Teste adaptativo computadorizado nas avaliações
educacionais e psicológicas

Thales Akira Matsumoto Ricarte

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Teste adaptativo computadorizado nas avaliações educacionais e psicológicas

Thales Akira Matsumoto Ricarte

Orientadora: Profa. Dra. Mariana Cúri

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

USP – São Carlos
Maio de 2013

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

R487t Ricarte, Thales Akira Matsumoto
Teste adaptativo computadorizado nas avaliações
educacionais e psicológicas / Thales Akira Matsumoto
Ricarte; orientadora Mariana Cúri. -- São Carlos,
2013.
67 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática
Computacional) -- Instituto de Ciências Matemáticas
e de Computação, Universidade de São Paulo, 2013.

1. Teste adaptativo computadorizado. 2. Teoria
de resposta ao item. 3. Modelo de resposta gradual.
4. Modelo de Samejima. I. Cúri, Mariana, orient.
II. Título.

Agradecimentos

Gostaria de agradecer

Aos meus pais, Juarez e Satie, e aos meus irmãos, Fabio e Rosana por todo o apoio moral e financeiro nestes 28 anos que os conheço.

À minha orientadora prof. Mariana, pelos conselhos e ensinamentos durante o mestrado, desejo à sua filha muita saúde.

À todos os meus professores do ICMC que transmitiram o conhecimento e me inspiraram a fazer o programa de mestrado.

Aos meus amigos de São José dos Campos, Jun, Duda e Cidão e aos meus amigos de São Carlos que conheci na graduação e no mestrado, por todos esses anos de convívio que passamos, estudando e saindo juntos, trocando conhecimento e experiências de vida.

Ao seu Ari e sua esposa Cidinha por toda ajuda que precisei, além das conversas que temos esporadicamente.

À CNPq pelo apoio financeiro.

E à todos que não citei que me ajudaram na execução deste trabalho.

Sumário

1	Introdução	1
2	Modelos da Teoria de Resposta ao Item	5
2.1	Modelo logístico de 3 parâmetros unidimensional	5
2.2	Modelo logístico de 4 parâmetros	8
2.3	Modelo de resposta gradual unidimensional	8
2.4	Modelo de escala gradual	10
2.5	Modelo de Crédito Parcial Generalizado	11
2.6	Modelo de Ogiva Normal Multidimensional	12
2.7	Modelo Logístico de 3 parâmetros Multidimensional	13
3	Calibração de itens	15
3.1	Máxima Verossimilhança Marginal (MVM)	15
3.2	Monte Carlo via Cadeia de Markov (MCMC)	19
4	Estimação do traço latente	21
4.1	Máxima verossimilhança	21
4.2	Métodos Bayesianos	22
5	Seleção de itens	24
5.1	Informação de Fisher	24
5.2	Informação de Kullback-Leibler	25
5.3	Informação a <i>posteriori</i> Ponderada	26

5.4	Abordagem <i>Shadow Test</i>	26
5.5	Critério de parada	32
6	Simulação	33
6.1	Banco Grande	33
6.2	Banco pequeno	36
7	Aplicação	41
7.1	Exame de Proficiência em Inglês	41
7.2	Introdução ao Inventário de Depressão de Beck	46
8	Programa	50
9	Conclusão e Estudos Futuros	59

Lista de Tabelas

5.1	Descrição do banco de itens do EPI.	28
5.2	Notação para descrever um banco de itens usado em um <i>Shadow test</i>	29
5.3	Notação para descrever um banco de itens usado em um <i>Shadow test</i> no exemplo EPI.	31
6.1	Percentual de seleção dos itens do BDI para um TAC com 10 itens	39
6.2	Percentual de itens do BDI para um TAC com 5 itens	39
7.1	Valores estimados dos parâmetros dos itens do EPI	44
7.2	Valores estimados dos parâmetros dos itens do BDI	47

Lista de Figuras

1.1	Proporção de resposta correta para itens do teste de Binet e Simon pela idade	2
2.1	Curva característica de um item sob o ML3.	6
2.2	Curva característica de um item sob o ML4.	8
2.3	Sob o MRG, P_k é a probabilidade de escolha da categoria k	10
2.4	Exemplo de superfície do M3PL considerando 2 traços latentes	14
6.1	Diferença entre os valores reais θ e estimados $\hat{\theta}$ dos traços latente para os TACs de 25 e 5 itens do banco grande (500 itens).	35
6.2	Diferença entre os valores reais (θ) e estimados ($\hat{\theta}$) dos traços latente para o teste com todos os itens e os TACs de 10 e 5 itens do banco pequeno (21 itens).	37
6.3	Boxplots dos traços latentes estimados para indivíduos subdivididos segundo o critério de Kendall.	38
7.1	Triângulo equilátero da MPA com a caracterização das alternativas de respostas de cada questão da prova.	42
7.2	Matriz das curvas características dos itens do EPI	43
7.3	Boxplot da proficiência em Inglês estimada sob o modelo de Samejima de 518 estudantes reprovados (1) ou aprovados (2) no EPI do ICMC.	45
7.4	Boxplot da proficiência em Inglês estimada sob o modelo de Samejima de 518 estudantes reprovados (1) ou aprovados (2) no EPI do ICMC utilizando a abordagem <i>Shadow test</i>	46
7.5	Matriz das curvas características dos itens do BDI.	47

7.6	Boxplot dos traços latentes estimados sob o modelo de Samejima de 591 estudantes que responderam todos os 21 itens do BDI. Grupo 1 = sem depressão , grupo 2 = com disforia e grupo 3 = com depressão	48
7.7	Boxplot dos traços latentes estimados sob o modelo de Samejima de 518 estudantes que responderam 10 itens do BDI. Grupo 1 = sem depressão, grupo 2 = com disforia e grupo 3 = com depressão	49
8.1	Janela inicial do programa.	51
8.2	Janela "Criar estrutura do banco".	53
8.3	Janela "Adicionar/Remover Item".	55
8.4	Janela "Criar TAC".	57
8.5	Janela inicial de "Aplicar TAC".	58
8.6	Janela final de "Aplicar TAC".	58

Abreviações

EAP - Esperança a posteriori

EP - Erro Padrão

EPI - Exame ao proficiência em inglês

EQM - Erro Quadrático Médio

ETS - Educational Testing Services

GRE - Graduate Record Examination

ICMC - Instituto de Ciências Matemáticas e de Computação

MAP - Moda a posteriori

MEG - Modelo de Escala Gradual

MCMC - Monte Carlo via Cadeia de Markov

MCPG - Modelo de Crédito Parcial Generalizado

ML1 - Modelo Logístico de 1 parâmetros

ML2 - Modelo Logístico de 2 parâmetros

ML3 - Modelo Logístico de 3 parâmetros

ML3M - Modelo Logístico de 3 parâmetros Multidimensional

ML4 - Modelo Logístico de 4 parâmetros

MPA - Medida de Probabilidade Admissível

MRG - Modelo de Resposta Gradual

MVM - Máxima Verossimilhança Marginal

NBME - National Board of Medical Examiners

NCLEX - National Council Licensure Examination

NCSBN - National Council of State Boards of Nursing

TAC - Testes Adaptativos Computadorizados

TOEFL - Test Of English as a Foreign Language

TRI - Teoria da Resposta ao Item

USP - Universidade de São Paulo

Resumo

Testes Adaptativos Computadorizados (TAC) são aqueles que selecionam gradativamente as questões (itens) a serem apresentadas ao indivíduo de acordo com o seu nível de conhecimento (traço latente). Um TAC pode se basear em um modelo da Teoria da Resposta ao Item (TRI) para a estimação do traço latente e escolha do item a ser apresentado em cada passo do teste. Este trabalho apresenta modelos da TRI utilizados em TAC encontrados na literatura e descreve alguns métodos de calibração de itens para a formação e manutenção do banco de questões do teste sob o modelo de Samejima (1969), estimação do traço latente, seleção de itens com restrições utilizando a abordagem *Shadow test* e critérios de parada normalmente utilizados. Foram realizadas simulações com um banco grande (500 itens) e com um banco pequeno (21 itens) e avaliada a qualidade das estimativas dos traços latente (através do cálculos dos vícios e erros quadráticos médios) de TACs com diferentes números de itens. Foi aplicado o modelo de Samejima às respostas de estudantes do Exame ao proficiência em inglês (EPI) do ICMC - USP, que é aplicado semestralmente no formato lápis e papel, para a formação de um banco de itens e posterior construção de um TAC. Também foi aplicado o modelo às respostas de pacientes clínicos do Hospital das Clínicas da Medicina da USP, cedido pelo doutor Yuang-Pang Wang, ao Inventário de Depressão de Beck (BDI) para os mesmos propósitos. Comparações com a atual metodologia para avaliação da proficiência em língua inglesa do EPI (Medida de Probabilidade Admissível, MPA) e para o diagnóstico de depressão do BDI (critério sugerido por Kendall et al., 1987) foram realizadas demonstrando as vantagens e maior riqueza dos resultados obtidos com a TRI e com os TACs implementados. Adicionalmente foi desenvolvido um programa Same-CAT que armazena bancos de itens e possibilita a criação e aplicação de TACs com restrições, através da abordagem *Shadow test*.

Abstract

Computerized Adaptive Tests (CAT) are those that select questions (items) gradually to be presented to an individual according to their proficiency (latent trait level). A CAT can be based on an Item Response Theory (IRT) model for estimation of the latent trait and selection of the next item to be presented in each step of the test. This paper presents IRT models used in CATs found in literature and describes some methods of item calibration for creation and maintenance of a test items bank under the Samejima's model (Samejima; 1969), estimation of latent trait, item selection with constraints using the Shadow test approach and usuals stopping criteria. Simulations were conducted with a large bank (500 items) and a small bank (21 items) and the quality of the estimatives of latent traits were evaluated (through calculations of mean squared errors and bias) TACs with different item numbers. Samejima's model were applied for the responses of students to the English Proficiency Exam (EPE) of ICMC - USP, a test applied twice a year in paper and pencil format, to create an item bank and subsequent construction of a CAT. The model was also applied to the responses of clinical patients from the Hospital das Clínicas - USP, given by Dr. Yuang-Pang Wang, to the Beck Depression Inventory (BDI) for the same purposes. Comparisons using the current methodology to evaluate the English Language Proficiency of EPE (Measure of Probability Allowable, MPA) and the BDI (criterion suggested by Kendall et al., 1987) were performed, and the CATs provided better and richer results. Furthermore a program, Same-CAT, that stores item banks and allows CAT's applications with constraints was created.

Capítulo 1

Introdução

Com o intuito de estimar a proficiência de um indivíduo em determinado assunto ou mensurar um traço de personalidade em uma avaliação psicológica, por exemplo, é usual a aplicação de testes convencionais (papel e caneta). Estes testes são constituídos de um número fixo de itens iguais para todos os indivíduos ou, pelo menos, para grupos de indivíduos. Geralmente, os itens são pré determinados sem levar-se em consideração o nível de conhecimento (também denominado proficiência ou traço latente) específico de um indivíduo. Por este motivo, é esperado que os indivíduos precisem responder a itens inadequados ao nível de seu traço latente, muito fáceis ou muito difíceis, causando desmotivação e cansaço no examinado e, possivelmente, aumentar o erro de estimação (Sands and Waters; 1997).

Um teste adaptativo computadorizado (TAC) procura otimizar o teste para cada indivíduo examinado. Para tanto, o traço latente é estimado a cada item respondido e a escolha do próximo item que será apresentado no teste é feita baseada nesta estimativa. Um banco de itens capaz de avaliar diferentes níveis de conhecimento deve ser formado para possibilitar a adequada seleção de um item. Geralmente, os itens selecionados são mais informativos no nível estimado do traço latente e podem ser baseados de acordo com um modelo da teoria de resposta ao item (TRI). A TRI é um conjunto de modelos estatísticos que relacionam a probabilidade de um indivíduo escolher uma determinada alternativa de resposta de um item com o nível do traço latente dos indivíduos. Os princípios da TRI sugerem várias maneiras de quantificar a informação do item ao longo dos níveis do traço latente, assim como diferentes modelos a serem adotados para a construção de um banco de itens. Algumas destas principais metodologias serão descritas neste trabalho e implementadas computacionalmente.

Geralmente, os TACs precisam de menos itens em relação aos testes convencionais para realizar a estimação do traço latente com a mesma precisão, o que resulta em testes menores (Sands and Waters; 1997). Entretanto, a construção do TAC requer consideráveis recursos financeiros e humanos por exigir constante manutenção do banco, especialmente

na área Educacional onde os indivíduos podem compartilhar respostas de certos itens. Por exemplo, alguns itens podem não mais avaliar o traço latente adequadamente porque já são conhecidos (“batidos”) ou muito expostos. Novos itens podem ser incluídos, aumentando o custo da implementação e operacionalização.

A origem da ideia intuitiva dos testes adaptativos é marcada com o trabalho de Binet and Simon (1905). Binet desenvolveu um teste cujo objetivo era medir o coeficiente intelectual (Q.I.) de crianças, onde a avaliação consistia na execução de tarefas como seguir comandos, nomear, ordenar e arranjar objetos.

Como o interesse era quantificar a inteligência individualmente, ao invés de em grupos, não existia a necessidade de que os testes fossem iguais para todos os indivíduos. Por esse motivo os itens foram ordenados pelo grau de dificuldade, e aplicou-se um subconjunto desses itens para os candidatos de acordo com o nível do traço latente que ele pressupunha. Se o candidato obtivesse sucesso, Binet aplicava itens mais difíceis até que o candidato falhasse. De maneira análoga, se o candidato falhasse, Binet aplicava itens mais fáceis até que o candidato obtivesse sucesso.

Thurstone (1925) propôs uma forma de apresentar os itens do teste em escala de idade. Com os dados da *London school children*, fornecidos por Burt (1949), Thurston apresentou o gráfico relacionando as idades das crianças com sucesso nas tarefas de Binet (Figura 1.1). Observe que as curvas apresentam muitas características dos modelos da TRI moderna. Mais informações sobre a história da TRI e do TAC encontram-se em Bock (1997), e do TAC na página do C.A.T. Central (David; 2009), respectivamente,

Figura 1.1: Proporção de resposta correta para itens do teste de Binet e Simon pela idade . Fonte: Bock (1997) , Adaptado de Thurstone (1925)

No entanto, o desenvolvimento e o aperfeiçoamento de técnicas psicométricas para a

implementação do TAC ainda demorou algumas décadas. As pesquisas do TAC se iniciaram na década de 70 por Lord (1971a, 1977), Owen (1975) e 80 por Weiss (1982, 1983). Os computadores da época não possuíam uma capacidade de processamento suficiente para estimar o traço latente em tempo imediato. Consequentemente, a maioria dos estudos daquela época foram voltados em desenvolver aproximações e novas formas que poderiam ser aplicados em testes convencionais. Exemplos são: teste de formato em dois estágios (Cronbach and Gleser; 1965), seleção Bayesiana de item com uma aproximação para a distribuição *a posteriori* do parâmetro (Owen; 1969), o método *up-and-down* para seleção de item (Lord; 1970), o algoritmo de *Robbins-Monro* (Lord; 1971a), o teste de nível flexível (Lord; 1971b), o teste *stradaptive* (Weiss; 1973) e o teste adaptativo piramidal (Larkin and Weiss; 1975).

Na década de 80, com o desenvolvimento da tecnologia que tornou os computadores pessoais acessíveis, viabilizando a aplicação dos testes adaptativos, foi desenvolvida a primeira versão em TAC do *Armed Services Vocational Aptitude Battery* (ASVAB). O resultado é normalizado de modo que se o candidato obteve pontuação de 50 significa que seu desempenho foi melhor do que 50% dos candidatos. O teste também fornece resultados para “Armed Forces Qualification Test” (AFQT), que consiste nas medidas de desempenho para as 4 áreas do ASVAB: raciocínio aritmético, conhecimento em matemática, interpretação de texto e conhecimento gerais. Esse teste foi desenvolvido em sua forma convencional na década de 60 para a qualificação para alistamento. Atualmente o teste é aplicado em 14.000 escolas de ensino médio e Military Entrance Processing Stations nos Estados Unidos.

Neste contexto, a migração dos testes convencionais para os TACs solidificou-se quando o *National Council of State Boards of Nursing* (NCSBN) e National Board of Medical Examiners (NBME) implementam a versão em TAC do *National Council Licensure Examination* (NCLEX) (NCSBN; 2005) em 1994. Este exame é requerido para a licença profissional na área de enfermagem nos Estados Unidos. Em seguida, em 1996, foi desenvolvida a versão TAC do *Graduate Record Examination*, GRE (2011), pelo *Educational Testing Services* (ETS), que avalia o candidato nas áreas de escrita analítica, raciocínio verbal e raciocínio quantitativo e fornece uma medida padronizada dos conhecimentos dos candidatos para as escolas de graduação e *business schools*. Um outro teste importante desenvolvido em TAC na década de 90, também pelo ETS, é o TOEFL, (*Test Of English as a Foreign Language*), que avalia a proficiência em inglês como segunda língua. O teste é dividido nas sessões leitura, gramática, escrita e compreensão oral e é reconhecido por mais de 7500 instituições em mais de 130 países (TOEFL; 2011).

Desde então, diversos outros TACs já foram implementados. Na área de psicologia podemos citar: Minnesota Multiphasic Personality Inventory-2 (Forbey and Ben-Porath; 2007), Pediatric Symptom Checklist (Gardner et al.; 2002). Na área da Educação no Brasil temos, por exemplo, o Computer-Aided Learning of English for Academic Purposes (Gonçalves; 2004) e a avaliação de proficiência em língua inglesa da UnB (Carvalho and Coelho; 2012).

Para a construção de um TAC baseado na TRI para sua implementação, várias etapas devem ser detalhadas. A primeira, já mencionada anteriormente, é a escolha do modelo da TRI. A segunda é a construção de um banco de itens cujo parâmetro de dificuldade de discriminação entre os níveis de conhecimento dos indivíduos devem ser conhecidos. A terceira diz respeito à estimação do traço latente individual. A quarta é referente a critérios para seleção de itens a cada passo do TAC. Finalmente, a última etapa corresponde à determinação do critério de parada do teste.

O objetivo deste trabalho é o desenvolvimento de um programa para a construção de TACs baseado no modelo de Samejima (1969). A teoria necessária para tal implementação será descrita em detalhes, o processo de desenvolvimento do programa, os testes de implementação através de simulações e aplicação a dados reais referentes a respostas de alunos de pós-graduação a questões de inglês para o Exame de Proficiência em Inglês (EPI) do Instituto de Ciência Matemáticas e Computação (ICMC-USP) e pacientes clínicos do Hospital das Clínicas da medicina da USP, cedido pelo doutor Yuang-Pang Wang, ao Inventário de Depressão de Beck (BDI).

Nos capítulos a seguir, serão descritos os modelos da TRI já implementados como base de algum TAC (capítulo 2), calibração de itens (capítulo 3) estimação dos traços latentes (capítulo 4), seleção de itens no TAC e critérios de parada do teste (capítulo 5), simulações de TACs sob o modelo de Resposta Gradual para um banco grande e um banco pequeno (capítulo 6), aplicações do modelo de Samejima às respostas ao EPI e ao BDI (capítulo 7) e o programa desenvolvido nestes projeto, Same-CAT (capítulo 8).

Capítulo 2

Modelos da Teoria de Resposta ao Item

Os modelos da TRI procuram descrever a probabilidade de um indivíduo obter um certo escore em um item em função do nível do traço latente (ou dos traços latentes). Essa probabilidade também é influenciada por parâmetros do item (dificuldade e capacidade de discriminação, por exemplo).

Diferentes modelos podem ser selecionados de acordo com certas características do teste: se os itens possuem duas ou mais categorias de respostas (itens dicotômicos ou politômicos), o número de traços latentes em interesse (modelos unidimensionais, no caso de apenas um traço latente por indivíduo, ou multidimensionais onde cada indivíduo possui um vetor de traços latentes) e número de populações envolvidas.

Em cada seção a seguir, descreveremos os modelos da TRI que já foram adotados como base para o desenvolvimento de um TAC encontrados na revisão bibliográfica realizada. Ao final de cada seção, citamos alguns dos artigos que se basearam nos respectivos modelos descritos para a criação do TAC.

A partir desta etapa, vamos considerar que $j = 1, \dots, n$ indexa os indivíduos e $i = 1, \dots, I$ os itens.

2.1 Modelo logístico de 3 parâmetros unidimensional

O modelo logístico de 3 parâmetros (ML3), é o modelo mais usual na literatura atualmente para dados dicotômicos unidimensionais e é dado por:

$$P(X_{ij} = 1 | \theta_j; a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1 + \exp[-Da_i(\theta_j - b_i)]}, \quad (2.1)$$

em que $P(X_{ij} = 1 | \theta_j)$ é a probabilidade do indivíduo j responder corretamente ao item i , X_{ij} é uma variável dicotômica que assume o valor 1 se o indivíduo j responder corretamente ao item i (resposta correta, por exemplo) e 0, em caso contrário, θ_j é o traço latente do indivíduo j , a_i e b_i são os parâmetros de discriminação e dificuldade do item i , respectivamente, c_i representa a probabilidade de acerto aleatório do item i e D é um fator de escala igual a 1, para modelo logístico e igual a 1.702 para que a função logística forneça valores parecidos com a da função ogiva normal. Outros modelos são definidos como casos particulares de (2.1). Se $c_i = 0$ temos o modelo logístico de 2 parâmetros (ML2) e se, além dessa restrição, o parâmetro de discriminação for igual para todos os itens, temos o modelo logístico de 1 parâmetro (ML1). Ainda no caso em que $c_i = 0$ e $a_i = 1$, para $i = 1, \dots, I$, temos o modelo de Rasch.

A curva característica de um item (CCI) sob o ML3, isto é, a probabilidade de resposta em função do traço latente, está representada na Figura 2.1.

Figura 2.1: Curva característica de um item sob o ML3.

Fonte: Reckase (2009)

Note que quanto maior o valor do parâmetro a mais acentuada será a inclinação da função, b equivale ao valor do traço latente no ponto de inflexão da função e representa a proficiência onde a probabilidade de acerto é de $(1 + c)/2$ e o valor de c determina a posição da assíntota inferior da curva.

No caso em que todos os parâmetros dos itens e o traço latente são desconhecidos, observe que se $\theta_j^* = \alpha\theta_j + \delta$, $a_i^* = \frac{a_i}{\alpha}$, $b_i^* = \alpha b_i + \delta$ e $c_i^* = c_i$, onde α e δ são constantes reais e $\alpha > 0$, temos

$$\begin{aligned} P_i(X_{ij} = 1 | \theta_j^*; a_i^*, b_i^*, c_i^*) &= c_i^* + \frac{(1 - c_i^*)}{1 + \exp[-Da_i^*(\theta_j - b_i^*)]} \\ &= c_i + \frac{(1 - c_i)}{1 + \exp[-D\frac{a_i}{\alpha}(\alpha\theta_j + \delta - \alpha b_i - \delta)]} \end{aligned} \quad (2.2)$$

$$= P_i(X_{ij} = 1 | \theta_j; a_i, b_i, c_i), \quad (2.3)$$

ou seja, diferentes valores dos parâmetros dos itens e do traço latente resultam na mesma probabilidade de acerto de um indivíduo para um certo item. Este *problema de falta de identifiabilidade* pode ser solucionado durante o processo de estimação, explicados nas seções seguintes.

Alguns TACs foram implementados com base no modelo de Rasch, ML1, e divulgados na literatura como, por exemplo: Triantafyllou et al. (2008) que desenvolveram um TAC para telefones móveis; Chang and Ying (2008) estudou porquê alguns indivíduos têm uma menor pontuação em TAC do que nos testes convencionais.

Também foram divulgados na literatura TACs baseados em modelos logísticos mais complexos. Utilizando o ML2 temos exemplos como: Eggen (1999) que comparou a eficiência de 2 tipos de informações (informações de Fisher e Kullback-Leibler, ver capítulo número 5), Eggen and Straetmans (2000) que estudaram a performance de um TAC, podemos citar para classificar examinados em categorias, Chang and Ying (2008). Com o ML3 foram feitos os seguintes trabalhos: McLeod et al. (2003) que usaram um método bayesiano para detectar indivíduos com pre-conhecimento dos itens, Wang and Chen (2004) que compararam TACs baseados em TRI multidimensionais com o ML3, Penfield (2007) que propôs medir o erro padrão do estimador de máxima verossimilhança usando uma posteriori ponderada da função de informação do teste; Belov, Armstrong e Weissman (2008)(Belov et al.; 2008) que apresentaram um novo algoritmo para TACs com restrições de conteúdo, Belov and Armstrong (2009) que apresentaram um método de Monte Carlo para obter um banco de itens Chang and Ying (2008), Cheng et al. (2009) implementaram um TAC usando *a-stratification*, Imai (2008) desenvolveu um TAC (J-CAT) que mede a proficiência em inglês como segunda língua e a avaliação de proficiência em língua inglesa da UnB (Carvalho and Coelho; 2012).

2.2 Modelo logístico de 4 parâmetros

Este modelo logístico de 4 parâmetros, ML4, (Barton and Lord; 1981) é uma extensão do ML3, que inclui um parâmetro $0 < d \leq 1$ definido pela assíntota superior na curva característica do item e possibilitando que esta seja menor que para 1. A expressão é dada por:

$$P(X_{ij} = 1|\theta_j; a_i, b_i, c_i, d) = c_i + \frac{(d - c_i)}{1 + \exp[-Da_i(\theta_j - b_i)]}.$$

A CCI sob o ML4 está representada na Figura 2.2.

Figura 2.2: Curva característica de um item sob o ML4.

Este modelo é interessante para TACs porque erros em itens fáceis, principalmente no começo do teste, podem prejudicar a estimação do traço latente, mesmo se o indivíduo responder corretamente aos itens mais difíceis. Neste sentido, a inclusão do parâmetro d cria uma pequena probabilidade de resposta errada ao item mesmo para indivíduos com um alto traço latente, diminuindo o impacto das primeiras questões do TAC no resultado do teste.

Rulison and Loken (2009), desenvolveram um TAC mostrando que o ML4 pode diminuir o Erro Quadrático Médio (EQM) do estimador do traço latente de estudantes com um alto-nível de proficiência que tiveram um baixo desempenho no início do teste.

2.3 Modelo de resposta gradual unidimensional

O MRG é adequado para itens politômicos cujas categorias de resposta possam ser ordenadas entre si.

Neste trabalho foi adotado o modelo de resposta gradual (MRG) proposto por Samejima (1969), porque os testes utilizados possuem respostas com múltipla categorias ordenadas de respostas, objetivo deles é tentar estimar apenas um traço latente e alguns outros motivos que adequam o modelo aos dados.

Suponha que o item i possui $(m_i + 1)$ categorias, $k = 0, \dots, m_i$, e que essas categorias estejam organizadas em ordem crescente, ou seja, quanto maior o valor de k , mais próxima ela estará da resposta totalmente correta. O modelo é uma extensão do ML2 cuja probabilidade de um indivíduo j escolher a categoria k ou alguma maior que ela do item i é dada por

$$P_{i,k}^+(\theta_j; a_i, b_{i,k}) = \frac{1}{1 + \exp[-Da_i(\theta_j - b_{i,k})]}, \quad (2.1)$$

sendo $b_{i,k}$ o parâmetro de dificuldade da k -ésima categoria do item i , $P_{i,0}^+ = 1$ e a_i , θ_j e D , como definidos anteriormente. Por definição, temos que:

$$b_{i,1} \leq b_{i,2} \leq \dots \leq b_{i,m_i}.$$

A discriminação de uma categoria específica não depende apenas do parâmetro a_i , mas também é influenciada pelos parâmetro de dificuldade das categorias adjacentes.

Para calcular a probabilidade de um indivíduo j escolher a categoria k no item i temos a expressão:

$$P_{i,k}(\theta_j) = P_{i,k}^+(\theta_j) - P_{i,k+1}^+(\theta_j) \quad (2.2)$$

com $P_{i,m_i+1}^+(\theta_j) = 0$ por definição.

Na Figura 2.3 podemos ver gráficos das curvas características de 4 itens sob o modelo de Samejima. Os itens 1 e 2 pertencem a um caso particular deste modelo que será explicado na seção seguinte.

No caso em que todos os parâmetros são desconhecidos, observe que se $\theta_j^* = \alpha\theta_j + \delta$, $a_i^* = \frac{a_i}{\alpha}$, $b_{i,k}^* = \alpha b_{i,k} + \delta$, onde α e δ são constantes reais e $\alpha > 0$, teremos um caso parecido com o (2.2) e um *problema de falta de identificabilidade*.

Na literatura é possível encontrar exemplos de TACs que utilizaram o MRG como em Gardner et al. (2004) que desenvolveu um TAC para medir depressão usando o BDI; Hol et al. (2007) que comparam a eficiência de um TAC com base no modelo de Samejima com outro tipo de teste não adaptativo (short Forms).

Figura 2.3: Sob o MRG, P_k é a probabilidade de escolha da categoria k . Fonte: Teoria da Resposta ao Item: Conceitos e Aplicações.

2.4 Modelo de escala gradual

O modelo de escala gradual, MEG, (Andrich; 1978) é um caso particular do MRG onde os escores das categorias possuem a mesma distância. Ele pode ser expresso por

$$P_{i,k}^+(\theta_j; a_i, b_i, d_k) = \frac{1}{1 + \exp[-Da_i(\theta_j - b_i + d_k)]},$$

sendo que b_i é chamado de parâmetro de localização do item i e d_k é o parâmetro de categoria.

Como $P_{i,k}(\theta_j) = P_{i,k}^+(\theta_j) - P_{i,k+1}^+(\theta_j) \geq 0$, então $d_k - d_{k+1} \geq 0$, ou seja:

$$d_1 \geq d_2 \geq \dots \geq d_{m_i}.$$

Considerando o parâmetro $b_{i,k}$ do MRG, ele pode ser decomposto:

$$b_{i,k} = b_i - d_k.$$

Como as categorias possuem a mesma distância, seja qual for o item (note que d_k independe de item), testes em que os itens possuem diferentes número de categorias, o MEG não é adequado.

A vantagem de impor a restrição de categorias dos itens equidistantes no MRG é a redução do número de parâmetros de itens. No MRG, considerando que cada item possui $m_i + 1$ categorias, o número de parâmetros dos itens é $(\sum_{i=1}^I m_i) + I$. No caso do MEG, o número total de parâmetros de itens (considerando que os itens tenham $m + 1$ categorias) é $2I + m$. Por exemplo, se um teste possui 5 itens com 5 categorias cada, no MRG teremos 30 parâmetros de itens, enquanto que no MEG, teremos 15 parâmetros de itens.

Na figura 2.3 da representação gráfica de 4 itens sob o MRG, cada um com 4 categorias de respostas, valor do parâmetro, a_i , igual 1 e diferentes valores para os parâmetros $b_{i,k}$.

Observe que os parâmetros $b_{i,k}$ do item 1 e do item 2 são equidistantes, e portanto, podem ser representados pelo modelo de escala gradual, enquanto os itens 3 e 4 só podem ser representados pelo modelo de resposta gradual.

Analisando os gráficos, percebemos que P_0 é igual para todos os itens e que, quanto menor o valor de $b_{i,3}$ mais a curva da última categoria é deslocada para a esquerda. Pelo gráfico do item 1, podemos perceber que todas as categorias, dependendo do intervalo de θ , possuem alta probabilidade de serem escolhidas. Enquanto que pelo gráfico do item 2, as categorias 1 e 2 serão pouco escolhidas. Comparando os gráficos dos itens 1 e 3, a redução do $b_{i,3}$ causou a diminuição na probabilidade da seleção da categoria 2 e o aumento na probabilidade da categoria 4. Comparando os gráficos dos itens 1 e 4, o aumento do $b_{i,2}$ causou um aumento na probabilidade da seleção da categoria 1 e uma diminuição na probabilidade da categoria 2.

Podemos encontrar na literatura TACs que foram implementados com base no modelo de Andrich. Hart, Cook, Mioduski, Teal and Crane (2006) desenvolveram um TAC para testes de pacientes com problemas nos ombros; Hart, Wang, Stratford and Mioduski (2006) programaram um TAC para avaliar o estado funcional da coluna lombar; Hart et al. (2008) construíram um TAC para pacientes com problemas de pé ou no tornozelo que buscam reabilitação em clínicas de tratamento ambulatoriais.

2.5 Modelo de Crédito Parcial Generalizado

O modelo de crédito parcial generalizado, MCPG, (Muraki; 1992) é outro modelo para itens politômicos ordinais. A probabilidade $P_{i,k}(\theta_j)$ de um indivíduo de traço latente θ_j escolher a categoria k do item, considerando $m_i + 1$ categorias:

$$P_{i,k}(\theta_j) = \frac{\exp[\sum_{u=0}^k Da_i(\theta_j - b_{i,u})]}{\sum_{u=0}^{m_i} \exp[\sum_{v=0}^u Da_i(\theta_j - b_{i,v})]},$$

onde $b_{i,k}$ é o parâmetro de item que regula a probabilidade de escolher a categoria k em vez da categoria adjacente $k - 1$ no item i . Cada parâmetro $b_{i,k}$ corresponde ao valor do traço latente em que o indivíduo tem a mesma probabilidade de responder à categoria k e à categoria $k - 1$, isto é, onde $P_{i,k}(\theta_j) = P_{i,k-1}(\theta_j)$, e a_i é o parâmetro que representa o poder de discriminação do item.

Podemos decompor o parâmetro $b_{i,k}$ em:

$$b_{i,k} = b_i - d_k.$$

Observe que d_k não é necessariamente ordenado, ao contrario do MEG, e esse parâmetro é interpretado como sendo o desvio da dificuldade de cada categoria em relação à locação do item, b_i . Devido ao problema de indeterminação no modelo, normalmente supõe-se $d_0 = 0$ e $\sum_{k=1}^{m_i} d_k = 0$.

Considerando $m_i + 1$ categorias de resposta do item i , há m_i parâmetros de categorias além dos parâmetro a e b para cada item, portanto há $\sum_{i=1}^I m_i + 2I$ parâmetros de item a serem estimados.

Haley et al. (2006) usaram um TAC baseado neste modelo com um banco de itens de funcionamento físico para comparar a precisão da estimação do traço latente em relação a um teste computadorizado com seleção de itens aleatória e Fliege et al. (2009) estudaram a viabilidade e validade da aplicação real do D-CAT, um TAC desenvolvido por alemães para avaliar o construto depressão.

2.6 Modelo de Ogiva Normal Multidimensional

O modelo multidimensional de ogiva normal (Linden and Hambleton; 1997) é um modelo do tipo compensatório para itens dicotômicos, ou seja, a probabilidade de resposta depende de alguma relação linear dos parâmetros, resultando que traços latentes de níveis elevado de um indivíduo podem compensar seus traços latentes de níveis baixos na probabilidade de resposta correta. Considerando $\theta_j = (\theta_{j1}, \dots, \theta_{jp})'$ o vetor de p traços latentes do indivíduo j , sua formulação é expressa por

$$p(X_{ij} = 1 \mid \theta_j) = \Phi(a'_i \theta_j - b_i) = \Phi(a'_{i1} \theta_{j1} - b_i + \dots + a'_{ip} \theta_{jp} - b_i),$$

em que $\Phi(\cdot)$ é a função distribuição normal padrão acumulada e $\mathbf{a}_i = (a_{i1}, \dots, a_{ip})'$ e b_i são os parâmetros do item i .

Petersen et al. (2006) programaram um TAC para medir traços latentes relacionados a qualidade de vida (funcionamento físico, funcionamento emocional, e as escalas de fadiga), usando um banco de dados de pacientes com cancer.

2.7 Modelo Logístico de 3 parâmetros Multidimensional

O modelo logístico de 3 parâmetros multidimensional, ML3M, (Linden and Glas; 1997) é a generalização do ML3 para o caso da probabilidade de acerto de um item depender de mais de um traço latente. O modelo é dado por:

$$P(X_{ij} = 1 \mid \boldsymbol{\theta}_j; \mathbf{a}_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp(-\mathbf{a}_i' \boldsymbol{\theta}_j - b_i)},$$

onde $\boldsymbol{\theta}_j = (\theta_1, \dots, \theta_p)$ é um vetor de traços latentes do indivíduo j , $\mathbf{a}_i$ é um vetor de parâmetros de discriminação do item i , b_i o parâmetro que representa a dificuldade do item i e c_i é o parâmetro de acerto ao acaso. A Figura 2.4 traz uma ilustração da função de probabilidade de um item sob o ML3M com 2 traços latentes.

Analogamente ao 3PL, se $c_i = 0$, define-se o modelo logístico de 2 parâmetros multidimensional (M2PL).

Veldkamp and Linden (2002) construíram um TAC, baseado no M2PL utilizando a abordagem Shadow Test (capítulo 5) para seleção de itens. Lee et al. (2008) propuseram uma estratégia (fornecendo validação teórica e empírica) baseada em estratificação e um funcional do vetor de discriminação do parâmetro usando o M2PL em seu TAC. Li and Schafer (2005) desenvolveram um TAC multidimensional para medir as habilidades em leitura e matemática e comparam o resultado com um TAC unidimensional.

Os capítulos seguintes terão foco no modelo de Samejima, pois é o modelo adotado para o desenvolvimento deste trabalho.

Figura 2.4: Exemplo de superfície do M3PL considerando 2 traços latentes
Fonte: Reckase (2009)

Capítulo 3

Calibração de itens

O processo de Calibração de itens nada mais é do que a etapa de estimação dos parâmetros de itens. Para a construção de um TAC eficiente é necessário um grande banco de itens calibrados de acordo com o modelo selecionado. Geralmente, para calibrar um banco de itens, um pré-teste é feito administrando-se subconjuntos de itens do banco para vários indivíduos, obtendo-se dados suficientes para a estimação de seus parâmetros. Também é comum utilizar os dados fornecidos do TAC para ajustar as estimações dos parâmetros dos itens feitas inicialmente.

Os métodos mais comuns para a calibração de itens na TRI são máxima verossimilhança marginal e métodos bayesianos como o Monte Carlo via Cadeia de Markov (MCMC), por exemplo. Muitos autores desenvolveram sua teoria para os modelos logísticos unidimensionais (Bock and Aitkin; 1981; Thissen; 1982; Rigdon and Tsutakawa; 1983), mas os métodos também podem ser aplicados em modelos multidimensionais (Bock et al.; 1988; Wilson et al.; 1991). Nas seções a seguir há uma breve descrição de cada um deles sob o enfoque do modelo de Samejima.

3.1 Máxima Verossimilhança Marginal (MVM)

Quando os traços latente são desconhecidos um processo para se estimar os parâmetros dos itens é a máxima verossimilhança marginal (MVM) usando a abordagem de Bock and Aitkin (1981).

Nesse processo, é considerado que pode-se atribuir, aos traços latentes dos indivíduos, uma distribuição de probabilidade na população cuja densidade será denotada por $g(\theta | \boldsymbol{\eta})$, com derivadas de segunda ordem conhecidas e finitas e $\boldsymbol{\eta}$ sendo os parâmetros da distribuição geralmente fixados a priori. É comum considerar que θ tem distribuição normal com o vetor de parâmetros $\boldsymbol{\eta} = (\mu = 0, \sigma^2 = 1)$.

Sejam $\mathbf{x}_{ij\cdot} = (x_{ij1}, \dots, x_{ijm_i})'$, em que x_{ijk} são variáveis que assumem o valor 1 se o indivíduo j escolhe a categoria k do item i e 0 caso contrario, $\mathbf{x}_{\cdot j} = (\mathbf{x}_{1j\cdot}, \dots, \mathbf{x}_{Ij\cdot})'$. A probabilidade de escolha da categoria k do item i pelo indivíduo j é dada por:

$$P(\mathbf{x}_{ij\cdot} = x_{ij\cdot} \mid \theta_j, \boldsymbol{\zeta}_i) = \prod_{k=1}^{m_i} [P_{i,k}(\theta)]^{x_{ijk}},$$

em que $P_{i,k}(\theta_j)$ é dado em (2.2) e $\boldsymbol{\zeta}_i = (a_i, b_{i1}, \dots, b_{im_i})'$.

Temos que a probabilidade marginal de um indivíduo j escolher as respostas $\mathbf{x}_{\cdot j}$ é dada por

$$P(\mathbf{X}_{\cdot j} = \mathbf{x}_{\cdot j} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}) \equiv P(\mathbf{X}_{\cdot j} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}) = \int P(\mathbf{X}_{\cdot j} \mid \theta, \boldsymbol{\zeta}) g(\theta \mid \boldsymbol{\eta}) d\theta.$$

Note que $P(\mathbf{X}_{\cdot j} \mid \theta, \boldsymbol{\zeta}) = \prod_{i=1}^I P(\mathbf{X}_{ij\cdot} \mid \theta, \boldsymbol{\zeta})$ não depende de $\boldsymbol{\eta}$. Pela independência entre as respostas dos indivíduos, a função de verossimilhança é dada por

$$L(\boldsymbol{\zeta}, \boldsymbol{\eta}) = P(\mathbf{X}_{\cdot\cdot} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}) = \prod_{j=1}^n P(\mathbf{X}_{\cdot j} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}),$$

e a função de log-verossimilhança é

$$l(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \sum_{j=1}^n \log P(\mathbf{X}_{\cdot j} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}). \quad (3.1)$$

A componente $\boldsymbol{\eta}$ é usada para solucionar o problema de falta de identificabilidade.

Diferenciando (3.1), para um modelo politômico temos que o vetor escore é dado por:

$$S(\boldsymbol{\zeta}_i) = \sum_{j=1}^n \left\{ \int_{\mathbb{R}} g_j^*(\theta) \left\{ \sum_{k=1}^{m_i} \left[\frac{x_{ijk}}{P_{ijk}} \left(\frac{\partial P_{ijk}}{\partial \zeta_i} \right) \right] \right\} d\theta \right\}, \quad (3.2)$$

onde $g_j^*(\theta) = \frac{P(\mathbf{X}_{\cdot j} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) g(\theta \mid \boldsymbol{\eta})}{\int_{\mathbb{R}} P(\mathbf{X}_{\cdot j} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) g(\theta \mid \boldsymbol{\eta})}$. Temos que a matriz Hessiana é dada por:

$$H(\zeta_i) = \sum_{j=1}^n \left\{ \int_{\mathbb{R}} g_j^*(\theta) \left\{ \sum_{k=1}^n \frac{x_{ijk}}{P_{ijk}} \left[\left(\frac{\partial P_{ijk}}{\partial \zeta_i \partial \zeta'_i} \right) - \left(\frac{P_{ijk}}{\zeta_i} \right) \left(\frac{P_{ijk}}{\zeta_i} \right)' \right] \right\} d\theta \right\}. \quad (3.3)$$

Para o modelo de Samejima (2.2) temos que as primeiras derivadas em relação aos parâmetros dos itens, são dadas por

$$\frac{\partial P_{ijk}}{\partial a_i} = \frac{\partial}{\partial a_i} \left\{ \frac{1}{1 + \exp^{-Da_i(\theta_j - b_{i(k-1)})}} - \frac{1}{1 + \exp^{-Da_i(\theta_j - b_{ik})}} \right\} \quad (3.4)$$

$$= D\{(\theta_j - b_{ik})W_{ijk}^+ - (\theta_j - b_{i(k+1)})W_{ij(k+1)}^+\}, \quad (3.5)$$

$$\frac{\partial P_{ijk-1}}{\partial b_{ik}} = Da_i P_{ijk}^+ Q_{ijk}^+, \quad (3.6)$$

$$\frac{\partial P_{ijk}}{\partial b_{ik}} = -Da_i P_{ijk}^+ Q_{ijk}^+, \quad (3.7)$$

onde $W_{ijk}^+ = P_{ijk}^+ Q_{ijk}^+$ e que

$$\frac{\partial W_{ijk}^+}{\partial \zeta_i} = (1 - 2P_{ijk}^+) \left(\frac{\partial P_{ijk}^+}{\partial \zeta_i} \right).$$

As segundas derivadas em relação aos parâmetros dos itens, são dadas por

$$\frac{\partial^2 P_{ijk}}{\partial a_i^2} = D^2 \left\{ (\theta_j - b_{ik})^2 \{1 - 2P_{ijk}^+\} W_{ijk}^+ - (\theta_j - b_{i(k+1)})^2 (1 - 2P_{ij(k+1)}^+) W_{ij(k+1)}^+ \right\}, \quad (3.8)$$

$$\frac{\partial^2 P_{ijk-1}}{\partial b_{ik}^2} = D^2 a_i^2 (1 - 2P_{ijk}^+) W_{ijk}^+, \quad (3.9)$$

$$\frac{\partial^2 P_{ijk-1}}{\partial b_{im} \partial b_{in}} = 0, m \neq n, \quad (3.10)$$

$$\frac{\partial^2 P_{ij(k-1)}}{\partial a_i \partial b_{ik}} = W_{ijk} \{1 + Da_i(\theta_j - b_{ik})(1 - 2P_{ijk}^2)\}, \quad (3.11)$$

$$\frac{\partial^2 P_{ij(k)}}{\partial a_i \partial b_{ik}} = -W_{ijk} \{1 + Da_i(\theta_j - b_{ik})(1 - 2P_{ijk}^2)\}, \quad (3.12)$$

onde $W_{ijk} = \frac{P_{ijk}^* Q_{ijk}^*}{P_{ijk} Q_{ijk}^*}$, $P_{ijk}^* = \{1 + \exp^{-Da_i(\theta - b_i)}\}^{-1}$, $Q_i^* = 1 - P_i^*$.

De (3.4), (3.6), (3.7) e (3.2) temos que as componentes do vetor escore em formas de quadratura, $\bar{\theta}_u$ com $u = 1, \dots, q$ pontos de quadratura, são dadas por

$$S(a_i) = D \sum_{u=1}^q \sum_{k=0}^{m_i} \left\{ \frac{\bar{r}_{iuk}}{P_{iuk}} [(\bar{\theta}_u - b_{iuk})P_{iuk}^+ - (\bar{\theta}_u - b_{iu(k+1)})P_{iu(k+1)}^+] \right\},$$

e

$$S(b_{ih}) = D \sum_{u=1}^q \left\{ W_{iuh}^+ \left[\frac{\bar{r}_{iu(h-1)}}{P_{iu(h-1)}} - \frac{\bar{r}_{iuh}}{P_{iuh}} \right] \right\},$$

com $\bar{r}_{iuk} = \sum_{j=1}^n y_{ijk} g_j^*(\bar{\theta}_j)$ e $g_j^*(\bar{\theta}_j) = \frac{P(X_{.j} | \zeta, \bar{\theta}_u) g(\bar{\theta}_u | \eta)}{\sum_{u=1}^q P(X_{.j} | \zeta, \bar{\theta}_u) g(\bar{\theta}_u | \eta)}$

Em relação a matriz Hessiana em formas de quadratura, de (3.8) a (3.12) e (3.3), temos que

$$\begin{aligned} H(a_i, a_i) = & \sum_{u=1}^n \sum_{k=0}^{m_i} \frac{r_{iuk}}{P_{iuk}} \{ \{ D^2[(\theta_u - b_{ik})^2(1 - 2P_{iuk}^+)W_{iuk}^+ \\ & - (\theta_u - b_{i(k+1)})^2(1 - 2P_{iu(k+1)}^+)W_{iu(k+1)}^+] \} P_{iuk} \\ & - D[(\theta_u - b_{ik})W_{iuk}^+ - (\theta_u - b_{i(k+1)})W_{iu(k+1)}^+]^2 \}, \end{aligned}$$

$$\begin{aligned} H(b_{ik}, b_{ik}) = & D \sum_{u=1}^q a_i \left\{ \left[\frac{r_{iu(k-1)}}{P_{iu(k-1)}} - \frac{r_{iuk}}{P_{iuk}} \right] [Da_i W_{iuk}^+ (1 - 2P_{iuk}^+)] \right. \\ & \left. - \left[\frac{r_{iu(k-1)}}{P_{iu(k-1)}^2} - \frac{r_{iuk}}{P_{iuk}^2} \right] Da_i (W_{iuk}^+)^2 \right\}, \end{aligned}$$

$$H(b_{ik}, b_{i(k-1)}) = D^2 a_i^2 \sum_{u=1}^q \left\{ \frac{r_{iu(k-1)} W_{iuk}^+ W_{iu(k-1)}^+}{P_{iu(k-1)} 2} \right\},$$

$$H(b_{ik}, b_{i(k+1)}) = D^2 a_i^2 \sum_{j=1}^q \left\{ \frac{r_{ij(k+1)} W_{ijk}^+ W_{ij(k+1)}^+}{P_{ij(k+1)} 2} \right\},$$

$$\begin{aligned} H(a_i, b_{ik}) = & \sum_{u=1}^q \left\{ D \left[\frac{r_{iu(k-1)}}{P_{iu(k-1)}} - \frac{r_{iuk}}{P_{iuk}} \right] W_{iuk}^+ [1 + D a_i (\bar{\theta}_u - b_{ik})(1 - 2P_{iuk}^+)] \right. \\ & - D^2 a_i W_{iuk}^+ \left\{ \frac{r_{iu(k-1)}}{P_{iu(k-1)}^2 [(\bar{\theta}_u - b_{i(k-1)}) W_{iu(k-1)}^+ - (\bar{\theta}_u - b_{ik}) W_{iuk}^+]} \right. \\ & \left. \left. - \frac{r_{iuk}}{P_{iuk}^2} [(\bar{\theta}_u - b_{ik}) W_{iuk}^+ - (\bar{\theta}_u - b_{i(k+1)}) W_{iu(k+1)}^+] \right\} \right\}, \end{aligned}$$

Para resolver as equações das estimações, Samejima (1997) propos uma extensão do uso do algoritmo tipo EM nos modelos dicotômicos com pequenas adaptações, o algoritmo está descrito em Azevedo (2003).

Passo E

Usar os pontos de quadratura $\bar{\theta}_l$, os pesos $A_u^{(t)}, u = 1, \dots, q$ e estimativas provisórias dos parâmetros dos itens, $\hat{\zeta}_i^{(t)}, i = 1, \dots, I$, para gerar $g_j^*(\bar{\theta}_u)^{(t)}$ e, posteriormente, $\bar{r}_{iuk}^{(t)}, i = 1, \dots, I, u = 1, \dots, q$ e $k = 0, 1, \dots, m_i$

Passo M

com $\bar{r}^{(t)}$ obtido no Passo E, resolver as equações de estimação para $\zeta_i, i = 1, \dots, I$, usando o algoritmo de Newton-Raphson.

As estimativas geradas dos parâmetros dos itens serão consistentes. Pela convexidade em quase toda superfície de verossimilhança, o vetor de pontos estacionários será único e portanto, máximo global da função (Samejima; 1997). As estimações serão convergentes a menos de padrões de resposta aberrantes (Bock and Zimowski; 1997).

3.2 Monte Carlo via Cadeia de Markov (MCMC)

A idéia do MCMC aplicada a modelos da TRI é definir uma cadeia de Markov $M_0, M_1, M_2, \dots, M_k$, onde o estado $M_k = (\boldsymbol{\theta}^k, \boldsymbol{\beta}^k)$, em que $\boldsymbol{\theta}^k$ representa o traço latente na k -ésima iteração e $\boldsymbol{\beta}^k$ denota os parâmetros dos itens na k -ésima iteração, e então simular

observações dela. Essas observações são usadas para fazer inferência sobre os parâmetros dos itens e dos traços latentes do modelo. Sob certas condições de regularidade (Tierney; 1994), M_k convergirá para uma distribuição estacionária $\pi(\boldsymbol{\theta}, \boldsymbol{\beta})$ conforme k cresce. Considerando o modelo bayesiano, $\pi(\boldsymbol{\theta}, \boldsymbol{\beta})$ é a distribuição a posteriori $p(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{X})$.

O comportamento da cadeia de Markov é determinado pelo seu núcleo de transição

$q[(\theta^k, \beta^k), (\theta^{k+1}, \beta^{k+1})] = P[M_{k+1} = (\theta^{k+1}, \beta^{k+1}) \mid M_k = (\theta^k, \beta^k)]$, a probabilidade de deslocar para um novo estado (θ^1, β^1) dado o estado atual (θ^k, β^k) . Chib and Greenberg (1995) descreveram a forma geral do MCMC sob enfoque da TRI como:

1. Para gerar $\boldsymbol{\theta}^k \sim p(\boldsymbol{\theta} \mid \boldsymbol{\beta}^{k-1}, \mathbf{X})$

(a) Gerar $\boldsymbol{\theta} \sim q_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{k-1}, \boldsymbol{\theta})$, um núcleo de transição proposto

(b) aceitar $\boldsymbol{\theta}^k = \boldsymbol{\theta}^*$ com probabilidade

$$\alpha(\boldsymbol{\theta}^{k-1}, \boldsymbol{\theta}^*) = \min \left\{ \frac{p(X \mid \boldsymbol{\theta}^*, \boldsymbol{\beta}^{k-1})p(\boldsymbol{\theta}^*, \boldsymbol{\beta}^{k-1})q_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{k-1})}{p(X \mid \boldsymbol{\theta}^{k-1}, \boldsymbol{\beta}^{k-1})p(\boldsymbol{\theta}^{k-1}, \boldsymbol{\beta}^{k-1})q_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{k-1}, \boldsymbol{\theta}^*)}, 1 \right\} \quad (3.1)$$

senão $\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1}$

2. Para gerar $\boldsymbol{\beta}^k \sim p(\boldsymbol{\beta} \mid \boldsymbol{\theta}^k, \mathbf{X})$

(a) Gerar $\boldsymbol{\beta} \sim q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{k-1}, \boldsymbol{\beta})$

(b) aceitar $\boldsymbol{\beta}^k = \boldsymbol{\beta}^*$ com probabilidade

$$\alpha(\boldsymbol{\beta}^{k-1}, \boldsymbol{\beta}^*) = \min \left\{ \frac{p(X \mid \boldsymbol{\theta}^k, \boldsymbol{\beta}^*)p(\boldsymbol{\theta}^k, \boldsymbol{\beta}^{k-1})q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{k-1})}{p(X \mid \boldsymbol{\theta}^k, \boldsymbol{\beta}^{k-1})p(\boldsymbol{\theta}^k, \boldsymbol{\beta}^{k-1})q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{k-1}, \boldsymbol{\beta}^*)}, 1 \right\}$$

senão $\boldsymbol{\beta}^k = \boldsymbol{\beta}^{k-1}$

A cadeia de Markov resultante terá distribuição estacionária $\pi(\boldsymbol{\theta}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{X}) \propto p(X \mid \boldsymbol{\theta}, \boldsymbol{\beta})p(\boldsymbol{\theta}, \boldsymbol{\beta})$. Se as distribuições $q_{\boldsymbol{\theta}}$ e $q_{\boldsymbol{\beta}}$ forem simétricas em seus argumentos, então serão cancelados no cálculo de α .

Capítulo 4

Estimação do traço latente

A seguir será apresentado o método de estimação bayesiano do traço latente pela média *a posteriori* sob o modelo de Samejima. A notação apresentada será considerada que alguns itens do teste já foram administrados ($l - 1$, mais especificamente) e outros não. Também adotaremos θ_j como θ para o traço latente, pois em um TAC não há necessidade de indexar indivíduos..

4.1 Máxima verossimilhança

Supondo o modelo de Samejima e considerando que $l - 1$ questões já foram administrados, temos que a função de verossimilhança é dada por:

$$L(\theta) = \prod_{i=1}^{l-1} \prod_{k=0}^{m_i} P_{i,k}^{x_{i,k}}(\theta) \quad (4.1)$$

onde $P_{i,k}$ é dado em (2.2), $k = 0, \dots, m_i$ as categorias dos itens e $x_{i,k}$ são variáveis que assumem o valor 1 se a categoria k do item i for escolhida. A log-verossimilhança é dada por

$$l(\theta) = \prod_{i=1}^{l-1} \prod_{k=0}^{m_i} x_{i,k} \ln(P_{i,k}(\theta)),$$

e o estimador de máxima verossimilhança é dado por

$$\hat{\theta}_{l-1} = \arg \max_{\theta} \{l(\theta)\},$$

a função do escore e a matriz Hessiana são dados a seguir:

$$S(\theta) = D \sum_{i=1}^I \sum_{k=0}^{m_i} \left\{ \frac{a_i x_{ik} (W_{ik}^+ - W_{i(k+1)}^+)}{P_{ik}} \right\}, \quad (4.2)$$

$$H(\theta) = D \sum_{i=1}^{l-1} \sum_{k=1}^{m_i} \left\{ a_i^2 x_{ik} \left[\frac{a_i (1 - 2P_{ik}^+) W_{ik}^+ - (1 - 2P_{i(k+1)}^+) W_{i(k+1)}^+}{P_{ik}} - \frac{(W_{i(k)}^+ - W_{i(k+1)}^+)^2}{P_{ik}^2} \right] \right\}, \quad (4.3)$$

4.2 Métodos Bayesianos

O método bayesiano para a estimação de θ para o modelo de Samejima (2.2), usando a distribuição a *posteriori* atualizada para cada item administrado é dado por:

$$\pi(\theta | \mathbf{x}) = \frac{L(\theta | \mathbf{x}) g(\theta | \boldsymbol{\lambda})}{\int L(\theta | \mathbf{x}) g(\theta | \boldsymbol{\lambda}) d\theta}, \quad (4.1)$$

onde, $g(\theta | \boldsymbol{\lambda})$ é a distribuição a *priori* de θ , $\boldsymbol{\lambda}$ são os hiperparâmetros e $\mathbf{x}$ a matriz de respostas aos itens formada pelos elementos $x_{i,k}$ definido em (4.1).

De (4.1) temos que a distribuição a posteriori para um indivíduo pode ser escrita como

$$\pi(\theta | \mathbf{x}) \propto L(\theta | \mathbf{x}) g(\theta | \boldsymbol{\lambda}) \quad (4.2)$$

Utilizando a moda a *posteriori* para estimar o traço latente, temos que o logaritmo de (4.2) é:

$$\log \pi(\theta) \propto l(\theta) + \log g(\theta | \boldsymbol{\lambda})$$

Com a função de estimação bayesiana, dada por

$$S(\theta)_B = \frac{\partial l(\theta)}{\partial \theta} + \frac{\log g(\theta | \boldsymbol{\lambda})}{\partial \theta}$$

Considerando como priori a distribuição $N(\mu_\theta, \sigma_\theta^2)$ temos que $S(\theta)_B$ pode ser escrito como

$$S(\theta)_B = D \sum_{i=1}^{l-1} \sum_{k=0}^{m_i} \left\{ \frac{a_i x_{ik} (W_{ik}^+ - W_{i(k+1)}^+)}{P_{ik}} \right\} - \frac{\theta - \mu_\theta}{\sigma_\theta^2} \quad (4.3)$$

Observe que a equação (4.3) não pode ser resolvida explicitamente, por isso é necessário aplicar algum método iterativo como o Newton-Raphson ou Escore de Fisher para encontrar uma solução.

Neste trabalho, primeiramente, foi aplicado o método de Newton-Raphson que é dado por

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - [H(\hat{\theta}^{(t)})]^{-1} S(\hat{\theta}^{(t)}), \quad (4.4)$$

onde $S(\theta)$ é o Vetor Escore e $H(\theta)$ é a matriz Hessiana que é dada por

$$H(\theta)_B = D \sum_{i=1}^{l-1} \sum_{k=1}^{m_i} \left\{ a_i^2 x_{ik} \left[\frac{a_i (1 - 2P^+ - ik) W_{ik}^+ - (1 - 2P^+ - i(k+1)) W_{i(k+1)}^+}{P_{ik}} - \frac{(W_{i(k)}^+ - W_{i(k+1)}^+)^2}{P_{ik}^2} \right] \right\} - \frac{1}{\sigma_\theta^2}, \quad (4.5)$$

e assim substituir em (4.4) as expressões (4.3) e (4.5).

Porém ao aplicar esse método, alguns resultados não convergiram. Mislevy and Stocking (1989) sugeriram que a estimativa pontual de θ seja obtida pela esperança a posteriori (EAP) porque o método pode ser calculado sem a necessidade de métodos iterativos e como não há problemas de convergência o método MAP nesse trabalho foi substituído pelo método EAP.

A esperança a posteriori é dada por

$$E(\theta \mid \mathbf{x}, \lambda) = \frac{\int^{\Re} \theta L(\mathbf{x} \mid \theta) g(\theta \mid \lambda) \partial \theta}{\int^{\Re} P(\mathbf{x} \mid \theta) g(\theta \mid \lambda) \partial \theta},$$

e aplicar um método numérico para solucionar as integrais.

E do mesmo modo, podemos calcular a variância

$$Var(\theta \mid \mathbf{x}, \lambda) = \frac{\int^{\Re} (\theta - E(\theta))^2 L(\mathbf{x} \mid \theta) g(\theta \mid \lambda) \partial \theta}{\int^{\Re} P(\mathbf{x} \mid \theta) g(\theta \mid \lambda) \partial \theta}.$$

Capítulo 5

Seleção de itens

A seleção de um item a ser apresentado ao indivíduo é feita segundo algum critério estatístico que quantifica a informação que o item traz para determinado nível do traço latente. A seguir, serão apresentados os critérios mais tradicionais presentes na literatura.

5.1 Informação de Fisher

É a medida mais usual de informação dos itens a partir de θ . Para o i -ésimo item, a informação de Fisher para um nível de traço latente θ é definida por:

$$I_i(\theta) = -E \left[\left(\frac{\partial}{\partial \theta} \log f(\mathbf{x}_{i\cdot} \mid \theta) \right)^2 \right], \quad (5.1)$$

com $f(x_{i\cdot} \mid \theta)$ e $\mathbf{x}_{i\cdot} = (x_{i,1}, \dots, x_{i,m_i})'$, em que $x_{i,k}$ foi definido em (4.1). A função de probabilidade da resposta dada pelo modelo (2.2), considerando o modelo de Samejima, a informação de Fisher é dada por:

$$I_i(\theta) = D^2 a_i^2 \sum_{k=0}^{m_i} \left\{ \frac{(W_{ik}^+ - W_{i(k+1)}^+)^2}{P_{i,k}(\theta)} \right\}.$$

onde $W_{ik}^+ = P_{i,k}^+ Q_{i,k}^+$, $Q_{i,k}^+ = 1 - P_{i,k}^+$, $P_{i,k}^+$ é dado em (2.1) e $P_{i,k}$ em (2.2).

Supondo independência local, a informação total de Fisher do conjunto $L = 1, \dots, l - 1$ itens já apresentados no teste é igual à soma das informações desses itens, ou seja:

$$I_L(\theta) = \sum_{i=1}^{l-1} I_i(\theta)$$

O critério determina que o item a ser apresentado ao indivíduo resume-se àquele com máximo valor de $I_i(\theta)$ em um certo intervalo centrado em 0, dentre todos os itens do banco que ainda não foram respondidos pelo indivíduo.

5.2 Informação de Kullback-Leibler

A informação de Kullback-Leibler mede a distância entre duas verossimilhanças para um mesmo espaço paramétrico. Quanto maior o valor da informação maior a discrepância entre as duas funções. A medida da informação de Kullback-Leibler para o i -ésimo item no teste entre os níveis de traço latente estimado, $\hat{\theta}$, e real, θ , é dado por.

$$K_i(\theta, \hat{\theta}) = E \left[\log \frac{f(\mathbf{X}_i. | \theta)}{f(\mathbf{X}_i. | \hat{\theta})} \right],$$

Supondo independência condicional entre as respostas, a informação de Kullback-Leibler após $l - 1$ itens administrados é escrita como

$$K_L(\theta, \hat{\theta}) = \sum_{i=1}^{l-1} K_i(\theta, \hat{\theta}). \quad (5.1)$$

No Modelo de Samejima, a informação de Kullback-Leibler é dada por:

$$K_i(\theta, \hat{\theta}) = \sum_{k=0}^{m_i} P_{i,k}(\theta) \log \frac{P_{i,k}(\theta)}{P_{i,k}(\hat{\theta})} \quad (5.2)$$

Note que o verdadeiro valor do traço latente (θ) é desconhecido. Sands and Waters (1996) propuseram considerar um intervalo de confiança do traço latente $[\hat{\theta} - \delta_L, \hat{\theta} + \delta_L]$, δ_L uma função decrescente em relação a L , e integrar (5.2). Assim o critério para selecionar o próximo item a ser apresentado no teste é dado por:

$$i_l \equiv \arg \max_i \left\{ \int_{\hat{\theta} - \delta_L}^{\hat{\theta} + \delta_L} K_i(\theta, \hat{\theta}) d\theta : i \in L \right\}$$

5.3 Informação a *posteriori* Ponderada

Esta abordagem é uma alternativa à estimação pontual da informação (em $\hat{\theta}$). Esse método considera o valor da função de informação do teste em relação a distribuição a *posteriori* de θ e é dado por.

$$I_S(PW) = \int_{-\infty}^{\infty} \pi(\theta | \mathbf{x}) I_T(\theta) d\theta, \quad (5.1)$$

tal que, $\pi(\theta | \mathbf{x})$ representa a função a *posteriori* de θ dado o vetor de resposta $\mathbf{x} = x_{Lk}$, $L = 1, \dots, l - 1$; $k = 0, \dots, m_i$ e $I_T(\theta)$ é o valor da função de informação do teste (como por exemplo a informação de Fisher ou Kullback-Leibler).

5.4 Abordagem *Shadow Test*

A abordagem consiste em apresentar a um indivíduo um item em cada passo do teste, escolhido de forma a maximizar a informação (de Fisher ou de Kullback-Leibler, por exemplo) para a estimativa corrente do seu traço latente, θ . A escolha é feita também levando em consideração algumas restrições definidas a priori, tais como, número de itens no teste, conteúdo mínimo (ou diversificado) a ser avaliado, tamanho máximo (em termos de número de páginas ou de palavras) do teste, nível de dificuldade mínimo a ser avaliado, etc.

A cada passo, a abordagem *Shadow test* (Linden and Hambleton; 1997) constrói um teste com todos os itens, incluindo todos aqueles já administrados anteriormente, que obedecem as restrições especificadas para o atual valor estimado de θ . A vantagem de testes com esse tipo de construção é a facilidade em incorporar restrições para seleção dos itens.

O próximo item a ser administrado no teste usando essa abordagem é um item do *Shadow test* não apresentado anteriormente e ótimo para o atual valor estimado de θ . Depois do item ser administrado, o restante dos itens não utilizados no teste voltam para o banco de itens como itens disponíveis a serem administrados nas próximas seleções.

Um pseudo-algoritmo pode ser usado para resumir a idéia:

Passo 1: Inicialização com uma estimativa para θ

Passo 2: Construção de um *shadow test*

Passo 3: Administração do item no *Shadow test* que otimiza o teste para o atual valor estimado de θ

Passo 4: Atualização da estimativa de θ

Passo 5: Atualização do modelo de construção do teste para incluir o item administrado no próximo *shadow test*

Passo 6: Retorno de todos os itens não utilizados para o banco de itens

Passo 7: Repetição dos passos 2 a 6 até que um número especificado de itens sejam administrados.

Um Exame de Proficiência em Inglês (EPI) aplicado aos alunos de pós-graduação em Ciência da Computação no ICMC-USP será utilizado para exemplificar as ideias básicas de um *Shadow test*. Acompanhe pela Tabela 5.1 a ilustração dessa estrutura do banco e, na Tabela 5.2, a definição da notação a ser utilizado no *Shadow test*. O banco de itens possui 27 itens divididos em sete conteúdos: 1-*introduction*, 2-*abstract*, 3-*settings*, 4-*review*, 5-*comprehension*, 6-GAP e 7-*purpose*. Cada um desses conteúdos possui um certo número de itens e alguns desses itens referem-se a um texto em comum, de um total de 9 textos no banco. Por exemplo, no *abstract* temos um texto e quatro itens que pergutam sobre ele.

Tabela 5.1: Descrição do banco de itens do EPI.

Atributos dos estímulos	Atributos dos estímulos							
	Categorizado (conteúdo)	Quantitativo (nº palavras)						
Atributos dos itens	1-Int.	374						
	2-Abs.	240						
		76						
	3-Set.	95						
		126						
	4-Rev.	281						
		121						
	5-comp.	214						
		585						
Atributos dos itens	1	1,2,3,4	Itens					
	2	5,6,7,8						
	3	9						
	4	10						
	5	11						
	6	12						
	7	13						
	8	14,15,16						
	9	17,18,19						
	nenhum(0)		20-25	26,27				
Atributos dos itens	Categorizado (Conteúdo)	1-Int.	2-Abs.	3-Set.	4-Rev.	5-comp.	6-GAP	7-Pur.
	Quantitativo (nº palavras)	11,11, 14,11	8,10, 8,5	5,5, 5	5,5	14,14,14, 14,14,14	12,12,12, 12,12,12	15,15

Na Tabela 5.2, encontra-se a notação geral para a descrição de um banco usado em um *shadow test* e seus valores particularizados no exemplo do EPI.

Tabela 5.2: Notação para descrever um banco de itens usado em um *Shadow test*.

Geral	no EPI
$i = 1, \dots, I$ (Itens no banco)	$I = 27$
$s = 1, \dots, S$ (estímulos no banco)	$S = 9$
$U_s = 1, \dots, S$ (conjunto de itens no banco com o estímulo s)	$U_1 = \{1, 2, 3, 4\}$ $U_2 = \{5, 6, 7, 8\}$ $U_3 = \{9\}$...
$c^{item} = 1, \dots, C^{item}$ (atributos categorizados dos itens)	$U_9 = \{17, 18, 19\}$ $C^{item} = \{1\}$ (conteúdo)
$d_c^{item} = 1, \dots, D_c^{item}$ (número de classes do atributos categorizados do c^{item})	$D_1^{item} = \{7\}$ (conteúdo)
$E_{c(dc)}^{item}$ (conjunto de itens da classe d_c^{item} do atributo categorizado c^{item})	$E_{1(1)}^{item} = \{1, 2, 3, 4\}$ $E_{1(2)}^{item} = \{5, 6, 7, 8\}$ $E_{1(3)}^{item} = \{9, 10, 11\}$...
$q^{item} = 1, \dots, Q^{item}$ (número de atributos quantitativos dos itens)	$E_{1(7)}^{item} = \{26, 27\}$ $Q^{item} = \{1\}$ (nº parágrafos do enunciado)
$T_{q,i}^{item}, q = 1, \dots, Q^{item}, i = 1, \dots, I$ (valor do atributo quantitativo q^{item} para o item i)	$T_{1,1}^{item} = \{11\}$ $T_{1,2}^{item} = \{11\}$...
$c^{est} = 1, \dots, C^{est}$ (atributos categorizados dos estímulos)	$T_{1,27}^{item} = \{15\}$ $C^{est} = \{1\}$ (conteúdo)
$d_c^{est} = 1, \dots, D_c^{est}$ (número de classes do atributos categorizados do c^{est})	$D_1^{est} = \{5\}$ (conteúdo)
$E_{c(dc)}^{est}$ (conjunto de estímulos da classe d_c^{est} do atributo categorizado c^{est})	$E_{1(1)}^{est} = \{1\}$ $E_{1(2)}^{est} = \{2\}$ $E_{1(3)}^{est} = \{3, 4, 5\}$...
$q^{est} = 1, \dots, Q^{est}$ (número de atributos quantitativos dos estímulos)	$E_{1(5)}^{est} = \{8, 9\}$ $Q^{est} = \{1\}$ (nº parágrafos do texto)
$T_{q,s}^{est}, q = 1, \dots, Q^{est}, s = 1, \dots, S$ (valor do atributo quantitativo q^{est} para o estímulo s)	$T_{1,1}^{est} = \{374\}$ $T_{1,2}^{est} = \{240\}$...
	$T_{1,9}^{est} = \{585\}$

onde i e s são os índices dos itens e dos estímulos no banco respectivamente, c^{item} e c^{est} são os índices dos atributos categorizados dos itens e dos estímulos no banco respectivamente, l_c^{item} e l_c^{est} são os índices das classes dos atributos categorizados dos itens e dos estímulos respectivamente, q^{item} e q^{est} são os índices dos atributos quantitativos dos itens e dos estímulos no banco respectivamente.

Nesse exemplo, os itens que se referem a um mesmo texto possuem o que genericamente denominamos de estímulo em comum, podemos definir os conteúdos como o que chamamos de atributo categorizado para os itens e estímulos, composto de 7 classes e 5 classes respectivamente. Tanto os itens quanto os estímulos também podem ter um atributo quantitativo associado. Neste exemplo definimos o número de palavras do texto e do enunciado do item como atributos quantitativos de estímulo e item, respectivamente.

Com base nesse banco de itens do EPI, suponha que as restrições para o teste com as seguintes características:

1. um número fixo de 10 itens e um número maior ou igual a 2 textos: 1 texto deverá aparecer no conteúdo *Introduction* e 1 texto deverá aparecer no conteúdo *Abstract*.
2. que o número de itens associados a cada texto não exceda 2.
3. que pelo menos 1 item de cada conteúdo esteja no teste,
4. que o número de palavras total dos enunciados dos 10 itens não exceda 150 e o número de palavras em todos textos (estímulos) não exceda 1000.
5. que os itens 20, 21 e 22 não apareçam simultaneamente no mesmo teste (ou porque, se 1 aparecer, o outro passa a ser redundante ou porque o enunciado de 1 pode fornecer informações sobre a resposta do outro, por exemplo)

A Tabela 5.3 traz a notação necessária para a construção do *shadow test* descrito acima. Note que o teste pode apresentar 1 ou mais restrições como a descrita em (5), (Na tabela 5.3 essa restrição é imposta por $\sum_{s \in V_{ex}^{item}}$ onde V_{ex}^{item} é o conjunto de itens que tenham algum tipo de conflito e z_s é o número de itens desse conjunto que pode aparecer simultaneamente no teste).

Tabela 5.3: Notação para descrever um banco de itens usado em um *Shadow test* no exemplo EPI.

Geral	no EPI
$x_i \in \{0, 1\}$, para $i = 1, \dots, I$ (= 1 se o item i é selecionado para o <i>Shadow test</i> e 0, c.c.)	
$z_s \in \{0, 1\}$, para $s = 1, \dots, S$ (= 1 se o estímulo s é selecionado para o <i>Shadow test</i> e 0, c.c.)	
$\sum_{i=1}^I x_i = n$; (números de itens <i>Shadow test</i>)	$n = 10$
$\sum_{s=1}^S z_s = m$; (número de estímulos <i>Shadow test</i>)	$n = 10$ $m \geq 2$
$\sum_{i \in I_{l-1}^*} x_i = l - 1$; (número de itens já administrados no <i>Shadow test</i>)	$l - 1$ itens já administrados
$\sum_{i \in U_s} x_i = n_s$, para $s = 1, \dots, S$; (números de itens do estímulo s no <i>Shadow test</i>)	$n_1 \leq 2, \dots, n_9 \leq 2$
$\sum_{i \in K_{c(lc)}^{item} x_i$; (números de itens da classe l_c^{item} do atributo categorizado no <i>Shadow test</i> c^{item})	≥ 1 para $K_{1,1}^{item}, \dots, K_{1,7}^{item}$
$\sum_{i \in K_{c(lc)}^{est} z_s$; (números de estímulos da classe l_c^{est} do atributo categorizado no <i>Shadow test</i> c^{est})	$= 1$ para $K_{1,1}^{est}$ e $K_{1,2}^{est}$
$\sum_{i=1}^I T_{qi}^{item} . x_i$; (valor total do atributo quantitativo q^{item} para os itens no <i>Shadow test</i>)	≤ 150
$\sum_{i=1}^S T_{qi}^{est} . z_s$; (valor total do atributo quantitativo q^{est} para os estímulos no <i>Shadow test</i>)	≤ 1000
V_{ex} ; (conjunto com itens mutuamente exclusivos, isto é, que não apareçam no mesmo TAC)	$V_{ex} = \{20, 21, 22\}$
$\sum_{i \in V_{ex}} x_i$; (número de itens de V_{ex} que aparecem no <i>Shadow test</i>)	≤ 1

Respeitando estas restrições, os itens selecionados no *shadow test* maximizam a informação, neste trabalho foi utilizada a Informação de Kullback-Leibler, pois em TACs a Informação de Kullback-Leibler possui uma performance melhor que a informação de Fisher no começo do teste (Passos et al.; 2008), isto é:

$$\max_{i \in A_l} \sum K_i(\theta, \hat{\theta}_{l-1}),$$

onde $l - 1$ é o número de itens apresentados no teste, $\hat{\theta}_{l-1}$ é a estimativa de θ depois de $l - 1$ itens respondidos, i_l é o l -ésimo item administrado no teste, $L_{l-1} \equiv \{i_1, \dots, i_{l-1}\}$ é o conjunto dos $l - 1$ itens administrados no teste, $T_l \equiv \{i_1, \dots, i_{l-1}, i'_l, \dots, i'_n\}$ são os itens ótimos do *Shadow test*, na l -ésima iteração, $A_l = \{i'_l, \dots, i'_n\}$ são itens ótimos não presentes no teste, mas presentes na solução do *shadow test* até o momento e $K_i(\theta, \hat{\theta})$ como definido em (5.2).

O próximo item a ser apresentado no teste é o item que possui a maior informação entre eles, ou seja

$$i_l \equiv \arg \max_i \{K_i(\hat{\theta}_{l-1}); i \in A_l\},$$

Neste trabalho, os itens a serem apresentados são sempre os que possuem a maior informação de Kullback-Leibler, em Linden and Hambleton (1997) quando um estímulo novo for aparecer no teste em uma certa iteração, todos os itens com aquele estímulo no *Shadow test* dessa iteração irão ser apresentados consecutivamente. A diferença é que no segundo caso, o texto do estímulo aparecerá apenas naquela etapa do teste; enquanto no primeiro caso, um texto lido por um examinado poderá aparecer de novo, porém a informação dos itens poderá ser maior que no segundo caso.

5.5 Critério de parada

O critério de parada do teste define o momento em que nenhum item mais é necessário ser respondido pelo indivíduo. Há dois critérios de paradas mais usados: o mais comum é o critério de parada por número de itens administrados e o outro é até ser atingido uma precisão da estimativa do traço latente, ou seja, um valor pré-determinado do erro padrão. Caso o objetivo do teste seja classificar um indivíduo de acordo com seu traço latente (Weiss and Kingsbury; 1984), o erro-padrão é usado para construir um intervalo de confiança. Se a estimativa do traço latente e seu intervalo de confiança estiverem acima ou abaixo do ponto de corte entre duas categorias, o indivíduo será classificado na respectiva categoria. Neste trabalho, um número fixo de itens foi adotado como critério de parada.

Capítulo 6

Simulação

Para fazer as simulações e aplicações a dados reais a informação de Kullback-Leibler foi utilizada como a função objetiva do *Shadow test*. Para as estimações do traço latente foi utilizado o método bayesiano EAP, pois a estimação de máxima verossilhança pode apresentar problemas em caso das respostas a todos os itens serem iguais ou em caso de respostas aberrantes, isto é, se um indivíduo de baixo trato latente seleciona a resposta de maior categoria e vise-versa (Barker; 1992). Neste trabalho, originalmente, foi utilizado o método MAP, porém algumas estimações não convergiram e, portanto, foi substituído pelo método do EAP que não apresentou esse problema.

As simulações realizadas neste capítulo foram divididas em duas sessões: uma para um banco grande (500 itens) e outra com um banco pequeno (21 itens do BDI). Para cada seção, vários TACs foram simulados com diferentes números de itens: 25 e 10 itens, para o banco grande, e 21, 10 e 5 itens, para o banco pequeno.

6.1 Banco Grande

Utilizando o software R, foram simulados um banco com 500 itens 3 categorias de respostas. O parâmetro a de cada item foi gerado através de uma distribuição $\log - normal(0, 7; (0, 1)^2)$ (Mislevy; 1986), simulando apenas bons valores para os parâmetros de dificuldade (valor mínimo de 1,51 e valor máximo de 2,67). Para simular os parâmetros b_1 e b_2 de cada item foram gerados 2 valores de uma distribuição $normal(0; 1, 2)$, para simular parâmetros de dificuldade abrangente para todos os valores do traço latente, o menor valor foi atribuído ao b_1 e o outro ao b_2 .

As definições do número de estímulos, atributos categorizados e quantitativos foram implicadas na estrutura do EPI. 380 itens do banco estavam ligados a algum estímulo (em particular neste trabalho, um estímulo foi considerado como um texto, totalizando 9 no

banco simulados). O número de itens por estímulo foi 80 para os dois primeiros, 20 para os estímulos 3 ao 7 e 60 para os dois últimos. Um atributo categorizado com 7 classes foi definido e o número de itens por classe foi de: 80, 80, 60, 40, 120, 90 e 30 para as classes 1 à 7, respectivamente. Cada estímulo foi associado a um atributo quantitativo, representando o número de palavras do texto, com valores de 100, 150, 90, 80, 100, 150, 125, 50 e 60 para os estímulos 1 ao 9, respectivamente.

Em seguida, foram fixados 900 valores do traço latente iguais a -3 , $-2,25$, $-1,5$, $-0,75$, 0 , $0,75$, $1,5$, $2,25$, 3 (100 de cada valor). Para simular as respostas foi utilizado o modelo de Smejima com 3 categorias de resposta, os valores gerados dos parâmetros de cada item e a probabilidade que o indivíduo selecionaria cada categoria de um item. Um valor entre 0 e 1 de uma distribuição uniforme padrão foi gerado para determinar a categoria de resposta do respectivo indivíduo ao item. Utilizando essas respostas, foram estimados os valores dos traços latentes. Assim sendo, temos os valores reais dos traços latentes (900 valores fixados de -3 à 3) e suas respectivas estimativas.

Conforme dito no início deste capítulo, dois TACs foram simulados através da abordagem *Shadow test*: um com 25 itens e outro com 5 itens. Em ambos os testes, o primeiro item foi escolhido aleatoriamente entre os 20 mais informativos para $\theta = 0$. Além das restrições de números de itens, para o banco com 25 foram impostas as seguintes restrições (para o TAC com 5 itens não foi imposta mais nenhuma restrição):

- número de estímulos no teste maior ou igual a 5.
- número de itens por estímulo maior ou igual ao mínimo entre 2 e o número de itens disponíveis no banco para o estímulo em questão.
- no mínimo 2 itens para as classes 3, 4 e 7 do atributo categorizado presentes no teste; e no mínimo 3 itens para as classes 1, 2, 5 e 6.
- a soma do número de palavras dos estímulos presentes no teste não pode ser superior a 904.

Lembre que para a seleção de um item, um *Shadow test* é construído para se determinar aquele com maior informação de Kullback-Leibler para a atual estimativa do traço latente e que respeita todas as restrições impostas acima. Para calcular as estimativas, foi utilizado o método bayesiano EAP.

As diferenças entre os valores reais e os estimados dos traços latentes estão apresentados na figura 6.1.

Note que para o teste maior, as diferenças são relativamente pequenas e aleatórias em torno do zero. Enquanto que para o teste menor, observa-se uma maior magnitude das diferenças e uma tendência a má estimação do traço latente nas simulações iniciais

Figura 6.1: Diferença entre os valores reais θ e estimados $\hat{\theta}$ dos traços latente para os TACs de 25 e 5 itens do banco grande (500 itens).

e finais. Isso se deve ao fato de que as simulações foram realizadas na ordem crescente dos valores do traço latente (de -3 até 3). Para valores altos e baixos de θ , portanto, a estimação no TAC é pior. Uma provável justificativa é que 5 itens são poucos para estimar com precisão o valor do θ , especialmente quando o valor inicial adotado ($\theta_0 = 0$) é distante do verdadeiro. As médias dessas 900 diferenças, denominamos vício, para os testes com 25 e 5 itens foram de 0,0049 e 0,0136, respectivamente. Os valores dos erros quadráticos médios ($EQM = \sum_{i=1}^{900} \frac{(\theta_i - \hat{\theta}_i)^2}{900}$) para os testes com 25 e 5 itens foram respectivamente de 0,0399 e 0,2595. Observe que o teste mais longo tem precisão consideravelmente maior do que o teste mais curto.

O item com a maior taxa de exposição no TAC com 25 itens foi o item 171, que apareceu em 70% dos testes e foi o único que esteve presente em mais da metade deles. 333 (67%) itens não apareceram em nenhum teste, nesta etapa.

O item com a maior taxa de exposição no TAC com 5 itens foi o item 43, que apareceu em 46% dos testes. O item com maior exposição do caso anterior não apareceu em nenhum teste neste caso. 460 (92%) itens não apareceram em nenhum teste, nesta etapa.

Não há nenhum algoritmo implementado para controlar a taxa de exposição dos itens. De fato, as imposições das restrições no *Shadow test* citadas no início desta seção, deve ter agravado o problema, pois há redução no espaço de itens que podem ser selecionados para o teste.

Para o teste com 25 itens, o tempo médio foi de 22.8 segundos, 0,9 segundos por item. O tempo máximo levado para a simulação de um teste foi de 26,4 segundos e o mínimo 16.6 segundos. Mesmo no pior caso, o tempo para selecionar o item subsequente no teste é menor que 1,5 segundos, bastante bom para a aplicação de um TAC.

O tempo médio para a simulação de cada traço latente para o teste com 5 itens foi de 1,9 segundos, 0,38 segundos por item. Uma possível razão para a diferença entre o tempo por item dos dois TACs pode ser que no teste maior é necessário maior tempo para eliminar os itens já presentes no teste, e calcular a função de verossilhança. O tempo máximo levado para a simulação de um teste foi de 2,4 e o mínimo 1,6 segundos.

6.2 Banco pequeno

Um banco de itens reais foi considerado nesta etapa do trabalho, baseado no BDI. O Inventário de Depressão de Beck (Beck et al.; 1961) é um teste que originalmente possui 21 itens referentes a sintomas e atitudes comumente presentes na depressão (por exemplo: tristeza, pessimismo, autodepreciação, etc.). Cada item possui 4 categorias de respostas de acordo com a intensidade, com que o sintoma está presente. O BDI é provavelmente a medida de auto-avaliação mais utilizada atualmente.

A aplicação do Modelo de Samejima às respostas de 591 pacientes clínicos do Hospital das Clínicas da medicina da USP ao BDI, cedido pelo doutor Yuang-Pang Wang do Instituto de Psiquiatria, gerou estimativas dos traços latentes individuais e dos parâmetros de localização b_{ik} , para $k = 0, 1, 2, 3$ e $i = 1, \dots, 21$, onde k é o número de categorias do modelo e i é o número de itens do teste. A tabela 7.2 apresenta as estimativas dos parâmetros dos itens, consideradas nesta seção como os valores verdadeiros para a formação do banco.

Da mesma forma que na seção anterior, 900 valores de θ iguais a $-3, -2, 25, -1, 5, -0, 75, 0, 0, 75, 1, 5, 2, 25, 3$ (100 de cada valor) foram fixados e usados para simular as respostas aos itens do banco, considerando o modelo de Samejima para itens com quatro categorias de resposta. Utilizando essas respostas, foram estimados os valores dos traços latentes individuais.

Na primeira simulação de um TAC, foram usados todos os 21 itens do banco. Nas outras simulações, foi aplicada a abordagem *Shadow test* para testes com 10 e 5 itens. Um valor inicial igual a zero foi atribuído ao traço latente e, portanto, o primeiro item selecionado foi o mesmo para todos os testes. As diferenças entre os valores reais e estimados dos traços latentes estão apresentados na Figura 6.2.

A não ser para valores altos e baixos de θ , as diferenças estão aleatórias em torno do zero, com amplitudes menores quanto maior for o tamanho do teste. A justificativa para a tendência observada nesses gráficos é a mesma que a da seção anterior. Os vícios para os testes com 21, 10 e 5 itens foram de 0,1351, 0,1600 e 0,2157, respectivamente. Se as observações de valores de $-3, -2, 25, 2, 25$ e 3 não forem consideradas, os vícios foram de $-0,0104, 0,0025, \text{ e } 0,0040$. Além disso, para os valores excluídos, as estimativas não foram tão próximos aos valores fixados dos traços latentes. Isso se deve ao fato de que os parâmetros de dificuldade dos itens do banco não são baixos ou altos o suficiente para que

Figura 6.2: Diferença entre os valores reais (θ) e estimados ($\hat{\theta}$) dos traços latente para o teste com todos os itens e os TACs de 10 e 5 itens do banco pequeno (21 itens).

os traços latentes estimados atinjam valores extremos (± 3 , ± 2 , 25, por exemplo) mesmo quando todo o banco é utilizado.

Com o intuito de comparar os resultados do TAC com a teoria clássica, os indivíduos foram divididos em classes diagnósticas a partir de seus escores totais no BDI. Há diversos critérios de diagnóstico de depressão em um paciente com base no BDI (Beck et al.; 1961; Gotlib; 1984; Oliver and Simmons; 1984; Tanaka-Matsumi and Kameoka; 1986; Louks et al.; 1989) para diferentes amostras e objetivos de estudo. Neste trabalho será utilizado o critério sugerido por Kendall et al. (1987), onde os indivíduos são divididos em 3 grupos: um grupo de indivíduos sem depressão (escore total igual a 15 ou menos), um grupo de indivíduos com disforia (escore total entre 16 e 20) e um grupo de indivíduos com depressão (escores total acima de 20). Este critério é utilizado para indivíduos não diagnosticados previamente com depressão.

Foram criados boxplots 6.3 utilizando os valores estimados dos traços latentes para o teste com todos os 21 itens e TACs de 10 e 5 itens divididos em três grupos. Para determinar

os grupos foi utilizado o sistema de classificação recomendado pelos Kendall et al. (1987). Este boxplot será apresentado a seguir.

Figura 6.3: Boxplots dos traços latentes estimados para indivíduos subdivididos segundo o critério de Kendall.

Observa-se nos 3 casos (TAC com o banco todo com 10 e 5 itens) que os indivíduos sem depressão (score menor que 16) obtiveram, em geral, estimativas menores que os indivíduos com disforia (score entre 16 e 20), que por sua vez obtiveram, também de maneira geral, estimativas menores que os indivíduos com depressão (score maior que 20). Adicionalmente, apesar das distribuições das estimativas de θ serem parecidas para os 3 TACs simulados, há uma leve tendência de melhor separação entre elas, quando maior o número de itens respondidos.

Os EQM's, para os 900 traços latentes fixados, com 21, 10 e 5 itens foram respectivamente de 0,1118, 0,1881, 0,2612. Sem considerar as observações -3, -2,25, 2,25 e 3 temos que os EQM's foram de 0,0680, 0,1437 e 0,2375.

Nas Tabelas 6.1 e 6.2 estão apresentadas as frequências (em porcentagem) de seleção de cada item para cada valor de θ (com 100 observações cada) e no total simulado para os testes de 10 e 5 itens, respectivamente.

Tabela 6.1: Percentual de seleção dos itens do BDI para um TAC com 10 itens

Item do BDI											
θ	1	2	3	4	5	6	7	8	9	10	11
<i>Todos</i>	80	37	59	100	33	4	40	61	41	70	82
-3	9	0	0	100	0	0	0	0	0	96	100
-2,25	41	0	0	100	0	0	0	0	0	84	100
-1,5	80	0	1	100	0	0	0	3	0	85	100
-0,75	99	17	29	100	2	0	0	51	0	99	100
0	100	95	99	100	82	9	52	99	0	82	54
0,75	93	100	100	100	100	30	98	99	72	3	5
1,5	100	98	100	100	94	0	25	99	100	2	82
2,25	100	26	100	100	17	0	84	97	100	76	100
3	100	0	100	100	0	0	100	100	100	100	100

Item do BDI										
θ	12	13	14	15	16	17	18	19	20	21
<i>Todos</i>	50	58	49	42	36	47	47	0	39	24
-3	0	0	100	100	100	100	98	0	100	97
-2,25	1	2	100	100	99	100	94	0	100	79
-1,5	8	16	99	100	92	100	83	0	97	36
-0,75	71	83	68	78	31	98	15	0	55	4
0	100	100	3	2	0	21	0	0	2	0
0,75	100	100	0	0	0	0	0	0	0	0
1,5	100	100	0	0	0	0	0	0	0	0
2,25	65	93	7	0	0	0	35	0	0	0
3	3	32	68	0	0	0	97	0	0	0

Tabela 6.2: Percentual de itens do BDI para um TAC com 5 itens

	Item do BDI										
θ	1	2	3	4	5	6	7	8	9	10	11
<i>Todos</i>	42	26	55	100	8	0	14	0	25	5	46
-3	0	0	0	100	0	0	0	0	0	0	100
-2,25	10	0	0	100	0	0	0	0	0	2	100
-1,5	47	0	0	100	0	0	0	0	0	18	100
-0,75	92	3	0	100	0	0	0	0	0	26	100
0	39	60	78	100	15	0	0	0	0	1	32
0,75	2	98	100	100	56	0	0	0	0	0	3
1,5	13	66	100	100	4	0	3	0	34	0	0
2,25	75	5	100	100	0	0	41	0	94	0	0
3	98	0	100	100	0	0	85	0	100	0	0

	Item do BDI										
θ	12	13	14	15	16	17	18	19	20	21	
<i>Todos</i>	26	39	0	27	0	43	0	0	32	0	
-3	0	0	0	100	0	100	0	0	100	0	
-2,25	0	2	0	90	0	100	0	0	96	0	
-1,5	0	8	0	53	0	100	0	0	74	0	
-0,75	18	57	0	5	0	82	0	0	17	0	
0	70	99	0	0	0	6	0	0	0	0	
0,75	42	99	0	0	0	0	0	0	0	0	
1,5	82	98	0	0	0	0	0	0	0	0	
2,25	22	63	0	0	0	0	0	0	0	0	
3	1	16	0	0	0	0	0	0	0	0	

No TAC com 10 itens, para todos os valores de θ , temos que apenas o item 19 não

apareceu em nenhum teste, o item 4 é o item que possui a maior informação de Kullback-Leibler para $\theta = 0$ e por isso apareceu em todos os testes. Mas se considerarmos os valores de cada θ , em todos eles há varios itens que aparecem em todos (ou quase todos os testes). Isso se deve o fato de que há apenas 21 itens no banco e nos casos extremos (valores -3 e 3 de θ) o número de itens com taxa de 100% de exposição do teste é ainda maior, porque os parâmetros de dificuldade dos itens sob este modelo para o BDI, em geral, não atingem valores tão extremos.

No TAC com 5 itens, para todos os valores de θ , temos que 7 itens (um terço dos itens no banco) não foram utilizados em nenhum teste, o item 4 aparece em todos os casos, como anteriormente. Individualmente, para as classes de valores de θ , há, novamente, varios itens que aparecem em 100% dos testes (ou quase 100%) e, para o caso $\theta = -3$, temos que sempre os mesmos 5 itens foram apresentados.

Porém esses fatos não caracterizam um problema, pois, por se tratar de um teste psicológico, a taxa de exposição de um item não costuma a ser relevante para o teste.

Capítulo 7

Aplicação

Uma aplicação real do TAC estudado no presente trabalho foi feita para as respostas de 518 estudantes de pós-graduação do ICMC ao EPI e às respostas de 591 universitários ao BDI utilizando o *software R* (R Development Core Team; 2008). Estimativas para os parâmetros de itens foram obtidas com o intuito de formar o banco de itens a ser utilizado no TAC. Adicionalmente, estimativas dos traços latentes foram calculadas com o objetivo de comparar os resultados da TRI com a forma de avaliação hoje utilizada, MPA para o EPI e um critério de classificação recomendado por Kendall et al. (1987) para o BDI.

7.1 Exame de Proficiência em Inglês

A atual forma de avaliação utilizada para o EPI, denominada MPA, classifica o conhecimento final do aluno como *totalmente informado*, *informado*, *parcialmente informado*, *mal informado*, *totalmente mal informado* e *não informado*, que possibilitam descrever o conhecimento ou desconhecimento sobre o conteúdo envolvido em cada questão da prova (Bruno; 1986). Cada questão possui 3 alternativas de resposta (representadas pelos vértices do triângulo da Figura 7.1). Se o examinando não estiver certo de qual resposta é a correta (por exemplo: o aluno está em dúvida entre duas alternativas), ele tem a opção de escolher entre 13 “posições” de resposta (um dos 13 pontos destacados na Figura 7.1), que representam o quão certo o estudante está de uma das 3 alternativas de resposta ser a correta. Um triângulo com letras de A à M é apresentado no início do teste para auxiliar o aluno a selecionar sua resposta. A Figura 7.1 ilustra três possíveis respostas em um teste de múltipla escolha, em que as alternativas *B* e *C* são incorretas (vértices inferiores do triângulo, em preto) e *A* é a alternativa correta (vértice superior do triângulo, em verde). Assim, cada questão é pontuada com uma das classes relativas ao conhecimento do estudante: totalmente informado (ponto verde), informado (pontos azuis), parcialmente informado (pontos roxos), mal informado (pontos vermelhos e pontos pretos na aresta inferior), totalmente

mal informado (vértices pretos) e não informado (ponto preto no centro do triângulo).

Figura 7.1: Triângulo equilátero da MPA com a caracterização das alternativas de respostas de cada questão da prova.

O critério de aprovação no EPI é baseado em duas avaliações: Avaliação Global e Avaliação por Módulos. A aprovação de um estudante ocorre quando, na Avaliação Global, obtiver 50% (ou mais) de respostas na classe Totalmente Informado, e 25% (ou menos) de respostas na classe Totalmente Mal Informado; ou 90% (ou mais) de respostas nas classes Totalmente Informado, Informado e Parcialmente Informado e 10% (ou menos) de respostas na classe Totalmente Mal Informado.

Para os estudantes que se aproximarem da nota de corte da Avaliação Global, ou seja, aqueles com uma ou duas questões abaixo de 50%, ou uma ou duas acima de 25%, ou ainda uma abaixo de 50% e uma acima de 25%, são considerados aprovados se obtiverem, com relação ao módulo que trata da Estrutura de Textos Científicos, 50% (ou mais) de respostas na classe Totalmente Informado, 25% (ou menos) de respostas nas classes Totalmente Mal Informado, Não informado, Mal Informado.

Primeiramente foram feitas estimações dos parâmetros dos itens seguindo o modelo de Samejima com 5 categorias de respostas utilizando o *software Multilog* (Toit; 2003). Porém, foi observado que as categorias que representam alguma indecisão nas respostas dos itens (respostas pertencentes às arestas do triângulo da Figura 7.1) eram incomumente selecionadas. Como consequência, no modelo de Samejima, as probabilidades de seleção dessas categorias intermediárias eram muito menores que as probabilidades das categorias das arestas do triângulo. Por essa razão, em vez de usar 5 categorias de resposta, nesse trabalho foram utilizados 3 categorias de resposta (certo, meio certo e errado). A categoria “certo” foi representada por $k = 2$, a categoria “meio certo” representada por $k = 1$ e a categoria “errado” representada por $k = 0$.

Utilizando o programa *Multilog*, a aplicação do Modelo de Samejima às respostas de 518 estudantes ao EPI gerou estimativas dos traços latentes individuais e dos parâmetros de localização b_{ik} , para $k = 0, 1, 2$ e $i = 1, \dots, 27$, onde k é o número da categoria do item e i é o número do item do banco, que se encontram na Tabela 7.1.

Figura 7.2: Matriz das curvas características dos itens do EPI

Tabela 7.1: Valores estimados dos parâmetros dos itens do EPI

Item	$a(\text{EP})$	$b_1(\text{EP})$	$b_2(\text{EP})$	Item	$a(\text{EP})$	$b_1(\text{EP})$	$b_2(\text{EP})$
1	0,30(0.12)	-0,12(0.41)	1,67(0.77)	15	4,87(0.56)	-1,33(0.07)	-0,87(0.05)
2	0,35(0.12)	0,03(0.34)	2,17(0.78)	16	2,46(0.26)	-1,38(0.13)	-0,94(0.09)
3	1,03(0.17)	-1,60(0.28)	-0,66(0.16)	17	0,74(0.15)	-1,72(0.40)	-0,23(0.19)
4	0,55(0.14)	-2,04(0.58)	-0,39(0.25)	18	2,13(0.25)	-0,79(0.09)	-0,38(0.08)
5	1,08(0.16)	-0,84(0.17)	-0,11(0.13)	19	2,04(0.23)	-1,41(0.15)	-0,78(0.10)
6	0,95(0.15)	-1,27(0.24)	-0,03(0.15)	20	1,60(0.20)	-1,03(0.13)	-0,36(0.10)
7	1,13(0.19)	-1,49(0.25)	-0,57(0.14)	21	1,53(0.22)	-1,43(0.19)	-0,65(0.12)
8	1,51(0.20)	-0,93(0.13)	-0,47(0.11)	22	1,13(0.17)	-0,96(0.17)	-0,19(0.13)
9	2,12(0.25)	-1,09(0.11)	-0,64(0.09)	23	1,36(0.17)	-1,22(0.17)	-0,46(0.12)
10	0,98(0.15)	-0,68(0.16)	0,26(0.15)	24	1,06(0.17)	-0,92(0.18)	-0,25(0.14)
11	0,62(0.13)	-0,33(0.21)	0,95(0.28)	25	1,28(0.19)	-1,09(0.17)	-0,35(0.12)
12	0,49(0.12)	-0,28(0.26)	1,09(0.36)	26	0,83(0.16)	-1,21(0.27)	-0,24(0.17)
13	1,03(0.17)	-1,53(0.28)	-0,60(0.16)	27	1,10(0.16)	-0,09(0.11)	0,47(0.15)
14	1,42(0.20)	-1,32(0.18)	-0,51(0.11)				

EP: erro padrão

Observando as curvas características dos itens (Figura 7.2) e os valores estimados dos parâmetros (Tabela 7.1), podemos perceber alguns problemas para a construção de um teste adaptativo computadorizado neste conjunto de dados. Nota-se que muitas curvas possuem baixa discriminação (estimativas de a_i) e que as curvas das categorias intermediária dos itens (linhas azuis) mostram estas não foram muito selecionadas. Porém, como há um critério de aprovação baseado nessas categorias na avaliação atual, ela foi mantida na análise.

Apesar desses itens possuírem algumas limitações no contexto de aplicar um TAC seguindo o MRG, a riqueza da estrutura do exame (presença de atributos qualitativo e quantitativo, vários estímulos e restrições) e o fato de ser um teste politômico na área educacional não são características muito comuns, fizeram com que ele fosse de grande ajuda para a construção do Same-CAT. A estrutura do EPI foi descrita na seção 4.2 pelas Tabelas 5.1, 5.2 e 5.3.

A Figura 7.3 traz o boxplot com os valores estimados dos traços latentes divididos em 2 grupos, segundo o critério atual baseado na avaliação MPA (reprovados e aprovados). Inicialmente, foram consideradas todas as respostas aos 27 itens.

Conforme o esperado, nota-se que os indivíduos reprovados obtiveram, em geral, estimativas menores do traço latente do que os indivíduos aprovados.

O primeiro e o terceiro quantil dos reprovados foram respectivamente -0,98 e -0,1258, para os aprovados os valores foram 0,1848 e 1,3100. Observe que os valores dentro do primeiro e terceiro quantis de um grupo não se sobrepõem aos valores dos quantis do outro grupo. O valor de corte do traço latente da TRI para classificar um estudante como reprovado ou aprovado poderia ser sugerido entre -0,1258 e 0,1848.

Temos que o valor médio do erro padrão das estimativas de θ foi de 0,1291, com mínimo

Figura 7.3: Boxplot da proficiência em Inglês estimada sob o modelo de Samejima de 518 estudantes reprovados (1) ou aprovados (2) no EPI do ICMC.

de 0,0358 e máximo de 0,4010, o que indica que as estimativas dos traços latente foram razoáveis para um banco de 27 itens.

Também foram estimados os traços latentes utilizando a abordagem do *shadow test* com as seguintes restrições:

- número de itens no teste igual a 13,
- número de estímulos no teste maior ou igual a 5,
- número mínimo de itens igual ao mínimo entre 2 e o número de itens disponíveis no banco para os estímulos 1, 2, 8 e 9; número mínimo de itens igual ao mínimo entre 1 e o número de itens disponíveis no banco para os estímulos 3 ao 7,
- no mínimo 1 itens para as classes 3, 4 e 7 do atributo categorizado presentes no teste; e no mínimo 2 itens para as classes 1, 2, 5 e 6.

O boxplot apresentado na Figura 7.4 traz as estimativas de θ sob a abordagem *Shadow test*.

As estimativas de θ dos indivíduos reprovados obtiveram, em geral, estimativas menores que os indivíduos aprovados como na análise do banco completo primeiro caso.

Figura 7.4: Boxplot da proficiência em Inglês estimada sob o modelo de Samejima de 518 estudantes reprovados (1) ou aprovados (2) no EPI do ICMC utilizando a abordagem *Shadow test*.

Temos que o valor médio do erro padrão das estimativas de θ foi de 0,1791, sendo mínimo de 0,0452 e máximo de 0,4127. Como era esperado, o erro padrão foi um pouco maior do que o teste completo, mas a diferença não foi muito grande e esse teste possui 13 itens, que é menos da metade do caso anterior.

Os resultados da aplicação da TRI aos dados do EPI fornecem muito mais informação do que a aplicação da MPA, que é bastante subjetiva. Pela TRI, a qualidade e a característica dos itens da prova podem ser averiguadas através das estimativas dos parâmetros de discriminação e localização. Por outro lado, um estudo mais detalhado sobre o ponto de corte nas estimativas do traço latente fornecidas pela TRI deve ser feito com o intuito de classificar os alunos em aprovado ou reprovado no EPI.

7.2 Introdução ao Inventário de Depressão de Beck

Pelas estimativas dos parâmetros dos itens do BDI (apresentadas na Tabela 7.2) aplicado a 591 indivíduos (citado na seção anterior) e pelas curvas características dos itens, representadas na Figura 7.5, podemos perceber que o grau de inclinação das curvas características dos itens indicam que os parâmetros de discriminação são altos, a menos do item 19. As probabilidades das categorias serem selecionadas são, em geral, significativas, justificando o número de categorias no modelo. Apesar de possuir apenas 21 itens é es-

perado que os resultados gerados por um TAC utilizando esses itens sejam melhores do que os resultados utilizando o EPI, porém as características do teste do BDI não permite a construção de restrições para a abordagem *shadow teste* tão complexo quanto as do EPI, por exemplo, os itens do BDI não possuem estímulos.

Figura 7.5: Matriz das curvas características dos itens do BDI.

Tabela 7.2: Valores estimados dos parâmetros dos itens do BDI

item	$a(\text{EP})$	$b_1(\text{EP})$	$b_2(\text{EP})$	$b_3(\text{EP})$	item	$a(\text{EP})$	$b_1(\text{EP})$	$b_2(\text{EP})$	$b_3(\text{EP})$
1	2,58(0.25)	-0,32(0.06)	1,27(0.13)	1,81(0.19)	12	2,75(0.25)	-0,07(0.06)	1,13(0.12)	1,50(0.16)
2	2,73(0.29)	0,17(0.06)	1,17(0.12)	1,49(0.15)	13	2,74(0.39)	-0,09(0.06)	1,06(0.11)	1,69(0.18)
3	3,00(0.32)	0,16(0.05)	1,07(0.11)	1,65(0.17)	14	1,83(0.20)	-0,42(0.08)	1,11(0.15)	1,92(0.24)
4	3,56(0.30)	-0,22(0.05)	0,89(0.09)	1,05(0.10)	15	1,80(0.19)	-0,60(0.09)	1,12(0.15)	1,70(0.21)
5	2,63(0.30)	0,36(0.07)	1,24(0.14)	1,50(0.17)	16	1,60(0.19)	-0,69(0.10)	1,27(0.17)	1,87(0.24)
6	2,11(0.30)	0,42(0.08)	0,73(0.11)	0,78(0.11)	17	2,14(0.24)	-0,75(0.08)	1,03(0.12)	1,49(0.17)
7	2,30(0.30)	0,31(0.07)	1,17(0.13)	2,09(0.26)	18	1,76(0.18)	-0,29(0.08)	1,44(0.19)	2,39(0.33)
8	2,39(0.25)	-0,02(0.06)	1,26(0.14)	1,76(0.20)	19	0,62(0.16)	0,27(0.23)	1,87(0.55)	2,84(0.79)
9	2,56(0.47)	0,98(0.10)	1,88(0.25)	2,22(0.29)	20	1,71(0.20)	-0,70(0.09)	1,21(0.16)	1,67(0.21)
10	2,33(0.35)	-0,27(0.07)	1,33(0.15)	1,83(0.21)	21	1,24(0.17)	-0,77(0.12)	0,72(0.15)	0,85(0.16)
11	2,40(0.21)	-0,56(0.07)	1,28(0.14)	1,81(0.21)					

EP: erro padrão

Foi criado um boxplot com os valores estimados dos traços latentes, com base nos 21 itens respondidos, dividido em três grupos. Para determinar os grupos foi utilizado o sistema de

classificação recomendado por Kendall et al. (1987). Este boxplot é apresentado na Figura 7.6. Nota-se que os indivíduos sem depressão (grupo 1) obtiveram, em geral, estimativas menores que os indivíduos com disforia (grupo 2), e por sua vez, os indivíduos com disforia obtiveram, genericamente, estimativas menores que os indivíduos com depressão (grupo 3).

Figura 7.6: Boxplot dos traços latentes estimados sob o modelo de Samejima de 591 estudantes que responderam todos os 21 itens do BDI. Grupo 1 = sem depressão , grupo 2 = com disforia e grupo 3 = com depressão

O primeiro e o terceiro quantis do grupo 1 foram, respectivamente, -1,29 e -0,23, para o grupo 2 os valores foram 0,13 e 0,21 e para o grupo 3 os valores foram 0,51 e 0,93. Observe que os valores dentro do primeiro e terceiro quantis de um grupo não se sobrepõem aos valores dos quantis dos outros grupos, indicando uma separação relativamente boa das 3 distribuições.

Temos também que o valor do erro padrão foi de 0,0819, com mínimo de 0,027 e o máximo de 0,28, o que indica que as estimativas dos traços latente foram razoáveis para um banco de 21 itens.

Em seguida foram estimados os traços latentes utilizando a abordagem do *shadow test* a restrição de apenas 10 itens do BDI serem apresentados aos indivíduos. A Figura 7.7 traz o boxplot dessas estimativas subdivididas em 3 grupos, da mesma forma que na Figura 7.6.

Nota-se (como no caso anterior) que os indivíduos sem depressão (grupo 1) obtiveram,

Figura 7.7: Boxplot dos traços latentes estimados sob o modelo de Samejima de 518 estudantes que responderam 10 itens do BDI. Grupo 1 = sem depressão, grupo 2 = com disforia e grupo 3 = com depressão

em geral, estimativas menores que os indivíduos com disforia (grupo 2), que por sua vez, obtiveram, em média, estimativas menores que os indivíduos com depressão (grupo 3).

O valor dos erros padrão foi de 0,1054, com mínimo de 0,0374 e o máximo de 0,2963. Como era esperado, o erro padrão foi um pouco maior do que no teste completo, mas considerando que foi obtida com apenas 10 itens respondidos, que é menos da metade do caso anterior, pode ser compensado na prática.

Capítulo 8

Programa

O programa desenvolvido neste trabalho foi implementado em Java 1.7.0, utilizando o *software NetBeans IDE 7.0.1* (Tulach; 2008), um software de desenvolvimento integrado (IDE) gratuito e de código aberto para desenvolvimento de softwares, e recebeu o nome de Same-CAT. O Same-CAT é um programa que permite a criação de testes adaptativos computadorizados calibrados seguindo o modelo de resposta gradual de Samejima (1969) da TRI (Bock; 1997) com qualquer quantidade razoável de categorias de respostas. A abordagem “Shadow Test” foi implementada, o que permite também a imposição de certos tipos de restrições que o teste deverá obedecer (por exemplo: número de itens, número de estímulos, número de itens por estímulo, etc).

Na janela inicial do programa (Figura 8.1), há uma barra com menus “Arquivos” e “Ajuda”. No menu “Ajuda”, poderão ser encontradas informações sobre o programa assim como explicações sobre cada uma das opções do menu “Arquivo”: Banco de itens (-Criar estrutura, -Adicionar/Remover Itens, -Adicionar/Remover Estímulos), Criar TAC, Aplicar TAC e Respostas.

Figura 8.1: Janela inicial do programa.

Selecionando no menu Arquivo/Banco de Itens/Criar estrutura (Figura 8.2), o usuário poderá criar a estrutura que seu Banco de itens segue. Todos os campos deverão estar preenchidos para efetivar a criação da estrutura conforme a descrição abaixo:

- Nome do Banco de itens: insira o nome do banco.
- Número de atributos categorizados (itens): insira um número natural (0,1,2...) para a quantidade de atributos categorizados dos itens. Um atributo categorizado particiona o banco de itens em subconjuntos. Por exemplo: uma prova possui itens com conteúdos de Álgebra, Trigonometria e Geometria. Conteúdo seria o nome do atributo categorizado e o número de classes seria 3 (com nome de Álgebra, Trigonometria e Geometria).
- Número de atributos quantitativos (itens): insira um número natural (0,1,2...) de atributos quantitativos dos itens no teste. Atributos quantitativos são atributos numéricos inteiros. Por exemplo: Número de palavras no enunciado do item.
- Número de estímulos: insira um número natural (0,1,2...) de estímulos no teste. Os estímulos podem ser textos que vários itens possuem em comum. Por exemplo, um banco com 5 itens, sendo que 4 itens se baseiam no mesmo estímulo e o outro item é baseado em um segundo estímulo possui 2 estímulos. Caso este último item não tenha estímulo associado a ele (e tenha apenas o seu enunciado), o número de estímulos no banco seria igual a 1. Um outro exemplo é um banco sem estímulos. Neste caso, defina o campo “número de estímulos” como zero.

- Número de atributos categorizados (estímulos): insira um número natural (0,1,2...) para a quantidade de atributos categorizados dos estímulos. Um atributo categorizado particiona os estímulos em subconjuntos. Por exemplo: uma prova possui estímulos com conteúdos de Álgebra, Trigonometria e Geometria. Conteúdo seria o nome do atributo categorizado e o número de classes seria 3. Os itens e estímulos podem possuir atributos categorizados semelhantes (como no caso do exemplo) e/ou diferentes.
- Número de atributos quantitativos (estímulos): insira um número natural (0,1,2...) de atributos quantitativos dos estímulos no teste. Atributos quantitativos são atributos numéricos inteiros. Por exemplo: Número de palavras no texto dos estímulos.
- Número de restrições lógicas: insira um número natural (0,1,2...) de restrições lógicas no banco. Estas restrições servem para que certos itens não apareçam juntos num mesmo teste ou para impor que alguns itens devem aparecer no teste. Por exemplo: um banco de itens que possui os itens 1,2,3,4 e 5 pode ter as seguintes duas restrições: 1) restrição de que o item 1 e o 5 não apareçam juntos no mesmo teste e 2) restrição de que aparecerá no mínimo dois dos itens 2, 3 e 4 num teste.
- Número de categorias de resposta: insira um número natural (2,3,4...) de categorias de resposta dos itens. Por exemplo, se o teste tiver as opções de respostas A,B, C ou D então o número de categorias de respostas dos itens é igual a 4. O número de categorias de respostas será igual para todos os itens.

Para terminar o cadastro, o usuário deve clicar no botão “Confirmar”.

Todos os nomes cadastrados podem possuir até 100 caracteres.

Nome do banco de itens:

Número de atributos categorizados (itens):

Nome das Classes (itens):

Número de atributos quantitativos (itens):

Número de restrições lógicas:

Número de categorias de resposta:

Número de estímulos:

Número de atributos categorizados (estímulos):

Nome das Classes (estímulos):

Número de atributos quantitativos (estímulos):

Criar

Figura 8.2: Janela "Criar estrutura do banco".

A janela Adicionar/Remover Itens, apresentada na Figura 8.3, é a opção que permite que o usuário cadastre ou remova itens em um banco. Primeiramente é necessário escolher o nome do banco de itens em "selecione o banco" e clicar no botão "ok".

O botão "adicionar" possibilita a inclusão de um novo item no banco. Após clicar em adicionar aparecerá campos para o cadastramento do item. Esses campos são: Número, Enunciado, Estímulo, Atributos categorizados, Atributos quantitativos, Alternativa de resposta, Escore da categoria e valores dos parâmetros "a" e "b's". Alguns desses campos poderão não aparecer, dependendo da estrutura do banco de itens previamente determinada. Após preencher todos os campos, clique em confirmar para efetuar o cadastro do item.

Para remover um item, clique na opção remover e informe no campo ao lado o número do item.

A seguir serão dadas mais informações sobre cada campo que poderá aparecer ao clicar em "adicionar" um item.

- Número: digitar um número natural maior que 0 (1,2,3...); itens em um banco não podem possuir números iguais.

- Enunciado: digitar texto com quaisquer caracteres (máximo de 10000 caracteres).
- Estímulo: selecionar na lista o número/nome do estímulo do item. Selecionando a opção “0-” sem estímulo” indicará que o item não possui estímulos.
- Atributo(s) categorizado(s) do item: selecionar na lista o número/nome das classes do(s) atributo(s) categorizado(s) do item.
- Atributo(s) quantitativo(s) do item: digitar um número natural indicando o valor (ou peso) do(s) atributo(s) quantitativo(s) do item.
- Alternativas de respostas: digitar o texto das alternativas.
- Valores das categoria: selecionar na lista o número dos valores das categorias. Escores mais baixos estão associados a valores mais baixos do traço latente em questão.
- a: digitar um número real positivo do parâmetro de discriminação.
- b(s): digitar um número real do(s) parâmetro(s) de dificuldade(s).

A janela da opção de Adicionar/Remover Estímulos é analoga à de itens.

A seguir serão dadas mais informações sobre cada campo que poderá aparecer ao clicar em “adicionar” um estímulo.

- Número: selecione o número do estímulo; dois estímulos não podem possuir números iguais.
- Texto: digitar texto com quaisquer caracteres (máximo de 10000 caracteres).
- Atributo(s) categorizado(s) do estímulo: selecionar na lista o número/nome das classes do(s) atributo(s) categorizado(s) do estímulo.
- Atributo(s) quantitativo(s) do estímulo: digitar um número natural indicando o valor (ou peso) do(s) atributo(s) quantitativo(s) do estímulo.

A janela Criar TAC, apresentada na Figura 8.4, é a opção que permite que o usuário cadastre um TAC com as restrições desejadas do teste. Primeiramente é necessário escolher o nome do banco de itens que o TAC irá se basear em “selecione o banco” e clicar no botão “ok”.

O botão “adicionar” possibilita a inclusão de um novo TAC associado ao banco selecionado. Após clicar em adicionar aparecerá campos para o cadastramento do TAC. Esses campos são: Nome, Número de itens, Número de estímulos, Número de itens por estímulo, Número de itens por classe do(s) atributo(s) categorizado(s), Número de estímulos por

Adicionar/Remover Itens

Selecione o banco:

testead

Selecione a tabela:

itens

OK

Numero	Enunciado	Acategor...	Acategor...	PesoAq1	PesoAq2	Alternati...	Categoria1	Alternati...	Categoria2	Alternati...	Categoria3	a	b1	b2
1	dsadsad	1	1	1	1	dsad	0	sadsa	1	dasdsadas	2	1	1	2
2	dsafsvgvv	1	1	2	3	cxavdv	1	vdscvc	0	cccc	2	3	-1	0
3	gvevaavd...	1	1	1	1	2dad	0	dsa21eds	1	ad121	2	2.5	0.5	0.6
4	bfbvfdsd...	2	3	2	2	vdsd	2	cdvczcz	1	zzzzzz	0	2	-2	-1
5	fdsadasdas	1	1	2	2	dsa	1	dsadsa	2	dsadsa	0	2	1	2

Adicionar

Remover

Número:

Enunciado:

Atributo categorizado "sadsadqd" do item:

1-) dsad

Atributo categorizado "dqwwdq" do item:

1-) qsqw

Atributo quantitativo "dqsd" do item:

Atributo quantitativo "sadsadasd" do item:

Alternativa de resposta 1:

Alternativa de resposta 2:

Alternativa de resposta 3:

Score da categoria 1:

0

Score da categoria 2:

0

Score da categoria 3:

0

Score do parametro a:

0

Score do parametro b1:

0

Score do parametro b2:

0

Confirmar

Figura 8.3: Janela "Adicionar/Remover Item".

classe do(s) atributo(s) categorizado(s), Restrição(ões) do(s) atributo(s) quantitativo(s) do itens, Restrição(ões) do(s) atributo(s) quantitativo(s) dos estímulos, Restrição(ões) lógicas. Em alguns campos, um número máximo, exato ou mínimo pode ser determinado selecionando-se a opção " $\leq$ ", " $=$ " ou " $\geq$ " respectivamente. O número de itens sempre será associado ao " $=$ ". Alguns desses campos poderão não aparecer dependendo da estrutura do banco de itens previamente determinada.

A seguir serão dadas mais informações sobre cada campo que poderá aparecer ao clicar

55

em “adicionar” um TAC.

- Nome do TAC: digitar o nome do TAC; um mesmo banco não pode possuir TACs com nomes iguais.
- Número de itens: digitar um número natural maior que 0 de itens no teste.
- Número de estímulos: digitar um número natural de estímulo no teste.
- Itens por estímulo: digitar um número natural de itens por estímulo no teste.
- Itens por classe do(s) atributo(s) categorizado(s): digitar um número natural de itens para uma classe do atributo(s) categorizado(s) no teste.
- Estímulos por classe do(s) atributo(s) categorizado(s): digitar um número natural de estímulos para uma classe do atributo(s) categorizado(s) no teste.
- Restrição(ões) do(s) atributo(s) quantitativo(s): digitar um número natural da soma dos pesos do(s) atributo(s) quantitativo(s).
- Restrição(ões) lógica(s): Digitar números dos itens desejados separados por “-” e por último o número da restrição. Por exemplo: no máximo 2 dos itens 1, 3, 5 e 8 podem aparecer no teste. A Restrição lógica deve ser dada por 1-3-5-8-2 com o sinal de desigualdade “<=” selecionado. Outro exemplo: apenas 1 dos itens 1 e 2 deverá aparecer no teste. A Restrição lógica deve ser dada por 1-2-1 com o sinal “=” selecionado.

Após preencher todos os campos, clique em confirmar para efetuar o cadastro do TAC.

Para remover um TAC, clicar na opção remover e informar no campo ao lado o nome do TAC.

Figura 8.4: Janela "Criar TAC".

A janela Aplicar TAC ilustrado na Figura 8.5 é a opção onde o usuário pode selecionar um dos bancos de itens cadastrados previamente e um TAC desse banco para aplica-lo a um indivíduo.

Primeiramente, selecione o nome do banco, digite o código do examinado (quaisquer caracteres), selecione o TAC desejado e clique em "confirmar".

Uma caixa de texto irá aparecer com o estímulo do item a ser respondido (caso exista um estímulo), uma caixa com o enunciado do item a ser respondido com suas respectivas alternativas de resposta. Ao lado, selecione a opção de resposta que deseja escolher e clique no botão "Responder". Nos itens seguintes siga o mesmo procedimento até o término do teste, quando uma mensagem aparecerá constando o código do examinado, o nome do banco de itens, o nome do TAC, o valor estimado do traço latente e seu erro padrão e as frequências das categorias selecionadas em valor absoluto e em porcentagem (Figura 8.6).

Figura 8.5: Janela inicial de "Aplicar TAC".

Figura 8.6: Janela final de "Aplicar TAC".

Na opção Respostas, um usuário que conheça a senha exigida poderá ver os resultados das aplicações de todos os TACs de todos os bancos de itens.

Capítulo 9

Conclusão e Estudos Futuros

De acordo com os resultados das simulações e aplicações dos TACs, podemos concluir que quanto melhor um banco de itens (ou seja, parâmetros a 's de valores altos e parâmetros b 's de valores abrangentes), mais precisas são as estimativas dos traços latentes. Para TACs com poucos itens, dependendo da diferença entre o valor inicial (no TAC) e valor real do traço latente há problemas na estimação. Em especial neste trabalho, observou-se este problema para valores extremos do traço latente.

Mesmo para um banco de itens grande, há uma alta taxa de exposição de alguns deles. Na Psiquiatria esse fato não representa um problema, mas na área da Educação levanta a importância de se incorporar algum método de controle de exposição de item (Barrada et al.; 2009; Lee et al.; 2008; Chang and Ying; 1999).

Uma aplicação real de um TAC utilizando um banco de itens grande poderá trazer mais informação futuramente, por exemplo: nas aplicações, foram se utilizadas as repostas dos itens que os indivíduos forneceram ao teste completo como as respostas para o TAC, mas as respostas dadas por um indivíduo para um teste com um número menor de questões podem ser diferentes para um teste longo, por causa da fadiga que este possa ocasionar.

Um estudo mais detalhado para sugerir um *cut-off* no traço latente para classificar um indivíduo como aprovado ou reprovado no EPI e, analogamente, para diagnosticar pacientes utilizando o BDI pode ser proposto como sequência dos resultados deste trabalho. Em estudos futuros, o programa Same-CAT pode ser melhorado com a implementação de um algoritmo para a estimação dos parâmetros dos itens, permitir a inserção de figuras e formulas matemáticas nos estímulos, nos enunciados dos itens e nas alternativas de respostas.

Referências Bibliográficas

- Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika* **43**: 561–573.
URL: <http://dx.doi.org/10.1007/BF02293814>
- Azevedo, C. L. N. (2003). *Métodos de estimação na teoria de resposta ao item*, Master's thesis, Universidade de São Paulo, São Carlos, SP.
- Barker, F. B. (1992). *Item Response Theory - Parameter Estimation Techniques*, New York: Marcel Dekker, inc.
- Barrada, J. R., Veldkamp, B. P. and Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing, *Applied Psychological Measurement* **33**(1): 58–73.
URL: <http://apm.sagepub.com/content/33/1/58.abstract>
- Barton, M. A. and Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model.
- Beck, A., Ward, C., Mendelson, M., Mock, J. and Erbaugh, J. (1961). An inventory for measuring depression., *Archives of general psychiatry* **4**.
- Belov, D. I. and Armstrong, R. D. (2009). Direct and inverse problems of item pool design for computerized adaptive testing, *Educational and Psychological Measurement* **69**(4): 533–547.
URL: <http://epm.sagepub.com/content/69/4/533.abstract>
- Belov, D. I., Armstrong, R. D. and Weissman, A. (2008). A monte carlo approach for adaptive testing with content constraints, *Applied Psychological Measurement* **32**(6): 431–446.
URL: <http://apm.sagepub.com/content/32/6/431.abstract>
- Binet, A. and Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel anormaux [new methods for the diagnosis of levels of intellectual abnormality], *Année Psychologique* **11**: 191–244.
- Bock, R. D. (1997). A brief history of item theory response, *Educational Measurement: Issues and Practice* **16**(4): 21–33.
URL: <http://dx.doi.org/10.1111/j.1745-3992.1997.tb00605.x>

- Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm, *Psychometrika* **46**: 443–459.
URL: <http://dx.doi.org/10.1007/BF02293801>
- Bock, R. D., Gibbons, R. and Muraki, E. (1988). Full-information item factor analysis, *Applied Psychological Measurement* **12**(3): 261–280.
URL: <http://apm.sagepub.com/content/12/3/261.abstract>
- Bock, R. and Zimowski, M. F. (1997). *Multiple group IRT In W.J. van der Linden and R. Hambleton (Eds)*, Springer Science+Business Media, LLC, 233 Spring Street, New York, USA.
- Bruno, J. E. (1986). Assessing the knowledge base of students: An information theoretic approach to testing., *Journal of Measurement and Evaluation in Counseling and Development* **19**.
- Burt, C. (1949). *Mental and scholastic tests*, London: P. S. King and Son.
- Carvalho, R. M. and Coelho, E. G. (2012). Sistema de testagem adaptativa computadorizada (cat) da universidade de brsilia com pr-testagem automatizada.
URL: <http://hdl.handle.net/123456789/351>
- Chang, H. H. and Ying, Z. (1999). a-stratified multistage computerized adaptive testing, *Applied Psychological Measurement* **23**(3): 211–222.
URL: <http://apm.sagepub.com/content/23/3/211.abstract>
- Chang, H.-H. and Ying, Z. (2008). To weight or not to weight? balancing influence of initial items in adaptive testing, *Psychometrika* **73**: 441–450.
URL: <http://dx.doi.org/10.1007/s11336-007-9047-7>
- Cheng, Y., Chang, H. H. and Douglas, J. G. F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints: Balancing measure efficiency and exposure control, *Educational and Psychological Measurement* **69**: 35–49.
- Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm, *The American Statistician* **49**(4): 327–335.
URL: <http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1995.10476177>
- Cronbach, L. J. and Gleser, G. C. (1965). *Psychological tests and personnel decisions*, 2nd ed. edn, Oxford, England: U. Illinois Press.
- David, J. W. (2009). C. a. t. central a global resource for computerized adaptive testing research and applications.
URL: <http://www.psych.umn.edu/psylabs/catcentral/>

- Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test, *Applied Psychological Measurement* **23**(3): 249–261.
URL: <http://apm.sagepub.com/content/23/3/249.abstract>
- Eggen, T. and Straetmans, G. (2000). Computerized adaptive testing for classifying examinees into three categories, *Educational and Psychological Measurement* **60**(5): 713–734.
URL: <http://epm.sagepub.com/content/60/5/713.abstract>
- Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. B. and Klapp, B. F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (d-cat) in clinical application, *International Journal of Methods in Psychiatric Research* **18**(1): 23–36.
URL: <http://dx.doi.org/10.1002/mpr.274>
- Forbey, J. D. and Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the mmpi-2 computerized adaptive version (mmpi-2- ca)., *Psychological Assessment* **19**: 14–24.
- Gardner, W., Kelleher, K. J. and Pajer, K. A. (2002). Multidimensional adaptive testing for mental health problems in primary care., *Medical Care* **40**: 812–823.
- Gardner, W., shear, K., Kelleher, K. J., Pajer, K., Mammen, O., Buysse, D. and Frank, E. (2004). Computerized adaptive measurement of depression: A simulation study., *BMC Psychiatry* **4**:13.
- Gonçalves, J. P. (2004). A integração de testes adaptativos informatizados e ambientes computacionais de tarefas para o aprendizado do inglês instrumental[dissertação]. São Carlos: Universidade de São Paulo, Instituto de Ciência Matemáticas e de Computação; [acesso 2012-11-19]. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-03052004-160334/>.
- Gotlib, I. (1984). Depression and general psychopathology in university students., *Journal of Abnormal Psychology* **93**.
- GRE (2011). Gre. educational testing service, web. 29 nov. 2012. Educational Testing Services.
URL: <http://www.ets.org/gre>
- Haley, S. M., Pengsheng, N., Hambleton, R. K., Slavin, M. D. and Jette, A. M. (2006). Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank., *Journal of Clinical Epidemiology* **59**: 1174–1182.
- Hart, D. L., Cook, K. F., Mioduski, J. E., Teal, C. R. and Crane, P. K. (2006). Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function, *Journal of Clinical Epidemiology* **59**(3): 290 – 298.
URL: <http://www.sciencedirect.com/science/article/pii/S0895435605002982>

- Hart, D. L., Wang, Y.-C., Stratford, P. W. and Mioduski, J. E. (2006). Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function, *Journal of Clinical Epidemiology* **59**: 947–956.
- Hart, D. L., Wang, Y.-C., Stratford, P. W. and Mioduski, J. E. (2008). Computerized adaptive test for patients with foot or ankle impairments produced valid and responsive measures of function, *Quality of Life Research* **17**: 1081–1091.
URL: <http://dx.doi.org/10.1007/s11136-008-9381-y>
- Hol, A. M., Vorst, H. C. M. and Mellenbergh, G. J. (2007). Computerized adaptive testing for polytomous motivation items: Administration mode effects and a comparison with short forms, *Applied Psychological Measurement* **31**(5): 412–429.
URL: <http://apm.sagepub.com/content/31/5/412.abstract>
- Imai, S. (2008). Development of japanese-computerized adaptive test, *Computer and Communication The 23rd International Technical Conference on Circuits/Systems*: 821–824.
- Kendall, P., Hollon, S., Beck, A., Hammen, C. and Ingram, R. (1987). Issues and recommendations regarding use of the beck depression inventory, *Cognitive Therapy and Research* **11**: 289–299.
URL: <http://dx.doi.org/10.1007/BF01186280>
- Larkin, K. C. and Weiss, D. J. (1975). *An empirical comparison of two-stage and pyramidal adaptive ability testing*, Minneapolis: Psychometrics Methods Program, Department of Psychology, University of Minnesota. (Research Report 75-1).
- Lee, Y. H., Ip, E. H. and Fuh, C. D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing, *Educational and Psychological Measurement* **68**(2): 215–232.
URL: <http://epm.sagepub.com/content/68/2/215.abstract>
- Li, Y. H. and Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics, *Applied Psychological Measurement* **29**(1): 3–25.
URL: <http://apm.sagepub.com/content/29/1/3.abstract>
- Linden, W. J. and Glas, C. A. W. (1997). *Elements of Adaptive Testing.*, Springer Science+Business Media, LLC, 233 Spring Street, New York, USA.
- Linden, W. J. and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.*, Springer Science+Business Media, LLC, 233 Spring Street, New York, USA.
- Lord, F. M. (1970). Some test theory for tailored testing. in w. h. holtzman (ed.), *Computer assisted instruction, testing, and guidance*. pp. 139–183.

- Lord, F. M. (1971a). Robbins-monro procedures for tailored testings, *Educational and Psychological Measurement* **31**(1).
- Lord, F. M. (1971b). The self-scoring flexilevel test1, *Journal of Educational Measurement* **8**(3): 147–151.
URL: <http://dx.doi.org/10.1111/j.1745-3984.1971.tb00918.x>
- Lord, F. M. (1977). Abroad-range tailored test of verbal ability, *Applied Psychological Measurement* **1**: 95–100.
- Louks, J., Hayne, C. and Smith, J. (1989). Replicated factor structure of the beck depression inventory., *Journal of Nervous and Mental Disease* **177**.
- McLeod, L., Lewis, C. and Thissen, D. (2003). A bayesian method for the detection of item preknowledge in computerized adaptive testing, *Applied Psychological Measurement* **27**(2): 121–137.
URL: <http://apm.sagepub.com/content/27/2/121.abstract>
- Mislevy, R. (1986). Bayes modal estimation in item response models, *Psychometrika* **51**: 177–195.
URL: <http://dx.doi.org/10.1007/BF02293979>
- Mislevy, R. J. and Stocking, M. L. (1989). A consumer’s guide to logist and bilog, *Applied Psychological Measurement* **13**(1): 57–75.
URL: <http://apm.sagepub.com/content/13/1/57.abstract>
- Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm, *Applied Psychological Measurement* **16**(2): 159–176.
URL: <http://apm.sagepub.com/content/16/2/159.abstract>
- NCSBN (2005). Fast facts about alternate item formats and the nclex examination, web. 29 nov. 2012. National Council Licensure Examination.
URL: <http://www.ncsbn.org>
- Oliver, J. and Simmons, M. (1984). Depression as measured by the dsm-iii and beck depression inventory in an unselected adult population., *Journal of Consulting and Clinical Psychology* **52**.
- Owen, R. J. (1969). *A Bayesian approach to tailored testing*, Princeton, NJ: Educational Testing Service. (Research Report 69- 92).
- Owen, R. J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing, *Journal of the American Statistical Association* **70**(350): 351–356.
URL: <http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1975.10479871>

- Passos, V. L., Berger, M. P. F. and S., T. F. E. (2008). The d-optimality item selection criterion in the early stage of cat: A study with the graded response model., *Journal of educational and behavial statistics* **33**.
- Penfield, R. D. (2007). Estimating the standard error of the maximum likelihood ability estimator in adaptive testing using the posterior-weighted test information function, *Educational and Psychological Measurement* **67**(6): 958–975.
URL: <http://epm.sagepub.com/content/67/6/958.abstract>
- Petersen, M. A., Groenvold, M., Aaronson, N., Fayers, P., Sprangers, M. and Bjorner, J. B. (2006). Multidimensional computerized adaptive testing of the eortc qlq-c30: Basic developments and evaluations, *Quality of Life Research* **15**: 315–329.
URL: <http://dx.doi.org/10.1007/s11136-005-3214-z>
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: <http://www.R-project.org>
- Reckase, M. D. (2009). *Multidimensional Item Response Theory*, Springer Science+Business Media, LLC, 233 Sptring Street, New Yourk, USA.
- Rigdon, S. E. and Tsutakawa, R. K. (1983). Parameter estimation in latent trait models, *Psychometrika* **48**: 567–574.
URL: <http://dx.doi.org/10.1007/BF02293880>
- Rulison, K. L. and Loken, E. (2009). I’ve fallen and i can’t get up: Can high-ability students recover from early mistakes in cat?, *Applied Psychological Measurement* **33**(2): 83–101.
URL: <http://apm.sagepub.com/content/33/2/83.abstract>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores., *Psychometrika Monograph Supplement* **34**.
- Samejima, F. (1997). *Graded response model* In W.J. van der Linden and R. Hambleton (Eds), Springer Science+Business Media, LLC, 233 Sptring Street, New Yourk, USA.
- Sands, W. A. and Waters, B. K. (1996). A global information approach to computerized adaptive testing., *Applied Psychological Measurement* **20**.
- Sands, W. A. and Waters, B. K. (1997). Introduction to asvab and cat, in Waters, Brian K. (Ed); McBride, James R. (Ed), (1997). *Computerized adaptive testing: From inquiry to operation.*, (pp. 3-9) .
- Tanaka-Matsumi, J. and Kameoka, V. (1986). Reliabilities and concurrent validities of popular self-report measures of depression, anxiety, and social desirability., *Journal of Consulting and Clinical Psychology* **54**.

- Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model, *Psychometrika* **47**: 175–186.
URL: <http://dx.doi.org/10.1007/BF02296273>
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests., *Psychometrika* **16**(7): 433–451.
- Tierney, L. (1994). Markov chains for exploring posterior distributions, *The Annals of Statistics* **22**(4): pp. 1701–1728.
URL: <http://www.jstor.org/stable/2242477>
- TOEFL (2011). Test of english as a foreign language, web. 29 nov. 2012. Educational Testing Services.
URL: <http://www.ets.org/toefl>
- Toit, M. D. (2003). *Irt from Ssi: Bilog-mg, Multilog, Parscale, Testfact*, Scientific Software International. Inc. Lincolnwood, IL.
- Triantafyllou, E., Georgiadou, E. and Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices, *Computers and Education* **50**(4): 1319 – 1330.
URL: <http://www.sciencedirect.com/science/article/pii/S0360131506001965>
- Tulach, J. (2008). Netbeans integrated development environment, netbeans. retrieved at may 25, 2008, from the website temoa : Open educational resources (oer).
URL: <http://www.temoa.info/node/2591>
- Veldkamp, B. P. and Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content, *Psychometrika* **67**: 575–588.
URL: <http://dx.doi.org/10.1007/BF02295132>
- Wang, W.-C. and Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing, *Applied Psychological Measurement* **28**(5): 295–316.
URL: <http://apm.sagepub.com/content/28/5/295.abstract>
- Weiss, D. J. (1973). *New horizons in testing: Latent trait test theory and computerized adaptive testing*, New York: Academic Press.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing, *Applied Psychological Measurement* **6**(4): 473–492.
URL: <http://apm.sagepub.com/content/6/4/473.abstract>
- Weiss, D. J. E. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*, New York: Academic Press.

Weiss, D. J. and Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems, *Journal of Educational Measurement* **21**(4): 361–375.

URL: <http://dx.doi.org/10.1111/j.1745-3984.1984.tb01040.x>

Wilson, D. T., Wood, R. and Gibbons, R. D. (1991). Testfact: Test scoring, item statistics, and item factor analysis (computer software), *Chicago: Scientific Software International, Inc* .