o item, $\mathbf{p}i$, e a proporção de indivíduos que passam o critério, $\mathbf{p}_x$, é possível mostrar que o coeficiente Phi pode ser expressado como:

$$\rho_{iX}\ (\text{phi}) = \frac{p_{iX} - p_i p_X}{\sqrt{p_i q_i p_X (1 - p_X)}}$$

Onde $P_{i\,X}$ é a proporção de indivíduos que passam o item, e também passam o critério; P_i é a proporção de indivíduos que passam o item i, e P_x é a proporção de indivíduos que passam o critério.

d) Índices de Confiabilidade e outros índices de Validez do item

Os índices confiabilidade e validez do item são funções conjuntas da variância do item e de sua correlação com o critério.
Se o critério usado é o escore total na mesma prova (critério interno) o índice se denomina *índice de confiabilidade* do item e se define como

$$\sigma_i \rho_{iX}$$

em que σ_i é o desvio padrão do item e ρ_{iX} é a correlação item-total.
Quando um critério é usado, o índice se denomina *índice de validez* do item e se define de modo similar como

$$\sigma_i \rho_{i\,Y}\ ,$$

em que ρ_{iY} é a correlação entre o item e um critério externo.

Estes índices são úteis pois sua combinação aditiva gera a variância do escore total, e o coeficiente de validez entre o teste e um critério externo pode ser expresso como a razão da suma dos índices de confiabilidade e validez, isto é:

$$\sigma_X^2 = \left(\sum_{i=1}^{k} \sigma_i \rho_{iX} \right)^2 \quad , \qquad \rho_{XY} = \frac{\sum_{i=1}^{k} \sigma_i \rho_{iY}}{\sum_{i=1}^{k} \sigma_i \rho_{iX}}$$

O índice de confiabilidade do item pode ser utilizado para estimar o valor do coeficiente alfa de Cronbach quando um novo item é retirado do teste. A expressão a usar é

$$\alpha = \frac{k}{k-1} \left[1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\left(\sum_{i=1}^{k} \sigma_i \rho_{iX} \right)^2} \right]$$

em que k representa o número de itens selecionados para entrar no teste até esse momento.

Quando o valor alfa é mais perto de 1 é melhor e significa que a soma das covariâncias dos itens em relação a variabilidade total é alta, indicando que os itens são consistentes entre si. Este índice também é chamado de consistência interna.

Cronbach's alpha	Internal consistency
α ≥ 0.9	Excellent (High-Stakes testing)
0.7 ≤ α < 0.9	Good (Low-Stakes testing)
0.6 ≤ α < 0.7	Acceptable
0.5 ≤ α < 0.6	Poor
α < 0.5	Unacceptable

https://en.wikipedia.org/wiki/Cronbach%27s_alpha

e) Crosvalidation ou Validação cruzada

Quando os itens são selecionados sobre a base de critérios estatísticos usando as respostas de uma amostra dada, o teste assim construído deveria

ser muito efetivo para essa amostra em particular, mas não necessariamente em uma outra amostra.

Num estudo de validação cruzada o criador do teste usa itens que tem escolhido considerando uma analise de itens, estes itens são aplicados a uma segunda amostra, independente da primeira, e a confiabilidade e validez dos escores são avaliados de novo.

Para obter informação relevante numa aplicação só do item, a amostra original - a qual se aplica todos os itens - é dividida em dois grupos aleatoriamente.
Num grupo é feito a analise dos itens. Logo, no outro grupo, analisasse o escore total do teste baseado no itens selecionado na analise dos itens do primeiro grupo.
Quando as analises dos itens no primeiro grupo se usam para escolher itens no segundo grupo, e ainda os resultados dos itens do segundo grupo se usam para selecionar os itens do primeiro grupo falamos de validação cruzada dupla.

5. CRITERIOS PARA INTERPRETACAO DE RESULTADOS OU DO ESPAÇO DE RESULTADOS

Critérios para a escolha de itens

Finalmente quando todos os analises estatísticos dos itens são completados precisasse de uma revisão critica dos mesmos. Esta revisão deve ser feita considerando:

a) Analisar a dificuldade de cada um dos itens de modo que posa se formar grupos de dificuldade e fazer uma ordem entre eles.
b) Analisar a discriminação dos itens e retirar aqueles que tenham valores muito baixo, inferiores ao critério empírico recomendado.
c) Analisar a validez dos itens, removendo aquele que não satisfazem o critério mínimo considerado.
d) Analisar para cada item de forma conjunta a dificuldade, discriminação e os outros critérios, e então escolher aqueles que satisfazem os três critérios ou boa parte deles ao mesmo tempo.

Geralmente logo de fazer as analises de itens quando construísse um teste pela primeira vez, são removidos uma grande quantidade de itens, porem precisasse de que no piloto sejam aplicadas uma grande quantidade dos mesmos. Mas se acontece que o numero de itens fica pequeno, então precisasse fazer itens adicionais e aplicar a uma nova amostra e volver a fazer os analises apresentados.

A ideia final é obter um teste com o qual obter a versão definitiva da validez e confiabilidade do teste e ainda estabelecer tabelas de interpretação

6. ANALISIS DE ITENS USANDO SOFTWARE

Objetivo
Utilizando os códigos em SPSS e R abaixo, faça uma análise de Itens para os dados do Teste de Matemática para alunos de 6ta séria, e dados da Escala de Atitudes frente a Estatística de professores.

1. Código de Analise de itens usando SPSS

Dados de Conhecimentos
Use o seguinte códigos para analisar os dados mathb usando SPSS.

```
**  Chama os dados dicotomicos

GET
  FILE='C:\Users\Jorge Luis\Dropbox\Eventos\SBPC\Dia2\mathb.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

```
DATASET ACTIVATE DataSet1.
RELIABILITY
  /VARIABLES=i01 i02 i03 i04 i05 i06 i07 i08 i09 i10 i11 i12 i13 i14
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /STATISTICS=DESCRIPTIVE SCALE CORR
  /SUMMARY=TOTAL MEANS VARIANCE CORR.
```

Dados de Atitudes

Use o seguinte códigos para analisar os dados baseunionfinal usando SPSS. Note que o código salva as analises e ainda exporta os resultados para um arquivo Word.

```
** Chama os dados politomicos

GET
  FILE='C:\Users\Jorge
Luis\Dropbox\Eventos\SBPC\Dia2\baseunionoriginal.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

```
*Calcula o analise de items

DATASET ACTIVATE DataSet1.
RELIABILITY
  /VARIABLES=pre1 pre2 pre3 pre4 pre5 pre6 pre7 pre8 pre9 pre10 pre11
pre12 pre13 pre14 pre15 pre16
    pre17 pre18 pre19 pre20 pre21 pre22 pre23 pre24 pre25
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.

*Organiza os proximos resultados segundo Pais

SORT CASES  BY pais$.
SPLIT FILE SEPARATE BY pais$.

* Calcula novamente a Analise de items por pais

DATASET ACTIVATE DataSet1.
RELIABILITY
  /VARIABLES=pre1 pre2 pre3 pre4 pre5 pre6 pre7 pre8 pre9 pre10 pre11
pre12 pre13 pre14 pre15 pre16
    pre17 pre18 pre19 pre20 pre21 pre22 pre23 pre24 pre25
```

```
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.

* Conclue os reportes por pais

SPLIT FILE OFF.

* Salva as resultados

OUTPUT SAVE NAME=Document1
 OUTFILE='C:\Users\Jorge
Luis\Dropbox\Eventos\SBPC\Dia2\AnaliseitemsAtitudes.spv'
 LOCK=NO.
```

* Exporta os resultados em Word

```
* Export Output.
OUTPUT EXPORT
  /CONTENTS  EXPORT=ALL  LAYERS=PRINTSETTING  MODELVIEWS=PRINTSETTING
  /DOC                                        DOCUMENTFILE='C:\Users\Jorge
Luis\Dropbox\Eventos\SBPC\Dia2\AnalisedeItemsAtitudes.doc'
     NOTESCAPTIONS=YES  WIDETABLES=WRAP
```

```
     PAGESIZE=INCHES(8.266535433070866,                 11.69015748031496)
TOPMARGIN=INCHES(1.0)
    BOTTOMMARGIN=INCHES(1.0)
     LEFTMARGIN=INCHES(1.0)       RIGHTMARGIN=INCHES(0.9999999999999991).
LEFTMARGIN=INCHES(1.0)  RIGHTMARGIN=INCHES(0.9999999999999991).
```

2. Código de Analise de itens usando R

Dados de Conhecimentos

Use o seguinte códigos para analisar os dados mathb usando R. Instale previamente os pacotes que são indicados.

```
#analise de items dicotómicos
require(foreign)
pasta="C:\\Users\\Jorge Luis\\Dropbox\\ICMC\\SME0876\\Aula3SME0876"
setwd(pasta)
mathb=read.spss("mathb.sav")
a=data.frame(mathb)
mathbitems=a[,2:15]
```

```
require(psych)
alpha(mathbitems)
help(alpha)

require(epicalc)
alpha(vars=c(i01:i14), mathbitems)


require(psychometric)
item.exam(mathbitems, y=mathb$puntaje, discrim=TRUE)
help(item.exam)

#intervalo de Confiança Bootstrap para alpha
require(ltm)
cronbach.alpha(mathbitems, CI=TRUE, B=500)

#Correlaçao ponto biserial

biserial.cor(rowSums(mathbitems), mathbitems[[1]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[2]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[3]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[4]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[5]], level = 2)
```

```
biserial.cor(rowSums(mathbitems), mathbitems[[6]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[7]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[8]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[9]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[10]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[11]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[12]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[13]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[14]], level = 2)

require(coefficientalpha)
alpha.mathb=cronbach(mathbitems)
plot(alpha.mathb,type="d")
summary(alpha.mathb)
```

Dados de Atitudes

Use o seguinte códigos para analisar os dados baseunionfinal usando R.

```
#analise de items politomicos
################################

atitudes=read.spss("baseunionfinal.sav")
a=data.frame(atitudes)
atitudesitems=a[,4:25]
alpha(atitudesitems)

#analises de items por paises
atitudesP=subset(a,pais.=="Perú")
atitudesP
atitudesitemsP=atitudesP[,4:25]
alpha(atitudesitemsP)

atitudesE=subset(a,pais.=="España")
atitudesE
atitudesitemsE=atitudesE[,4:25]
alpha(atitudesitemsE)
```

3. Atividade 2

- Salve os resultados encontrados usando SPSS e R numa folha Excel para os dados dicotômicos e politômicos.
- Usando os critérios mostrados acima, compare os resultados obtidos por ambos programas. Indique quais resultados são obtidos em ambos programas, quais são diferentes.
- Compare seus resultados com os obtidos em IRTPRO

4. Alcances finais

As estatísticas mais apropriadas apresentadas pelo módulo SPSS e os diferentes softwares como R e IRTPRO são a média do teste se o item foi eliminado, a variância do teste se o item foi eliminado, a correlação item-teste corrigida e o alfa se o item é eliminado. Porém, também é considerado conhecimento do pesquisador para decidir quais itens serão eliminados. Por isso, afirma-se que a análise estatística dos itens, consiste em técnicas, mais ou menos adequados.

Estes análises estatística dos itens sob perspectiva clássica, especialmente com o cálculo da média, variância e alfa de Cronbach se o item é eliminado, e a correlação item-total corrigida são os mais comuns e são válidos para o caso dicotômico e politômico. Más eles são analise básicas.

Referencias

EZCURRA, L (1995) Análisis Estadísitico de Items. Separata del curso Seminario de Construcción de Pruebas I. UNMSM. Facultad de Psicología. 3 p.

GUILFORD, J. P. (1954) *Psychometrics Methods*, New York Mc Graw Hill.

KLINE P. (1986) *A Handbook of Test Construction: Introduction to Psychometric Desig*, New York, Methuen And. Co., Ltd.

LAZARTE, A (1995) Análisis de Ítems. Separata del curso PSB234. PUCP. Facultad de Psicología 3p.

MAGNUSSON, D (1990) *Teoría de los Test.* Edit. Trillas México

NUNNALLY, J. (1987) *Teoría Psicométrica*, México. Ed. Trillas.

NURIA CORTADA DE KOHAN, (1968) *Estadística Aplicada.* Bs. Aires. Argentina

Aula 3. Introduçao à TRI

Mariana Cúri - ICMC/USP

mcuri@icmc.usp.br
www.icmc.usp.br/∼mcuri

julho de 2015

Conteúdo da Apresentação

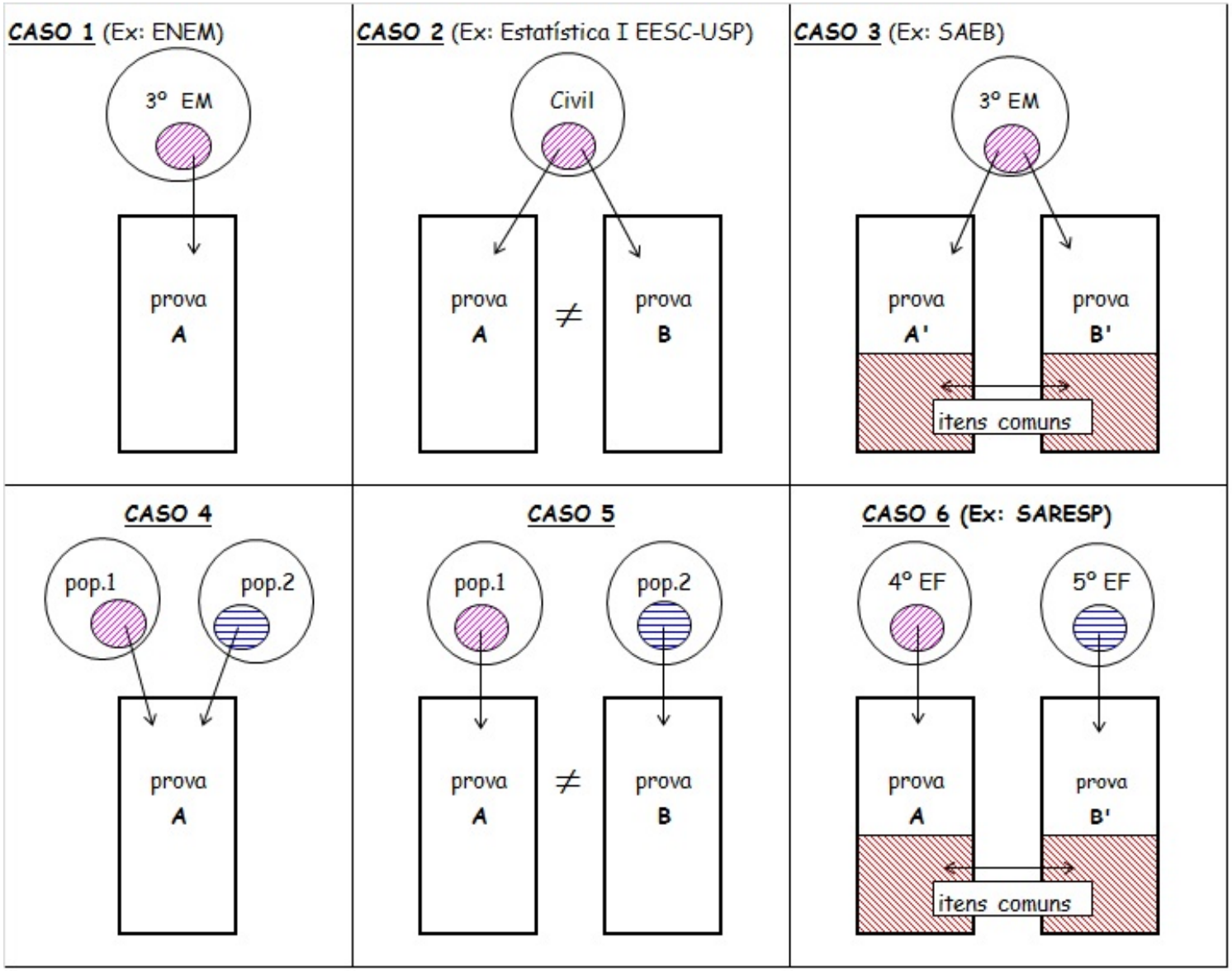
1. Introdução: Avaliações Educacionais

- Objetivo: classificação (Vestibular), certificação (aprovação em um curso), políticas educacionais (SARESP, SAEB)
- Construto: traço(s) latente(s) não observável(eis)
  - proficiência em língua estrangeira
  - habilidade em Matemática
  - outras áreas: intensidade de depressão, nível de qualidade de vida, grau de aceitação de um novo produto no mercado, predisposiçao para desenvolver determinada doença
- Instrumento de avaliação: prova composta por itens

1. Introdução: Avaliações Educacionais

- Número de itens (dicotômicos, nominais, ordinais ou abertos)
- Número de categorias de resposta
- Auto-aplicativo ou entrevistador
- Número de dimensões
  (traços latentes - uni ou multidimensional)
- Grau de dificuldade dos itens/prova
- Número de provas (paralelas?)
- Indivíduos realizam a prova ao mesmo tempo?
- Número de populações
- Tipo de prova: via lápis e papel, teste informatizado, teste adaptativo

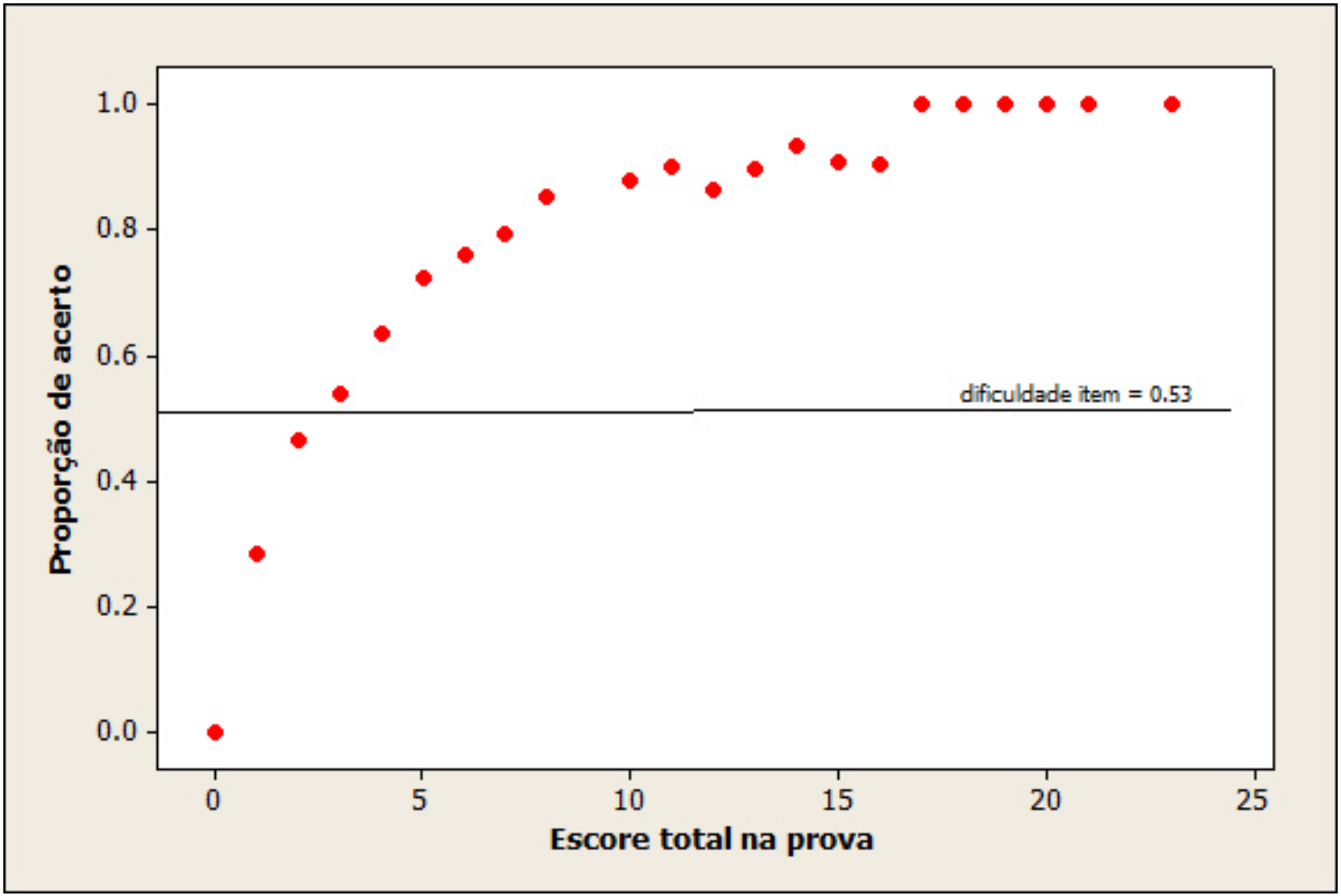
1. Introdução: Avaliações Educacionais



1. Introdução: Teoria Clássica

- atribuição de escore às alternativas de resposta dos itens:
  $\downarrow$ escore $\Leftrightarrow \uparrow$ traço latente ou $\uparrow$ escore $\Leftrightarrow \uparrow$ traço latente
- em testes de múltipla escolha (0=incorreta e 1=correta):
- Escore total (indivíduo): estimativa do traço latente
  número de itens corretos, varia de 0 a I
  (ou % de acerto, varia de 0 a 100%)
- Dificuldade (item): % de acertos, varia de 0 a 1 (ou 100%)
- Discriminação (item):
  % acertos grupo superior – grupo inferior, varia de -1 a 1
  Grupo superior: 27% com os escores mais altos.
  Grupo inferior 27% com os escores mais baixos.

1. Introdução: Teoria Clássica



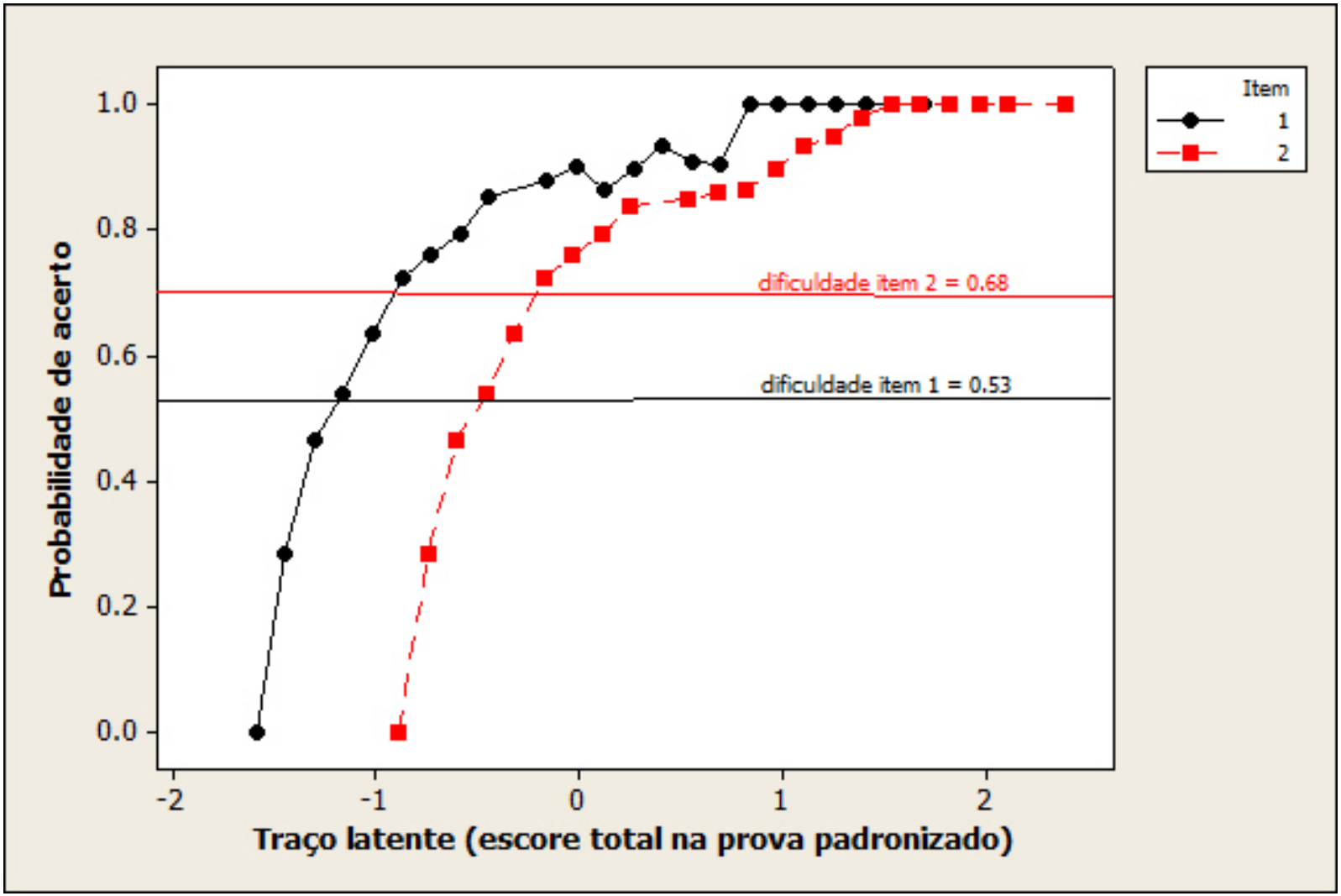
1. Introdução: Teoria Clássica

- Resultados dependem do particular conjunto de itens da prova (Prova - elemento central)
- Resultados dependentes do grupo de respondentes
- Compração entre indivíduos: somente com mesma prova ou provas paralelas
- Comparação proporção acertos entre séries?

1. Introdução: TRI

- Surgiu formalmente a partir dos trabalhos de Lord (1952) e Rasch (1960)
- Item - elemento central
- Permite a comparação entre indivíduos, mesmo submetidos a provas diferentes
- Analisa itens com diferentes escores para as categorias sem desbalancear a estimativa do traço latente
- 2 tipos de parâmetros: de itens e individuais (traços latentes)
- Modelos: probabilidade de determinada resposta ao item = f(parâmetros do item, traço latente)

1. Introdução: TRI



1. Introdução: TRI

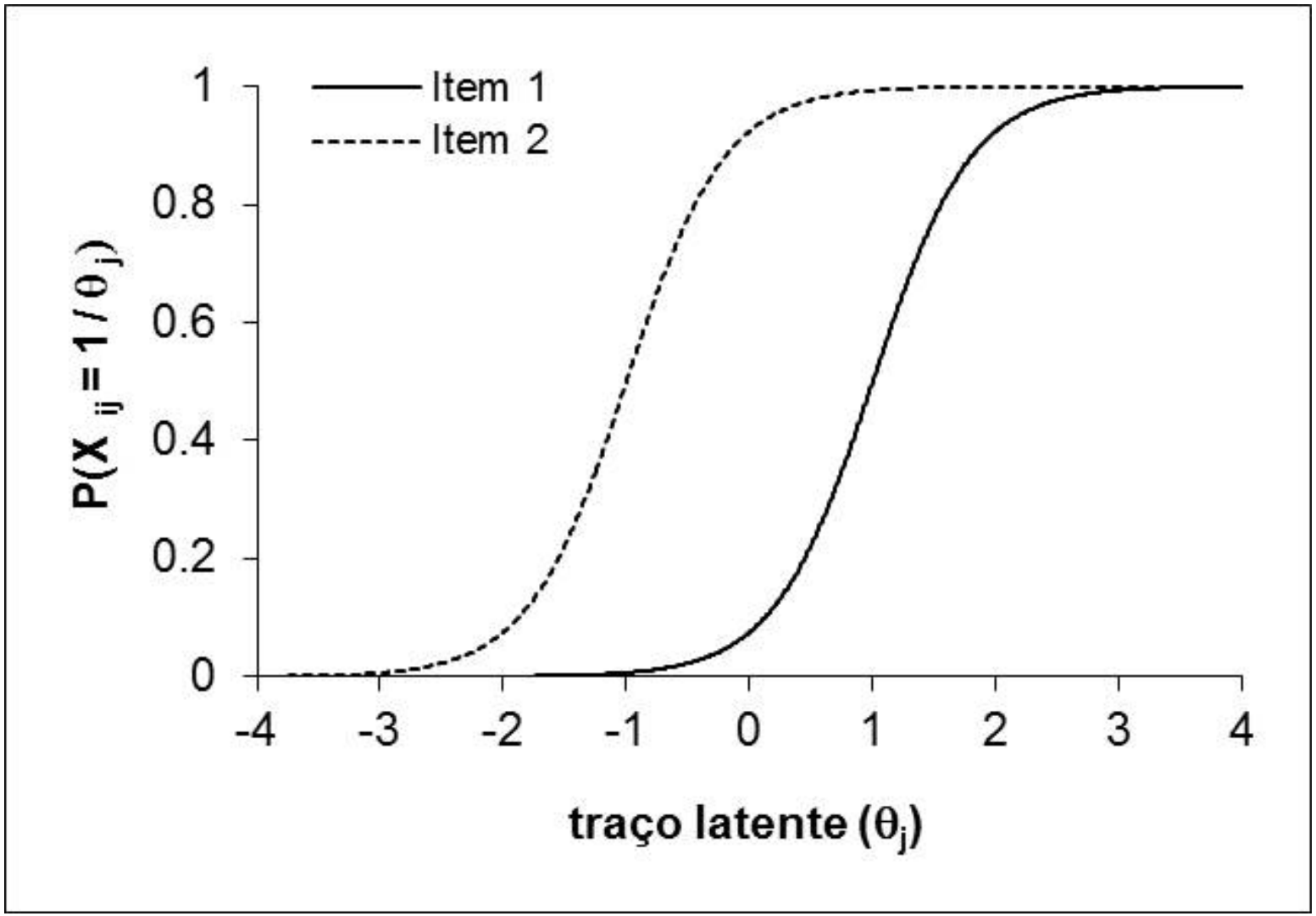
$X_i = 0$ ou 1: resposta do indivíduo ao item i (incorreta ou correta)
$X_i \sim$ Bernoulli (P_i)
$P_i = P(X_i = 1) = f(\theta, b_i)$,
sendo b_i a dificuldade do item i e θ, o traço latente do indivíduo.

Definições comuns de $f(\theta, b_i)$ na literatura:

$\Phi(.)$: fda da $N(0,1)$ - Modelos de ogiva normal
$\frac{1}{1+e^{-(\theta - b_i)}}$: função logística - Modelo (logístico) de Rasch

$$log \frac{P_i}{1 - P_i} = -(\theta - b_i)$$

1. Introdução: TRI



1. Introdução: TRI

Os modelos propostos dependem:

1. da natureza do item: dicotômicos, ordinais ou nominais
2. do número de populações envolvidas: apenas uma ou mais de uma população
3. da quantidade de traços latentes considerados: apenas um ou mais de um
4. Mais usual: Modelos logísticos unidimensionais para itens dicotômicos
   Se diferenciam pelo número de parâmetros utilizados para descrever o item:
   - 1 parâmetro = somente a dificuldade do item (modelo de Rasch);
   - 2 parâmetros = a dificuldade e a discriminação;
   - 3 parâmetros = a dificuldade, a discriminação e a probabilidade de acerto por indivíduos de baixo traço latente ("chute").

1. Introdução: TRI

Avaliações Educacionais que usam a TRI (nacionais e internacionais)

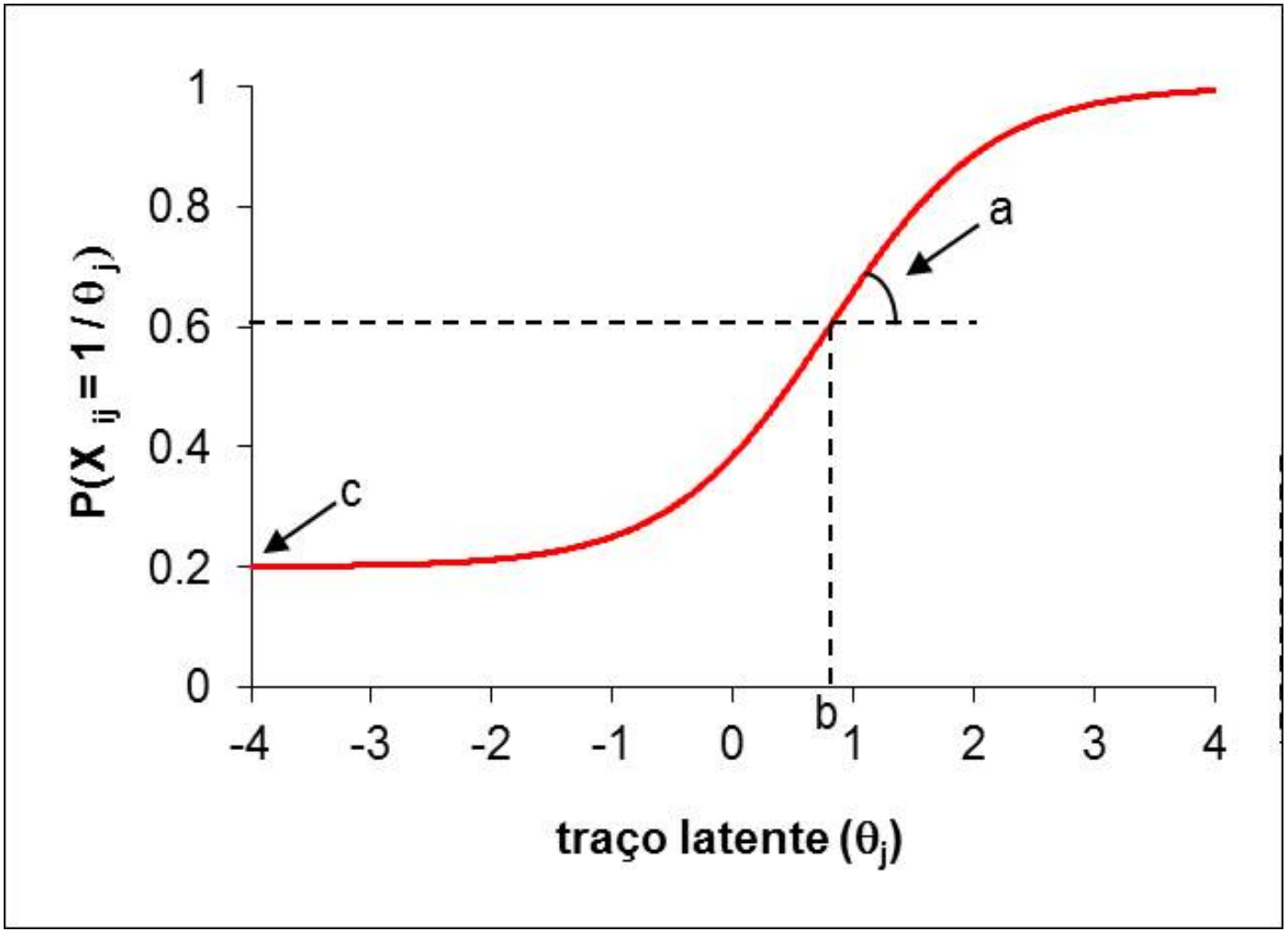
- ENEM
- SAEB
- ENCCEJA
- SARESP
- TOEFL
- GRE
- PISA

2. Modelos da TRI: ML3

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1 + e^{-a_i(\theta_j - b_i)}},$$

i=1, ..., I (itens)
j=1, ..., n (indivíduos)
X_{ij}=1, se indiv j acerta o item i, e X_{ij}=0, c.c.
θ_j é o nível do traço latente do indiv j
a_i parâmetro de discriminação do item i,
derivada no ponto de inflexão
b_i parâmetro de dificuldade do item i,
se $b_i = \theta_j$, $P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = (1 + c_i)/2$
c_i parâmetro de acerto ao acaso ("chute") do item i

Curva Caractrística de Item (CCI)



Função de Informaçao do Item

Pelas c.r. (devido à família exponencial):

$$I_i(\theta) = \frac{\left(\frac{\partial P_i(\theta)}{\partial \theta}\right)^2}{P_i(\theta)(1 - P_i(\theta))}$$

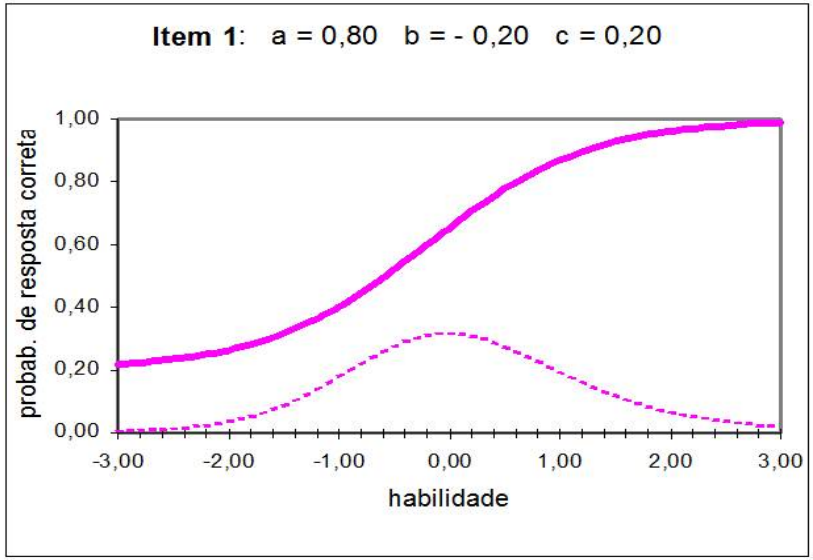
e

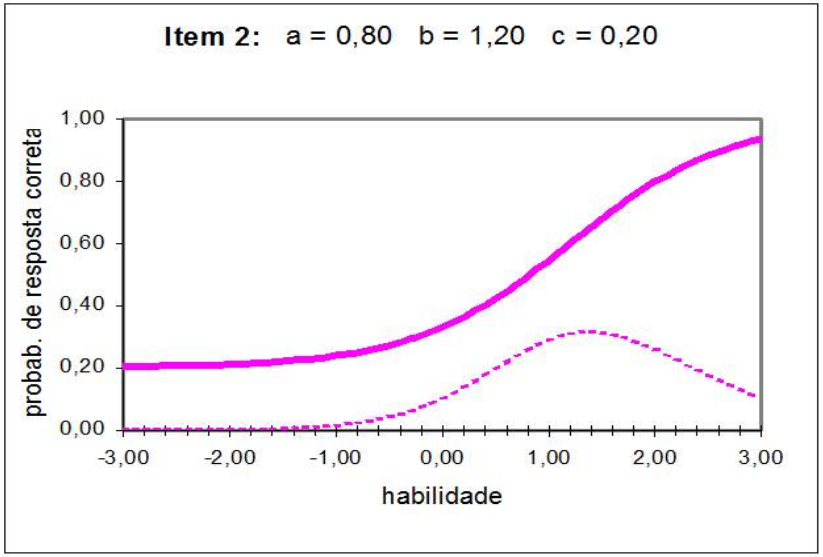
$$I(\theta) = \sum_{i=1}^{I} I_i(\theta),$$

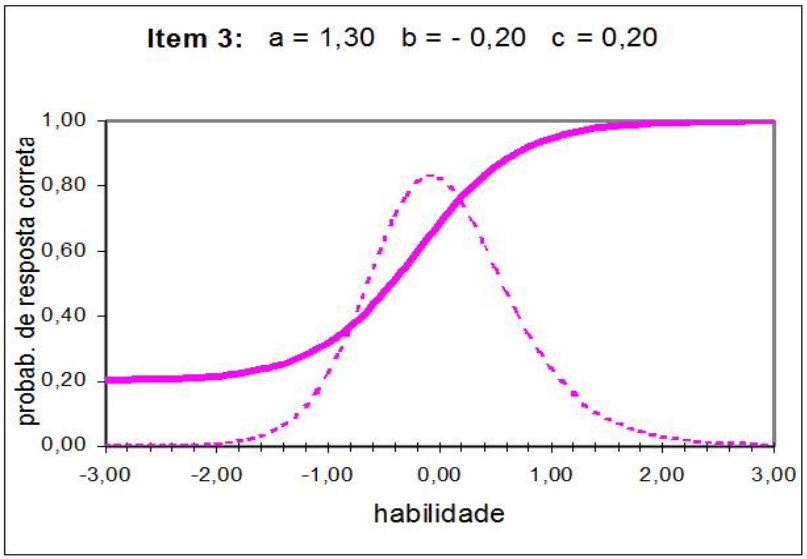
em que $P_i(\theta) = P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i)$, para $\theta_j = \theta$.

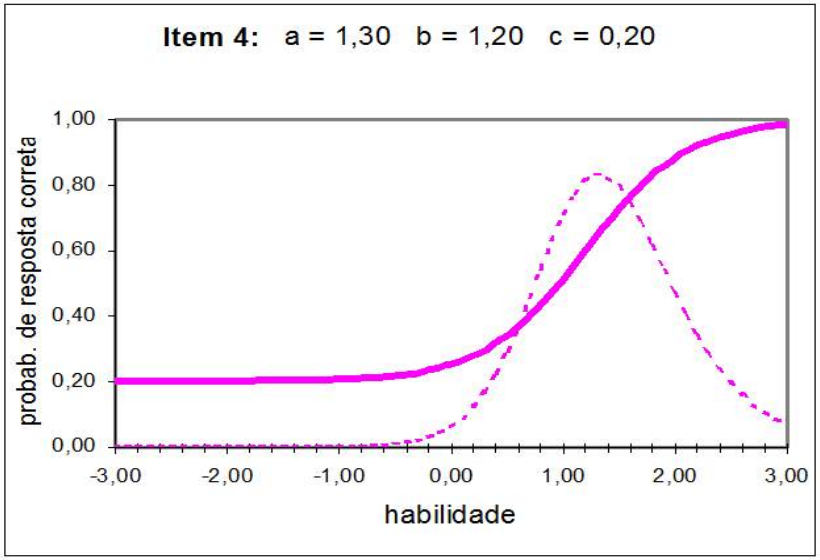
Em particular, para o ML3: $I_i(\theta) = a_i^2 \frac{(1-P_i(\theta))}{P_i(\theta)} \left[\frac{P_i(\theta)-c_i}{1-c_i}\right]^2$.

2. Modelos da TRI: Função de Informação do ML3

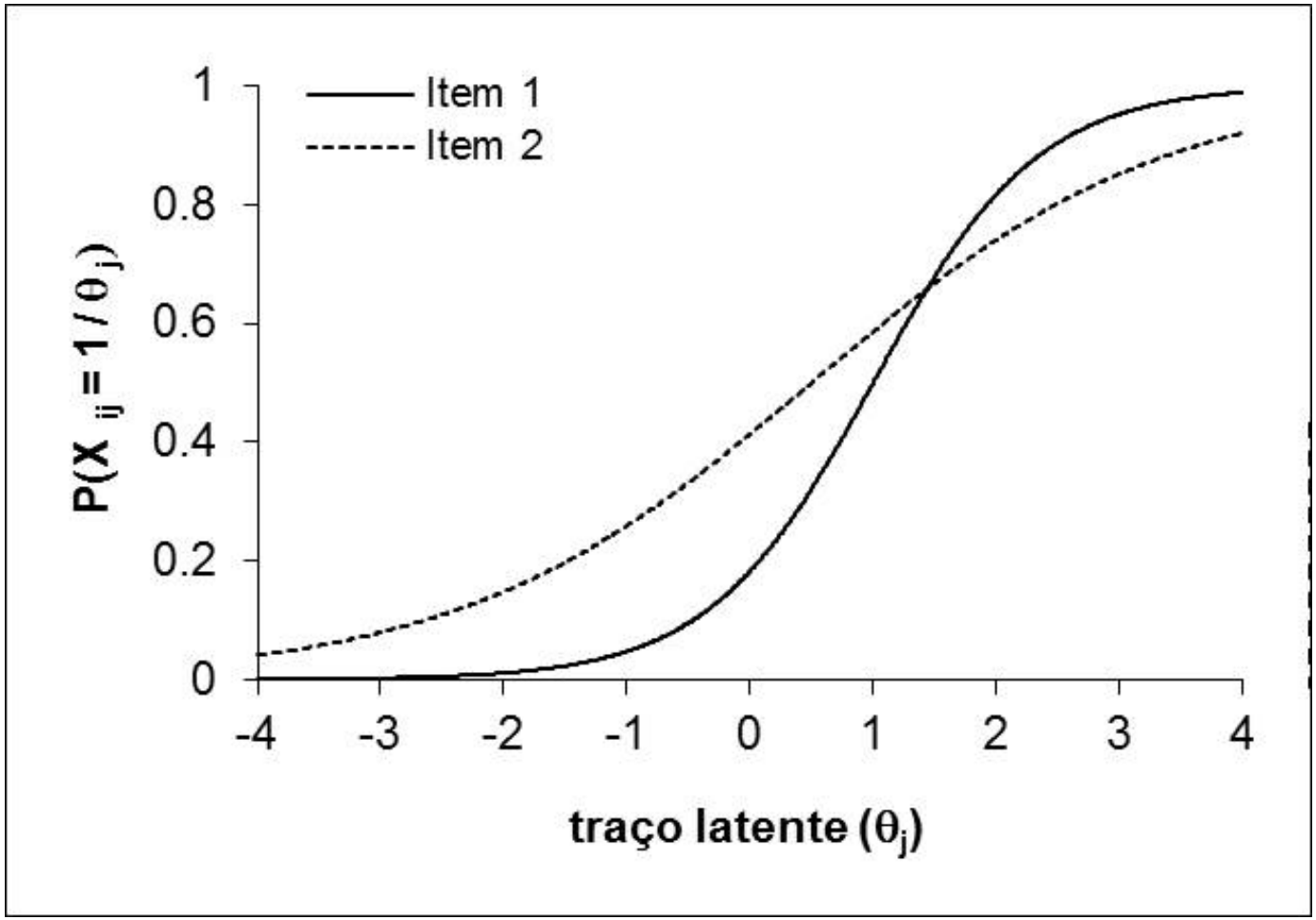






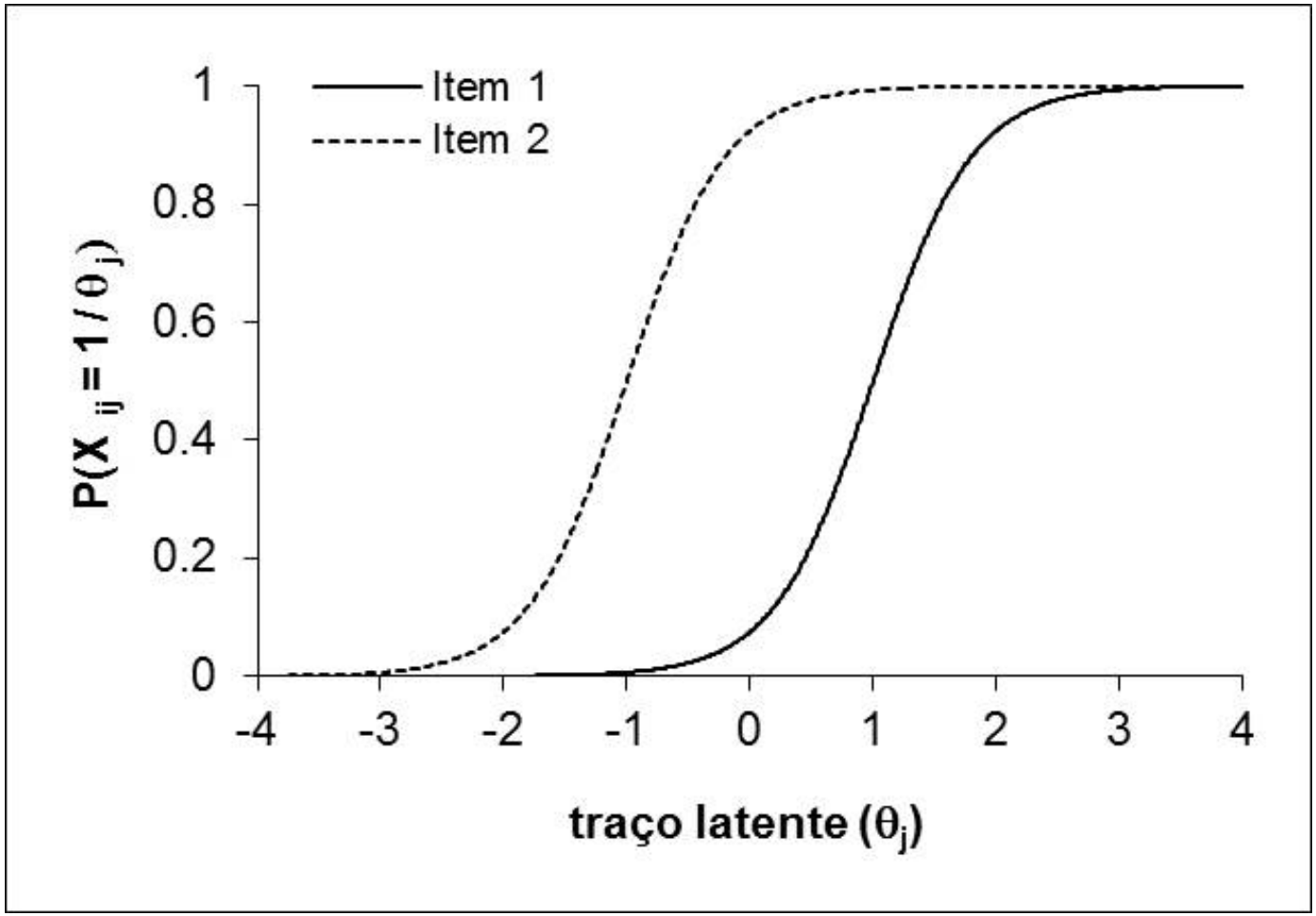


$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$



2. Modelos da TRI: ML1 (Rasch)

$$P(X_{ij} = 1 \mid \theta_j, b_i) = \frac{1}{1 + e^{-a(\theta_j - b_i)}}$$



$$P(X_{ij} = 1 \mid \eta_{ij}) = \int_{-\infty}^{\eta_{ij}} \frac{1}{\sqrt{2\pi}} e^{\left(\frac{-t^2}{2}\right)} dt.$$

equivale ao modelo logístico

$$P(X_{ij} = 1 \mid \eta_{ij}) = \frac{1}{1 + e^{(-\eta_{ij})}}.$$

com

$\eta_{ij} = \theta_j - b_i$ no modelo de Rasch,

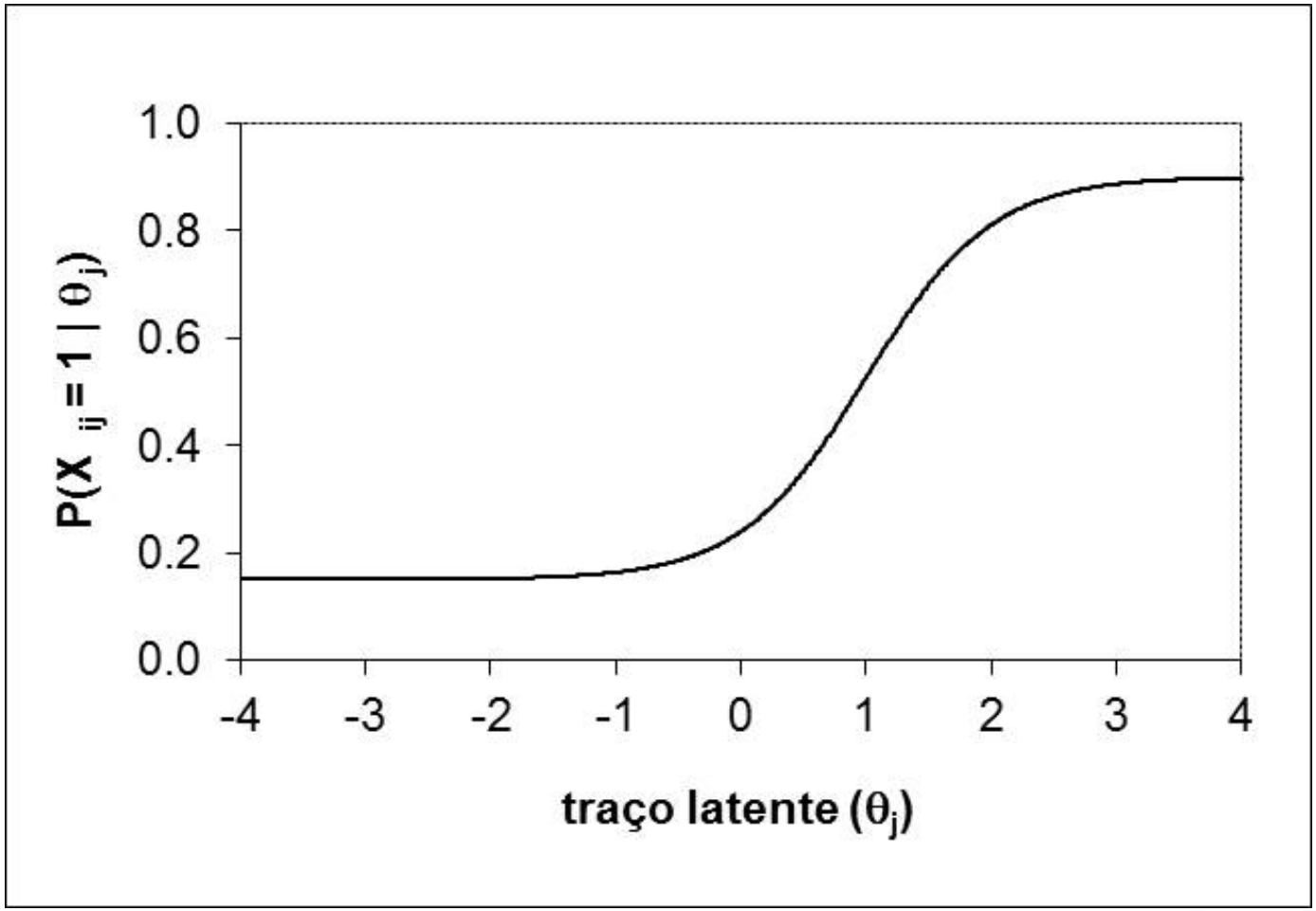
$\eta_{ij} = a_i(\theta_j - b_i)$ nos modelos de 2 e 3 parâmetros e

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = \int_{-\infty}^{a_i(\theta_j - b_i)} c_i + (1 - c_i) \frac{1}{\sqrt{2\pi}} e^{\left(\frac{-t^2}{2}\right)} dt,$$

no ML3

2. Modelos da TRI: ML4

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i, \gamma_i) = c_i + \frac{(\gamma_i - c_i)}{1 + e^{-Da_i(\theta_j - b_i)}}$$



2. Modelos da TRI: Samejima - modelo de resposta gradual

$$P_{ik}^{+}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_{ik})}}$$

k = 0,1, ..., m_i

m_i+1: n^o categorias do item i

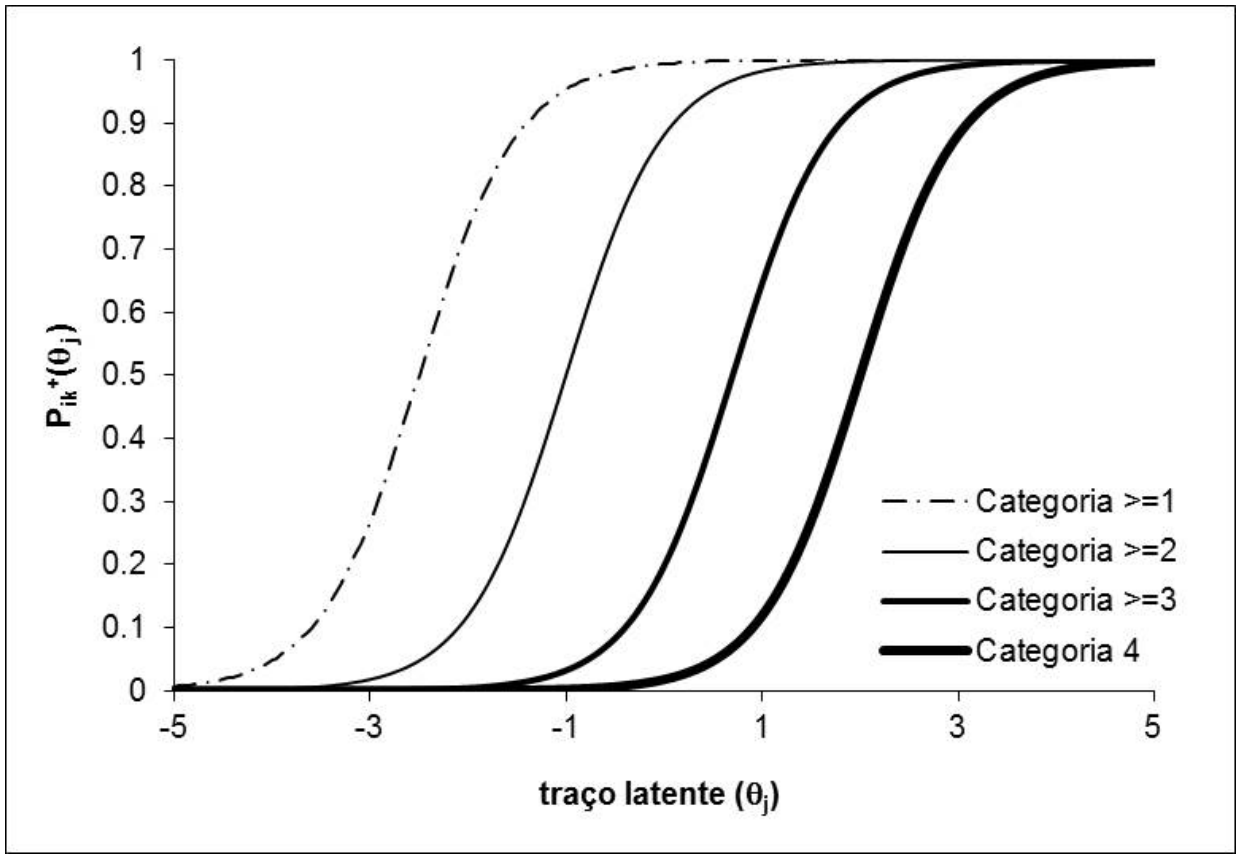
$P_{ik}^{+}(\theta_j)$: prob. de um indivíduo com traço latente θ_j escolher a categoria de resposta k ou qualquer outra de ordem acima de k no item i

a_i: parâmetro de discriminação comum a todas as categorias do item i

b_{ik}: parâmetro de gravidade que representa o nível latente necessário para a escolha da categoria de resposta acima de k com probabilidade igual a 0.50 $(b_{i1} \leq b_{i2} \leq ... \leq b_{im_i})$

$P_{i0}^{*}(\theta_j) = 1$ e $P_{ik+1}^{*}(\theta_j) = 0$.

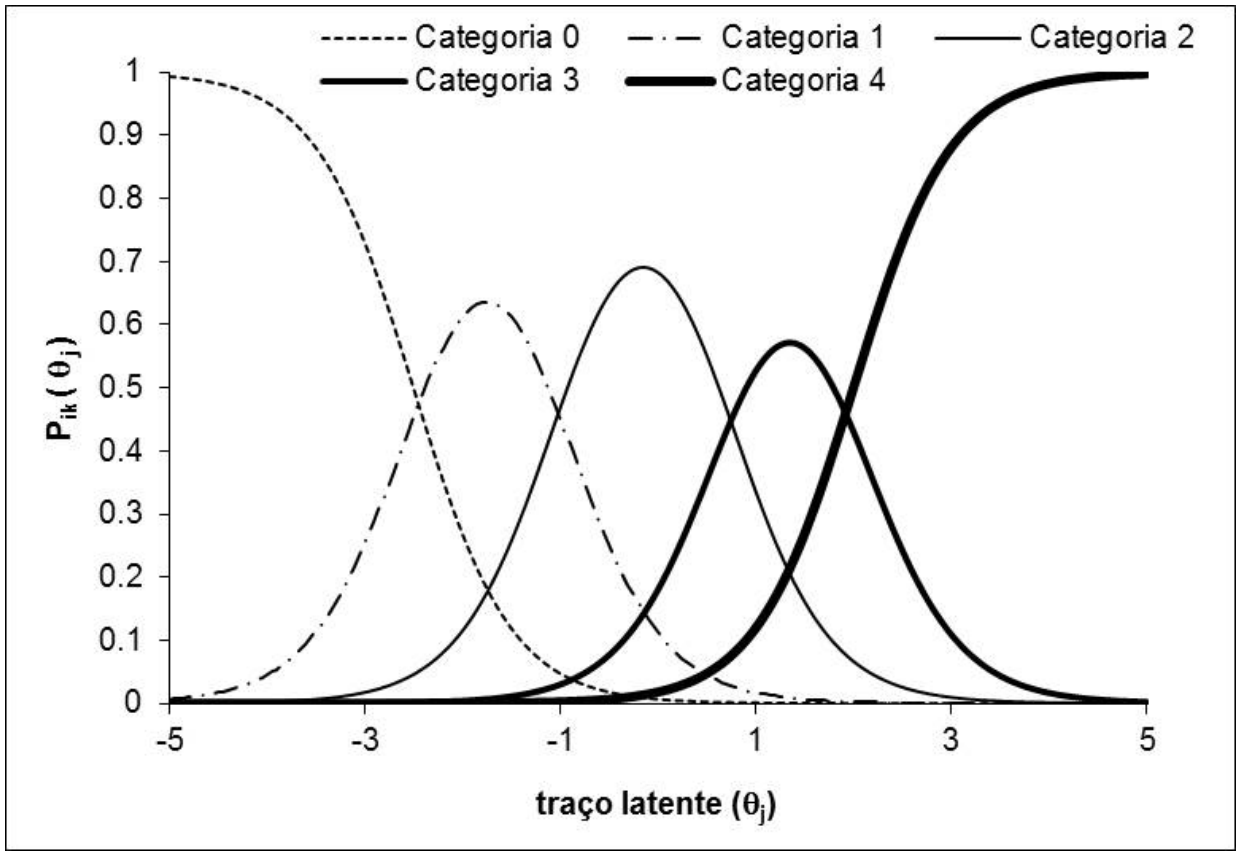
2. Modelos da TRI: Samejima - modelo de resposta gradual



$$P_{ik}(\theta_j) = P_{ik}^{+}(\theta_j) - P_{ik+1}^{+}(\theta_j)$$

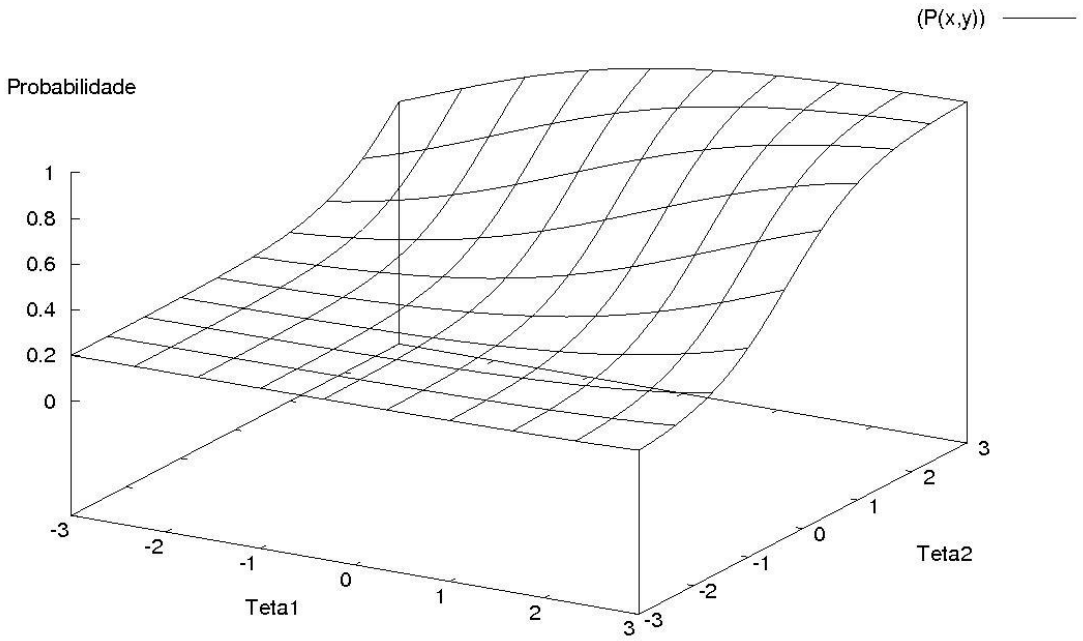
2. Modelos da TRI: Samejima - modelo de resposta gradual

$$P_{ik}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_{ik})}} - \frac{1}{1 + e^{-Da_i(\theta_j - b_{ik+1})}}$$



$$P(X_i = 1|\boldsymbol{\theta}, \mathbf{a}_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + \exp\left[-\sum_{k=1}^{p} a_{ki}\theta_k + b_i\right]},$$

com $\mathbf{a}_i = (a_{1i}, \ldots, a_{pi})$, p: número de traços latentes e
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.
Para $a_1 = 0,8$, $a_2 = 1,4$, $b = -2,0$ e $c = 0,2$



2. Modelos da TRI: Outros modelos

Ordinais

- Modelo de Escala Gradual
- Modelo de Crédito parcial
- Modelo de Crédito Parcial Generalizado
- Modelo Nominal

Referência: Andrade, Tavares e Cunha (2000)

Multidimensionais

- logísticos
- ogiva
- não compensatórios
- bifatorial

Referência: Reckase (1997), Li and Lissitz (2000), Rost and Carstensen (2002) e Gardner et al (2002)

$$P(X_{ijk} = 1 \mid \theta_{jk}, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1 + e^{-a_i(\theta_{jk} - b_i)}},$$

i=1, ..., I (itens)
j=1, ..., n_k (indivíduos no grupo k)
k=1, ..., g (grupos)

Referência: Bock, R.D., Zimowski, M.F. (1997). *Multiple group IRT. In Handbook of Modern Item Response Theory*. W.J. van der Linden and R.K. Hambleton Eds. New York: Springer-Verlag

Tipos de parâmetro $\begin{cases} \text{indivíduos: } \theta_j \\ \text{itens: } \zeta_i = (\ a_i, \quad b_i, \quad c_i\ )^t, \text{no ML3, por exemplo} \end{cases}$

Suposições:

- indep entre respostas de $\neq$ indiv
- indep entre respostas de $\neq$ itens condicionada a θ_j
- mesma probabilidade de seleção amostral
- dados omissos são não informativos

	θ_j	ζ_i
MV	X	conhecido
MV	conhecido	X
MV conjunta	X	X
MV marginal		X
MCMC	X	X

MV ou Bayesiano EAP ou MAP

Função de verossimilhança dos modelos unidimensionais dicotômicos:

$$L(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \prod_{i=1}^{I} \prod_{j=1}^{n} P_{ij}^{x_{ij}} (1 - P_{ij})^{(1-x_{ij})}$$

- Parâmetros de itens conhecidos
  $L(\boldsymbol{\theta}) : \frac{\partial log L(\boldsymbol{\theta})}{\partial \theta_j} = 0$

- Traços latentes conhecidos
  $L(\boldsymbol{\zeta}) : \frac{\partial log L(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_i} = \mathbf{0}$

Há necessidade de uso de processo iterativo
Não está definido para alguns padrões de resposta

Falta de identificabilidade: θ, ζ desconhecidos

Exemplo : indivíduo com θ = 1,20 na escala (0 ; 1) .
Qual sua habilidade na escala (200 ; 40) ?

$$a(\theta - b) = (a / 40) [(40 \times \theta + 200) - (40 \times b + 200)] = a^* (\theta^* - b^*)$$

Resposta : $\theta^* = 248$

1. $\theta^* = 40 \times \theta + 200$
2. $b^* = 40 \times b + 200$
3. $a^* = a / 40$
4. $P(X_i=1| \theta) = P(X_i=1| \theta^*)$

Solução: fdp para θ: $g(\theta/\tau)$

Etapa 1: Tornar a verossimilhança independente de θ_j e estimar ζ_i
Etapa 2: Estimar θ_j, considerando-se ζ_i conhecidos

População de indivíduos $\rightarrow$ seleção aleatória: $\theta_j \sim g(\theta \mid \boldsymbol{\eta})$
$\theta_j \sim N(0,1), \boldsymbol{\eta} = (\mu = 0, \sigma^2 = 1)$: define a métrica

Função de verossimilhança marginal:

$$L(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \prod_{j=1}^{n} \int_{\mathbb{R}} \prod_{i=1}^{I} P(X_{ij} = x_{ij} \mid \theta, \boldsymbol{\zeta})\, g(\theta \mid \boldsymbol{\eta})\, d\theta$$

$$\frac{\partial log L(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial \boldsymbol{\zeta}_i} = 0$$

Proposta Bock & Aitkin: estimar itens individualmente
Reestruturação EE + Hermite-Gauss: nós $\bar{\theta}_k$, k=1, ..., q.

3. Estimaçao: MVM

Derivação das fórmulas para o modelo ML1:
Slides Prof. Caio Lucidius Naberezny Azevedo - UNICAMP
http://www.ime.unicamp.br/∼cnaber/Material_TRI.htm

Estimação Frequentista - pag 43 à 47
Estimação Bayesiana - pag 1 à 7

Função de verossimilhança:

$$L(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \prod_{i=1}^{I} \prod_{j=1}^{n} P_{ij}^{x_{ij}} (1 - P_{ij})^{(1-x_{ij})}$$

Função de distribuição *a posteriori*:

$$f(\boldsymbol{\theta}, \boldsymbol{\zeta}) \propto L(\boldsymbol{\theta}, \boldsymbol{\zeta}) g(\boldsymbol{\theta} \mid \boldsymbol{\eta}) h(\boldsymbol{\zeta} \mid \boldsymbol{\tau})$$

$\uparrow$

distribuição estacionária de uma cadeia de Markov

com $g(\boldsymbol{\theta} \mid \boldsymbol{\eta})$ e $h(\boldsymbol{\zeta} \mid \boldsymbol{\tau})$ distribuições *a priori*

3. Estimação: Múltiplos Grupos

- Diferentes grupos: séries, turnos, países
- Grupos definidos previamente

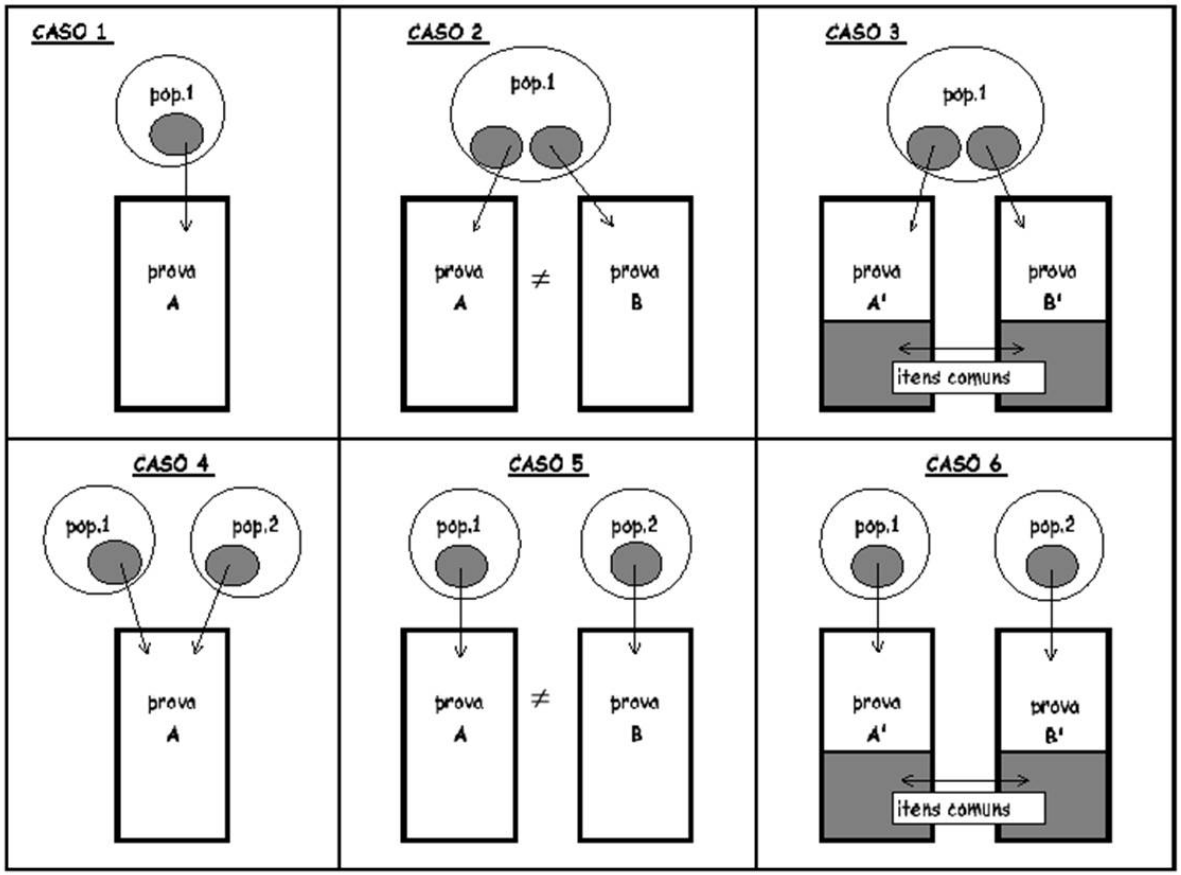
$$P(X_{ijk} = 1 \mid \theta_{jk}, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1 + e^{-a_i(\theta_{jk} - b_i)}},$$

i=1, ..., I (itens)
j=1, ..., n_k (indivíduos no grupo k)
k=1, ..., g (grupos)

- Estimação: $\theta_{jk} \mid \boldsymbol{\eta}_k \sim N(\mu_k, \psi_k)$
- Identificabilidade: $\mu_1 = 0, \psi_1 = 1$
- Estimam-se (μ_k, ψ_k), para $k = 2, ..., g$

4. Equalização

- Colocar itens de provas distintas ou habilidades de populações diferentes numa mesma escala, podendo ser comparados



Provas: - apenas itens novos
- apenas itens já calibrados
- itens novos e já calibrados

4. Equalização: itens novos

- via TRI: estimação de todos os dados conjuntamente equaliza em todos os casos MENOS no 5
- caso 5: resultados em métricas diferentes; sem comparações
- casos 4 e 6: modelos para múltiplos grupos
- caso 6: representa o melhor exemplo do uso da equalização e o maior avanço da TRI sobre a Teoria Clássica

4. Equalização: caso 6

- Quantos itens comuns?
- Depende do tipo de equalização e da qualidade dos itens comuns
- Quanto maior o parâmetro de discriminação dos itens e quanto mais próximos estiverem os parâmetros de dificuldade dos itens da média da população avaliada, menor o número de itens comuns para uma boa equalização
- Ex: 2 provas de 30 itens - pelo menos 6 em comum

4. Equalização: SAEB

SAEB: Sistema Nacional de Avaliação da Educação Básica

- bienal desde 1995
- séries: 4a. e 8a. do EF e 3a. do EM
- uma análise para cada disciplina
- itens de múltipla escolha (95: itens 0,1,2)
- um grande número de itens para cobrir a grade curricular
- provas diferentes para uma mesma série/disciplina (BIB)
- aluno faz somente uma das provas de uma das disciplinas
- http://www.inep.gov.br/basica/saeb/

4. Equalização: SAEB

- O número de itens requerido pelos especialistas, para cada série e disciplina, é maior do que um estudante pode responder em 2 horas.
- Equalização: obter resultados comparáveis (mesma escala) para as 4a., 8a. and 3a. séries e também ao longo do tempo.
- Matemática, 3a. série: 169 itens.
  - 13 blocos com 13 itens cada (169=132)
  - Provas: 26 cadernos de provas com 3 blocos
    total de 39=3x13 itens
  - 130 itens "não apresentados"a cada um dos alunos
  - Cadernos possuem itens comuns com alguns dos outros
  - Blocos comuns e/ou itens já aplicados em anos anteriores
  - Blocos da 4a. série na 8a. série
  - Blocos da 8a. série na 3a. série do ensino médio

Blocos Incompletos Balanceados - BIB

Cadernos de provas	Conjuntos de itens			Cadernos de provas	Conjuntos de itens		
1	1	2	5	**14**	1	3	8
2	2	3	6	**15**	2	4	9
3	3	4	7	**16**	3	5	10
4	4	5	8	**17**	4	6	11
5	5	6	9	**18**	5	7	12
6	6	7	10	**19**	6	8	13
7	7	8	11	**20**	7	9	1
8	8	9	12	**21**	8	10	2
9	9	10	13	**22**	9	11	3
10	10	11	1	**23**	10	12	4
11	11	12	2	**24**	11	13	5
12	12	13	3	**25**	12	1	6
13	13	1	4	**26**	13	2	7

> Cada conjunto de 13 itens aparece em 6 cadernos de provas
> Cada conjunto de itens aparece duas vezes em cada uma das 3 posições nos cadernos de provas
> Um par de conjuntos de itens aparece somente uma vez em um caderno de provas

4. Equalização: itens já calibrados

- Desejamos apenas estimar as habilidades dos indivíduos
- Situação comum devido à criação de Bancos de Itens
- conjunto de itens que já foram testados e calibrados a partir de um número significativo de sujeitos de uma dada população
- parâmetros "conhecidos"
- As habilidades estimadas a partir de itens do banco estarão na mesma métrica do grupo de indivíduos utilizados na calibração inicial

4. Equalização: itens novos + já calibrados

- Situação comum devido à ampliação de Bancos de Itens
- continuamente em formação / itens saem e itens entram no banco
- Problema:
  itens novos devem ser calibrados na mesma métrica de itens do banco: programas computacionais específicos
- Objetivos :
  criar e testar itens novos
  comparar o desempenho da rede pública estadual de São Paulo com o desempenho nacional, por ex

4. Equalização: *a posteriori*

- Pode ser feita quando há itens comuns entre 2 populações
- Calibra-se separadamente 2 conjuntos de itens, que foram submetidos a 2 populações de interesse
- Para os itens comuns, teremos 2 conjuntos de estimativas, cada uma na métrica de suas respectivas populações

4. Equalização: *a posteriori*

- Estabelece-se algum tipo de relação (preferencialmente simétrica) que permita colocarmos os parâmetros de um dos conjuntos de itens na escala do outro
- Utiliza-se essa relação para transformar os parâmetros de todos os itens (comuns e não comuns) de um conjunto na escala do outro
- Com todos os itens na mesma métrica, pode-se estimar as habilidades de todos os respondentes, que também estarão na mesma escala

4. Equalização: *a posteriori*

- Pela propriedade de invariância, temos:
  $b_1 = \alpha * b_2 + \beta$ e $a_1 = \frac{1}{\alpha} * a_2$
- Uma vez determinados os coeficientes α e β, as estimativas dos parâmetros dos itens do grupo 2 podem facilmente ser colocados na mesma escala das estimativas do grupo 1

4. Equalização: Método Média-Desvio

É um método simétrico

$\alpha = \frac{S_1}{S_2}$ e $\beta = \bar{X}_1 - \bar{X}_2$, em que

S_1 e S_2 são os desvios padrão e
$\bar{X}_1$ $\bar{X}_2$ são as médias amostrais

das estimativas dos parâmetros de dificuldade dos itens comuns nos grupos 1 e 2, respectivamente.

Para a equalização das habilidades:

$$\theta_1 = \alpha * \theta_2 + \beta$$

em que θ_i é a habilidade na escala do grupo i.

5. Simulações

Perguntas frequentes:

1. Quem acerta mais itens tem sua estimativa de habilidade maior?
2. Como a presença do parâmetro de "acerto ao acaso" (c) influencia na estimativa da habilidade?
3. Responder "fora do padrão esperado" (acertar as difíceis e erras as fáceis) diminui a estimativa da habilidade?
4. Duas estimativas de habilidade de um mesmo indivíduo feitas a partir de respostas a provas diferentes geram valores equivalentes?

5. Simulações: Ogiva Normal 2 parâmetros e ML2

Samejima, F. (2000). Logistic positive exponent family of models: virtue of asymmetric item characteristic curves. **Psychometrika, 65**. 319-335

- EE para θ: $\sum_{i=1}^{I} a_i X_i = \sum_{i=1}^{I} a_i P_i(\theta)$
- $\hat{\theta} \uparrow$ com $\sum_{i=1}^{I} a_i$
- 5 itens: **a** =(1 1 1 1 1), **b** =(-3 -1,5 0 1,5 3)
- No modelo de ogiva normal:
  - para apenas 1 item correto: $\hat{\theta} \uparrow$ quando b item acertado $\uparrow$
  - para apenas 1 item incorreto: $\hat{\theta} \uparrow$ quando b item errado $\uparrow$
  - não há uma regra simples determinante da posição relativa de duas estimativas de habilidades para diferentes padrões de respostas
- No modelo logístico:
  - essa contradiçao não ocorre
  - quanto maior for a do item correto, maior será $\hat{\theta}$ - para itens com mesma discriminação, $\hat{\theta} \uparrow$ com o número de acertos
  - porém, a dificuldade do item não é levada em consideração para estimar θ

5. Simulações: Ogiva Normal 2 parâmetros e ML2

	Response Pattern	Normal Ogv.	Logistic
1	00000	neg. infinity	neg. infinity
2	10000	−2.28385	−2.28753
3	01000	−2.27016	−2.28753
4	00100	−1.84831	−2.28753
5	00010	−1.34811	−2.28753
6	01100	−1.15759	−0.75260
7	00001	−0.86577	−2.28753
8	11000	−0.75034	−0.75260
9	10100	−0.75021	−0.75260
10	01010	−0.75013	−0.75260
11	00110	−0.75011	−0.75260
12	00101	−0.36062	−0.75260
13	10010	−0.34310	−0.75260
14	01001	−0.27309	−0.75260
15	00011	−0.19116	−0.75260
16	01110	−0.15292	0.75260
17	10001	0.15292	−0.75260
18	00111	0.19116	0.75260
19	01101	0.27309	0.75260
20	10110	0.34310	0.75260
21	01011	0.36062	0.75260
22	10011	0.75011	0.75260
23	10101	0.75013	0.75260
24	11010	0.75021	0.75260
25	11100	0.75034	0.75260
26	01111	0.86577	2.28753
27	11001	1.15759	0.75260
28	10111	1.34811	2.28753
29	11011	1.84831	2.28753
30	11101	2.27016	2.28753
31	11110	2.28385	2.28753
32	11111	pos. infinity	pos. infinity

Minhas simulações: Resultados Simulações no Excel
Artigo Caio

Perguntas frequentes:

1. Quem acerta mais itens tem sua estimativa de habilidade maior? SIM
2. Como a presença do parâmetro de "acerto ao acaso" (c) influencia na estimativa da habilidade? $c \uparrow$ implica $\hat{\theta} \downarrow$
3. Responder "fora do padrão esperado" (acertar as difíceis e erras as fáceis) diminui a estimativa da habilidade? depende do modelo: no ML2 NÃO, mas no ML3 SIM
4. Duas estimativas de habilidade de um mesmo indivíduo feitas a partir de respostas a provas diferentes geram valores equivalentes? SIM

6. Interpretação da Escala

- métrica arbitrária para parâmetros dos itens e habilidades
- define a ordem, mas não o significado prático
- ex: na escala (0,1), qual a interpretação de $\theta = -0,8$ versus $\theta = 1,5$

Para interpretação:

- criação de escalas de conhecimento que tornam possível a interpretação pedagógica dos resultados
- definição de níveis âncora e itens âncora

Níveis âncora: pontos selecionados na escala da habilidade para serem interpretados pedagogicamente

Item âncora

Considere 2 níveis âncora consecutivos θ_1 e θ_2, com $\theta 1 < \theta 2$

Um item i é âncora no nível θ_2 se, e somente se:

- $P(X_i = 1 \mid \theta = \theta_2) \geq 0,65$
- $P(X_i = 1 \mid \theta = \theta_1) < 0,50$
- $P(X_i = 1 \mid \theta = \theta_2) - P(X_i = 1 \mid \theta = \theta_1) \geq 0,30$

	Níveis Âncora									
Item	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
1							sim			
2					sim					
3									sim	
4						sim				
5		sim							sim	
6				sim						
7						sim				
8								sim		
9							sim			
10				sim						

6. Interpretação da Escala: um exemplo

Slides Raquel da Cunha Valle
Fundação Carlos Chagas

7. Aplicação: *Softwares Computacionais*

- SAS, SPSS, Stata? Não
- Programas individuais em Splus, R e SAS
- No R: ltm, mirt
- Testfact
- Bilog-MG
- Xcalibre
- Multilog
- Parscale
- Noharm
- WinBUGS
  Bayesian Modeling - Jorge L. Bazàn
  (http://argos.pucp.edu.pe/∼jlbazan/software.html)

7. Aplicação: PISA

- Programa Internacional de Avaliação dos Estudantes (PISA) é aplicado a alunos na faixa dos 15 anos, idade na qual a maioria dos estudantes finalizam a escolaridade básica obrigatória
- realizado a cada 3 anos desde 2000
- disciplinas: Leitura, Matemática e Ciências
- planejamento BIB
- amostragem complexa: estrato (UF) e 3 subestratos (pública/privada, rural/urbana e IDH)
- modelo de Rasch

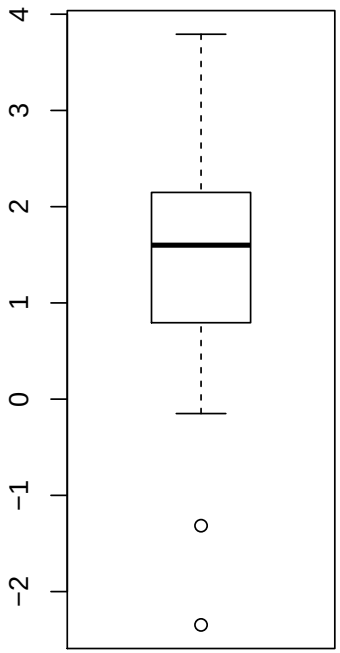
7. Aplicação: PISA

- PISA 2009 - Brasil
- 20127 estudantes brasileiros participaram (destes, 4000 prova informatizada)
- Apenas as questões de matemática (35 questões)
- 6 provas diferentes (B8=B12 e B10=B27)

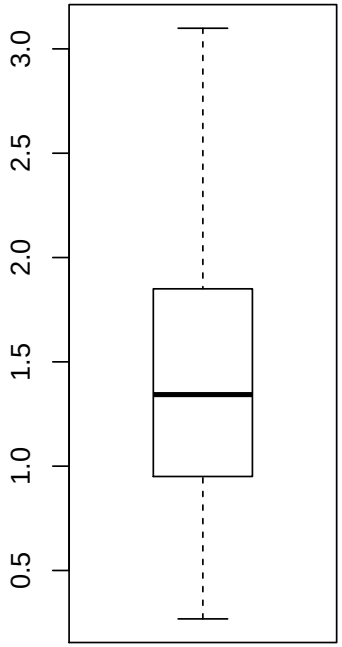
Tabela: Estimativas dos parâmetros de itens - PISA 2012 Brasil

Item	b	Erro Padrão	a	Erro Padrão
MAT01	-0,150	0,0304	1,020	0,0456
MAT02	1,773	0,0533	1,633	0,0754
MAT03	0,791	0,0279	1,707	0,0662
MAT04	0,772	0,0266	1,830	0,0713
MAT05	2,281	0,0809	1,437	0,0749
MAT06	1,097	0,0439	1,136	0,0488
...	...	...	...	...
MAT35	1,793	0,0479	2,095	0,1034
Média	1,469		1,434	

7. Aplicação: PISA



Dificuldade



Discriminação

7. Aplicação: PISA

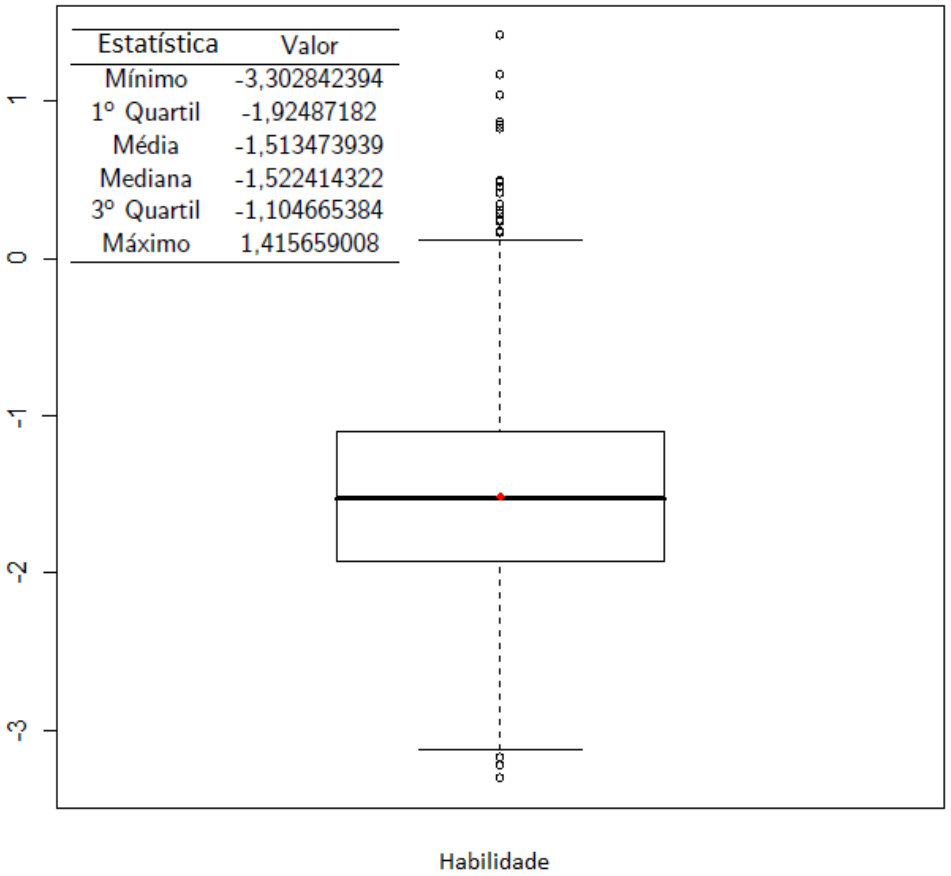


Tabela: Itens Âncora

Item	Níveis Âncora (θ)						
	-3	-2	-1	0	1	2	3
MA1							
MA2						sim	
MA3					sim		
MA4							sim
MA5						sim	
MA6							sim
...							
MA35						sim	

7. Aplicação: Outras Aplicações

- PorSimples
- BDI
- CAT - BDI

FIM