

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, análises 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 aprendiz. 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.

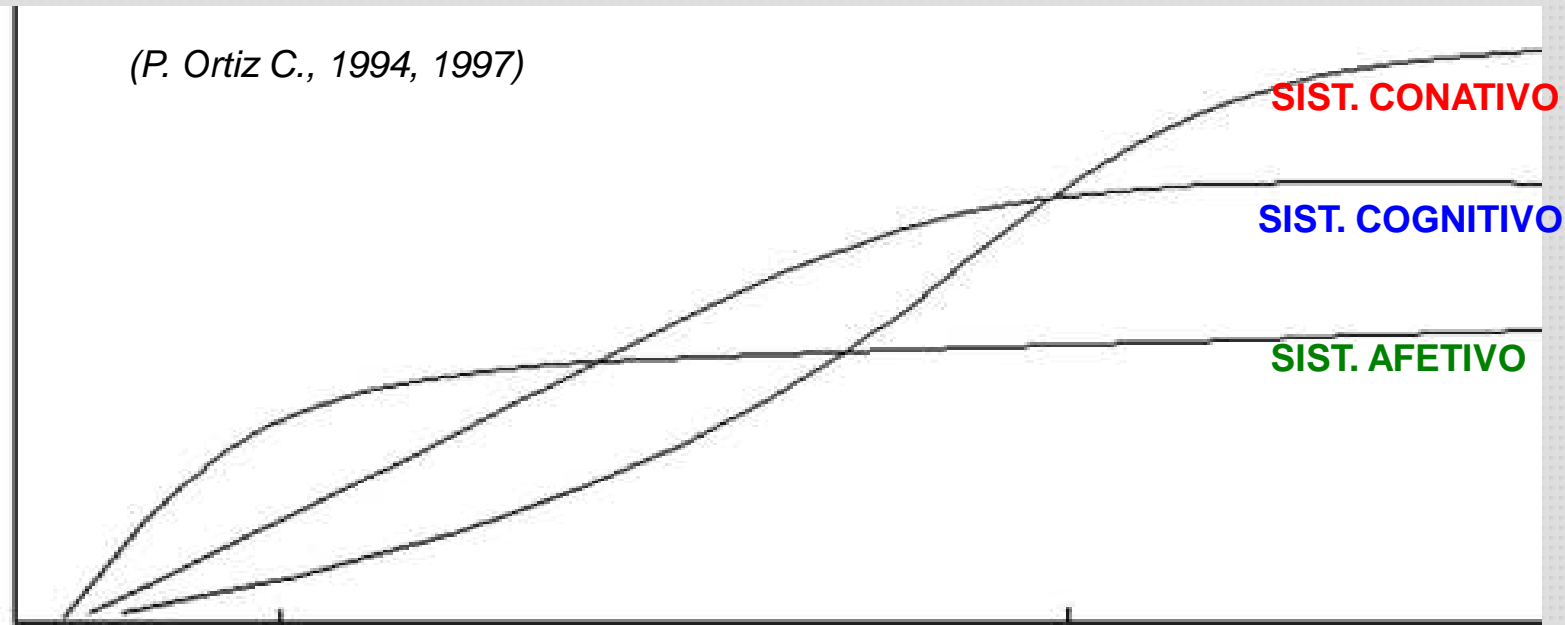
Estas avaliações têm objetivos e procedimentos diferenciados das avaliações tradicionais de salas de aula. Por exemplo para certificação, o credenciamento, o diagnóstico e a prestação de contas.

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

(P. Ortiz C., 1994, 1997)



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?

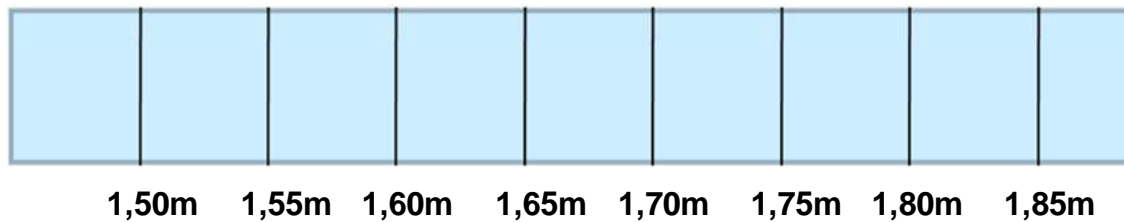
- É o campo de estudo relacionado com a teoria e técnica da medição psicológica, incluindo a medição de conhecimentos, habilidades, atitudes e traços de personalidade e a medição educacional.
- A área está principalmente associada com a construção e validação de instrumentos de medição, como questionários, provas, escalas, inventários e testes, entre outros.
- A psicometria tem duas 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 psicometria devem ter pelo menos um Maestría, e a maioria tem Doutorado.
- Por causa de que a Psicometria 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 é infrequente encontrar muitos especialistas com 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 psicometria 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 país 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.

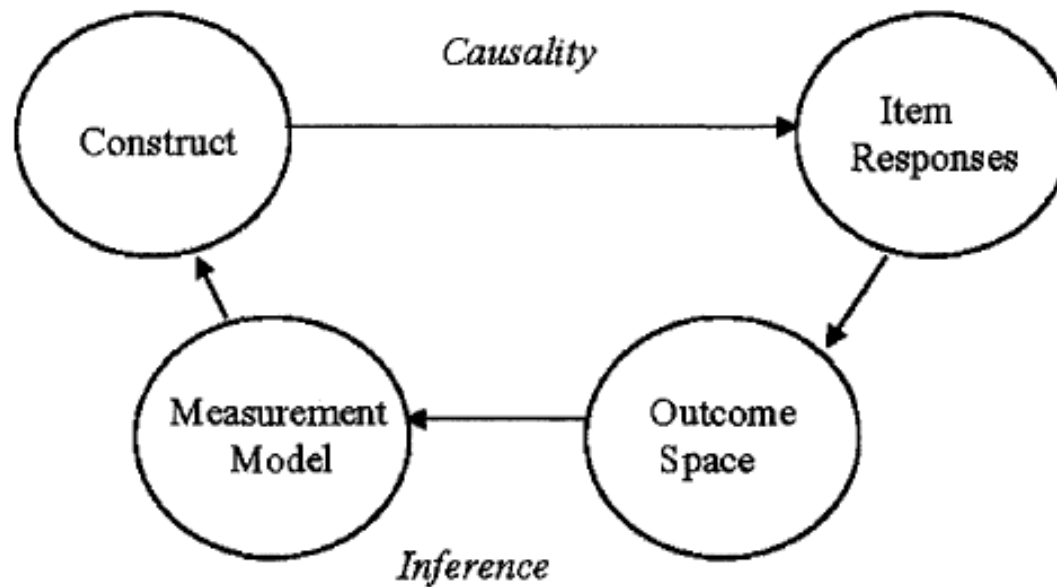


FIG. 1.8 The “four building blocks” showing the directions of causality and inference.

Wilson, Mark (2005). Constructing Measures: An Item Response Modeling Approach. Lawrence 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?

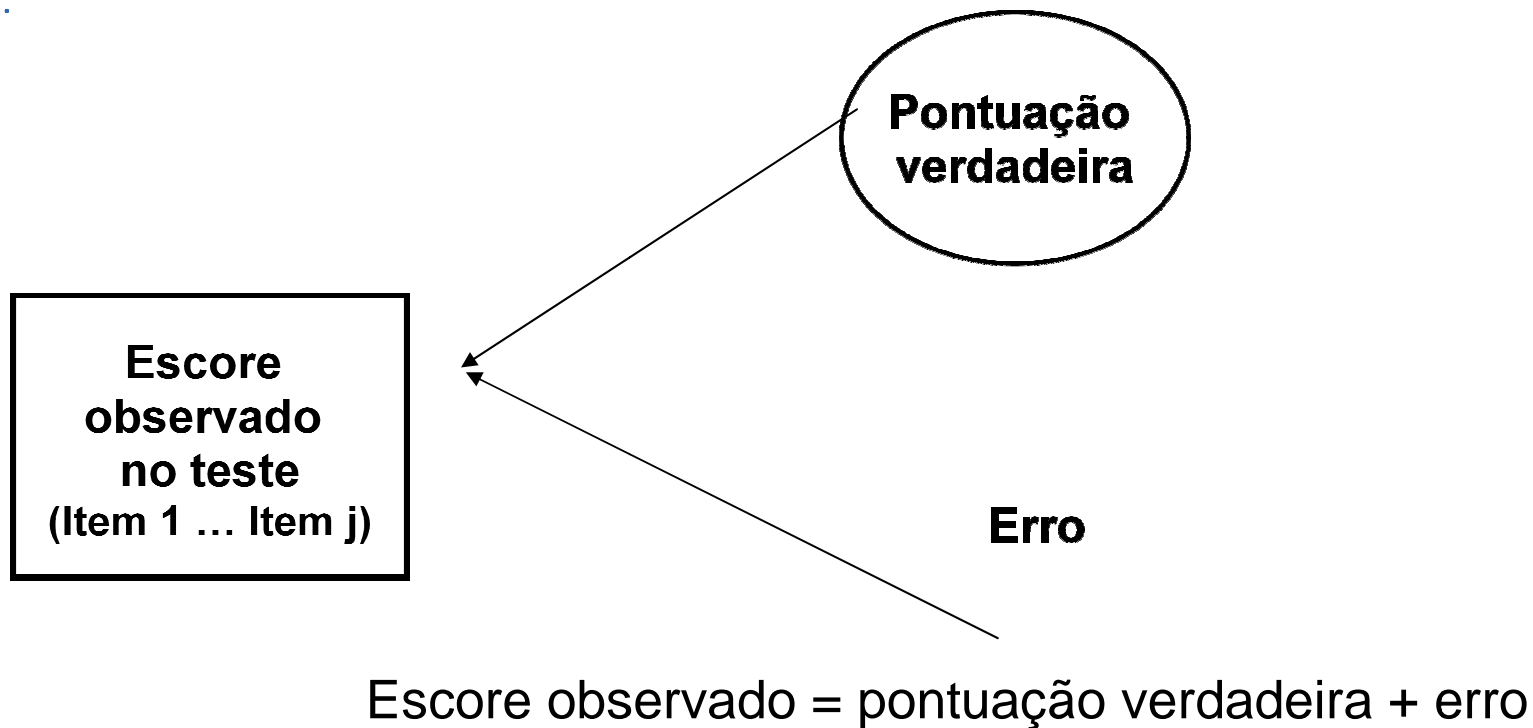
Utilizado para relacionar as variáveis observadas, registradas e medidas (respostas aos itens) com as variáveis latentes (habilidade).

- 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.



- 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 análise 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 análises de itens como proporção de acerto, porcentagem de omissão, discriminação, correlação pergunta-prova, alfa de cronbach 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.

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

- Esta teoria é válida 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-tot al 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 estatísticos (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

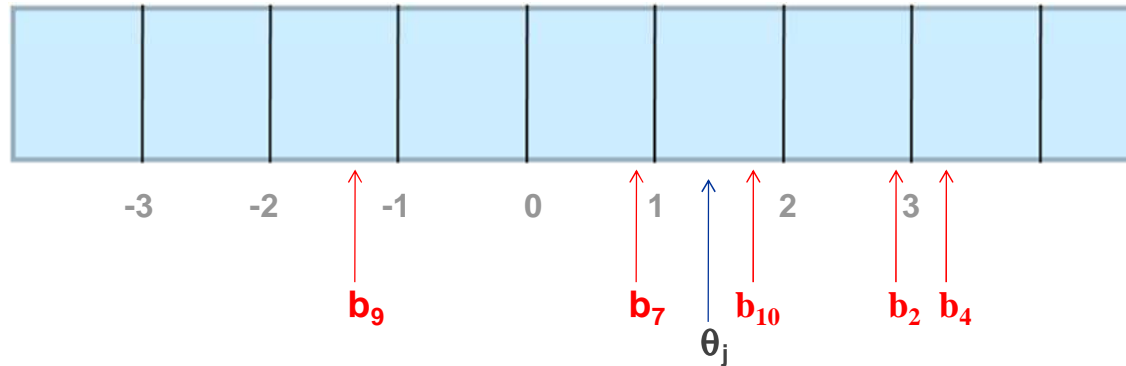
5. MODELOS DE RESPUESTA AL ITEM

De acordo com a chamada Teoria de Resposta ao Item (TRI), o interesse primário está 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 mais 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")

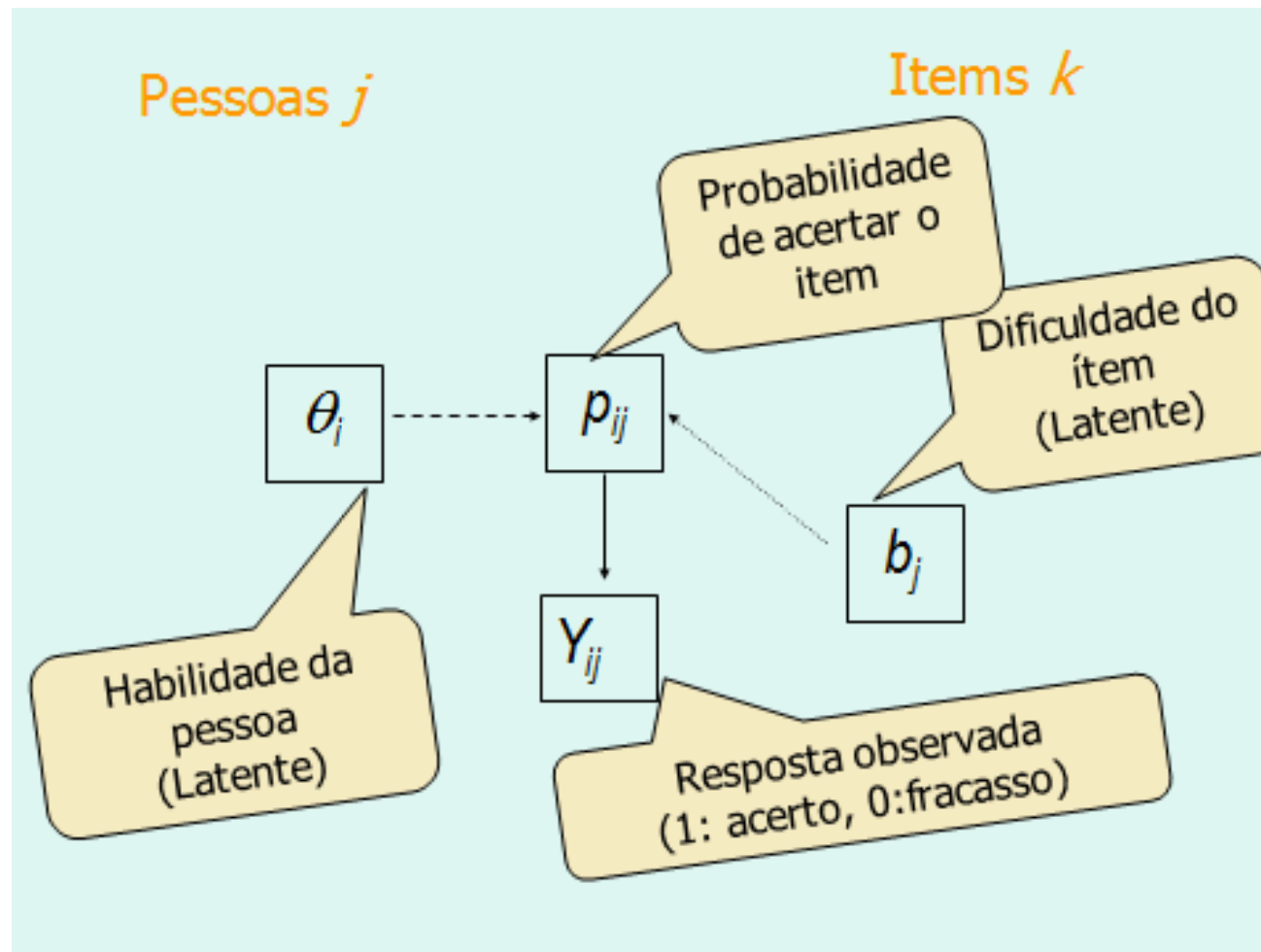
b_i : dificuldade do item "traço latente" (parâmetro do item)

> 0 examinado está "acima" do item

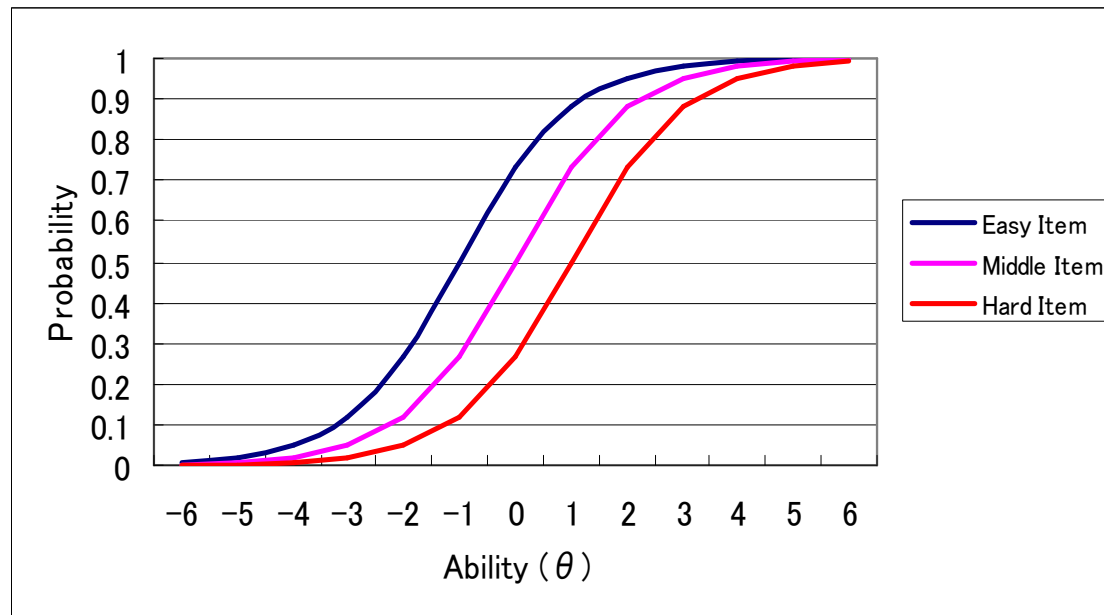
$(\theta_j - b_i) \approx 0$ é considerado "próximo" do item

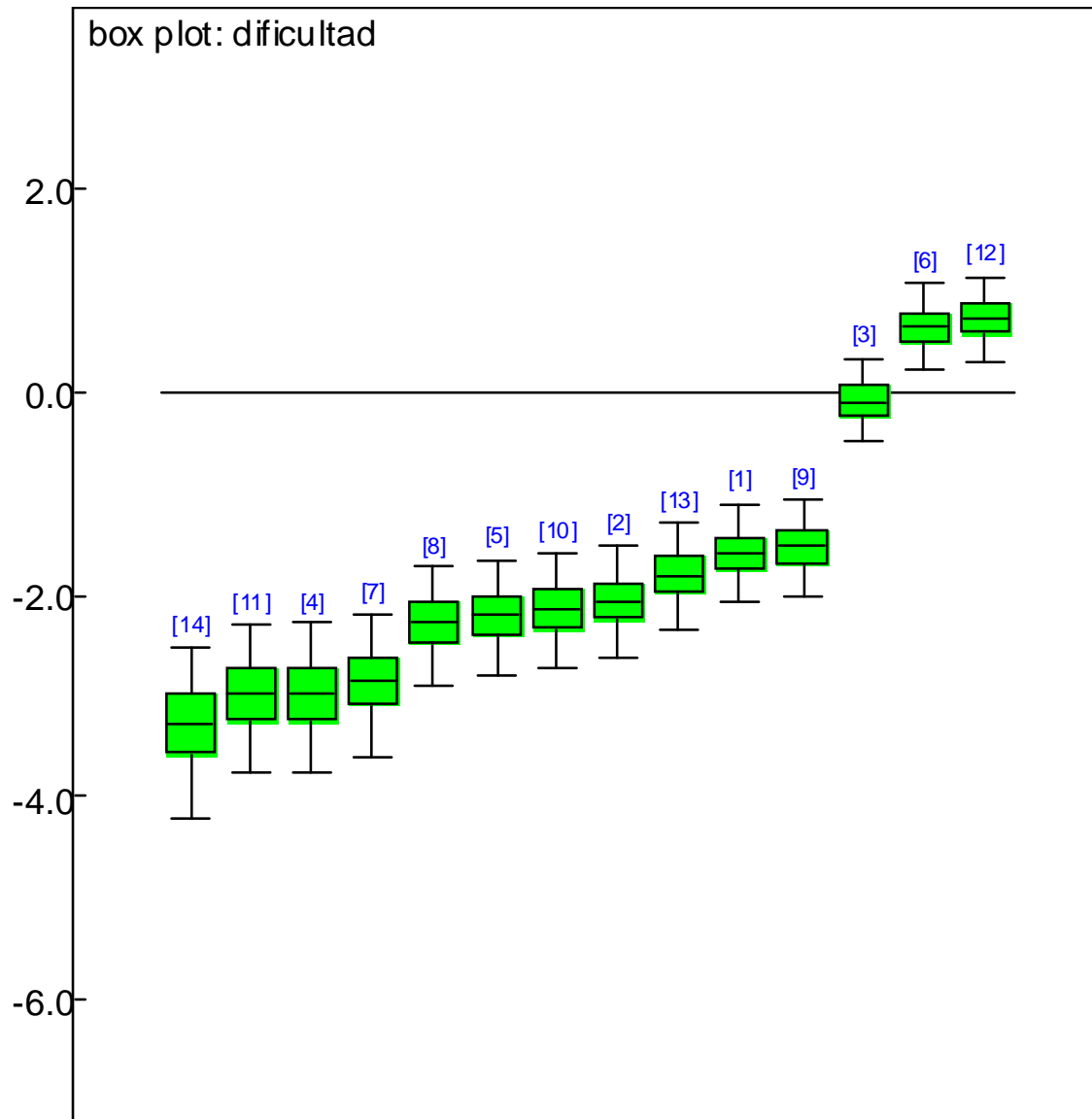
< 0 é considerado "abaixo" do item

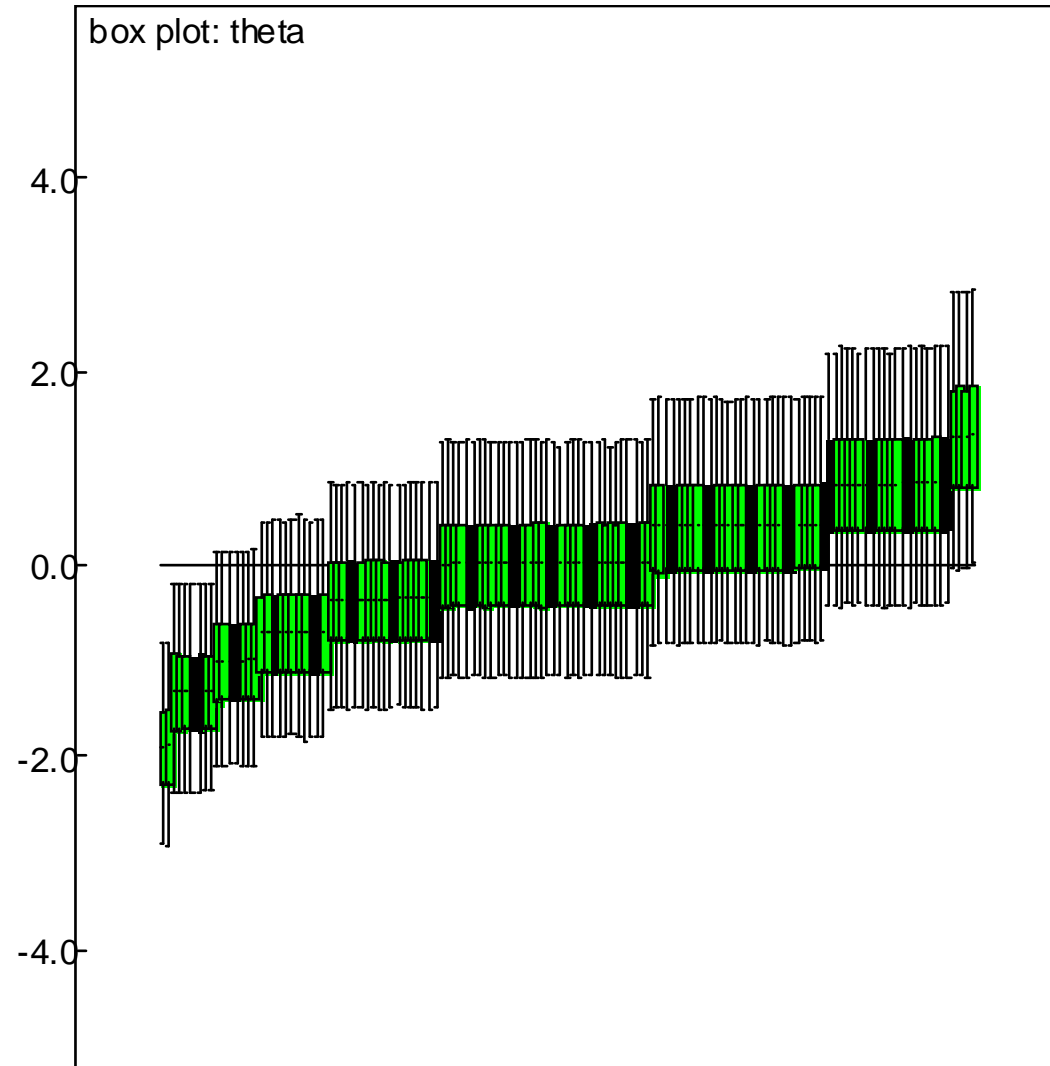
- Baseado nas respostas dos itens de um teste desejasse estimar:
 - parâmetros dos Itens (calibração)
 - Traços latentes dos examinados
 - Parâmetros da população (distribuição dos traços latentes): média, desvio padrão, etc
- A probabilidade de uma resposta "correta" para um item é modelada como função da habilidade do examinando e os parâmetros do item.



Curvas características dos itens







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 computerized 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 of 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 personal 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 mais predominante
- A TRI trata novos problemas como multidimensionalidade, diferenciabilidade, testes de adaptação, teste de velocidades (speededness), 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) $\frac{1}{2} > \frac{3}{4}$ | B) $\frac{7}{6} < \frac{1}{2}$ | C) $\frac{3}{4} > \frac{7}{6}$ | D) $\frac{1}{2} < 1\frac{1}{4}$ |
|--------------------------------|--------------------------------|--------------------------------|---------------------------------|

3. Um metro de pano custa S.l. 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: $960 \div 87$ 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: $\frac{2}{3} + \frac{3}{4}$

- A) $\frac{5}{7}$ B) $\frac{17}{7}$ C) $\frac{6}{12}$ D) $\frac{17}{12}$

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

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

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

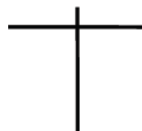


Figura 2

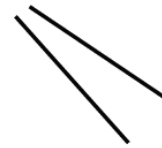


Figura 3



Figura 4

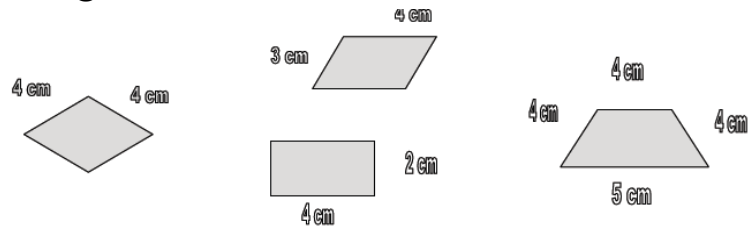
A) A figura
4

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>