Avaliação da proficiência em inglês acadêmico através de um teste adaptativo informatizado

Vanessa Rufino da Silva

Avaliação da proficiência em inglês acadêmico através de um teste adaptativo informatizado

Vanessa Rufino da Silva

Orientadora: **Profa. Dra. Mariana Cúri**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

USP/UFSCar – São Carlos
Abril de 2015

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

Rufino da Silva, Vanessa
R586a Avaliação da proficiência em inglês acadêmico através de um teste adaptativo informatizado / Vanessa Rufino da Silva; orientadora Mariana Cúri. -- São Carlos, 2015.
33 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015.

1. Teste Adaptativo Informatizado. 2. Exame de Proficiência em Inglês. 3. Teoria de Resposta ao Item . 4. Shadow test. I. Cúri, Mariana, orient. II. Título.

VANESSA RUFINO DA SILVA

AVALIAÇÃO DA PROFICIÊNCIA EM INGLÊS ACADÊMICO ATRAVÉS DE UM TESTE ADAPTATIVO INFORMATIZADO

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

Aprovada em 09 de abril de 2015.

COMISSÃO JULGADORA:

Profa. Dra. Mariana Cúri
ICMC-USP
Presidente

Prof. Dr. Heliton Ribeiro Tavares
UFPA
Membro

Prof. Dr. Ricardo Primi
USF
Membro

Agradecimentos

Primeiramente ao meu pai, Altair Rufino da Silva pelo amor e apoio para que eu conseguisse concluir essa jornada deixando que eu nunca me sentisse sozinha.

Ao meu irmão, Felipe Rufino da Silva pelos fins de semana jogando video game e pelo seu sorriso e alegria para que pudesse relaxar e dar mais motivação para continuar.

À minha mãe, Gislene Maria de Lima Silva que mesmo não estando mais fisicamente presente tenho certeza que sempre esteve ao meu lado me dando força.

Ao meu namorado, George Lucas Moraes Pezzott, meu grande incentivador, por sempre me ajudar quando necessário, pela compreensão, pelo amor, carinho e companheirismo.

À minha orientadora, Mariana Cúri, pela sabedoria na orientação, por compartilhar seus conhecimentos, por sua compreensão e auxílio para alcançarmos mais essa vitória.

À minha amiga Nayara Pegoraro Reina por sempre ouvir minhas reclamações, pelas férias maravilhosas e por me apoiar sempre.

À CAPES pelo apoio financeiro.

Não poderia deixar de agradecer aos professores, funcionários, amigos e colegas da UFSCar e USP.

Enfim, à todos que de alguma forma me ajudaram na concretização deste sonho.

Resumo

Este trabalho descreve as etapas de transformação de um exame de proficiência em inglês acadêmico, aplicado via lápis-e-papel, com itens de múltipla escolha administrados segundo o método de Medida de Probabilidade Admissível (Shuford Jr et al, 1966), utilizado no programa de pós-graduação do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP), em um teste adaptativo informatizado (TAI-PI) baseado em um modelo da Teoria de Resposta ao Item (TRI). Apesar do programa aceitar diversos exames que atestam a proficiência em inglês para indivíduos não-nativos de abrangência e reconhecimento internacionais, como o TOEFL (Test of English as a Foreign Language), IELTS (International English Language Testing System) e CPE (Certificate of Proficiency in English), por exemplo, a sua obrigatoriedade é incoerente em universidades públicas do Brasil devido ao custo que varia de 200 a 300 dólares por exame. O software TAI-PI (Teste Adaptativo Informatizado para Proficiência em Inglês), que foi desenvolvido em Java e SQLite, será utilizado para a avaliação da proficiência em inglês dos alunos do programa a partir do segundo semestre de 2013, de forma gratuita. A metodologia estatística implementada foi definida considerando a história e objetivos do exame e adotou o modelo de resposta gradual unidimensional de Samejima (Samejima, 1969), o critério de Kullback-Leibler para seleção de itens, o método de estimação da esperança a posteriori para os traços latentes (Baker , 2001) e a abordagem Shadow test (Van der Linden & Pashley, 2010) para imposição de restrições (de conteúdo e tamanho da prova) na composição do teste de cada indivíduo. Uma descrição da estrutura do exame, dos métodos empregados, dos resultados das aplicações do TAI-PI a alunos de pós-graduação do ICMC e estudos de classificação dos alunos em aprovados e reprovados, são apresentados neste trabalho, evidenciando a boa qualidade da nova proposta adotada e aprimoramento do exame com a utilização dos métodos de TRI e TAI.

Palavras-chave: teste adaptativo computadorizado, teoria de resposta ao item, *shadow test*.

Abstract

This work describes the steps for converting a linear paper-and-pencil English proficiency test for academic purposes, composed with multiple choice items that are administered following the admissible probability measurement procedure (Shuford Jr et al, 1966), adopted by the graduate program of Institute of Mathematical Sciences and Computing of University of São Paulo (ICMC-USP), Brazil, to a computerized adaptive test (CAT) based on an item response theory model (IRT). Despite the Institute recognizes reliable international English-language exams for academic purposes and non-native speakers, as TOEFL (Test of English as a Foreign Language), IELTS (International English Language Testing System) and CPE (Cambridge English: Proficiency), for instance, it is inconsistent that public universities in Brazil require them as certification because of the cost of approximately US$ 200.00 to US$ 300.00 per exam. The software TAI-PI (computerized adaptive test for English proficiency) was implemented in Java language, used SQLite as database engine, and it shall be offered free of charge for English proficiency assessment of the graduate students from October 2013. The statistical methodology employed for TAI-PI construction was defined considering the history and the aims of the evaluation and adopted the Samejima's graded response model (Samejima, 1969), the Kullback-Leibler information criterion for item selection, the expected a posteriori Bayesian estimation for latent trait (Baker, 2001) and shadow test approach (Van der Linden & Pashley, 2010) for test constraints (content and size of the test, for example). A description of the test design, the employed statistical methods, study results of a real application of TAI-PI to graduate students are presented in this work and the validation studies of the new methodology for pass/fail classification, highlighting the good quality of the new evaluation system and examination of improvement with the use of the methods of IRT and CAT.

Keywords: computerized adaptive test, item response theory, *shadow test*.

Sumário

Lista de Abreviaturas

TOEFL	*Test of English as a Foreign Language*
IELTS	*International English Language Testing System*
CPE	*Certificate of Proficiency in English*
CCMC	Ciências da Computação e Matemática Computacional
PIPGEs	Programa Interinstitucional de Pós Graduação em Estatística
ICMC	Instituto de Ciências Matemáticas e de Computação
USP	Universidade de São Paulo
UFSCar	Universidade Federal de São Carlos
EPI	Exame de Proficiência em Ingçês
TAI	Teste Adaptativo Informatizado
TRI	Teoria de Resposta ao Item
GRE	*Graduate Record Examination*
ETS	*Educational Testing Service*
ASBAV	*Armed Services Vocational Aptitude Test Battery*
TAI-PI	Teste Adaptativo Informatizado de Proficiência em Inglês
MPA	Medida de Probabilidade Admissível
CALEAP-WEB	Computer-Aided Learning of English for Academic Purposes
MRG	Modelo de Resposta Gradual
EAP	Esperança a Posteriori
CAT	*Computerized Adaptive Test*
IRT	*Item Response Theory*
CCI	Curva Característica do Item

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O domínio da língua inglesa por alunos de pós-graduação em diversas áreas da ciência é fundamental, tanto para uma adequada absorção do conteúdo do curso, quanto para o desenvolvimento e divulgação do trabalho de pesquisa realizado. Neste contexto, muitas instituições de pós-graduação no Brasil exigem uma comprovação do domínio do inglês por seus alunos, seja no ato de ingresso no programa, seja ao longo de seu curso.

Há diversos exames de abrangência e reconhecimento internacionais normalmente aceitos, que atestam a proficiência em inglês para indivíduos não-nativos. Dentre os mais tradicionais, podemos citar o Test of English as a Foreign Language (TOEFL), o International English Language Testing System (IELTS) e o Certificate of Proficiency in English (CPE). No entanto, um dos maiores inconvenientes destes exames é o custo, que varia de 200 a 300 dólares americanos, aproximadamente. Isto torna a sua obrigatoriedade incoerente em universidades públicas de nosso país. Essa é uma das principais razões dos programas de pós-graduação em Ciências da Computação e Matemátia Computacional (CCMC) e em Estatística (PIPGEs) do ICMC-USP, sendo este último em conjunto com a UFSCar, oferecerem um Exame de Proficiência em Inglês (EPI) como uma alternativa gratuita a seus alunos.

Até o ano de 2013, o EPI era aplicado no formato lápis-e-papel e este trabalho tem como objetivo torná-lo um teste adaptativo informatizado (TAI), no qual os itens são selecionados gradativamente para o indivíduo, de acordo com sua proficiência. O valor da proficiência do indivíduo é atualizado após cada resposta, com base em um modelo da Teoria de Resposta ao Item (TRI). Assim, itens muito fáceis ou muito difíceis para um determinado indivíduo nem chegam a ser apresentados a ele, diminuindo o tamanho do teste, seu tempo de realização e tornando-o mais eficiente, objetivo e com resultado imediato (Weiss, 1982). Na Figura 1.1 são apresentados os procedimentos de seleção e administração de itens no teste e atualização das estimativas das proficiências, iterativamente, até que algum critério de parada seja satisfeito.

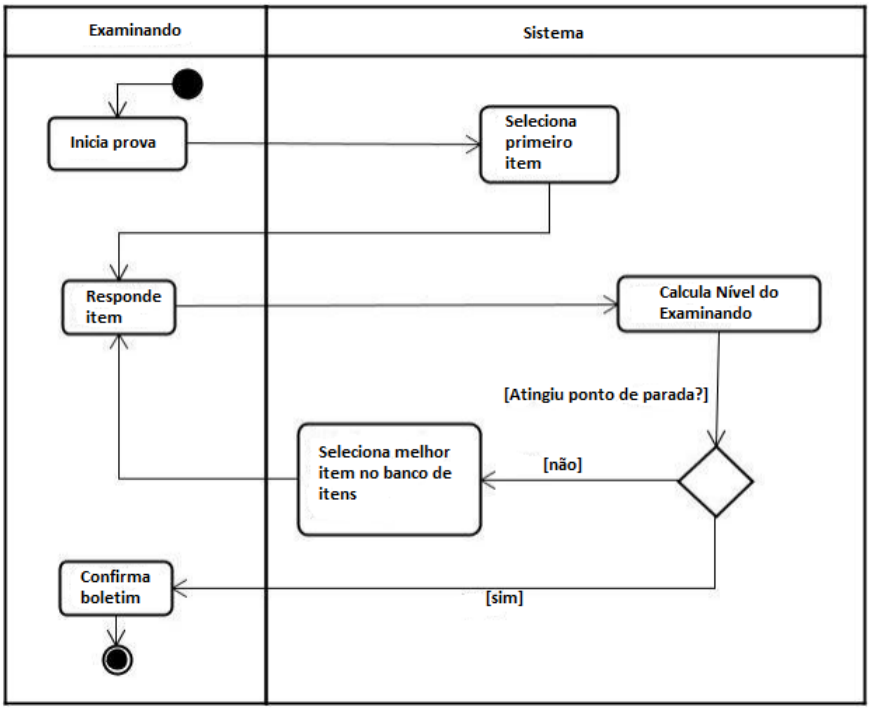

Figura 1.1: *Diagrama de um teste adaptativo informatizado*

Além das vantagens acima mencionadas, os testes adaptativos informatizados (TAI), em combinação com a TRI, possibilitam o cálculo de proficiências comparáveis entre indivíduos que responderam conjuntos diferentes de itens, e em momentos diferentes (Wainer et al, 2000). Isto facilita bastante a avaliação de construtos em larga escala, resultando no seu uso em importantes exames, tais como *Graduate Record Examination* (GRE) (Bridgeman & Cline, 2000; Eignor, 1993) , desenvolvido pelo *Educational Testing Service* (ETS) em 1996; *Test of English as a Foreign Language* (TOEFL)(Eignor et al, 1998; Kirsch et al, 1998; Wainer & Wang 2000), também desenvolvido pelo ETS; *Armed Services Vocational Aptitude Test Battery* (ASBAV)(Sands et al, 1997; Segall, 1993), desenvolvido pelo Departamento de Defesa dos Estados Unidos para selecionar potenciais recrutas para o serviço militar.

Para justificar a escolha da metodologia estatística adotada neste trabalho, apresentamos a seguir o formato com que o EPI foi aplicado no programa CCMC entre 2002 e 2013, servindo como um alicercee para a construção do Teste Adaptativo Informatizado de Proficiência em Inglês (TAI-PI).

Desde o ano 2002, o EPI consiste em uma prova composta por itens de múltipla escolha, aplicada semestralmente em formato lápis-e-papel. Para a confecção da prova, selecionam-se de 25 a 30 itens de um banco com 167 itens no total, subdivididos em 3 módulos da seguinte forma:

1. No módulo 1 são apresentados resumos de artigos de periódicos das áreas de Computação, Matemática Aplicada e Estatística e perguntas sobre seus componentes de **estrutura científica**.

2. No módulo 2 tem-se um trecho da introdução de um artigo científico, das mesmas áreas, e perguntas (em português) sobre **compreensão do texto** e relação entre ideias contidas nele.

3. No módulo 3 as perguntas estão relacionadas a **convenções da língua inglesa**, tais como conjunções, tempos verbais, *relative clauses* e artigos.

Os itens do banco foram elaborados, na época, por Andrea Jessica Borges Monzón, atual docente do Instituto Federal do Rio Grande do Sul, em conjunto com a professora Sandra Maria Aluísio, do Departamento de Computação do ICMC, sua orientadora de mestrado (Monzón, 2010). A elaboração

baseou-se na taxonomia de Bloom (Bloom, 1984), objetivando avaliar a competência na leitura e compreensão de artigos científicos das áreas de Computação, Matemática Aplicada e Estatística.

Com o intuito de minimizar a possibilidade de "chute"da resposta correta e de não pressupor que o conhecimento do aluno sobre um determinado item é apenas binário (correto ou incorreto), a classificação da resposta do aluno segue a metodologia da Medida de Probabilidade Admissível (MPA) (Klinger, 1997). A Figura 1.2 ilustra as possibilidades de resposta do aluno a um item do EPI. São três as alternativas de resposta em cada item de múltipla escolha (A, B e C), representadas nos vértices de um triângulo equilátero, sendo apenas uma delas a correta, no entanto, o aluno possui 13 possibilidades de resposta, que expressam o seu grau de certeza sobre qual das 3 afirmativas é a correta. Que podem ser selecionadas da maneira a seguir:

- Se estiver totalmente certo/seguro use as opções A, B ou C.
- Se estiver totalmente incerto/inseguro use a opção M.
- Se uma das opções A, B, C parece definitivamente errado, escolha entre as 5 opções da linha oposta.
- Indique preferência entre duas opções A, B ou C escolhendo as opções D, F, G, I, J ou L.
- Se duas opções A, B, ou C parecem iguais escolha as opções E, H ou K .

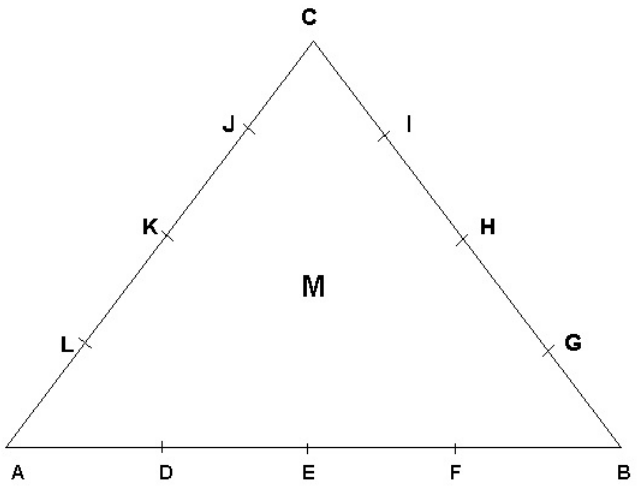

Figura 1.2: *Exemplo de um item e sua possíveis alternativas de respostas no EPI.*

Na Figura 1.3 tem-se a classificação da resposta do aluno de acordo com sua escolha em: totalmente informado, informado, parcialmente informado, mal informado, totalmente mal informado e não informado (Aluísio et al, 2003).

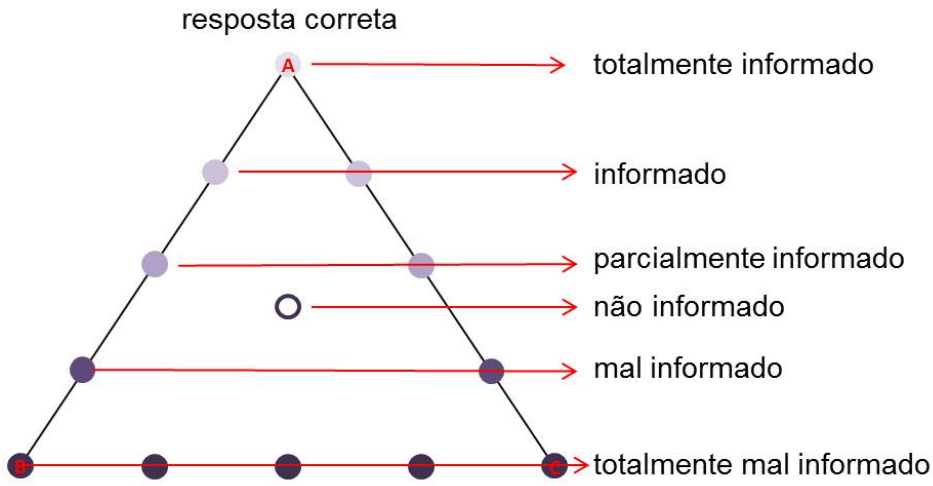

Figura 1.3: *Triângulo equilátero da MPA com a caracterização das alternativas de respostas de cada item da prova supondo a alternativa A como a correta.*

O aluno é aprovado se o percentual de respostas classificadas como “totalmente informado"for maior ou igual a 50% e se o percentual de respostas “totalmente mal informado"for menor ou igual a 25% ou se 90% ou mais de respostas estiverem nas classes “totalmente informado", “informado"e “parcialmente informado"e 10% ou menos de respostas na classe “totalmente mal informado", conforme Figura 1.4. Para os alunos não aprovados segundo esse critério, mas quase aprovados se não fossem as respostas de um ou dois itens (no máximo), um critério alternativo é disponibilizado considerando-se apenas o módulo 1,“Estrutura de Textos Científicos"(Figura 1.4).

<table>
<tr><td rowspan="2">Pontuação</td><td colspan="3">% de respostas</td></tr>
<tr><td>Critério 1</td><td>Critério 2</td><td>Módulo: Estrutura de Textos Científicos</td></tr>
<tr><td>Totalmente Informado</td><td>>=50%</td><td rowspan="3">>=90%</td><td>>=50%</td></tr>
<tr><td>Informado</td><td></td><td rowspan="2"><25%</td></tr>
<tr><td>Parcialmente Informado</td><td></td></tr>
<tr><td>Mal Informado</td><td></td><td></td><td rowspan="3"><25%</td></tr>
<tr><td>Não Informado</td><td></td><td></td></tr>
<tr><td>Totalmente Mal Informado</td><td><25%</td><td><10%</td></tr>
</table>

Figura 1.4: *Critério de classificação com base em MPA.*

Na perspectiva de auxiliar o estudo para o EPI, o aluno tem acesso a uma preparação orientada via *web* através de um ambiente computacional de aprendizado do inglês instrumental, denominado *Computer-Aided Learning of English for Academic Purposes* (CALEAP-Web). No ambiente, seu conhecimento de inglês é avaliado através de um teste e, com base nesse resultado, o aluno é orientado a realizar tarefas direcionadas a suprir as deficiências identificadas pelo teste anterior (Goncalves, 2004; Piton-Gonçalves et al, 2009).

Capítulo 2

Metodologia TAI-PI

O TAI-PI foi desenvolvido em Java 1.7.0 e SQLite versão 3.7.2 (Lanş 2006), para armazenamento do banco de dados.

O programa Same-CAT, criado por Thales Akira M. Ricarte ao longo de seu programa de mestrado, finalizado em 2012 pelo CCMC-ICMC, foi usado como base para a construção do TAI-PI (Ricarte, 2013). O Same-CAT é um programa que permite a aplicação de testes adaptativos informatizados, com itens do banco calibrados de acordo com o modelo de resposta gradual de Samejima (1969) da TRI, critério de Kullback-Leibler para seleção de itens a cada passo do teste, estimação da proficiência pelo método bayesiano de esperança *a posteriori* e com abordagem *Shadow test* para imposição de restrições.

A versão atual do TAI-PI contou com a colaboração do aluno do Bacharelado em Ciências da Computação do ICMC, Leonardo Tadashi Myiake, para a inclusão de algumas facilidades na interação usuário-máquina.

Na Seção 2.1, apresentamos os estudos realizados para a definição de quantas categorias de respostas serão consideradas no modelo de Samejima. Na Seção 2.2, é definido o modelo de Samejima. Na Seção 2.3 descrevemos o método de estimação para as habilidades do TAI-PI. A Seção 2.4 traz o método de seleção de itens para o teste adaptativo. A Seção 2.5 apresenta o método Shadow Test para imposição de restrições dessa seleção. Finalmente, a Seção 2.6 descreve os critérios de início e parada do teste.

2.1 Número de categorias

As possibilidades de resposta a cada item do EPI podem ser consideradas ordinais com 13 possibilidades de respostas classificadas em 6 categorias de acordo com a Figura 1.3. Como o intuito deste trabalho é aprimorar o EPI mantendo e aproveitando o máximo possível sua história, o modelo de resposta gradual (MRG), proposto por Samejima (1969), foi adotado para o TAI-PI. No entanto, uma análise descritiva e inferencial sobre o número de categorias a ser considerado apontou para a necessidade de redução a 3 possiveis respostas em cada item.

Essa decisão foi baseada nas análises de uma parte dos 167 itens criados e aplicados no periodo de 2002 a 2012 no EPI formato lápis-e-papel, os quais foram filtrados 74 itens para a calibração dos parâmetros que tinham ao menos 80 respondentes ao longo do tempo (para obterem-se estimativas relativamente confiáveis dos parâmetros) e presentes em provas comuns (para garantir a equalização simultânea com a estimação dos parâmetros do modelo da TRI), no entanto devido ao pequeno tamanho amostral, algumas categorias de respostas não tinham frequências suficiente para estimar os parâmetros de itens.

Juntando isso com a análise exploratória das curvas característica de item (CCI), decidiu-se juntas as respostas em 3 categorias, Nas Figuras 2.1 e 2.2 são apresentados as CCI do item 25 antes e depois do agrupamento de categorias.

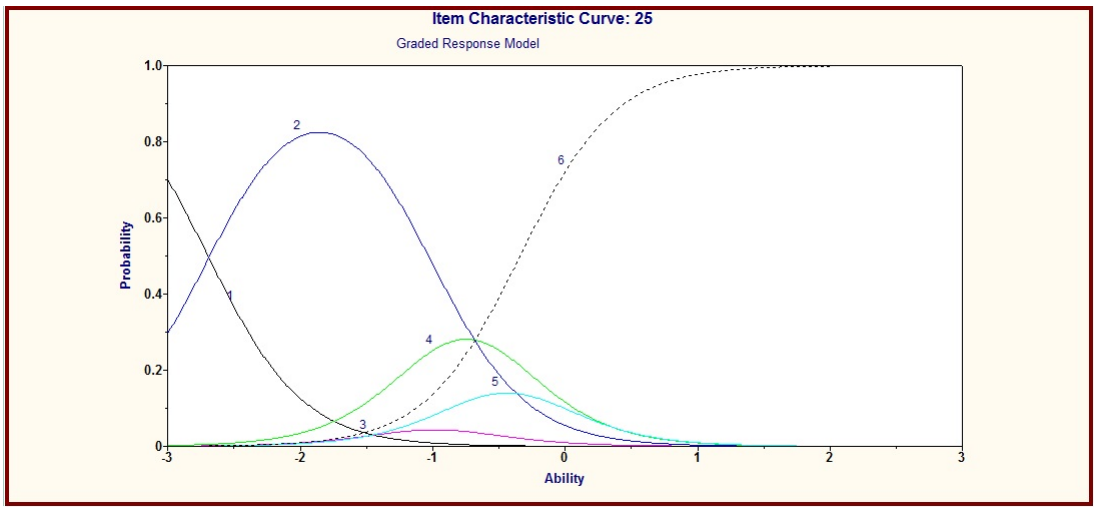

Figura 2.1: *Curva Característica do Item 25 com as 6 categorias iniciais*

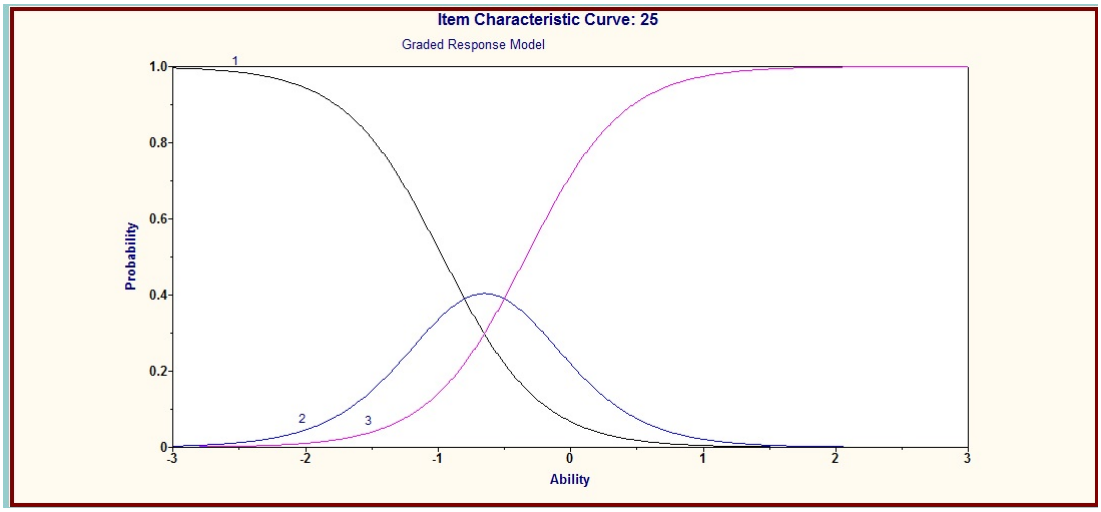

Figura 2.2: *Curva Característica do Item 25 com agrupamento em 3 categorias*

O agrupamento das categorias encontra-se definido na Figura 2.3.

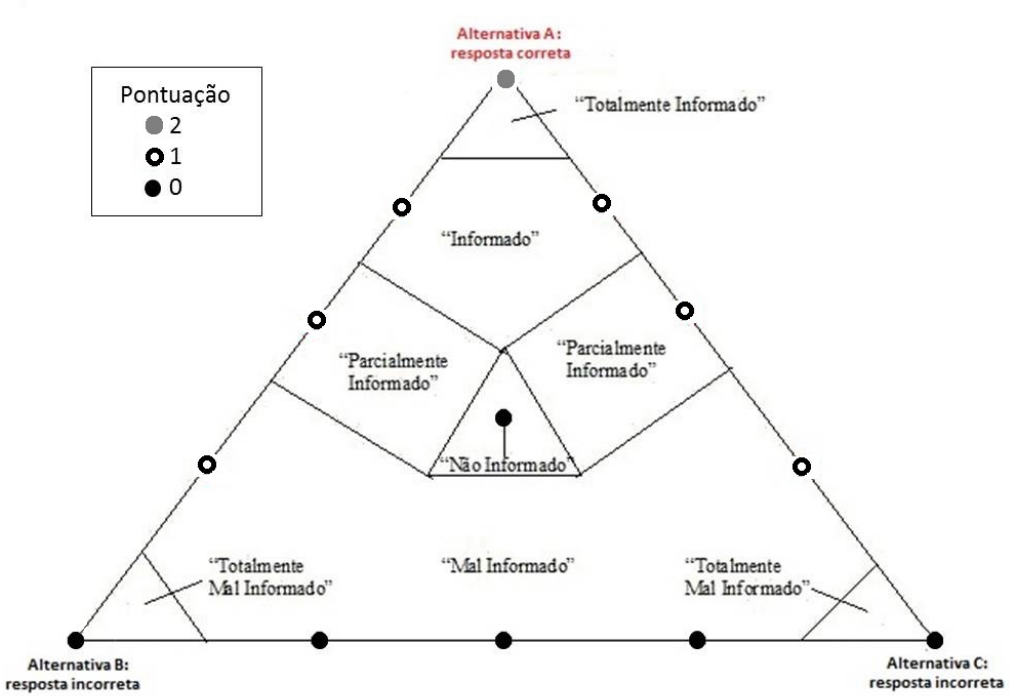

Figura 2.3: *Agrupamento das categorias de resposta aos itens adotado para o TAI-PI.*

2.2 Modelo de Samejima

Suponha $k = 0, 1, 2$ denotanto as 3 categorias da Figura 2.3 categorias organizadas em ordem crescente, ou seja, quanto maior o valor de k, mais próxima ela estará da resposta totalmente correta. A probabilidade de um indivíduo, com valor de proficiência θ, escolher a categoria k ou alguma maior que ela no item i é dada por:

$$P^{+}_{i,k}(\theta) = \frac{1}{1 + exp[-a_i(\theta - b_{i,k})]}, \tag{2.1}$$

sendo $b_{i,k}$ o parâmetro de dificuldade da categoria k do item i, $P^{+}_{i,0} = 1$ e a_i o parâmetro de dificuldade do item i . Além disso, $b_{i,0} \leq b_{i,1} \leq b_{i,2}$.

A Figura 2.4 ilustra o gráfico da probabilidade acumulada sob o modelo de Samejima com 3 categorias de resposta.

A probabilidade de um individuo, com valor de proficiência θ, escolher a categoria k no item i é:

$$P_{i,k}(\theta) = P^{+}_{i,k}(\theta) - P^{+}_{i,k+1}(\theta), \tag{2.2}$$

com $P^{+}_{i,3}(\theta) = 0$, por definição. Na Figura 2.5 tem-se um exemplo da curva das probabilidades de escolha da categoria sob o modelo de Samejima para 3 categorias de resposta.

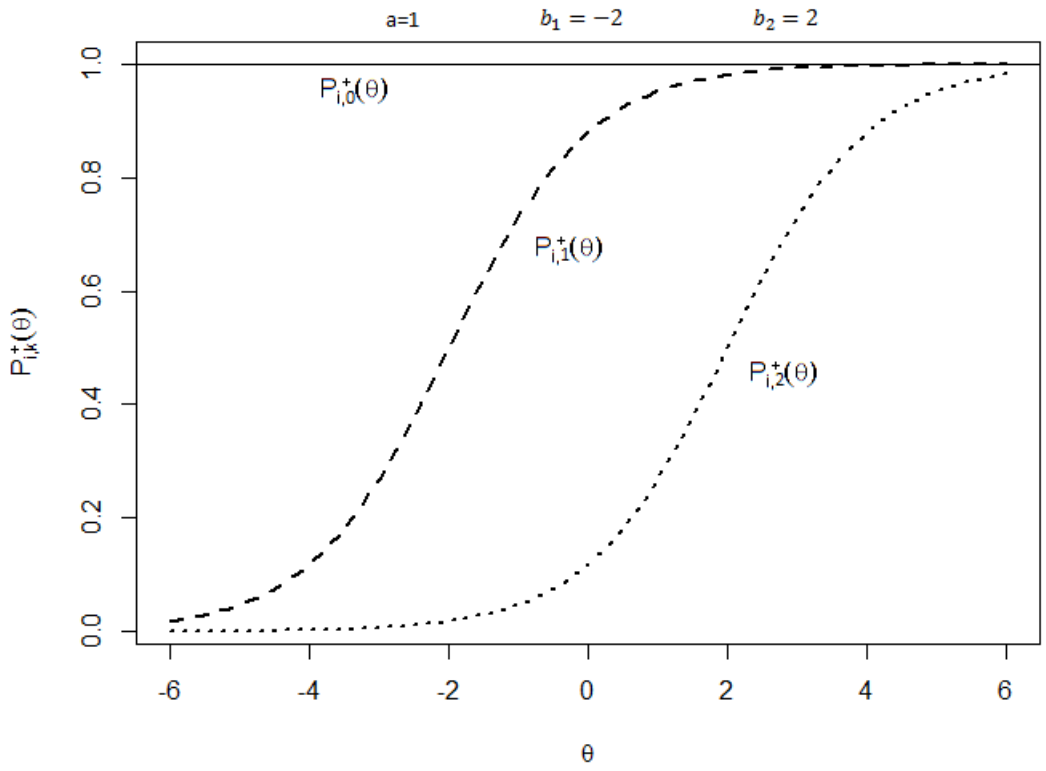

Figura 2.4: *Exemplo de uma curva de probabilidade acumulada sob o modelo de Samejima com 3 categorias de resposta.*

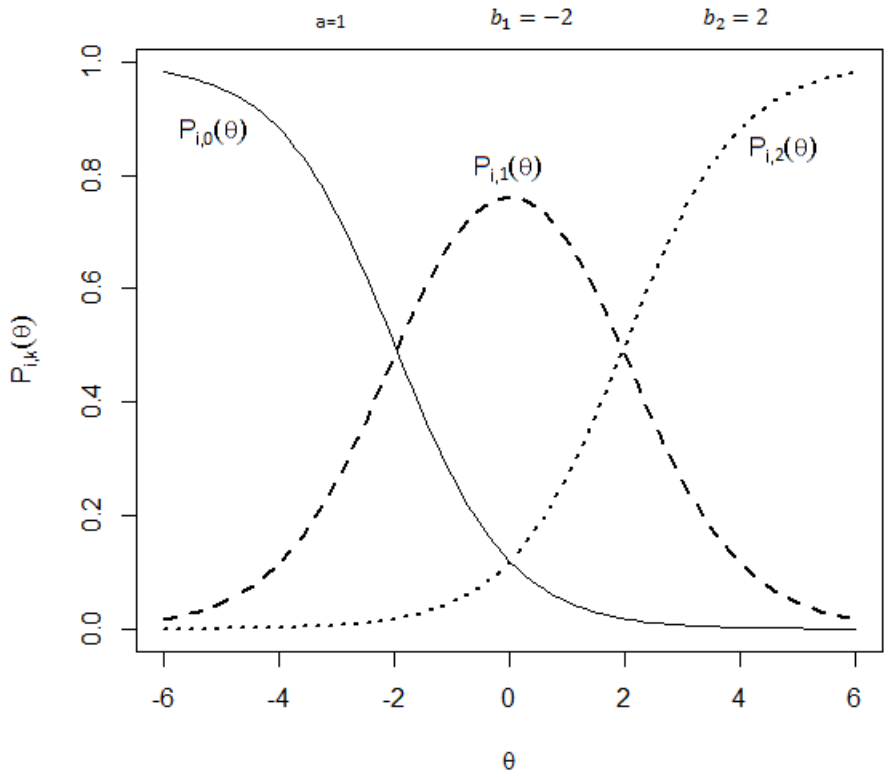

Figura 2.5: *Exemplo de uma curva de probabilidade de escolha da categoria sob o modelo de Samejima com 3 categorias de resposta.*

Supondo $l-1$ itens já administrados, a função de verossimilhança do modelo de Samejima é:

$$L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{l-1} f(\boldsymbol{x}_i \mid \theta) = \prod_{i=1}^{l-1} \prod_{k=0}^{2} P_{i,k}^{x_{ik}}(\theta), \tag{2.3}$$

em que $P_{i,k}(\theta)$ é dado em 2.2, $k = 0, 1, 2$ representa as categorias dos itens, $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_{l-1}), \boldsymbol{x}_i = (x_{i0}, x_{i1}, x_{i2})$ e x_{ik} são variáveis que assumem o valor 1 se a categoria k do item i for escolhida e 0, caso contrário.

2.3 Estimação do traço latente

O método EAP foi usado para a estimação de θ, seguindo o conhecido fato da estimação bayesiana funcionar melhor em TAI. Baker (2001), Baker & Kim (2004), Mislevy & Stocking (1989) sugeriam que a estimativa pontual de θ seja obtida por EAP porque o método pode ser calculado sem a necessidade de métodos iterativos, o que diminiu tempo de estimação. A distribuição, esperança e variância *a posteriori* de θ, condicionadas aos dados observados, são dadas, respectivamente, pelas expressões 2.4, 2.5 e 2.6.

$$\pi(\theta) \propto L(\theta \mid \boldsymbol{x}) g(\theta \mid \boldsymbol{\lambda}), \tag{2.4}$$

$$E(\theta \mid \boldsymbol{x}, \lambda) = \frac{\int_{\mathbb{R}} \theta L(\theta \mid \boldsymbol{x}) g(\theta \mid \lambda) d\theta}{\int_{\mathbb{R}} L(\theta \mid \boldsymbol{x}) g(\theta \mid \lambda) d\theta}, \tag{2.5}$$

$$Var(\theta \mid \boldsymbol{x}, \lambda) = \frac{\int_{\mathbb{R}} (\theta - E(\theta))^2 L(\theta \mid \boldsymbol{x}) g(\theta \mid \boldsymbol{\lambda}) d\theta}{\int_{\mathbb{R}} L(\theta \mid \boldsymbol{x}) g(\theta \mid \boldsymbol{\lambda}) d\theta}, \tag{2.6}$$

em que $g(\theta \mid \boldsymbol{\lambda})$ é a distribuição a *priori* de θ e $\boldsymbol{\lambda}$ são os hiperparâmetros.

O método numérico de quadratura gaussiana (Dehghan et al, 2006) foi utilizado para o cálculo das integrais.

2.4 Seleção de itens

Para a seleção dos itens a cada passo do TAI-PI, não foi adotado o critério mais tradicional da literatura (informação de Fisher). Isto porque se a estimativa corrente de θ for distante do seu valor real, este critério pode ser inadequado. Num TAI, especialmente nos passos iniciais, é muito possível que isso ocorra.

Para o i-ésimo item, a informação de Fisher para um valor de traço latente θ é definida por:

$$I_i(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \log f(\boldsymbol{X_{i.}}|\theta)\right)^2\right] \tag{2.7}$$

com $f(X_{i.}|\theta)$ sendo a função de probabilidade da resposta dada pelo modelo 2.2, ou seja, especificamente para o modelo de Samejima. Pode-se observar que 2.7 é calculada em função de θ, que é substituido pela estimativa corrente. Se essa estimativa estiver muito distante do verdadeiro valor, a informação de Fisher será maximizada em um ponto distante do real e, portanto, pode estar muito equivocada.

A informação de Kullback-Leibler, a qual mede a distância entre duas verossimillhanças para um mesmo espaço paramétrico, é que foi escolhida para implementação no TAI-PI, conforme sugerido por Chang & Ying (1996). Quanto maior o valor dessa 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, $\widehat{\theta}$, e real θ, no modelo de Samejima, é dada por:

$$K_i(\theta, \widehat{\theta}) = E\left[\log \frac{f(\boldsymbol{X}_i|\theta)}{f(\boldsymbol{X}_i|\widehat{\theta})}\right] = \sum_{k=0}^{2} P_{i,k}(\theta) \log \frac{P_{i,k}(\theta)}{P_{i,k}(\widehat{\theta})} \tag{2.8}$$

Vale lembrar que, como temos independência condicional entre as respostas, a informação de Kullback-Leibler após *l-1* itens administrados é escrita como:

$$K_{l-1}(\theta, \widehat{\theta}) = \sum_{i=1}^{l-1} K_i(\theta, \widehat{\theta}). \tag{2.9}$$

Uma vez que o verdadeiro valor do traço latente (θ) é desconhecido, Sands and Waters (1996) propuseram considerar (2.8) em um intervalo de confiança do traço latente $[\widehat{\theta} - \delta_l, \widehat{\theta} + \delta_l]$, sendo δ_l uma função decrescente em relação a l (passo corrente do TAI). Assim o critério para selecionar o próximo item a ser apresentado no teste é dado por:

$$i_l \equiv arg\max_i \left\{ \int_{\widehat{\theta}-\delta_l}^{\widehat{\theta}+\delta_l} K_i(\theta, \widehat{\theta}) d\theta : i \epsilon L \right\} \tag{2.10}$$

sendo $\delta_l = z_\gamma/\sqrt{l}$ e L, o conjunto dos itens ainda não apresentados ao indivíduo no teste.

2.5 Shadow test

O critério de seleção expresso em (2.10) não leva em consideração nenhuma restrição subjetiva para a composição do teste. Por exemplo, uma avaliação adequada da proficiência em inglês deve contemplar itens associados a cada um dos três modulos: estrutura de texto científico, compreensão de texto e convenções gramaticais da língua. A não imposição dessa restrição pode gerar testes com nenhum item de um dos módulos e um número excessivo de itens de outro, distorcendo a avaliação desejada. Outra necessidade de ordem prática diz respeito ao número de itens e de textos no teste. Dois indivíduos que realizam testes com números (e tamanhos) de textos muito distoantes ou com números de itens muito diferentes podem gerar a sensação de injustiça na comparação das estimativas das respectivas proficiências. O indivíduo que respondeu um teste com muitos textos ou muito longo pode reclamar que o cansaço foi maior e, portanto, o desgaste ao longo da realização da prova pode ter influenciado de forma não equiparável entre os indivíduos. Por outro lado, aquele que respondeu um teste curto, com poucos itens, pode reclamar que não houve "tempo"(i.e., número de itens) suficiente para que sua real proficiência seja evidenciada.

Uma das propostas mais interessantes na literatura para uma fácil inclusão de restrições desse tipo é a abordagem *Shadow test* (Van der Linden, 2000). O método de otimização de programação linear inteira é usado para obtenção de um subconjunto (de tamanho n, previamente fixado) de itens do banco que maximiza a informação (2.9), tomando $l - 1 = n$, sujeita às restrições de interesse, também previamente definidas. Um conjunto de n itens é obtido em cada passo do algorítmo, gerando um teste como conjunto solução. É importante destacar que uma das restrições do teste no passo l é que todos os itens já administrados anteriormente estão presentes no teste corrente. O item a ser administrado no passo l do teste é aquele não apresentado nos passos anteriores, que pertença ao conjunto solução do passo corrente e que apresente a máxima informação (2.8) para o atual valor estimado $\hat{\theta}$. 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 do *Shadow test*. Uma representação bastante ilustrativa do funcionamento do *Shadow test* encontra-se na Figura 2.6.

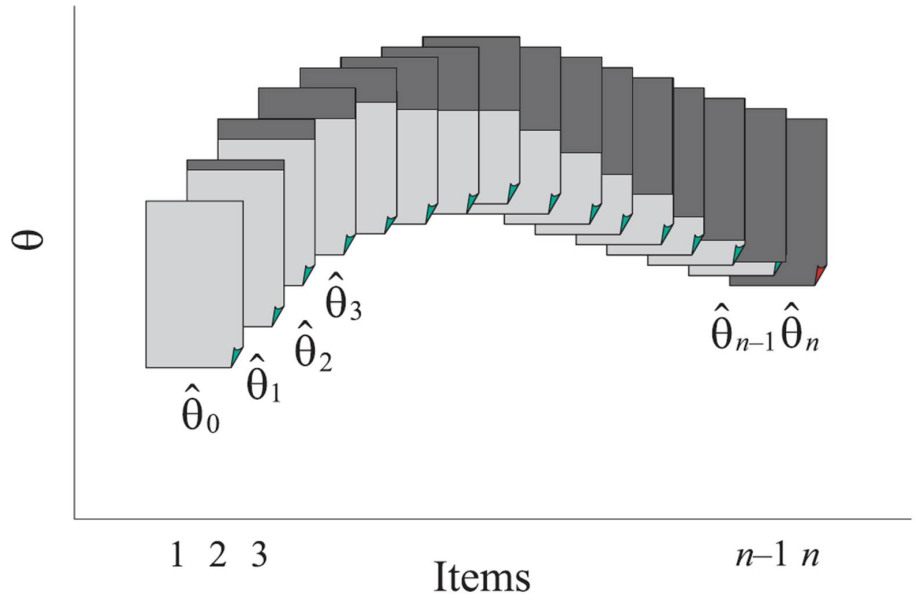

Figura 2.6: *Representaçao do Shadow test. Fonte: van der Linden, 2010.*

Um especial cuidado deve ser tomado na presença de itens associados a um mesmo texto ou figura (genericamente denominado "estímulo"). Em situações reais, o ideal é apresentar todos os itens associados ao mesmo estímulo sequencialmente (Van der Linden & Glas, 2010; Van der Linden & Pashley , 2010). Neste trabalho, quando um estímulo novo aparece em uma certa iteração do *Shadow test*, todos os itens associados àquele estímulo no *Shadow test* dessa iteração irão ser apresentados consecutivamente ao indivíduo.

2.6 Critério de início e parada

O nível de conhecimento inicial é a estimativa inicial provisória da proficiência, necessária para o início do teste e relacionada com o nível de dificuldade da primeira questão. Conejo et al. (2001) considera que o nível de dificuldade da primeira questão deve ser escolhido de forma a possibilitar uma redução no tempo de teste. Sukamolson (2002) salienta que usualmente a primeira questão possui um nível de dificuldade médio. Como a princípio não se sabe qual a habilidade do respondente, um item médio é a melhor opção inicial.

Como nenhuma informação prévia da proficiência dos alunos é conhecida decidiu-se iniciar com um valor médio $\widehat{\theta_0} = 0$ para todos.

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 por numéro 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. Neste trabalho foi considerado a prova com um número fixo de itens. Isso é exigido pelo Shadow test e é bastante coerente com o EPI, uma vez que a prova tem tempo máximo de 2 horas para ser aplicada e que não se deseja que os alunos respondam a um número muito diferente de itens para evitar questionamentos.

Capítulo 3

Estudo de simulação

Este estudo de simulação foi executado em *software* R e delineado para ilustrar o funcionamento de um TAI, em termos da eficiência na estimação do traço latente, comparando distribuições a prioris na utilização do EAP e números diferentes de itens no teste (10, 20, 25 e 30). Para o banco de itens foram simulados 500 itens com 3 categorias de respostas, com $a \sim lognormal(0.7, 0.1)$, gerando apenas bons valores para os parâmetros de dificuldade (Mislevy, 1986), e $b_i \sim N(0, 1.2)$ com $b_1 < b_2$, com base no modelo de Samejima com 3 categorias de resposta. O método EAP foi utilizado para estimar o traço latente do mesmo modelo e o critério de Kullback-Leibler para a seleção de itens. Como critério de início do TAI foi adotado que todos com $\hat{\theta}_0 = 0$ e como critério de parada número de itens fixos(10, 20, 25 e 30 itens).

No primeiro estudo de simulação, para cada $\theta \in \{-3, -2, -1, 0, 1, 2, 3\}$ foram simulados 200 testes. Adicionalmente para cada teste simulado, três tipos de distribuições a priori para as habilidades foram adotadas para o método EAP: N(0,1), que é usualmente adotada na literatura, N(0,2), para dar menos informação sobre o parâmetro em questão e $N(\hat{\theta}_{i-1}, \sigma^2(\hat{\theta}_{i-1}))$ para aproveitar a estimativa corrente como informação a priori da habilidade do indivíduo. Totalizando assim $7 \times 3 \times 200 = 4200$ testes simulados.

Para avaliar a qualidade do modelo, foram calculadas as medidas a seguir:

$$Vicio = \sum_{j=1}^{200} \frac{\hat{\theta}_{jl} - \theta_l}{200},$$

$$EQM = \frac{\sum_{j=1}^{200} (\hat{\theta}_{jl} - \theta_l)^2}{200},$$

em que $j = 1, ..., 200$ representa as repetições de cada valor fixado de θ, $l = 1, ..., 7$ indexa os 7 valores fixados para θ, $\theta_l \in \{-3, -1, -2, 0, 1, 2, 3\}$ são os valores fixados para θ, $\hat{\theta}_{jl}$ é a estimativa de θ_l obtida pelo método EAP na simulação. Os valores encontrados são apresentados nas Tabelas 3.1 e 3.2, nas quais pode-se observar que o vício se aproxima de zero e o erro quadrático médio diminui conforme aumenta o número de itens e que o ganho em precisão de estimação aumentando o tamanho do teste de 25 e 30 talvez não seja importante e 20 itens já são suficientes em um teste. Com exceção da priori $N(\hat{\theta}_{i-1}, \sigma^2(\hat{\theta}_{i-1}))$, as estimativas foram muito próximas dos valores reais para as demais prioris apresentadas, o que pode ser confirmado pelos baixos valores do vício e do erro quadrático médio. No entando aumentam consideravelmente quando se é usado N(0,1) como priori para valores de habilidades muito distantes da média (ou seja, próximos a -3,3).

Tabela 3.1: *Vício dos estimadores para diferentes prioris.*

Priori	Itens no teste	Valores reais de θ						
		-3	-2	-1	0	1	2	3
N(0,1)	10	0.3978	0.1478	0.0636	0.0057	-0.0707	-0.1050	-0.6543
	20	0.1756	0.0768	0.0373	0.0180	-0.0423	-0.0246	-0.5341
	25	0.1403	0.0451	0.0175	0.0104	-0.0256	-0.0112	-0.5029
	30	0.1200	0.0497	0.0168	0.0045	-0.0154	-0.0093	-0.4754
N(0,2)	10	0.1020	0.0569	-0.0270	-0.0149	-0.0199	0.1536	-0.1768
	20	0.0467	0.0222	-0.0144	-0.0110	-0.0032	0.1192	-0.0814
	25	0.0482	0.0237	-0.0086	-0.0134	-0.0043	0.1064	-0.0636
	30	0.0486	0.0241	-0.0100	-0.0099	-0.0037	0.1153	-0.0558
$N(\hat{\theta}_{i-1}, \sigma^2(\hat{\theta}_{i-1}))$	10	0.5851	0.2588	0.0557	-0.0056	-0.0599	-0.1677	-0.7912
	20	0.3225	0.1229	0.0404	-0.0010	-0.0311	-0.0718	-0.6824
	25	0.2759	0.0971	0.0228	-0.0038	-0.0337	-0.0550	-0.6565
	30	0.2391	0.0820	0.0210	-0.0014	-0.0287	-0.0536	-0.6352

Tabela 3.2: *Erro Quadrático Médio dos estimadores para diferentes prioris.*

Priori	Itens no teste	Valores reais de θ						
		-3	-2	-1	0	1	2	3
N(0,1)	10	0.2498	0.1295	0.0760	0.0735	0.0790	0.1015	0.4478
	20	0.0795	0.0525	0.0430	0.0395	0.0427	0.0662	0.3105
	25	0.0568	0.0397	0.0355	0.0314	0.0352	0.0663	0.2800
	30	0.0463	0.0397	0.0325	0.0249	0.0304	0.0608	0.2559
N(0,2)	10	0.0899	0.1164	0.0896	0.0778	0.0962	0.2381	0.0923
	20	0.0516	0.0580	0.0550	0.0438	0.0400	0.1551	0.0675
	25	0.0426	0.0462	0.0492	0.0339	0.0321	0.1350	0.0667
	30	0.0371	0.0394	0.0395	0.0287	0.0308	0.1391	0.0730
$N(\hat{\theta}_{i-1}, \sigma^2(\hat{\theta}_{i-1}))$	10	0.4508	0.1217	0.0872	0.1041	0.0764	0.0920	0.6352
	20	0.1641	0.0539	0.0433	0.0487	0.0416	0.0502	0.4755
	25	0.1251	0.0446	0.0392	0.0358	0.0356	0.0474	0.4408
	30	0.0953	0.0427	0.0339	0.0305	0.0280	0.0473	0.4133

Contemplando apenas os tamanhos de teste de 20, 25 e 30 itens e aumentando o número de habilidade geradas de 200 para 1000, realizou-se uma segunda simulação com o intuito de constrastar graficamente os valores estimados e verdadeiros. Nas Figuras 3.1 e 3.2 são apresentadas a proficiência verdadeira em função da estimada e os obtidos respectivamente, pode-se observar que para todas as distribuições a prioris consideradas e número de itens no teste as estimações são bem próxima dos valores reais. Este fato auxiliará na escolha do número de itens a serem apresentados na prova.

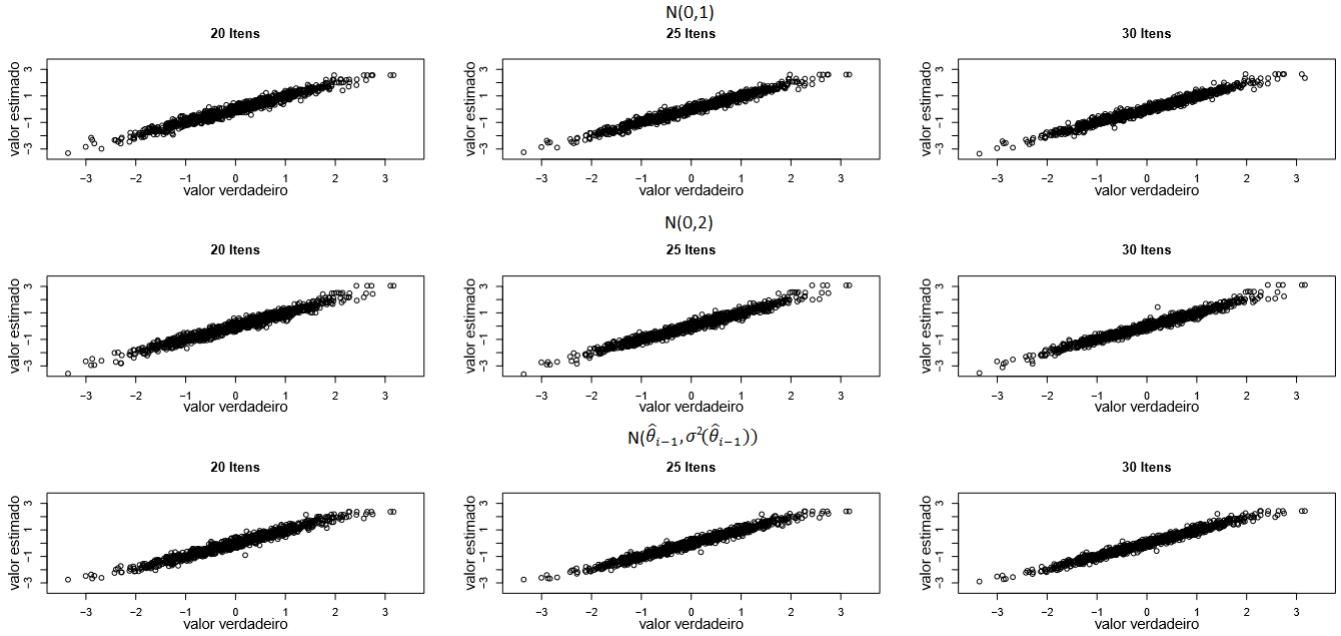

Figura 3.1: *Valor verdadeiro em função do valor estimado de θ por 3 prioris, para diferentes tamanhos de teste (20, 25 e 30 itens)*

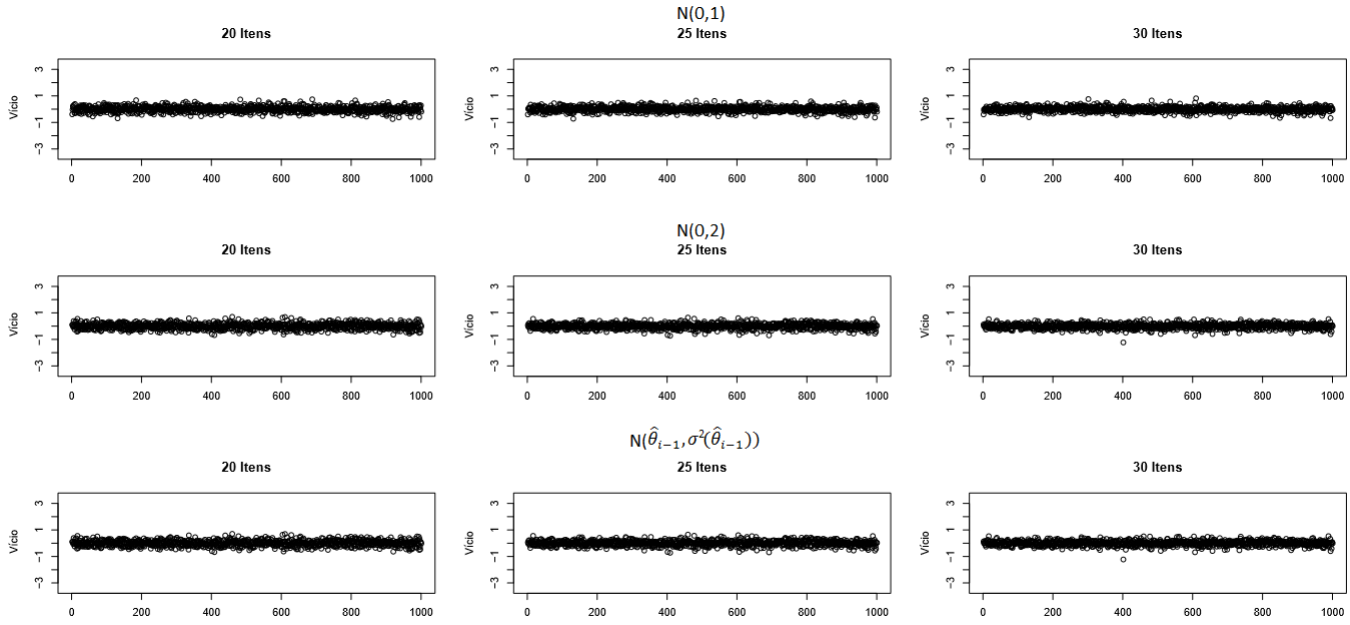

Figura 3.2: *Vício de 1000 respondentes simulados estimado por 3 prioris, para diferentes tamanhos de teste(20, 25 e 30 itens)*

Capítulo 4

Aplicação real

As aplicações reais do TAI-PI foram baseadas a partir dos 167 itens criados e aplicados no periodo de 2002 a 2012 no EPI formato lápis-e-papel. No entanto para as aplicações foram filtrados 74 itens para a calibração dos parâmetros, os quais tinham ao menos 80 respondentes ao longo do tempo (para obterem-se estimativas relativamente confiáveis dos parâmetros) e presentes em provas comuns (para garantir a equalização simultânea com a estimação dos parâmetros do modelo da TRI). Desses itens foram selecionados os que tinham parâmetros de discriminação acima de 0,7, associados a nenhum texto ou com pelo menos 3 itens associados a um texto. As estimativas dos parâmetros dos itens que compõem esse banco foram obtidas na métrica(0,1) (de Andrade et al. , 2000) utilizando-se as respostas resultantes das aplicações desde 2002. Sobrando asssim depois de várias filtragem 40 itens para o banco que foram utilizados em aplicações reais e também usados para o estudo do número de categorias que deveriam ser consideradas no modelo de Samejima.

A primeira aplicação do TAI-PI ocorreu em novembro de 2013 a 26 alunos de pós-graduação do CCMC. Como o método de avaliação proposto no TAI-PI ainda não havia sido validado e não se sabia nem mesmo um ponto de corte adequado na escala do traço latente para aprovação ou reprovação do aluno, optou-se por usar a ocasião apenas para um teste do funcionamento do *software*. Na realidade, os alunos responderam todos os mesmo 25 itens (pré-escolhidos dentre todos os 40 calibrados do banco), com variação na ordem de apresentação segundo a maximização da informação de Kullback-Leibler, conforme explicado anteriormente. A aplicação indentificou um problema de autorização para salvamento das respostas em alguns computadores devido à forma incorreta de efetuar o login na rede da sala por alguns alunos. O problema foi solucionado inserindo uma verificação de autorização de acesso já no início do teste, nas versões subsequentes do TAI-PI, e emitindo um aviso se o acesso for negado.

A segunda aplicação do TAI-PI foi realizada em maio de 2014 por 59 alunos da pós graduação do CCMC-ICMC e do Programa Interinstitucional de Pós Graduação em Estatística (PIPGES). Foram selecionados para compor o banco 40 itens, com as mesmas restrições de parâmetro de discriminação, associadas a texto citadas acima na primeira aplicação. Nessa fase queria-se testar o funcionamento do TAI-PI em uma prova real composta de 25 itens selecionados de um banco de 40 itens, pelo método explicado no Capítulo 2 e obedecendo a algumas restrições de conteúdo. Todos os alunos responderam os 40 itens (pois o critério de correção ainda seria pelo MPA), sendo que apenas os 25 primeiros itens foram selecionados de forma adaptativa (conforme metodologia no Capítulo 3 e os demais 15 itens foram apresentados sequencialmente. A estrutura dos 40 itens é apresentada na Tabela 4.1

Tabela 4.1: *Banco de itens do TAI-PI, maio de 2014*

Módulo	Estímulo	Número de Itens
1	Abstract 1	3
	Abstract 2	4
	Abstract 3	3
	Abstract 4	4
2	Introduction 5	7
	Introduction 6	4
3	Nenhum	15

Para a seleção dos 25 primeiros itens de forma adaptativa, foram impostas as seguintes restrições de conteúdo e tamanho de textos, implementados via Shadow test (Van der Linden, 2000):

- Exatamente 3 textos do módulo 1 (com no mínimo 3 e no máximo 4 itens cada texto)
- Exatamente 1 texto do módulo 2 (com no mínimo 4 e no máximo 7 itens)
- Nó máximo 11 itens do módulo 3.
- No máximo 2015 palavras contidas nos textos contemplando pelos 25 itens.

A prova foi realizada sem maiores intercorrências. Os dados obtidos correspondem, além das respostas a todos os itens do banco, ao traço latente estimado e respectivo erro padrão em cada um dos 25 passos do TAI e, também, após os 40 itens do banco serem respondidos. A Tabela 4.2 traz as estimativas do traço latente dos alunos após o teste adaptativo de 25 itens e após a resposta a todos os 40 itens do banco, assim como a classificação (em aprovado ou reprovado) com base no MPA aplicado às 40 respostas. Como pode-se observar destacado em cinza, 4 alunos estão com estimativas baixas e foram aprovados. Note, pela Tabela 4.3, que os resultados desses indivíduos correspondem ao mínimo necessário (ou quase isso, no caso do aluno 17) para aprovação pelo MPA. Pode-se interpretar tal fato como uma amenização dessa contradição entre os métodos (MPA e TAI)

Tabela 4.2: *Estimativas dos traços latentes (após 25 e 40 itens respondidos) e classificação pelo MPA na aplicaçao do TAI- de maio de 2014.*

Número	Indivíduo	$\widehat{\theta}_{25}$	$\widehat{\theta}_{40}$	MPA
1	10	-2,07	-1,68	REFAZER
2	1	-1,62	-1,66	REFAZER
3	6	-1,29	-1,60	REFAZER
4	35	-1,39	-1,59	REFAZER
5	37	-1,-19	-1,44	REFAZER
6	26	-1,34	-1,41	APROVADO
7	52	-1,09	-1,37	REFAZER
8	23	-1,37	-1,28	REFAZER
9	29	-0,90	-1,16	REFAZER
10	9	-1,06	-1,15	REFAZER
11	20	-1,07	-1,15	REFAZER
12	12	-0,96	-1,11	REFAZER
13	38	-0,92	-1,08	APROVADO
14	27	-1,01	-0,99	REFAZER
15	3	-0,28	-0,76	REFAZER
16	54	-0,29	-0,75	REFAZER
17	22	-0,58	-0,73	REFAZER
18	47	-0,34	-0,69	APROVADO
19	49	-0,55	-0,69	REFAZER
20	18	-0,46	-0,68	REFAZER
21	28	-0,81	-0,66	REFAZER
22	51	-0,62	-0,65	REFAZER
23	25	-0,52	-0,65	REFAZER
24	2	-0,61	-0,64	REFAZER
25	39	-0,71	-0,60	REFAZER
26	5	-0,62	-0,56	REFAZER
27	11	-0,29	-0,54	REFAZER
28	19	-0,17	-0,52	REFAZER
29	17	-0,56	-0,48	APROVADO
30	40	-0,28	-0,41	REFAZER
31	53	-0,58	-0,41	REFAZER
32	44	-0,32	-0,40	APROVADO
33	42	-0,01	-0,34	APROVADO
34	45	-0,29	-0,24	APROVADO
35	48	-0,03	-0,22	APROVADO
36	41	-0,36	-0,15	APROVADO
37	31	-0,07	-0,15	APROVADO
38	34	-0,22	-0,14	APROVADO
39	57	-0,29	-0,12	APROVADO
40	50	0,19	-0,10	APROVADO
41	58	0,00	-0,09	APROVADO
42	33	-0,23	-0,04	APROVADO
43	43	-0,01	0,08	APROVADO
44	13	0,44	0,09	APROVADO
45	55	0,11	0,13	APROVADO
46	32	0,11	0,27	APROVADO
47	30	0,46	0,35	APROVADO
48	15	0,21	0,38	APROVADO
49	4	0,70	0,45	APROVADO
50	14	0,51	0,47	APROVADO
51	16	0,49	0,51	APROVADO
52	56	0,30	0,53	APROVADO
53	8	0,62	0,58	APROVADO
54	21	0,42	0,67	APROVADO
55	36	0,59	0,70	APROVADO
56	24	0,72	0,93	APROVADO
57	7	0,89	0,94	APROVADO
58	59	0,85	0,95	APROVADO
59	46	1,17	1,08	APROVADO

Tabela 4.3: *Porcentagem de acertos de 4 alunos em cada categoria na aplicação de maio de 2014.*

Aluno	% Total.Informado	%Informado	% Parc. Informado	% Total. Mal Informado
26	0%	0%	90 %	10%
38	50%	0%	0%	25%
47	50%	12,5%	5%	25%
17	52,5%	5%	7,5%	25%

A Figura 4.1 traz o *boxplot* das proficiências estimadas no TAI-PI para os 59 alunos classificados em reprovados ou aprovados, segundo o MPA, pode-se verificar um possível valor de corte para a classificação do aluno na escala traço latente em torno de -0.5 e que as proficiências parecem ser bem maiores no grupo aprovado do que no reprovado, com pouca sobreposição entre as duas distribuições do traço latente, mais presente na estimação dos 25 itens do que com 40.

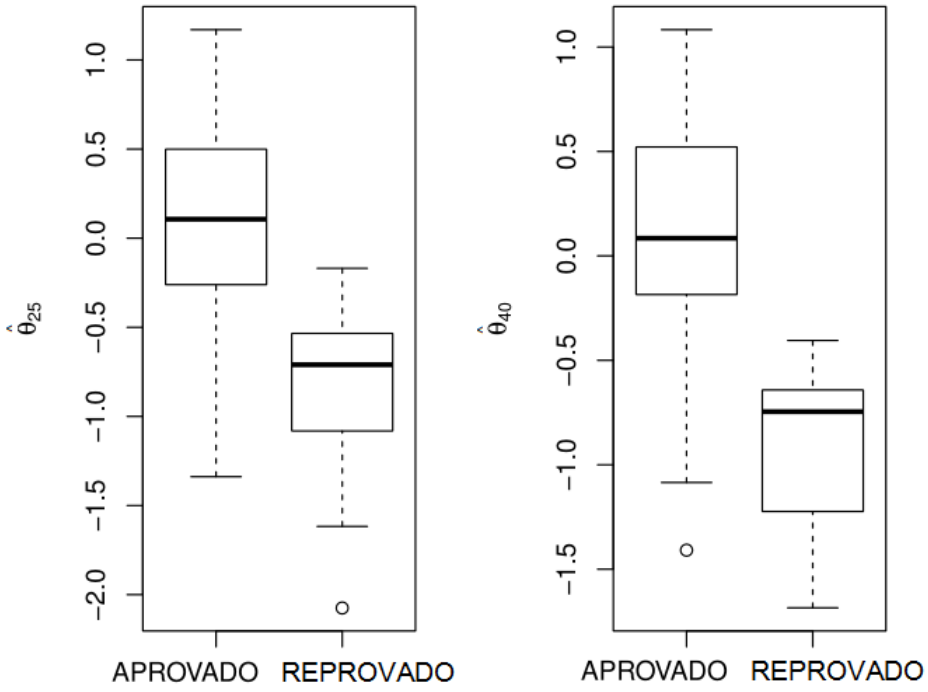

Figura 4.1: *Boxplot da proficiência estimada após 25 e 40 itens respondidos dos 59 alunos na aplicação do TAI-PI, separadamente para o grupo reprovado e aprovado pelo critério MPA.*

Nas Figuras 4.2, 4.3 e 4.4 são apresentadas proficiências estimadas e seus erros padrão de alguns alunos (dois deles, com respostas descritas na Tabela 4.3 em cada passo do teste adaptativo, pode-se observar que quando o aluno erra o item (representado pela categoria 0), a proficiência subsequente é estimada num valor menor. Enquanto que, quando o aluno acerta (categoria 2), a proficiência estimada aumenta.

Além disso, em testes adaptativos informatizados classificatórios encontrados na literatura é comum usar um intervalo de credibilidade para o traço latente no último passo para classificar o aluno como aprovado (se o intervalo todo estiver acima do corte) ou reprovado (se o intervalo estiver abaixo do corte). No caso em que o intercalo contiver o corte, seguir com o teste para ter-se uma estimativa mais precisa de $\widehat{\theta}$. Por esse critério, tanto o aluno 26, quanto 46, provavelmente estariam classificados após 25 itens respondidos (se considerar-se intervalos com amplitudes de 4 vezes o erro padrão, aproximadamente).

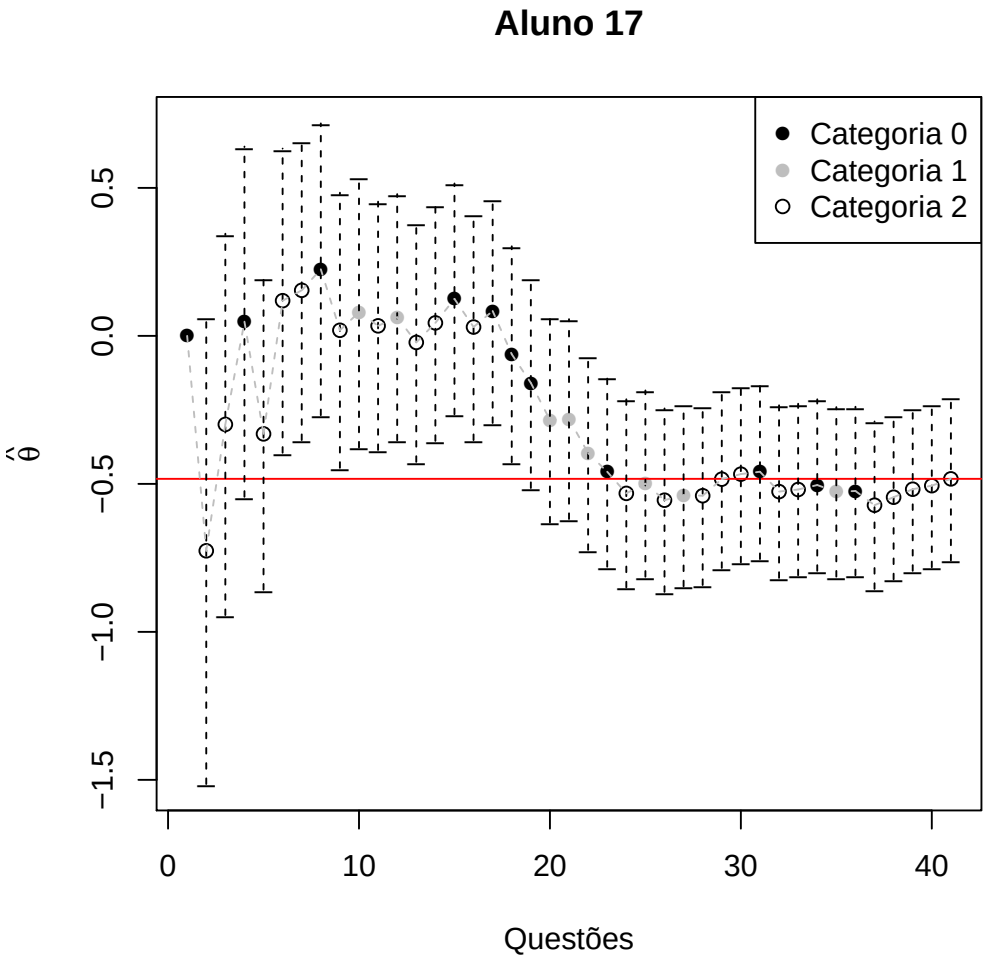

Figura 4.2: *Proficiências estimadas do aluno 17 em cada passo do TAI-PI, a linha vermelha representa a estimativa final após 40 itens.*

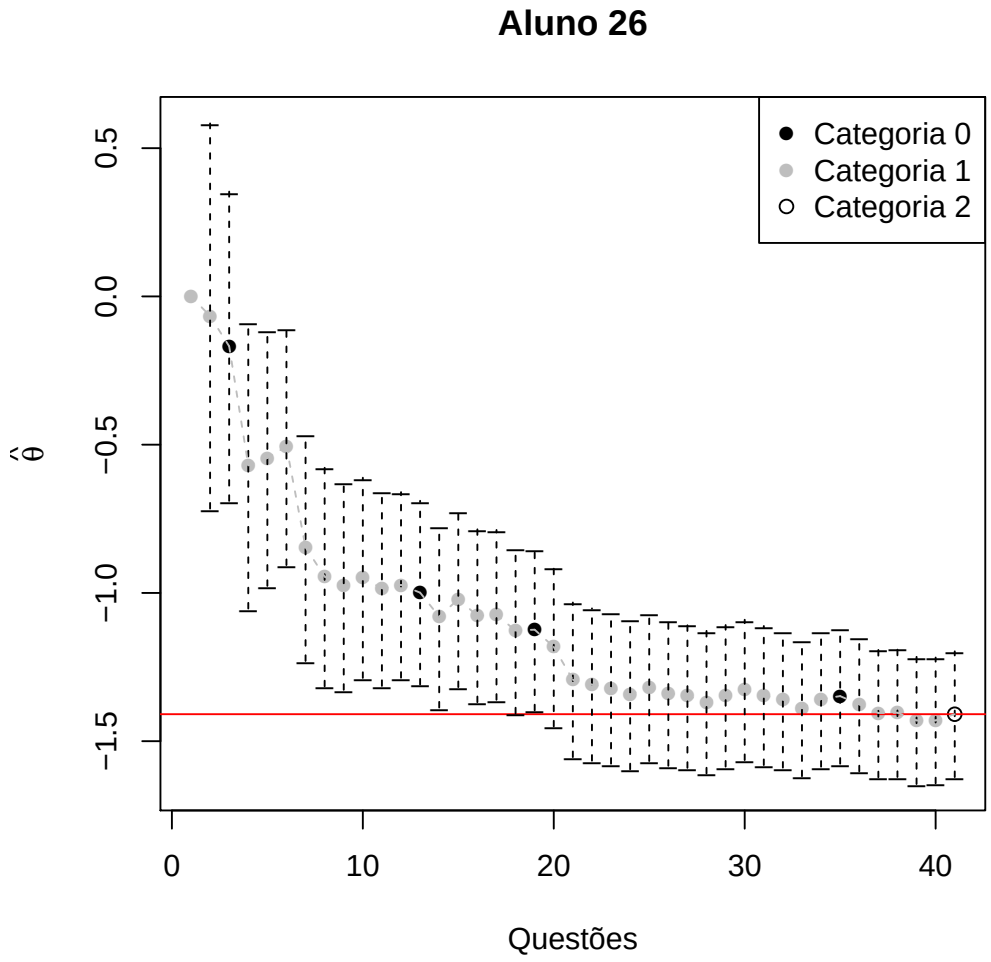

Figura 4.3: *Proficiências estimadas do aluno 26 em cada passo do TAI-PI, a linha vermelha representa a estimativa final após 40 itens.*

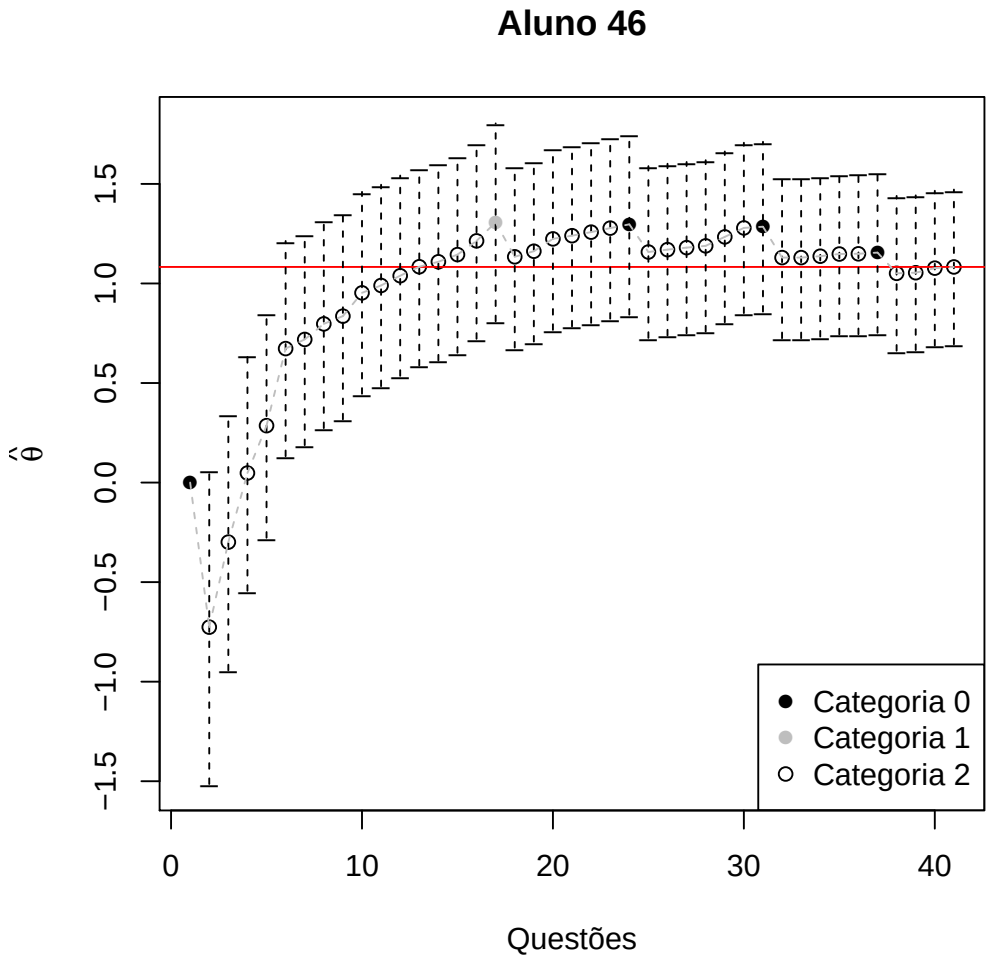

Figura 4.4: *Proficiências estimadas do aluno 46 em cada passo do TAI-PI, a linha vermelha representa a estimativa final após 40 itens.*

Capítulo 5

Estudos de classificação

O objetivo do EPI não é estimar a proficiência dos alunos, mas classificá-los em aprovados ou reprovados. Assim, é preciso que encontrar um ponto de corte, no qual os alunos com proficiências estimadas acima deste ponto sejam classificados em aprovados e abaixo em reprovados. Para isso foram realizados estudos nos resultados da aplicação no Capítulo 4. Pelo fato de quer-se manter a história e estrutura do EPI, o ponto de corte será baseado no $\widehat{\theta}_{40}$. Mas para medir qualidade e eficiência deste ponto os estudos serão realizados no $\widehat{\theta}_{25}$.

Na Figura 5.1 é apresentado a curva ROC para o $\hat{\theta}_{40}$ baseado no critério MPA, a qual é muito utilizada na área da saúde para medir e especificar problemas no desempenho de testes diagnósticos. Neste trabalho a curva ROC será utilizada para obter o ponto de corte da proficiência.

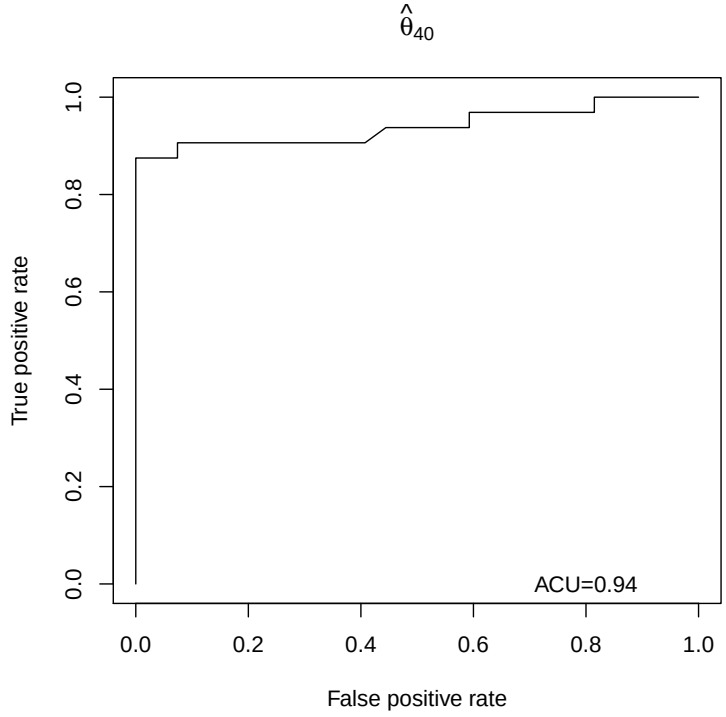

Figura 5.1: *Curva ROC para a proficiência de acordo com a classificação MPA*

O método custo-benefício (McNeil et al. (1975), Metz (1978)) foi utilizado para encontrar o ponto de corte ótimo, o qual é obtido quando a inclinação da curva ROC é dada por

$$S = \frac{1-p}{p} CR = \frac{1-p}{p} \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}} \tag{5.1}$$

onde p é a prevalência de reprovados no teste, C_{FP} o custo do falso positivo, C_{TN} custo do verdadeiro

negativo, C_{FN} custo do falso negativo e C_{TP} custo do verdadeiro positivo. Foi utilizado CR=1, o valor encontrado para o $\widehat{\theta}_{40}$ é -0.40. Na Figura 5.2 são apresentadas as densidades estimadas para a proficiência separadas em aprovados e reprovados, na qual é possível observar pelas caudas sobrepostas que há uma pequena probabilidade de se cometer erros na classificação.

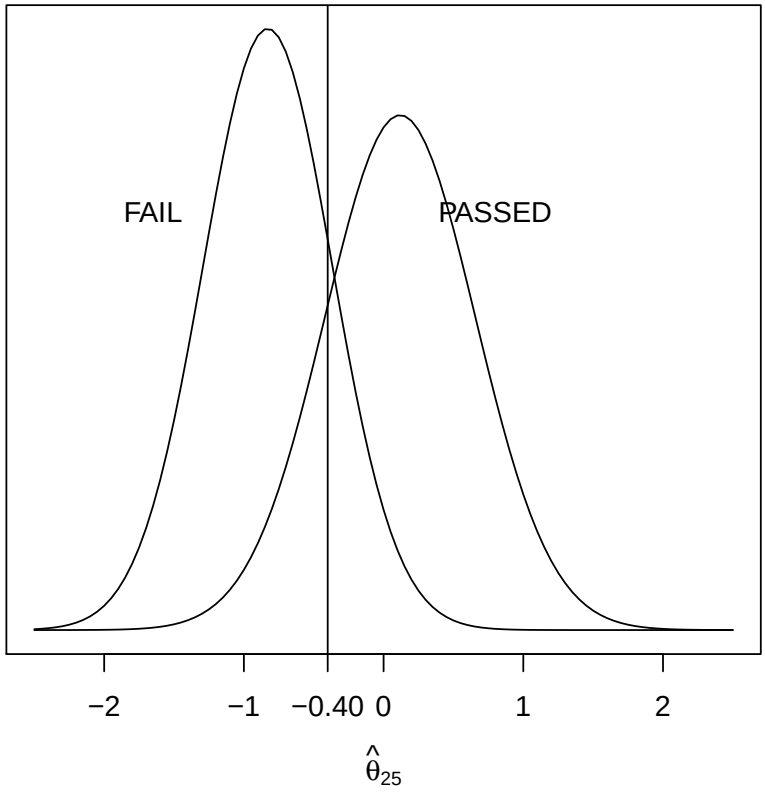

Figura 5.2: *Densidades estimadas para a proficiência por grupo*

Esse erro de classificação pode ser medido por meio do método Acurácia Esperada da Classificação proposto por Rudner (2005). A acurácia mede o quanto o teste consegue obter resultados verdadeiros, ou seja, no nosso caso quando o aluno classificado em aprovado pelo ponto de corte é aprovado pelo critério MPA ou o caso oposto, reprovado pelo ponto corte e reprovado pelo critério MPA. A acurácia esperada é dada por

$$AE = \frac{\sum_{\theta_i<\theta_c} P(\hat{\theta} < \theta_c|\theta_i)}{n} + \frac{\sum_{\theta_i>\theta_c} P(\hat{\theta} > \theta_c|\theta_i)}{n} \tag{5.2}$$

onde $\theta_c = -0.40$. A acurácia esperada encontrada foi de 0,85, ou seja, espera que se cometa um erro de 15% na hora de classificar um aluno, um valor razoávelmente pequeno pelo fato que geralmente o aluno tem uma segunda chance de realizar a prova.

Um estudo de simulação também foi realizado para verificar a eficácia de estimação da proficiência do aluno próximo ao ponto de corte a partir das simulações apresentadas na Seçao 3 e adotando a distribuição a priori para a proficiência do indivídio N(0,1), ou seja, os mesmo métodos utilizados no TAI-PI. Foram simulados testes de 1000 alunos com proficiências adotadas a partir de uma sequência de -1,4 a 0,6 com espaçamento de 0.002, e foi calculada a acúrácia apresentadas na Tabela 5.1. Obteve-se um erro de classificação de aproximadamente 6%, um número pequeno e geralmente aceitável, o que pode evidenciar a validade do ponto de corte adotado e dos métodos

utlizados próximo a este ponto.

Tabela 5.1: *Acurácia para o ponto de corte no estudo de simualção*

	$\theta < -0.4$	$\theta \geq -0.40$
$\hat{\theta} < -0.40$	0.4585	0.024
$\hat{\theta} \geq -0.40$	0.040	0,4755

Capítulo 6

Conclusões

Pode-se observar por meio dos resultados das simulações (baixo vício e erro quadrático médio para 20 ou 30 itens e estimaçao por EAP) e da aplicação real do TAI-PI que o teste adaptativo informatizado desenvolvido é viável para a avaliação da proficiência em língua inglesa dos alunos de pós-graduação do ICMC, além de possibilitar a compilação imediata do resultado do teste. Pelo nível educacional e formação dos examinados, não se esperava encontrar ninguém com dificuldade de realização do teste via computador. Isso se confirmou no dia das aplicações do TAI-PI e, portanto, supõe-se que a mudança do formato papel-e-caneta para informatizado não ocasiona nenhum confundimento relevante para a avaliação da proficiência em inglês. Os resultados obtidos evidenciam que um teste de 25 itens parece ser suficiente para estimar satisfatoriamente o traço latente em questão (Van der Linden & Pashley, 2010). Com os estudos de classificação será possível determinar um ponto de corte para a proficiência do indivídio e em uma próxima aplicação poder usá-lo para classificar o aluno em aprovado ou reprovado. Portanto o TAI-PI poderá ser uma nova forma de avaliação a ser utilizada para os alunos da pós-graduação do ICMC, em que cada aluno terá uma prova individual e que melhor o represente. Como já evidenciado um TAI é uma forma mais eficiente de avaliar os alunos, o que irá acrescentar muito mais agilidade e confiabilidade do teste.

6.1 Sugestões para Pesquisas Futuras

Vários estudos devem ser realizados para o aperfeiçoamento do TAI-PI. Pode-se dividí-los e três aspectos: (i) o aprimoramento do banco de itens e da interpretação da escala, que é um aspecto de trabalho a ser realizado pelo especialista em língua inglesa, (ii) o aperfeiçoamento do *software* propriamente dito, que é tarefa do especialista em computação, e (iii) a validação do TAI-PI, tanto sob o aspecto de correta classificação do indivíduo em aprovado/reprovado, quanto sob o aspecto da metodologia estatística implementada. Este último, tarefa específica da área da estatística.

O aprimoramento do banco, será focado na revisão dos itens que tiveram estimativas dos parâmetros ruins (estimativas de a baixas, por exemplo). Além disso, itens novos serão criados, tanto na forma tradicional, quanto na forma interativa, ou seja, usando do fato do item ser aplicado via computador e viabilizar a exibição de imagem, som e interatividade com o examinado. Para esta tarefa, a pesquisadora Andrea Jéssica Borges Monzón, colaboradora deste projeto de pesquisa, contará com uma aluna bolsista de iniciação científica da área de Letras, especialidade em inglês.

Referências Bibliográficas

Aluísio, S., de Aquino, V., Pizzirani, R., & de Oliveira, O. (2003). Assessing high-order skills with partial knowledge evaluation: Lessons learned from using a computer-based proficiency test of english for academic purposes. *Journal of Information Technology Education: Research*, *2*(1), 185–201. 3

Baker, F. (2001). *The basic of Item Response Theory. EUA: ERIC Clearinghouse on Assessment and Evaluation.* , 2 ed. iii, v, 9

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques.* CRC Press. 9

Bloom, B. (1984). *Taxonomy of Educational Objectives.* New York, Longman. 3

Bridgeman, B., & Cline, F. (2000). Variations in mean response times for questions on the computer-adaptive gre® general test: Implications for fair assessment. *ETS Research Report Series*, *2000*(1), i–29. 2

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229. 9

Conejo, R., Millán, E., Perez-de-la Cruz, J.-L., & Trella, M. (2001). Modelado del alumno: un enfoque bayesiano. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, *5*(12), 50–58. 11

de Andrade, D. F., Tavares, H. R., & da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo.* 17

Dehghan, M., Masjed-Jamei, M., & Eslahchi, M. (2006). On numerical improvement of open newton–cotes quadrature rules. *Applied mathematics and computation*, *175*(1), 618–627. 9

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). Development of a scale for assessing the level of computer familiarity of toefl examinees. *ETS Research Report Series*, *1998*(1), i–32. 2

Eignor, D. R. (1993). Deriving comparable scores for computer adaptive and conventional tests: An example using the sat1, 2. *ETS Research Report Series*, *1993*(2), i–16. 2

Goncalves, S. M. O. L. H. M. d. . O. J. O. N., J. P.; Aluísio (2004). A learning environment for english for academic purposes based on adaptive tests and task-based systems. *Intelligent Tutoring Systems: 7th International Conference, ITS 2004, Maceió, Alagoas, Brazil, August 30-September 3, 2004, Proceedings*, *175*(1), 1–11. 4

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). Computer familiarity among toefl examinees. *ETS Research Report Series*, *1998*(1), i–23. 2

Klinger, A. (1997). Experimental validation of learning accomplishment. In *Frontiers in Education Conference, 1997. 27th Annual Conference. Teaching and Learning in an Era of Change. Proceedings.*, vol. 3, (pp. 1367–1372). IEEE. 3

Lans, R. F. (2006). Introduction to sql: mastering the relational database language. 5

McNeil, B. J., Keeler, E., & Adelstein, S. J. (1975). Primer on certain elements of medical decision making. *New England Journal of Medicine*, *293*(5), 211–215. 25

Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, vol. 8, (pp. 283–298). Elsevier. 25

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195. 13

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to logist and bilog. *Applied psychological measurement*, *13*(1), 57–75. 9

Monzón, A. J. B. (2010). Construçao de banco de questoes para exames de proficiência em inglês para programas de pós-graduaçao. 2

Piton-Gonçalves, J., Borges Monzón, A. J., & Aluísio, S. M. (2009). Métodos de avaliação informatizada que tratam o conhecimento parcial do aluno e geram provas individualizadas. In *Anais do Simpósio Brasileiro de Informática na Educação*, vol. 1. 4

Ricarte, T. A. M. (2013). *Teste adaptativo computadorizado nas avaliações educacionais e psicológicas*. Ph.D. thesis, Universidade de São Paulo. 5

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation*, *10*(13), 1–4. 26

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*. iii, v

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.*. American Psychological Association. 2

Segall, D. (1993). Score equating verification analyses of the cat-asvab. *Briefing presented to the Defense Advisory Committee on Military Personnel Testing. Williamsburg, VA, USA*. 2

Shuford Jr, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, *31*(2), 125–145. iii, v

Sukamolson, S. (2002). Computerized test/item banking and computerized adaptive testing for teachers and lecturers. *Information Technology and Universities in Asia–ITUA*. 11

Van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. *Computerized adaptive testing: Theory and practice*, (pp. 27–52). 10, 18

Van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. Springer. 11

Van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing*, (pp. 3–30). Springer. iii, v, 11, 29

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. 2

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score toefl. *Journal of Educational Measurement*, *37*(3), 203–220. 2

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492. 1