TESTES ADAPTATIVOS COMPUTADORIZADOS

Juliana Guimarães LABARRÈRE[1]
Cibele Queiroz DA-SILVA[2]
Denise Reis COSTA[3]

- RESUMO: Os Testes Adaptativos Computadorizados (CAT) são aqueles aplicados, por meio eletrônico, no qual os itens são selecionados de acordo com o examinando que está realizando o teste e, com isso, a proficiência do examinando é mensurada iterativamente. Para que o CAT seja implementado, faz-se necessário a construção de um banco de itens, com qualidade pedagógica e psicométrica, sendo imprescindível o uso de modelos da Teoria de Resposta ao Item (TRI). O modelo de TRI relaciona a proficiência do examinando com a resposta dada ao item, de modo que, quanto maior a proficiência da pessoa avaliada, maior é a probabilidade de acertar o item. Com o auxílio da TRI calibra-se os itens. As estimativas obtidas nesse passo irão alimentar o CAT, procedimento com o qual estima-se as habilidades dos examinandos durante o teste e, torna possível, a comparação das habilidades de diferentes examinandos ao final do processo. Utilizando dados simulados, em conjunto com dados reais, neste trabalho analisam-se os erros envolvidos nas estimativas das habilidades (proficiências) de examinandos avaliados tanto utilizando-se o CAT quanto provas tradicionais.

- PALAVRAS-CHAVE: Teste Adaptativo computadorizado; teoria da resposta ao item; avaliação computadorizada.

1 Introdução

Não se pode falar em atualidade sem que venha à mente o avanço tecnológico e a informatização. Hoje, ao acessar um computador, os recursos são inúmeros.

[1]Universidade de Brasília – UnB, Centro de Seleção e de Promoção de Eventos da UnB (CESPE), CEP: 70910-900, Brasília, DF, Brsil. E-mail: *ju.labarrere@gmail.com*
[2]Universidade de Brasília – UnB, Departamento de Estatística, CEP: 70910-900, Brasília, DF, Brasil. E-mail: *cibeleqs@unb.br, cibeleqs@gmail.com*
[3]Universidade de Campinas – UNICAMP, Departamento de Estatística, CEP: 13083-859, Campinas, SP, Brasil. E-mail: *denise.reis@gmail.com*

Desde ações simples, tais como pesquisar, comprar e pagar contas, até a realização de questões mais burocráticas, como declarar o Imposto de Renda e fazer um boletim de ocorrência. Segue-se a tendência de simplificar as diversas ações do cotidiano pelo meio eletrônico.

Na Educação, esse processo de informatização está cada vez mais desenvolvido. Existem projetos de inclusão digital para escolas públicas, que prometem melhorar o ensino ao dispor o computador em sala de aula. Fato esse que também sugere a aplicação de testes virtuais, podendo-se fazer uso até de recursos multimídia para a elaboração de questões.

Uma forma interessante de informatizar testes se faz por meio da utilização dos testes adaptativos computadorizados (CAT — do inglês *Computerized Adaptive Test*), que são aqueles aplicados em meio eletrônico, no qual os itens são selecionados de acordo com o examinando que está realizando o teste: para alunos de maior proficiência, um teste com itens mais difíceis; para os de menor proficiência, itens mais fáceis serão selecionados. Desse modo, a proficiência do examinando é mensurada iterativamente. O teste é totalmente adaptado ao indivíduo avaliado, podendo até mesmo ser de tamanho diferente para os diferentes alunos.

Para que o CAT seja implementado, faz-se necessário banco de itens com qualidade pedagógica e psicométrica. Uma sugestão para preparar o banco é fazer uso da Teoria de Resposta ao Item, TRI, com a qual se pode inferir a dificuldade e a discriminação[1] de cada item.

Seguem abaixo a descrição de alguns exemplos da aplicação do CAT.
(1) *Epidemiologia/Medicina* - WARE JR. *et al.* (2003) comparam a eficácia do CAT versus da aplicação de questionários convencionais (estáticos) na avaliação do impacto, prognóstico e possível prevenção de episódios dores de cabeça e enxaquecas em diferentes grupos de pessoas nos EEUU. Neste estudo os dados foram obtidos através de questionários aplicados por meio de ligações telefônicas ou pela Internet. O questionário aplicado por telefone, contendo 120 itens, foi elaborado pelo National Survey of Headache Impact (NSHI). Tal questionário é considerado muito longo, e responder ao mesmo pode ser uma tarefa árdua, em especial quando o entrevistado está, no momento, enfrentando um episódio de enxaqueca. O items do referido questionário compuseram o banco de itens utilizados no CAT, que foi administrado via Internet. Os autores concluíram que as sondagens feitas via CAT reduziram drasticamente o número de itens que precisavam ser respondidos pelos entrevistados, sem, no entanto, comprometer a validade da pesquisa. HARTA *et al.* (2006) apresenta um estudo semelhante ao de WARE JR. *et al.* (2003), mas no contexto específico da percepção do estado de saúde, pelo paciente, das boas condições de sua região lombar e ombros.
(2) *Psicologia* - FLIEGE *et al.* (2005) desenvolveram uma aplicação do CAT para medir os sintomas da depressão. Segundo os autores, a depressão é um dos problemas mentais mais prevalentes. A avaliação dos sintomas depressivos, e capacidade de localizar o paciente numa escala de depressa, utilizando-se os

[1]Discriminação é a capacidade do item de diferenciar os examinandos que têm conhecimento sobre o conteúdo avaliado dos que não têm, de distinguir os diferentes níveis de proficiência.

métodos de TRI tem se tornado cada vez mais importante na ciência e prática médica. Na pesquisa de FLIEGE *et al.* (2005), três mil e duzentos pacientes psicossomáticos, ligados ao Hospital Universitário de Berlim, responderam a 11 questionários de saúde mental. Dos 320 items ao todo, com o auxilio de três especialistas, trabalhando independentemente, selecionou-se 144 items para o banco de items. Os autores concluem que, com o auxílio do CAT, é possível apresentar ao paciente um número reduzido de items sem, no entanto, comprometer a precisão das estimativas da medida do grau de depressão apresentado pelo paciente.
(3) *Educação* - TRIANTAFILLOU *et al.* (2008), desenvolveram desenhos amostrais e a implementação do CAT em dispositivos móveis (CAT-MD) para a telefonia celular. Na atualidade, o uso de diferentes produtos móveis, tais como o telefone celular e do Assistente Digital Personalizado (ADP), têm crescido rapidamente. Além disso, a disponibilização de outros tipos de tecnologias móveis de última geração, tais como as de redes sem fio, propiciam o ambiente ideal para o chamado "e-learning". Por exemplo, utilizando-se um aparelho celular conectado à rede sem fio é possível que o usuário possa ser submetido a algum tipo de avaliação ou sondagem.

Este artigo está organizado como a seguir: na Seção 2 introduz-se alguns conceitos sobre o CAT e a necessidade do uso conjunto da técnica de TRI. Nas Seções 3 e 4 apresentam-se, respectivamente, as técnicas de TRI e CAT. Na Seção 5 apresentam-se estudos de simulação e discussões dos resultados.

2 CAT e TRI

Ao corrigir um teste utilizando a análise clássica (a forma comum, mais conhecida), as características dos indivíduos testados e da própria prova se confundem. Dessa forma, um só pode ser analisado e interpretado a partir da análise e interpretação do outro. Na análise clássica, a habilidade do examinando é expressa pelo seu escore simples, ou seja, pela quantidade de itens que foram respondidos corretamente. Nesse contexto, um examinando que realiza um teste de dificuldade alta, isto é, que exige maior conhecimento, pode apresentar um nível de habilidade menor do que se tivesse realizado um teste de dificuldade baixa, que exige menor conhecimento.

Segundo a análise clássica, a "dificuldade" de um item é definida pela proporção de examinandos em um grupo de interesse que responderam corretamente ao item (HAMBLETON *et al.*, 1991). Essa característica depende tanto do teste no qual o item está inserido, quanto dos examinandos que realizaram esse teste. E, como o nível de dificuldade da prova é considerado diferente para os examinandos com distintas habilidades, a comparação entre grupos se torna complicada, já que os escores ficam em escalas desiguais e sem relação funcional entre elas.

É como no caso de dois alunos que realizaram a mesma matéria na faculdade. Ambos passaram com cinco, a nota de corte, porém o professor de um era bem mais rigoroso na elaboração e correção do teste do que o professor do outro. Ambos têm a mesma menção no currículo, porém podem não possuir o mesmo nível de

conhecimento na matéria. Pode-se até dizer que o cinco de um "vale mais" do que o cinco do outro.

Outro fato que deve ser observado é que, ao utilizar a teoria clássica do teste, a mensuração da nota é dada ao nível do teste e não do item. Dessa maneira, não se pode saber como será o desempenho de examinandos em relação a determinado item, o que não facilita na comparação entre grupos.

A TRI é uma família de modelos probabilísticos que visa a inferir, ou explicar, a proficiência de um examinando a partir da probabilidade de ele marcar determinada resposta para uma questão. O modelo descreve os traços latentes[2] do indivíduo, no caso, sua proficiência (ou habilidade), relacionando-os à resposta dada ao item. E, considerando que o conhecimento de uma pessoa não é uma característica que se altera ao passar do tempo, de forma a perder o conhecimento antigo ao adquirir um novo, pode-se dizer que essa habilidade possui um caráter cumulativo. Quanto maior a proficiência da pessoa avaliada, maior é a probabilidade de acertar o item.

Essa teoria surge visando a acabar com a dependência do grupo e do teste. O modelo expressa o teste ao nível do item e proporciona medidas mais precisas sobre os escores dos examinandos, além de não exigir exames idênticos e aplicados ao mesmo tempo, para que haja confiabilidade. A habilidade e os parâmetros do item — dificuldade e discriminação - são ditos invariáveis, pois as estimativas sobre as habilidades, obtidas em diferentes grupos de itens, são as mesmas, e, de forma semelhante, as estimativas sobre os parâmetros do item, obtidas por diferentes grupos de examinandos, também são as mesmas. Dessa forma, examinandos de mesma habilidade têm a mesma probabilidade de responder a um item corretamente.

Vale enfatizar que a habilidade do indivíduo examinado não depende do item. No entanto, para estimar essa habilidade, faz-se necessário que os itens sejam elaborados de forma a captar adequadamente essa habilidade. E, como a habilidade é discutida pelo parâmetro do item, esse também é o mesmo para os diferentes grupos. O problema de comparação entre grupos é solucionado.

A TRI possibilita expressar todos os escores dos examinandos em uma mesma escala, à qual pode ser atribuída uma interpretação prática. Conhecendo-se o nível de proficiência de um indivíduo, pode-se determinar, com precisão, o conjunto de conhecimentos adquiridos e também os que ainda não o foram.

A par das informações expostas, conclui-se que a TRI está presente não somente como um pré-requisito do CAT, na elaboração do banco de itens, mas também para a seleção dos itens no teste adaptativo, uma vez que, com o auxílio de tal técnica, pode-se inferir iterativamente a habilidade do examinando e escolher itens mais próximos a essa habilidade ou proficiência. Além disso, no final do teste, o examinando pode ser informado do valor estimado de sua proficiência em uma escala geral, na população de examinandos.

Para melhor entender o CAT e o contexto em que está inserido, é dado, a seguir, um exemplo de sua aplicação: considere um banco de itens já calibrado, ou

[2]Traços latentes são características que não podem ser observadas diretamente, devendo ser inferidos a partir da observação de outras variáveis.

seja, que já teve os parâmetros dos itens estimados pela TRI. Verificam-se cinco níveis de proficiência do examinando:

0 → não possui conhecimentos em Matemática;

1 → sabe somar e subtrair;

2 → sabe utilizar as quatro operações básicas da Matemática: somar, subtrair, multiplicar e dividir;

3 → domina as quatro operações básicas e sabe trabalhar com frações;

4 → além do descrito pelo nível 3, sabe utilizar números negativos.

Um examinando vai realizar o teste. De início, supõe-se que ele não tenha conhecimentos em Matemática. O nível de proficiência do examinando é zero. O computador seleciona aleatoriamente um item, digamos, uma operação de multiplicação, e o aluno acerta. A estimativa da proficiência é atualizada para 2. Com isso, supõe-se que o aluno também saiba somar e subtrair. Então, é selecionado outro item para verificar se o nível de proficiência do aluno é realmente 2 ou superior. É selecionado um item de fração, e o aluno também acerta. Seu nível de proficiência é novamente atualizado, passando a ser estimado como 3. Mais um item é selecionado, agora, utilizando números negativos, porém o aluno erra, e seu nível de proficiência estimado permanece em 3. Assim, o teste prossegue, adaptando-se ao conhecimento do aluno e convergindo para uma estimativa mais real da proficiência, até que um critério de parada pré-estabelecido seja alcançado, que pode ser o número de itens do teste ou um erro de mensuração da proficiência, entre outros.

Há diversas vantagens em aplicar um teste em versão CAT no lugar de um na forma tradicional (papel-e-caneta). São elas:

1. **Ecologicamente correto:** Um teste em versão CAT dispensa a impressão das provas. Portanto, economiza-se papel e tinta, contribuindo com a sustentabilidade do planeta.

2. **Redução do teste:** Em um teste tradicional, são apresentados todos os itens para todos os examinandos. Já na versão CAT, não. Como o teste é adaptativo, a tendência é que sejam selecionados itens compatíveis com o nível de habilidade do examinando. Por exemplo, no exame de Matemática previsto acima, o examinando não precisa responder a todos os itens de somar e subtrair, considerados fáceis para o aluno, nem a todos os itens de números negativos, considerados difíceis. Se outro aluno acerta todos os itens correspondentes ao nível 4, supõe-se que o nível de proficiência dele é 4. Assim sendo, ele não precisa responder às demais questões, já que itens de nível 1 são irrelevantes para um aluno de proficiência 4.

3. **Flexibilidade dos itens:** Com um teste aplicado eletronicamente, há possibilidade de se utilizarem itens em formato multimídia, como um vídeo que ajude na compreensão do item.

4. **Flexibilidade da aplicação das provas:** Pode-se aplicar o teste em versão CAT em diferentes dias e horários. As instruções podem ser passadas virtualmente e, como os testes são distintos para os diferentes examinandos, diminui o risco de cópia entre os alunos.

5. **Segurança do teste:** Da mesma forma que diminui o risco de cola, diminui qualquer risco de que uma pessoa que, por acaso, tenha visto o banco de itens beneficie algum examinando. Claro que se supõe um banco de itens suficientemente grande. Ainda existe possibilidade de criptografar os dados armazenados no computador, dando mais um ponto à segurança do CAT.

6. **Rapidez e precisão na correção:** O CAT dispensa corretores para os testes, o que demanda tempo, e também dispensa transcrições e leituras ópticas, que podem acarretar erros. O resultado pode sair logo em seguida à conclusão da prova.

7. **Enriquecimento do banco de dados:** Além de saber se o item foi marcado corretamente, há como saber quanto tempo o examinando gastou para responder o mesmo.

8. **Precisão das estimativas:** Em um teste tradicional, pode ocorrer que muitos itens sejam elaborados de modo a cobrir apenas uma faixa da escala de proficiência. Dessa forma, o teste torna-se mais adequado aos examinandos daquela faixa de proficiência e não distingue os demais. Por exemplo: se no teste de Matemática só houvesse questões de frações, não seria possível identificar os alunos que sabem apenas somar e subtrair dos que sabem fazer uso das quatro operações, e tampouco seria possível saber quais desses alunos vão além e sabem trabalhar com números negativos.

Mesmo com todos esses pontos positivos, não se pode deixar de mencionar que o CAT também possui desvantagens. O banco de dados para a elaboração da prova deve ser consideravelmente grande e deve sofrer atualizações constantemente, o que demanda recursos humanos e financeiros, encarecendo a informatização dos testes.

Outra barreira, porém temporária, é a fase em que os estudos sobre esse assunto encontram-se no Brasil. Observa-se o déficit de profissionais de estatística capacitados para trabalhar com o CAT. Há pouquíssimos trabalhos sobre o tema e, dentre esses, há os que são focados na parte computacional, deixando a desejar na teoria estatística. As aplicações no Brasil estão começando somente neste ano, sendo uma delas a prova de proficiência em Inglês Instrumental 1 da Universidade de Brasília, promovida pelo CESPE — Centro de Seleção e Promoção de Eventos.

No exterior, além de livros e outras publicações, já existem alguns testes em que o método CAT é adotado. Um caso importante a ser mencionado é o TOEFL (*Test of English as a Foreign Language*), no qual estudantes cujo idioma nativo não

é Inglês têm o seu conhecimento, ou proficiência, na língua Inglesa mensurado por meio de tal exame. O TOEFL é um teste obrigatório para o ingresso em programas de pós-graduação em países de língua Inglesa.

Ao combinar o CAT e a TRI na avaliação do ensino público e privado no Brasil, pode-se dizer que os meios de avaliação estão tornando-se cada vez mais informatizados. Utilizando dados simulados, em conjunto com dados reais, neste trabalho analisa-se os erros envolvidos nas estimativas das habilidades (proficiência) de examinandos avaliados tanto utilizando-se o CAT quanto provas tradicionais.

3 Teoria da Resposta ao Item

Nesta seção, o primeiro passo para o CAT será discutido: a calibração do banco de dados. O modelo matemático utilizado, seus parâmetros e a melhor forma de estimá-los. Para as fórmulas, seguiu-se notação semelhante a de Andrade *et al.* (2000).

3.1 O modelo e seus parâmetros

Os modelos matemáticos expressos pela teoria de resposta ao item, TRI, especificam que a probabilidade de um examinando responder a um item corretamente depende tanto de sua habilidade, quanto das características do item. Esses modelos levam em consideração a natureza do item, se ele é ou não dicotômico; o número de populações envolvidas; e a quantidade de traços latentes que serão mensurados.

Neste trabalho serão estudados apenas casos em que os itens são dicotômicos, corrigidos exclusivamente como certos ou errados, somente uma população envolvida e um traço latente ou habilidade a ser estimada. Há vários modelos que podem ser utilizados na TRI. Um dos mais aplicados é o modelo logístico unidimensional para dados dicotômicos.

A probabilidade de um examinando j responder corretamente ao item i, condicionado a seu traço latente θ_j, é dada por:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \quad (1)$$

com i identificando o item e variando entre 1 e I, e j identificando o examinando e variando entre 1 e n. Nota-se que:

- U_{ij} é a variável dicotômica, que assume valor igual a 1 quando o examinando j responde corretamente ao item i, ou 0, em caso contrário;
- θ_j representa a habilidade (proficiência ou traço latente) do *j-ésimo* indivíduo;
- c_i é o parâmetro do item que representa a probabilidade de acerto casual ao item i, isto é, a probabilidade de que alunos de baixa proficiência respondam corretamente ao item i;

- D é um fator escala constante, usualmente igual a 1. Porém, se igualado a 1,7, o modelo fornecerá resultados análogos ao da função ogiva normal;
- a_i é o parâmetro que representa a discriminação do item i; e
- b_i é o parâmetro que representa a dificuldade, ou posição, do item i, medido na mesma escala da habilidade.

O modelo descrito possui três parâmetros para descrever o item — dificuldade, discriminação e acerto casual — e é conhecido por modelo logístico de três parâmetros. Tal modelo expressa uma relação não-linear, porém estritamente crescente entre a probabilidade de acerto a um item e os seus parâmetros. Essa relação pode ser observada pelas curvas características do item — CCI, grafadas abaixo (Figura 1 (a)) As curvas mostram a relação citada sobre um item respondido corretamente: quanto maior a habilidade do examinando, maior a probabilidade de acerto do item. Tal relação é estritamente crescente, porém, logística.

(a) (b)

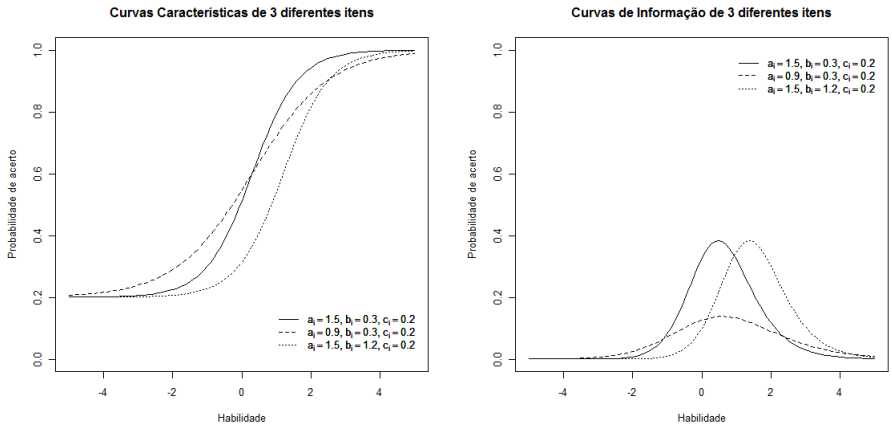

Figura 1 - (a) Curva Característica de 3 diferentes itens. (b) Curva de Informação de 3 diferentes itens.

O parâmetro c determina o deslocamento vertical da CCI e representa o acerto casual, isto é, a probabilidade de um indivíduo de baixa habilidade acertar o item. Como se trata de uma probabilidade, assume valores entre 0 e 1. No caso da curva acima, a probabilidade de acerto casual tende a 0,2 à medida que a habilidade do examinando diminui, aproximando-se de $-\infty$.

O parâmetro b é medido na mesma escala da habilidade estimada. Essa escala, que pode variar de $-\infty$ a $+\infty$, é elaborada subjetivamente e os valores numéricos que esse parâmetro assume não têm grande importância, pois o que realmente importa é a ordem dos pontos da escala e a diferença entre a habilidade estimada do examinando e o parâmetro b. Se essa diferença for positiva, interpreta-se que o item

i de dificuldade b_i antecede a habilidade θ, na escala de habilidades. Dessa forma, examinandos que acertem o item têm sua habilidade estimada como sendo superior à habilidade dada como conhecida no cálculo da probabilidade. O parâmetro b representa a proficiência necessária para que a probabilidade de acerto do item seja igual a $(1 + c)/2$. É o parâmetro de dificuldade, no qual, quanto maior seu valor, mais difícil é o item.

O parâmetro a, por sua vez, proporciona a inclinação da CCI no ponto b_i. Isto é, descreve o quão íngreme é o "S" da curva característica. É esperado que esse parâmetro assuma sempre valores positivos, uma vez que a probabilidade de acertar determinado item aumenta com o aumento da habilidade dos indivíduos. Quanto mais alto o valor assumido pelo parâmetro a, mais restrita é a faixa de habilidades dos alunos que têm aproximadamente a mesma probabilidade de acertar determinado item, ou seja, mais discriminativo é o item.

Outra forma de analisar o item é por sua 'função de informação', com a qual é possível analisar o quanto um item tem de informação para que a habilidade seja mensurada. A função de informação de Fisher do item é dada por $I_i(\theta) = \frac{\left[\frac{d}{d\theta}P_i(\theta)\right]^2}{P_i(\theta)Q_i(\theta)}$, em que $P_i(\theta) = P(U_{ij} = 1|\theta_j)$ e $Q_i(\theta) = 1 - P_i(\theta)$. Substituindo a probabilidade $P_i(\theta)$ pela determinada pelo modelo logístico de 3 parâmetros (equação 1), temos:

$$I(\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i}\right]^2. \qquad (2)$$

Pode-se extrair da função acima que a informação do item é maior quando a sua dificuldade se aproxima da habilidade do examinando, ou quanto maior for a discriminação do item, ou ainda quanto menor for a probabilidade de acerto casual. A informação do teste é dada pela soma das informações dos itens que compõem o teste, isto é, $I(\theta) = \sum_{i=1}^{I} I_i(\theta)$.

Podem-se analisar os parâmetros de três itens distintos ao representar as curvas características e de informação de cada um, como segue:

Considere o item 1 como sendo o de linha sólida, o item 2 como sendo o de linha tracejada e o 3, pontilhada. Com base na figura 1 (a) e (b), pode-se verificar que o item 1 é mais discriminativo que o item 2, porém, da mesma discriminação, se comparado ao item 3. Isso deve-se pela acentuação do "S" da curva característica dos itens. Pode-se perceber que tanto o item 1, quanto o item 3, possuem a mesma inclinação, porém, o "S" relativo ao item 2 é menos acentuado, ou seja, a curva demora mais para crescer. Com o item 2, demora-se mais para estimar as habilidades dos examinandos, já que uma faixa maior de alunos com distintas habilidades possui probabilidades próximas de acerto. Esse parâmetro é observado na curva de informação do item pela altura e largura que possui. No caso, a curva do item 2 é a mais achatada, e as demais possuem alturas e larguras idênticas. Diferentemente, ao analisar a dificuldade, verifica-se que os itens 1 e 2 possuem o mesmo nível nesse quesito, porém são mais fáceis que o item 3. Isso é percebido pela curva característica do item 3, que é mais baixa que as demais, isto é, mesmo com habilidades mais elevadas, a probabilidade de marcar o item 3 corretamente

é inferior. Na curva de informação, por sua vez, o item 3 está mais à direita em relação aos outros, indicando sua maior dificuldade. Os três itens possuem a mesma probabilidade de acerto casual.

3.2 Estimação dos parâmetros da TRI

Uma das questões mais importantes da TRI é a estimação dos parâmetros. Afinal, visamos ao cálculo da probabilidade de um indivíduo acertar determinado item, e essa probabilidade é devida aos parâmetros dos itens e à habilidade do indivíduo em questão, informações essas que usualmente desconhecemos. A única informação conhecida é a quantidade de itens marcados como corretos, de cada examinando, e quais são esse itens. A precisão do cálculo da probabilidade em questão depende da precisão obtida na estimação dos parâmetros.

Uma solução para a estimação dos parâmetros, considerando que nenhum deles é conhecido, é a estimação em duas etapas, na qual se supõe uma distribuição latente, associada às habilidades. Dessa forma, podem-se estimar, primeiramente, os parâmetros dos itens por meio do 'método de máxima verossimilhança marginal'. Nesse caso, considera-se uma função de probabilidade condicionada aos parâmetros de habilidade observados na população de interesse, aplica-se a função de verossimilhança marginal obtida, integrando-se em θ e obtendo-se o máximo da função. Posteriormente, estima-se a habilidade de cada examinando, uma a uma, por máxima verossimilhança, pela moda ou média da distribuição condicional suposta para θ, dados os parâmetros estimados dos itens e as respostas, corretas ou não, dadas a cada item respondido.

Para que essa técnica possa ser aplicada, deve-se supor independência entre os itens, condicionada à habilidade, visto que a estimação é feita por métodos numéricos que dependem das derivadas segundas da log-verossimilhança com relação aos parâmetros dos itens.

O método de estimação dos itens via máxima verossimilhança marginal foi proposto por BOCK e LIEBERMAN (1970). Nele se assume que os examinandos no estudo são elementos de uma amostra aleatória, na qual a população de onde foram extraídos possui características em relação às habilidades segundo uma função de densidade $g(\theta)$. O artifício proporciona a obtenção de uma verossimilhança que não depende das habilidades desconhecidas.

Seja $G_n(.)$ a função distribuição empírica relativa às proficiências de uma amostra de n indivíduos. Para n for suficientemente grande, $G_n(\theta)$ pode ser aproximada por uma distribuição contínua. Assim, $g(\theta)$ pode ser considerada como a função de densidade para θ. Vale ressaltar que aplicar uma distribuição para θ não consiste em fazer uso de inferência bayesiana.

Para que a verossimilhança independa das habilidades, ela tem sua escala fixada e é marginalizada, integrando-se em relação à distribuição das habilidades. Dessa forma, verifica-se como sendo a probabilidade marginal de $U_{j.}$ a expressão

$$P(U_{j.} = u_{j.} \mid \zeta, \eta) = \int_{\mathbb{R}} P(U_{j.} = u_{j.} \mid \zeta, \eta, \theta) g(\theta \mid \eta) d\theta,$$

na qual ζ representa o vetor de parâmetros dos itens ($\zeta_i = (a_i, b_i, c_i)$), e η, o conjunto de parâmetros conhecidos e finitos associados à população Π. $g(\theta \mid \eta)$ é a função de densidade de probabilidade de θ, que possui distribuição contínua. Supõe-se que θ tem distribuição Normal, logo $\eta = (\mu, \sigma^2)$, sendo μ a média, e σ^2 a variância da população em questão. Assumindo-se a independência entre as respostas de diferentes examinandos, pressuposto citado anteriormente, observa-se que:

$$P(U_{..} = u_{..} | \zeta, \eta) = \prod_{j=1}^{n} P(U_{j.} = u_{j.} | \zeta, \eta).$$

Considerando r_l como sendo o número de ocorrências dos diferentes padrões de resposta l e $s \leq min(n, S)$, em que $S = 2^I$ representa a quantidade total de possíveis respostas, como sendo o número de padrões de resposta com $r_l > 0$, nota-se que $\sum_{l=1}^{s} r_l = n$. E, considerando-se a suposição de independência entre respostas, verifica-se uma distribuição Multinomial como segue:

$$L(\zeta, \eta) = \frac{n!}{\prod_{l=1}^{s} r_l!} \prod_{l=1}^{s} [P(U_{l.} = u_{l.} | \zeta, \eta)]^{r_l}.$$

Onze anos depois, Bock e Aitkin (1981) propuseram uma modificação na abordagem de Bock e Lieberman (1970). Tal modificação é baseada na suposição de que diferentes itens são independentes entre si, isto é, a resposta dada a um determinado item independe do que foi respondido a qualquer outro item do teste. Em outras palavras, não se pode ter um item cuja resposta só possa ser obtida a partir da resposta dada a outro item. Assim, os itens podem ser estimados um a um.

Para a obtenção das estimativas de máxima verossimilhança dos parâmetros, é sugerida a utilização de um processo iterativo, como a aplicação do algoritmo EM, no qual cada iteração do processo é feita em dois passos: E - Esperança e M - Maximização. O algoritmo é naturalmente aplicado à TRI, pois visa à estimativa, por meio da máxima verossimilhança, de parâmetros de modelos de probabilidade na presença de variáveis latentes.

Suponha que a distribuição das habilidades seja discretizada, de modo que as habilidades sejam expressas por $\overline{\theta}_k$, com $k = 1, \cdots, q$ e tenham probabilidades $\pi_k, k = 1, \cdots, q$. Considere f_{ki} o número de examinandos de habilidade $\overline{\theta}_k$ que responderam ao item i e r_{ki} o número de examinandos com habilidade $\overline{\theta}_k$ que responderam corretamente ao item i. Assim, observa-se $f_i = (f_{1i}, \cdots, f_{qi})^{'}$, com $\sum_{k=1}^{q} f_{ki} = N$ e $f = (f_1, \cdots, f_I)$. Da mesma forma, $r_i = (r_{1i}, \cdots, r_{qi})^{'}$, com $r = (r_1, \cdots, r_I)$. Sendo que f_{ki} e r_{ki} podem ser tratados como quantidades não observadas.

Se são selecionados n indivíduos da população para responderem ao item i, a probabilidade conjunta de que os f_{ki} examinandos tenham habilidades iguais a $\overline{\theta}_k$ é dada pela distribuição multinomial

$$P(f_i|\pi) = \frac{n!}{\prod_{k=1}^{q} f_{ki}!} \prod_{k=1}^{q} \pi_k^{f_{ki}}, i = 1, \cdots, I.$$

A probabilidade de ocorrerem r_{ki} acertos no item i, dado que houve f_{ki} examinandos de habilidade $\overline{\theta}_k$ respondendo ao item, segue distribuição Binomial:

$$P(r_{ki}|f_{ki}, \overline{\theta}_k) = \binom{f_{ki}}{r_{ki}} P_{ki}^{r_{ki}} Q_{ki}^{f_{ki}-r_{ki}},$$

em que P_{ki} é a função adotada com θ_j e substituída por $\overline{\theta}_k$. A probabilidade conjunta de f e r dados $\overline{\theta} = (\overline{\theta}_1, \cdots, \overline{\theta}_q)$ e π é dada por:

$$\begin{aligned} P(f, r|\overline{\theta}, \pi) &= P(f|\overline{\theta}, \pi) P(r, f, \overline{\theta}, \pi) \\ &= P(f|\pi) P(r|f, \overline{\theta}) \\ &= P(f|\pi) \left[\prod_{i=1}^{I} \prod_{k=1}^{q} P(r_{ki}|f_{ki}, \overline{\theta}_k) \right]. \end{aligned}$$

Assim,

$$\begin{aligned} \log L(\zeta) &= \log \left(P(f, r|\overline{\theta}, \pi) \right) \\ &= \log \left(P(f|\pi) \right) + \sum_{i=1}^{I} \sum_{k=1}^{q} \log P(r_{ki}|f_{ki}, \overline{\theta}_k) \\ &= \log \left(P(f|\pi) \right) + \sum_{i=1}^{I} \sum_{k=1}^{q} \left[\log \binom{f_{ki}}{r_{ki}} + r_{ki} \log P_{ki} + (f_{ki} - r_{ki}) \log Q_{ki} \right] \\ &= C + \sum_{i=1}^{I} \sum_{k=1}^{q} [r_{ki} \log P_{ki} + (f_{ki} - r_{ki}) \log Q_{ki}]. \end{aligned}$$

C é constante em relação a ζ e (f, r) são não-observáveis. Porém, usando a notação $\overline{r}_{ki} = E(r_{ki}|u_{..}, \zeta)$, $\overline{f}_{ki} = E(f_{ki}|u_{..}, \zeta)$ e $\overline{C} = E(C|u_{..}, \zeta)$, pode-se obter a esperança da log-verossimilhança condicionada a $u_{..}$ e ζ:

$$E[\log L(\zeta)] = \overline{C} + \sum_{i=1}^{I} \sum_{k=1}^{q} [\overline{r}_{ki} \log P_{ki} + (\overline{f}_{ki} - \overline{r}_{ki}) \log Q_{ki}]. \tag{3}$$

Aplicados na TRI, os passos do algoritmo EM são:

- **Passo E:** Maximizar $E[\log L(\zeta)]$ em relação a ζ_i. Usar o método de quadratura, que será explicado em seguida, para obter $\overline{\theta}_k$ e estimativas iniciais dos parâmetros dos itens. Dessa forma, pode-se obter $g_j^*(\overline{\theta}_k)$ e, assim, $\overline{r}_{ki}$ e $\overline{f}_{ki}$.

- **Passo M:** Maximizar a função do passo E. Com r e f obtidos pelo passo anterior, estimar ζ_i, pelo algoritmo de Newton-Raphson ou "Scoring" de Fisher, que também serão explicados em seguida.

Quando acontece de um indivíduo responder a todos os itens corretamente ou incorretamente, ocorre um problema na estimação por máxima verossimilhança (pois o máximo da função de log-verossimilhança vai para $+\infty$ ou $-\infty$, respectivamente), e os parâmetros não podem ser maximizados. Da mesma forma, podem ser obtidos parâmetros fora dos limites, como discriminação negativa ou probabilidade de acerto casual fora do intervalo $[0, 1]$. Para que esses problemas não ocorram, uma metodologia bayesiana é proposta para a estimação dos parâmetros. A estimação bayesiana também é feita em duas etapas, sendo a primeira delas a 'estimação bayesiana marginal', que trata de uma extensão da 'máxima verossimilhança marginal'. Distribuições a priori são estabelecidas para os parâmetros e, a partir delas, uma distribuição a posteriori é obtida de forma a possibilitar, com base em alguma característica dessa distribuição, a estimativa dos parâmetros dos itens.

Considere que a distribuição da habilidade θ é função de um vetor de parâmetros η, cuja densidade pode ser expressa por $g(\theta|\eta)$, e que a distribuição do parâmetro dos itens ζ_i é função de um vetor de parâmetros τ, com densidade $g(\zeta|\tau)$. Seja $f(\tau)$ e $g(\eta)$ as distribuições a priori estabelecidas para τ e η, respectivamente. Dessa forma, a densidade a priori conjunta dos parâmetros é esta:

$$\begin{aligned} f(\theta, \zeta, \eta, \tau) &= f(\zeta|\tau)g(\theta|\eta)f(\tau)g(\eta) \\ &= \left[\prod_{i=1}^{I} f(\zeta_i|\tau)\right]\left[\prod_{i=1}^{I} g(\theta_j|\eta)\right] f(\tau)g(\eta). \end{aligned}$$

A distribuição *a posteriori*, obtida pelo Teorema de Bayes, é dada por:

$$f(\theta, \zeta, \eta, \tau|u_{..}) \propto L(u_{..}; \theta, \zeta)f(\zeta|\eta)g(\theta \mid \eta)f(\tau)g(\eta).$$

3.2.1 Estimação dos parâmetros dos itens

Da mesma forma como na abordagem clássica, será utilizada a probabildade marginal, porém, desta vez, será marginalizada a posteriori, integrando-se em relação a θ e τ e obtendo a distribuição a posteriori de ζ e η. Será utilizada uma posteriori apenas em função dos parâmetros de interesse.

$$
\begin{aligned}
f^*(\zeta,\eta|u_{..}) &\propto \int\int P(u_{..};\theta,\zeta)f(\zeta|\tau)g(\theta|\tau)f(\tau)g(\eta)d\theta d\tau \\
&\propto g(\eta)\left[\int f(\zeta|\tau)f(\tau)d\tau\right]\left[\int P(u_{..};\theta,\zeta)g(\theta|\tau)d\theta\right] \\
&\propto L(\zeta,\eta)f(\zeta)g(\eta),
\end{aligned}
$$

em que $L(\zeta,\eta) \equiv P(u_{..};\theta,\eta)$ e $f(\zeta) = \int f(\zeta|\tau)f(\tau)d\tau$.

Será utilizada a moda como sendo a característica de $f^*(\zeta,\eta|u_{..})$, que ajudará a estimar os parâmetros dos itens. Isto é, a 'moda a posteriori' será o estimador de ζ. A estimação dos parâmetros do item i é dada pela equação

$$
\frac{\partial f^*(\zeta,\eta|u..)}{\partial\zeta_i} = \frac{\partial \log L(\zeta,\eta)}{\partial\zeta_i} + \frac{\partial \log f(\zeta)}{\partial\zeta_i} = 0. \tag{4}
$$

Distribuição a priori para a_i

É esperado que o parâmetro de discriminação do item assuma valores positivos, já que ele proporciona a inclinação da CCI e esta é estritamente crescente. Assim, pode-se modelar o parâmetro a_i por uma distribuição cujos valores de x sejam maiores (ou iguais) a zero. Nesse caso, será utilizada a distribuição Log-normal.

Distribuição a priori para b_i

O parâmetro de dificuldade é mensurado na mesma escala da proficiência, portanto, assume qualquer valor pertencente ao conjunto dos reais. Uma boa escolha de *priori* para o parâmetro b_i é a distribuição com parâmetros $\tau = (\mu_b, \sigma_b^2)$.

Distribuição a priori para c_i

O parâmetro c_i representa uma probabilidade, logo é definido no intervalo [0,1]. Portanto, será assumida uma *priori* de distribuição Beta com parâmetros $\alpha - 1$ e $\beta - 1$. Dessa forma, a segunda parcela de (4) é expressa por

$$
\frac{\partial \log f(a_i|\mu_a,\sigma_a^2)}{\partial a_i} = -\frac{1}{a_i}\left(1 + \frac{\log a_i - \mu_a}{\sigma_a^2}\right). \tag{5}
$$

$$
\frac{\partial \log f(b_i|\mu_b,\sigma_b^2)}{\partial b_i} = -\frac{(b_i - \mu_i)}{\sigma_b^2}. \tag{6}
$$

$$
\frac{\partial \log f(c_i|\alpha,\beta)}{\partial c_i} = \frac{\alpha - 2}{c_i} - \frac{\beta - 2}{1 - c_i}. \tag{7}
$$

Com base na equação (4), e nas componentes obtidas em (5), (6) e (7), obtem-se as equações para estimar cada parâmetro dos itens. Os estimadores dos parâmetros do item não possuem soluções explícitas.

3.2.2 Estimação das habilidades

Segue a segunda etapa da estimação dos parâmetros: a estimação das habilidades, considerando os parâmetros dos itens conhecidos. Pela suposição de

independência das habilidades dos diferentes examinandos, esses parâmetros são estimados um a um para cada indivíduo.

Será assumido para θ_j uma priori de distribuição Normal com parâmetros $\eta = (\mu, \sigma^2)$. Desse modo, a posteriori é dada por:

$$g_j^*(\theta_j) \equiv g(\theta_j|u_{j.}, \zeta, \eta) \propto P(u_{j.}|\theta_j, \zeta)g(\theta_j|\eta).$$

Algumas características da posteriori, como a média ou a moda, são tomadas para estimar θ_j. Neste trabalho utilizamos o método de estimação pela média da posteriori, EAP, para estimar tal parâmetro. Considere a distribuição a posteriori de θ_j dada por

$$g(\theta_j|u_{j.}, \zeta, \eta) = \frac{P(u_{j.} \mid \theta_j, \eta, \zeta)g(\theta_j|\eta)}{P(u_{j.}|\zeta, \eta)},$$

a esperança a posteriori de θ_j é dada por

$$\hat{\theta}_j \equiv E[\theta_j|u_{j.}, \zeta, \eta] = \frac{\int_{\mathbb{R}} \theta_j g(\theta_j|\eta)P(u_{j.}|\theta_j, \zeta)\partial\theta_j}{\int_{\mathbb{R}} g(\theta_j|\eta)P(u_{j.}|\theta_j, \zeta)\partial\theta_j}.$$

Porém, considerando o método de quadratura (vide GRAY, 2001) o estimador pela média da posteriori, EAP, de θ_j é dado pela expressão

$$\begin{aligned}\hat{\theta}_j &= \frac{\int_{\mathbb{R}} \theta_j L(\theta_j; u_1, \cdots, u_{k-1})g(\theta_j)d\theta_j}{\int_{\mathbb{R}} L(\theta_j; u_1, \cdots, u_{k-1})g(\theta_j)d\theta_j} \\ &\approx \frac{\sum_{t=1}^{q} \overline{\theta}_t L(\overline{\theta}_t; u_1, \cdots, u_{k-1})A_t}{\sum_{t=1}^{q} L(\overline{\theta}_t; u_1, \cdots, u_{k-1})A_t},\end{aligned} \tag{8}$$

em que $\hat{\theta}_t$ representa os pontos de quadratura e A_t, o peso associado a $\hat{\theta}_t$ (vide GRAY, 2001). A variância a posteriori associada ao EAP é dada por

$$\begin{aligned}Var[\theta_j|u_1, \cdots, u_{k-1}] &= \frac{\int_{\mathbb{R}} [\theta_j - \hat{\theta}_j]^2 L(\theta_j|u_1, \cdots, u_{k-1})g(\theta_j)d\theta_j}{\int_{\mathbb{R}} L(\theta_j|u_1, \cdots, u_{k-1})g(\theta_j)d\theta_j} \\ &\approx \frac{\sum_{t=1}^{q} [\overline{\theta}_t - \hat{\theta}_j]^2 L(\overline{\theta}_t; u_1, \cdots, u_{k-1})A_t}{\sum_{t=1}^{q} L(\overline{\theta}_t; u_1, \cdots, u_{k-1})A_t}.\end{aligned} \tag{9}$$

Não ser necessário métodos iterativos para o cálculo da EAP é sua grande vantagem.

4 CAT - seleção adaptativa de itens

Em um teste adaptativo computadorizado, são apresentados itens distintos para cada examinando, selecionados de modo a serem os mais apropriados possíveis para estimar cada proficiência. Nesta seção detalharemos o CAT a fim de diferenciá-lo do método de teste tradicional, tipo papel-e-caneta. Para tanto, descreveremos um algoritmo de seleção dos itens de uma prova.

Com o banco de dados suficientemente grande e todos os parâmetros dos itens já estimados via TRI, pode-se aplicar algum algoritmo para selecionar itens, personalizando a prova, de cada examinando, para, por fim, estimar as habilidades. A importância da TRI se evidencia porque, mesmo que cada indivíduo realize uma prova diferente, há possibilidade de comparação entre eles. Para a seleção, assume-se que os parâmetros dos itens sejam conhecidos e considerados verdadeiros, o que não implica grandes consequências, desde que número de respostas aos itens seja grande o bastante para a calibração do banco.

O CAT pode ser esquematizado por uma estrutura cíclica, que se inicia com uma primeira estimativa da proficiência do examinando. Um item é selecionado e apresentado, o examinando responde e é avaliado, tendo sua marcação para o item dada como correta ou incorreta. A estimativa inicial é recalculada, considerando a resposta dada. Um novo item é selecionado e apresentado. O ciclo prossegue até que um critério de parada seja acionado, e, tendo sido satisfeito, o teste é finalizado, e o examinando é avaliado pela última proficiência estimada. Esse esquema pode ser observado na Figura 2.

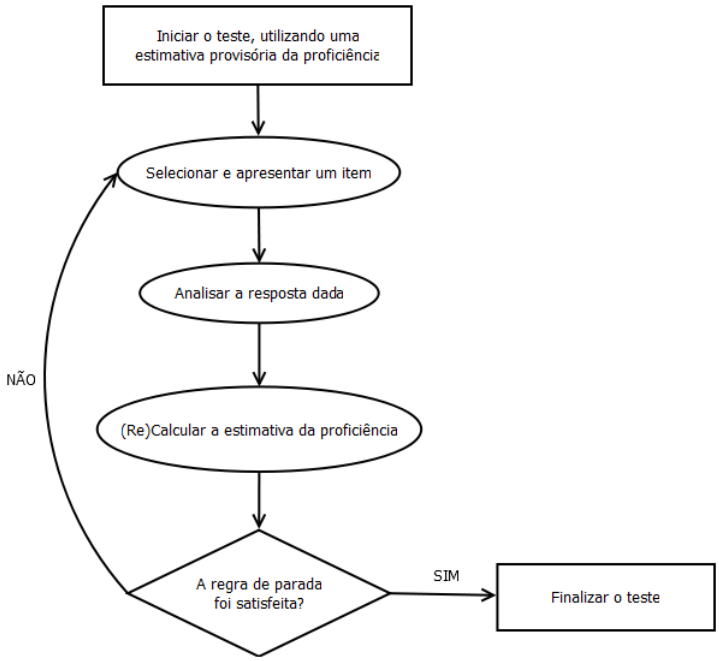

Figura 2 - Diagrama do CAT.

A primeira estimativa da habilidade, para se dar início ao cíclo supracitado, não acarreta erros no resultado. Assim, a escolha inicial fica por conta do pesquisador. Diferentes escolhas implicarão em quantidades distintas de questões para que o critério de parada seja acionado.

Um ponto importante a ser mencionado é que, em um teste adaptativo, se o critério de parada utilizado for o erro da estimativa da habilidade, a precisão da estimativa será a mesma para todos os examinandos. Portanto, se os examinandos forem classificados, posicionando-os na escala de habilidades, os erros cometidos nessa classificação terão distribuição uniforme.

Os métodos de seleção dos itens apresentados a cada examinando dependem tanto dos parâmetros dos itens, dados como conhecidos, quanto dos valores iniciais e atualizados das proficiências. Com isso, pode-se determinar uma medida de informação e, assim, avaliar qual item contribuirá da melhor forma para estimar a habilidade do indivíduo. Vale ressaltar que é desejável que um item seja discriminativo, de forma a minimizar a faixa de habilidades plausíveis para o examinando que o acerte, ou erre. Também se aspira a itens de dificuldade mediana ao examinando, pois itens muito difíceis ou muito fáceis não são de grande eficácia para detectar com precisão o conhecimento alcançado. COSTA (2009) discute alguns critérios para a seleção adaptativa dos itens. O quadro-resumo ilustrado pela Tabela 1 sumariza algumas características dos três métodos de seleção adaptativa de itens revisados. O critério da Máxima Informação foi proposto por LORD (1980), CHANG e YING (1996) sugerem substituir a medida de Informação de Fisher pela Informação de Kullback-Leibler (KL). A motivação para o uso de KL é que a aplicação da Informação de Fisher pode ser pouco eficiente se a estimativa da prociência não estiver próxima ao valor verdadeiro, especialmente na fase inicial do CAT, quando a quantidade de itens do teste ainda é muito pequena para se avaliar com acurácia o valor verdadeiro da proficiência, θ.

Tabela 1 - Quadro-resumo - comparativo entre três métodos de seleção adaptativa de itens

Método	**Medida**	**Motivação**
Máxima Informação	Informação de Fisher	Facilidade Computacional
Máxima Informação Global	Informação de Kullback-Leibler	Ideal para a seleção de itens quando a amostra das respostas do examinando ainda é pequena.
Máxima Informação Esperada	Informação Observada	Baseia-se na análise preditiva, isto é, deseja-se prever a resposta aos itens ainda não administrados ao CAT.

4.1 Composição da prova

Selecionar itens para compor o CAT não é tão "simples" quanto apenas calcular estatísticas. Profissionais da educação podem questionar, e com certa razão, a diversidade das questões. Em uma prova de Inglês, por exemplo, o professor deseja medir o conhecimento dos alunos tanto na interpretação de texto, quanto na gramática, e se os itens forem selecionados apenas por algum critério citado, pode ser que as questões de interpretação sejam escolhidas com muito mais frequência, desequilibrando a prova. Outra questão a ser analisada é que um item pode ter tamanha qualidade, que será selecionado em todas as provas, assim, esse item é conhecido pelos examinandos, e o teste é comprometido.

Algumas restrições precisam ser adotadas. Afinal, o teste deve ter qualidade pedagógica bastante semelhante, ou até mesmo superior a um teste papel-e-caneta, e a redução da quantidade de itens apresentados não deve influenciar negativamente nessa condição.

4.2 Controle de exposição de itens

Mais do que diversidade dos itens, a segurança do teste pode ser comprometida se itens passam a ser conhecidos pelos indivíduos que realizarão o teste. Afinal, como as habilidades serão mensuradas com precisão se a resposta dada a certo item não for dada em função de sua habilidade, mas em função de já se ter conhecimento prévio sobre a pergunta?

Se, por exemplo, em um teste, o ponto de partida do CAT é dado supondo a mesma proficiência para todos os examinandos e o critério de máxima informação for utilizado, dessa forma, haverá apenas um item mais informativo para todos os examinandos, e isso significa que todos responderão ao mesmo primeiro item. O segundo item, por sua vez, será escolhido apenas entre duas opções: se o examinando errou ou acertou o primeiro. E assim por diante. Pessoas que realizarem o teste em tempo futuro já saberão as primeiras perguntas que os aguardam, e o erro das estimativas de suas proficiências será inflacionado.

Usualmente, no teste papel-e-caneta, os itens são aplicados e posteriormente descartados. Porém, para o CAT, o banco de itens deve ser bastante extenso, de forma que descartar itens é uma ação quase que proibida. Dessa forma, deve-se evitar que itens do CAT sejam superexpostos, perdendo, assim, sua qualidade psicométrica. Ante o exposto, pode-se controlar a frequência de exposição dos itens de forma probabilística, por meio de seleção condicional. O algoritmo de Sympson-Hetter (HETTER e SYMPSON, 1997), calcula parâmetros de exposição do item, a fim de reduzir a quantidade de itens superexpostos, e pode ser resumido em 8 passos descritos a seguir.

- **Passo 1:** Especificar a taxa de exposição máxima esperada de um item para o teste, representada por r.
- **Passo 2:** Construir uma tabela de informação cujo conteúdo são listas dos itens do banco por habilidade e, em cada lista, os itens são ordenados do maior para o menor, de acordo com uma função de informação para cada habilidade.
- **Passo 3:** Gerar os primeiros conjuntos de parâmetros dos itens. Se o banco contém I itens, criar um vetor de tamanho I com todos os elementos iguais a 1. Esse vetor representa o parâmetro de exposição dos itens, a probabilidade do item ser administrado (A), dado que foi selecionado (S), e será representado por $P(A|S)$.

Passos aplicados iterativamente até que algum critério de parada seja acionado:

- **Passo 4:** Simular um teste adaptativo para uma amostra aleatória de examinandos. Para cada item do CAT, identificar o item mais informativo

da tabela de informação que seja o mais próximo possível da estimativa da proficiência do examinando. Gerar um número x pseudo-aleatório da distribuição Uniforme (0,1). Administrar o item i, se x for menor ou igual a $P_i(A|S)$. Independentemente, se o item i foi ou não administrado, excluí-lo dos possíveis futuros itens selecionados para o teste do mesmo examinando. Observa-se que, para a primeira simulação, $P_i(A|S) = 1, \forall i$. Portanto, todos os itens selecionados serão administrados.

- **Passo 5:** Acompanhar o número de vezes que cada item é selecionado (NS) e a quantidade de vezes que ele é administrado (NA) no total de amostras simuladas. Quando a amostra completa for testada, calcule, para cada item, a probabilidade $P(S) = NS/NE$ do item ser selecionado e $P(A) = NA/NE$ de ser administrado, em que NE representa o número total de examinandos.

- **Passo 6:** Utilizar o valor de r, especificado no Passo 1, e $P(S)$ para calcular os novos valores de $P_i(A|S)$, como segue:

$$P_i(A|S) = \begin{cases} \frac{r}{P(S)} & \text{se} \quad P(S) > r \\ 1 & \text{se} \quad P(S) \leq r \end{cases} .$$

- **Passo 7:** Para um CAT de tamanho n, deve-se assegurar que ao menos n itens tenham nova $P_i(A|S)$ igual a 1. Esses itens sempre serão administrados quando selecionados, visto que o número aleatório gerado sempre será menor ou igual a 1. Caso isso não ocorra, deve-se garantir que haja n itens cujo maior valor de $P_i(A|S)$ seja igual a 1. Dessa forma, assegura-se que o banco não será desgastado antes que os examinandos possam completar o teste.

- **Passo 8:** Calculados os novos valores de $P_i(A|S)$ e utilizando a mesma amostra de examinandos, refazem-se os Passos 4 a 7 até que o máximo valor de $P(A)$ (calculado no Passo 5) aproxime-se do limite um pouco acima de r e, então, oscile em sucessivas simulações.

Para a aplicação real do CAT, os valores $P_i(A|S)$ obtidos ao final da simulação serão utilizados, e os itens serão administrados, ou não, de acordo com o Passo 4, que será repetido até que um critério de parada seja acionado.

Como em qualquer restrição, limitar o algoritmo de seleção dos itens implica perder em informação do teste. Portanto, os itens do banco devem ser elaborados com alta qualidade, para que, quando um item não puder ser administrado, o algoritmo tenha opções um pouco menos informativas, mas também interessantes.

Esse tipo de restrição, apesar da grande contribuição para a segurança do teste, tem suas limitações. Itens com baixa probabilidade de seleção vão permanecer pouco expostos, e os parâmetros de controle de exposição devem ser atualizados a cada alteração do banco ou da distribuição das proficiências, exigindo atenção e esforços computacionais.

A estratificação do banco de itens é sugerida para complementar o procedimento probabilístico explicado. Estratifica-se o banco de dados baseando-se nos parâmetros dos itens e, com isso, o teste é dividido em estágios. O

primeiro estrato é composto por itens menos discriminativos, e o último, pelos mais discriminativos (CHANG e YING, 1996).

Ao dividir o banco pela discriminação dos itens, por exemplo, no primeiro estágio itens menos discriminativos são selecionados do primeiro estrato e, nos próximos estágios, itens mais discriminativos dos outros estratos serão selecionados. Nos estágios iniciais do teste, a estimativa da proficiência dos examinandos ainda é bastante imprecisa, e, dessa forma, a utilização de itens muito discriminativos é um desperdício. Itens mais discriminativos devem ser deixados para serem apresentados nos estágios finais do teste.

Outra forma de estratificação é dividir o banco pelo parâmetro de discriminação, utilizando um bloco formado pela dificuldade dos itens (CHANG *et al.*, 2001). Forma que se justifica, pois, estratificando apenas pela discriminação, pode ser que não haja itens de baixa dificuldade no último estrato. O banco é dividido em pequenos níveis por dificuldade. Em cada nível, os itens são classificados em ordem ascendente por discriminação, e itens de menor discriminação são agrupados no primeiro estrato, e assim por diante.

Mais informações sobre estratégias de controle para a frequência de exposição dos itens podem ser encontradas no artigo de GEORGIADOU *et al.* (2007).

4.3 Balanceamento do conteúdo do teste

Uma forma de controlar o balanceamento do conteúdo de um CAT é fazer uma modificação do critério de Máxima Informação, considerando, além da informação do item, o descritor (a categoria do conteúdo) no qual o item se encaixa (KINGSBURY e ZARA, 1989). Caso o item selecionado seja de um descritor que ainda não tenha sido aplicado no teste, não há problemas, o item é administrado. Porém, se o descritor já foi apresentado por algum item, o de segunda maior informação é avaliado de acordo com seu conteúdo, e se decide se esse será ou não aplicado.

Uma forma alternativa de controlar o balanceamento do conteúdo de um CAT é através do chamado "shadow tes" (VAN DER LINDEN e REESE, 1998; VAN DER LINDEN e PASHLEY, 2000), que consiste em uma abordagem de aplicação do exame utilizando um CAT com conteúdos sujeitos a restrições, i.e., quando um grande número de especificações precisa ser levado em conta no processo de seleção dos items.

De modo a aumentar o número de items disponíveis para o CAT e reduzir o custo na elaboração dos mesmos, GLAS e VAN DER LINDEN (2003) discutem o uso da técnica de clonagem. Uma forma de clonagem é feita através da especificação de items-pai, cuja descrição sintática deixa uma ou mais lacunas a serem preenchidas por um conjunto específico de possibilidades ("replacement sets").

4.4 Critério de parada do teste

Em um teste adaptativo computadorizado, itens são aplicados até que um critério de parada seja atendido. Mas como escolher um critério de parada? Isso

depende dos objetivos do teste, se será utilizado para seleção ou apenas classificação dos indivíduos.

Com o objetivo de classificar os indivíduos, deve-se determinar um escore de corte que irá definir se o examinando foi ou não aprovado. Um indivíduo poderá ser classificado quando o intervalo de confiança de 95% (calculado por $\hat{\theta} \pm 2\epsilon$, em que ϵ representa o erro-padrão da estimativa) da sua proficiência estimada estiver acima ou abaixo do ponto de corte. Com essa classificação, o teste poderá ser finalizado e cada indivíduo terá uma margem de erro de pelo menos 5% da estimativa de sua habilidade. Vale ressaltar que a confiança do intervalo pode ser alterada, modificando, a taxa de erro da medida.

Quando não se objetiva classificar, mas selecionar alguns indivíduos dentre os que realizam o teste, a estimativa das proficiências deve possuir nível de precisão fixo, para que não existam injustiças. Dessa forma, o erro-padrão deve ser fixado e, quando atingido, o teste será finalizado, e proficiências equiprecisas serão obtidas.

Também existe possibilidade de fixar a quantidade de itens administrados e finalizar o teste quando essa quantidade for alcançada. Pode-se, ainda, impor um tempo limite. Porém, esses critérios são utilizados apenas por conveniência do aplicador e não são recomendados para esse tipo de avaliação. Afinal, o erro-padrão da estimativa da habilidade de cada examinando decresce de forma desigual, dependendo da proficiência inicial suposta para cada indivíduo, do padrão de respostas observado e dos itens aplicados.

Um assunto correlato ao presente relaciona o tempo necessário para a conclusão de um teste. Examinandos com a mesma habilidade podem necessitar de tempos distintos para a realização de um exame. Dessa forma, um aluno mais vagaroso pode não conseguir terminar a prova em uma prova que demanda rapidêz nas respostas. A incorporação da informação dos items não respondidos, devido à falta de tempo, no escore final de habilidade é muito complexa. VAN DER LINDEN *et al.* (1999) propuseram algumas técnicas na tentativa de neutralizar possíveis efeitos diferenciais causados pela pressão de tempo, sofrida pelo examinando, na realização de um exame via CAT.

5 Estudos de simulação

5.1 Dados

A Universidade de Brasília, por intermédio de CESPE – Centro de Seleção e Promoção de Eventos, oferece, semestralmente a seus alunos, uma prova de proficiência em Inglês. Os alunos interessados podem realizar a prova e, caso sejam aprovados, recebem os créditos referentes às disciplinas Inglês Instrumental 1 e/ou Inglês Instrumental 2. No primeiro semestre de 2010, houve uma novidade: a prova foi aplicada por meio do CAT.

O banco de itens para o CAT foi elaborado a partir de nove provas tradicionais anteriormente aplicadas e identificadas neste trabalho pelos códigos 105, 205, 106, 206, 207, 108, 208, 109, 209. Haviam itens em comum entre as provas e, assim, eles foram calibrados conjuntamente: os três parâmetros do modelo foram estimados

pelo modelo Normal e os itens de pouca qualidade foram retirados do banco, resultando em 383 itens distintos.

5.2 Processo de simulação do CAT e estatísticas utilizadas

Para cumprir com o objetivo do estudo, foram realizadas algumas simulações do CAT. Para isso, foram utilizados os itens, com seus respectivos parâmetros, referentes à prova CAT de Inglês 1 e 2 mencionada acima e simulou-se as proficiências de 200 alunos, aleatoriamente atribuídas, a partir da distribuição Normal padrão, isto é, $\theta_j \sim N(0,1)$, $i = 1, \ldots, 200$.

A aplicação foi implementada a partir de um programa desenvolvido na linguarem R. Na primeira parte do programa, é criada uma função para cálcular os pontos de quadratura (GRAY, 2001) e seus respectivos pesos. Esses comandos foram retirados de GRAY (2001). Então a função é utilizada e 30 pontos são gerados.

As habilidades dos 200 alunos são geradas, os parâmetros dos itens são importados e, com essas informações, é calculada uma matriz com a probabilidade de acerto de cada item por indivíduo. A função utilizada para esse cálculo é o modelo logístico de três parâmetros, descrito pela Equação 1, sendo que o fator escala D deve ser igualado a $1,7$ para que os resultados sejam análogos à função Normal e, assim, fique equivalente ao modelo utilizado para estimar os parâmetros dos itens. A partir das probabilidades geradas, aplica-se a distribuição Bernoulli para se obter uma matriz de zeros e uns, definindo o acerto ou erro de cada item por indivíduo.

As habilidades iniciais de todos os alunos são igualadas a zero, a média da distribuição de θ. Para cada aluno, o programa começa um *loop*, que é encerrado quando o critério de parada for atingido. Na primeira rodada[3] do *loop*, um item é selecionado aleatoriamente. Já nas demais rodadas, a informação de cada item é calculada pelo Critério de Máxima Informação (Equação 2), e o item de maior informação, dada a atual habilidade estimada do examinando, é selecionado. Vale ressaltar que não há repetição de itens para um mesmo aluno e, dessa forma, os itens que já foram expostos são retirados do banco antes do referido cálculo. Busca-se na matriz de acertos e erros se a resposta do examinando àquele item é correta ou incorreta e esse dado é guardado em um vetor cujo comprimento é igual a quantidade de itens respondidos pelo aluno. A habilidade do examinando é estimada pelo método EAP, levando em consideração o método da quadratura (Equação 8). Para a mensuração da habilidade, considera-se todos os itens, com seus respectivos parâmetros e respostas previamente estimadas, já expostos aos examinando. Junto com o cálculo da proficiência, também é calculada a variância a posteriori associada à estimativa obtida (Equação 9). Verifica-se o critério de parada, que caso tenha sido alcançado, finaliza o programa e $\hat{\theta}$ fica sendo igual à última habilidade estimada. Caso não tenha sido, o programa recomeça com a escolha de um novo item e habilidade é estimada novamente.

[3] A rodada da simulação é definida pela quantidade de itens selecionados ao examinando. Por exemplo: na primeira rodada da simulação, o primeiro item é selecionado ao aluno; na segunda rodada, um segundo item é selecionado, sendo que o aluno já respondeu um item.

Quatro estudos foram feitos e utilizou-se dois critérios de parada diferentes. Um dos critérios de parada é definido pelo erro, que é dado pela raiz quadrada da variância a posteriori associada à EAP (Equação 9). Dessa forma, o programa é interrompido com a convergência de $\hat{\theta}$. O outro critério de parada utilizado é a quantidade de itens expostos aos examinandos, ou seja, todos os alunos respondem a mesma quantidade fixa de itens. Deve-se advertir que, quando esse último critério é utilizado, não há controle sobre os erros das estimativas, que são distintos entre os examinandos, podendo levar a conclusões precipitadas em um exame cuja classificação dos indivíduos faça diferença. Ressalta-se que, para todas as simulações, supõe-se distribuição Normal Padrão para as habilidades.

5.3 Estudo I

Um estudo inicial foi feito para que o processo do CAT fosse melhor compreendido. Definiu-se o erro como critério de parada do CAT e foram feitas seis simulações com *erros predefinidos* em 30%, 25%, 20%, 15%, 10% e 5%, isto é, a habilidade de cada examinando foi estimada iterativamente até que o desvio-padrão do cálculo dessa estimativa fosse inferior ao erro estipulado. Cada examinando teve sua habilidade estimada uma única vez, para cada um dos erros, e se analisou a quantidade de itens que cada um respondeu para que o critério de parada fosse alcançado. A probabilidade de acerto de cada item, por indivíduo, foi mantida nas diferentes simulações. Desse modo, pode-se observar o comportamento da seleção dos itens, a quantidade média de itens respondidos para cada erro estipulado e o comportamento da estimação iterativa da habilidade. A seguir descreve-se algumas caracteristicas do CAT para cada um dos erros em estudo.

Erro = 30% - A média de itens respondida pelos examinandos foi de $7,8$. Um aluno respondeu apenas 5 itens, a menor quantidade aplicada pela simulação CAT. A maior quantidade respondida foi de 17 itens, e isso ocorreu com um examinando de habilidade alta, aproximadamente igual a $2,13$ (aluno194).

Erro = 25% - A média de itens respondida pelos examinandos foi de $13,4$. Três alunos responderam apenas 8 itens nessa simulação, a menor quantidade para atingir o erro estipulado. Eles tinham habilidades iguais a $0,35$; $0,45$ e $0,80$. A maior quantidade respondida foi de 43 itens, e isso ocorreu com um examinando de $\theta = 2,18$.

Erro = 20% - Um examinando (aluno35), com $\theta = -2,84$, não alcançou o erro estipulado e precisou responder a todos os itens da prova. O aluno35 permaneceu com erro final de aproximadamente $0,23247$. A média de itens respondida pelos examinandos foi de $28,5$, porém, se for retirado o examinando que não alcançou o erro estipulado, a média cai para $26,7$ itens. Dois examinandos ($\theta = 0,33; \theta = 0,41$) atingiram o critério de parada do CAT ao responderem 14 itens, a menor quantidade nessa simulação. Dentre os alunos que atingiram o erro estabelecido, a maior quantidade de itens respondidos foi de 184 e, como era esperado, ocorreu com um aluno de habilidade extrema, o mesmo que respondeu mais itens na primeira simulação, aluno194 ($\theta = 2,13$).

Erro = 15% - Cinco examinandos não alcançaram o erro estipulado e precisaram responder a todos os itens disponíveis na prova. Um desses examinandos é o aluno35, que também não atingiu o erro de 20% na simulação anterior. Dentre os alunos que não alcançaram o erro estipulado, o menor e o maior erro atingidos foram de $0,17172$ e $0,23247$ (aluno35), respectivamente. A média de itens respondida pelos examinandos foi de $67,5$, porém, se forem retirados todos os examinandos que não alcançaram o erro estipulado, a média cai para $59,4$ itens. Dois alunos, de habilidades $0,35$ e $0,45$, responderam apenas 23 itens, a menor quantidade para atingir o erro estipulado. Dentre os alunos que atingiram o erro estabelecido, a maior quantidade de itens respondidos foi de 302, e isso ocorreu com um examinando de $\theta = 1,14$.

Erro = 10% - 34 examinandos não alcançaram o erro estipulado e precisaram responder a todos os itens, sendo que o aluno35 está entre eles. Dentre os alunos que não alcançaram o erro estipulado, o menor e o maior erro atingidos foram de $0,10995$ e $0,23247$ (aluno35). A média de itens respondida pelos examinandos foi de $141,7$, porém, se forem retirados todos os examinandos que não alcançaram o erro estipulado, a média cai para $92,2$ itens. Três examinandos, de habilidades iguais a $-0,85$; $0,35$ e $0,51$, responderam a apenas 40 itens, a menor quantidade para atingir o erro estipulado. Dentre os alunos que atingiram o erro estabelecido, a maior quantidade de itens respondidos foi de 340, e isso ocorreu, surpreendentemente, com um examinando de $\theta = -0,06$.

Erro = 5% - 80 examinandos não alcançaram o erro estipulado e precisaram responder a todos os itens disponíveis na prova. Novamente o aluno35 se enquadra nessa situação. Dentre os alunos que não alcançaram o erro estipulado, o menor e o maior erro atingidos foram de $0,05211$ e $0,23247$ (aluno35). A média de itens respondida pelos examinandos foi de $234,2$, porém, se forem retirados todos os examinandos que não alcançaram o erro estipulado, a média cai para $134,9$ itens. Um único examinando, de habilidade $\theta = -0,57$, respondeu apenas 67 itens, a menor quantidade para atingir o erro estipulado. Dentre os alunos que atingiram o erro estabelecido, a maior quantidade de itens respondida foi de 341, e isso ocorreu com um examinando de $\theta = 0,68$.

A Tabela 2 sumariza as observações feitas acima. Verifica-se que o número de itens respondidos cresce à medida que o erro requerido na estimação de θ diminui. Para um erro de 30%, um número médio de 7,8 questões tiveram que ser respondidas para que fosse possível estimar a proficiência. Para um erro de 5%, tal número médio de questões foi de 234,2. Analogamente, o número de alunos que não alcançaram o erro estipulado cresce à medida que o erro requerido na estimação de θ diminui.

Não se deve esquecer que, na prática, não se deixa ilimitada a quantidade de itens que serão respondidos por cada examinando. Então, por mais que algum indivíduo de habilidade extrema não alcance o erro estipulado, será visível que sua habilidade é ainda superior (ou ainda inferior) à última estimada e, portanto, esse indivíduo será, sem problemas, aprovado (ou reprovado) no exame CAT.

Outra questão a ser lembrada é que, pelo fato de o primeiro item ser selecionado aleatoriamente, há diferenças nas estimativas das habilidades e na consequente

Tabela 2 - Análise da quantidade de itens nas simulações do Estudo I. *Valores desconsiderando os alunos que não alcançaram o erro. ** Apenas entre os alunos que não alcançaram o erro

Simulação	30%	25%	20%	15%	10%	5%
Média de Itens	7,8	13,4	28,5	67,5	141,7	234,2
Mínimo de Itens	5	8	14	23	40	67
Quantidade de alunos que não alcançaram o erro	-	-	1	5	34	80
Média*	-	-	26,7	59,4	92,2	134,9
Máximo*	17	43	184	302	340	341
Erro Mínimo**	-	-	23,20 %	17,20%	11,00%	5,20%
Erro Máximo**	-	-	23,20 %	23,20 %	23,20 %	23,20 %

seleção de novos itens, tornando incerta a repetição de indivíduos que responderam a mais ou menos itens. Apesar disso, após as primeiras rodadas do CAT, os itens selecionados para os mesmos indivíduos permanecem bastante similares.

Ao tomar como exemplo três alunos de habilidades reais $-0,58$ (aluno5), $0,16$ (aluno187) e $0,86$ (aluno154), que alcançaram o erro estipulado em todas as simulações, e o aluno35, que não alcançou o erro predefinido a partir da simulação de erro 20%, pode-se observar os gráficos (Figura 3) dos processos de estimação das habilidades e seus respectivos erros.

Pelo eixo das abscissas, pode-se verificar a quantidade de itens que o aluno precisou responder para atingir o critério de parada, ou seja, quantas rodadas do programa foram necessárias. Já no eixo da ordenadas, verifica-se as habilidades (gráficos à esquerda) ou os erros (gráficos à direita) estimados. Cada ponto, por sua vez, representa a habilidade ou o erro estimado a cada rodada do CAT.

Nota-se, pelos gráficos da Figura 4, que os pares de pontos grafados se aproximam de uma linha diagonal, observa-se correlação entre as habilidades reais e estimadas dos examinandos, indicando a expressiva qualidade das estimativas realizadas.

Ao analisar os itens selecionados para cada examinando, observa-se que, apesar de existirem itens de alta qualidade, que são expostos para a grande maioria dos alunos, a significância do item realmente é relativa à habilidade do aluno. O item 140, por exemplo, foi selecionado para 90% dos alunos (na simulação de erro 30%), porém, foi dispensável, por não trazer muita informação, para o cálculo da habilidade de 20 indivíduos.

5.4 Estudo II

Para o segundo estudo, testes CAT foram simulados para cada examinando, e o parâmetro $\hat{\theta}$ foi estimado $n = 50$ vezes, para cada uma das n amostras simuladas respostas aos itens em estudo. O critério de parada adotado foi o erro. Optou-se por erros de 30%, 25% e 20%, pois, apesar de serem os maiores do estudo anterior, são os

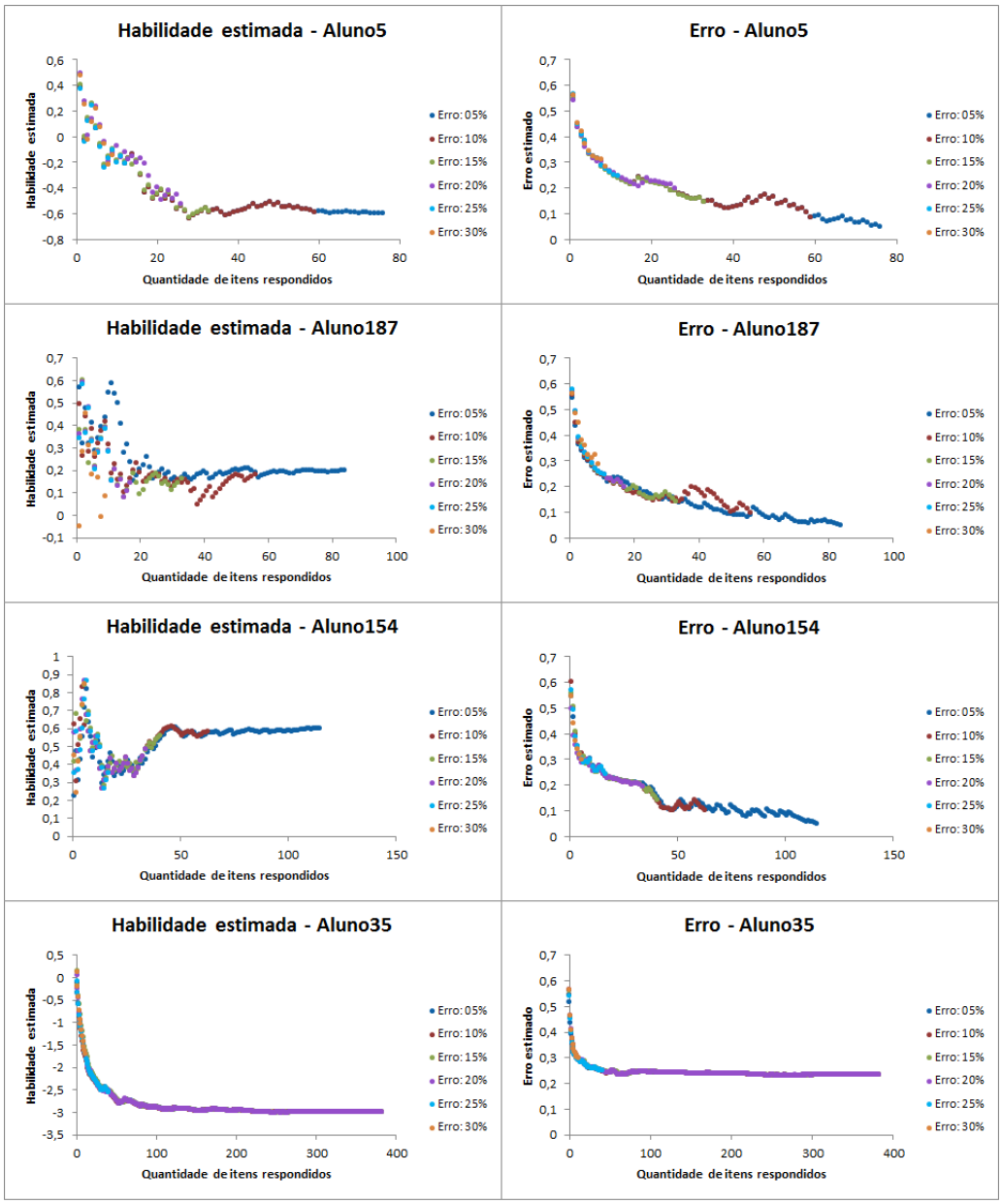

Figura 3 - Habilidades estimadas e respectivos erros do aluno5 ($\theta = -0,58$), aluno187 ($\theta = 0,16$), aluno154 ($\theta = 0,86$) e aluno35 ($\theta = -2,84$) a cada rodada do CAT por erro predefinido.

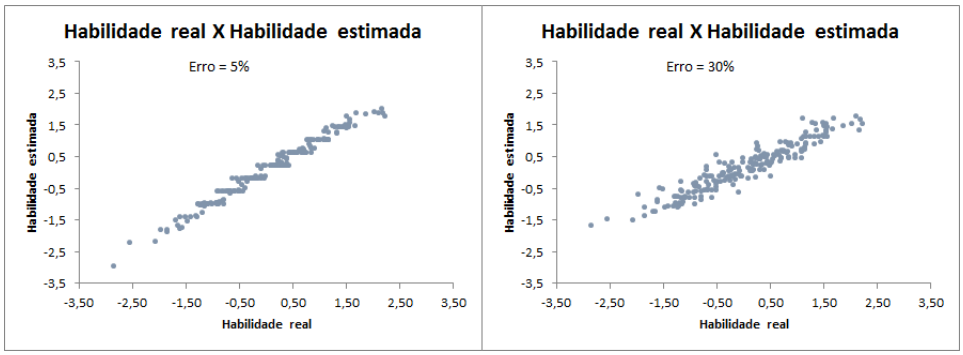

Figura 4 - Habilidades reais e estimadas de cada examinando nas simulações CAT de erros iguais a 5% e 30%.

apresentados na literatura (como em BOCK e MISLEVY, 1982), além de próximos aos erros gerados nas estimativas das provas tradicionais (tabela não inclusa), das quais os itens foram retirados.

Dos 10 mil cálculos feitos para estimar 50 vezes as habilidades dos 200 indivíduos, apenas em 22 vezes o indivíduo precisou responder a todos os itens e, mesmo assim, não atingiu o erro estipulado de 20%. As 50 habilidades estimadas dos 4 indivíduos utilizados como exemplo no 1º estudo podem ser visualizadas nos gráficos a seguir (Figura 5). A abscissa representa as 50 simulações feitas, já as ordenadas, as habilidades estimadas. Cada ponto grafado indica a habilidade estimada para o aluno em cada simulação, e a reta traçada, a habilidade real do aluno.

Na Tabela 3, verifica-se a quantidade de alunos e a média de itens respondidos para cada nível de erro, por categorias de proficiências. Nota-se que indivíduos de habilidades extremas tendem a responder mais itens que indivíduos de habilidades medianas.

Tabela 3 - Quantidade média de itens selecionados por habilidade nas simulações CAT de erro predefinido

θ_j	**Nº de alunos**	**30%**	**25%**	**20%**
$< -1,5$	12	9,1	17,1	40,0
$-1,5 \vdash -0,5$	51	7,7	12,5	22,9
$-0,5 \vdash 0,5$	70	7,3	11,4	20,8
$0,5 \vdash 1,5$	52	6,9	11,4	25,0
$> 1,5$	15	10,4	24,8	76,0
Média total	200	7,7	13,0	27,7

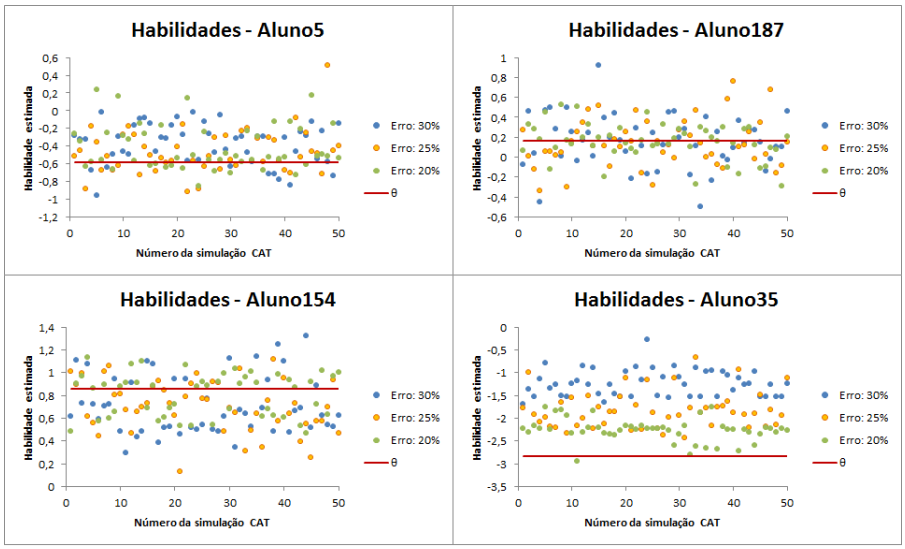

Figura 5 - Habilidades estimadas a cada repetição da simulação CAT por erro predefinido e habilidade real.

5.5 Estudo III

Analisando a quantidade de itens das provas tradicionais, também foi feito um estudo fixando a quantidade de itens expostos para cada examinando. Como no estudo anterior, cada examinando teve sua proficiência estimada 50 vezes via CAT, a partir de novas probabilidade de acerto para cada item. Desta vez, porém, o erro foi desconsiderado.

Na Figura 6 podem-se visualisar as proficiências estimadas em cada repetição. A abscissa representa as 50 simulações feitas, já as ordenadas representam as habilidades estimadas. Cada ponto grafado indica a habilidade estimada para o aluno em cada simulação, e a reta traçada, a habilidade real do aluno.

5.6 Estudo IV

Neste último estudo, as habilidades dos alunos foram estimadas como se cada um deles tivesse sido examinando por cada uma das provas papel-e-caneta 50 vezes. Para cada uma das 50 vezes que as habilidades foram mensuradas, a probabilidade de acerto de cada item foi recalculada. Com base nas simulações realizadas para cada indivíduo, pode-se avaliar o vício (B), a variância (Var) e o erro quadrático médio (EQM) das estimativas de θ nas provas tradicionais.

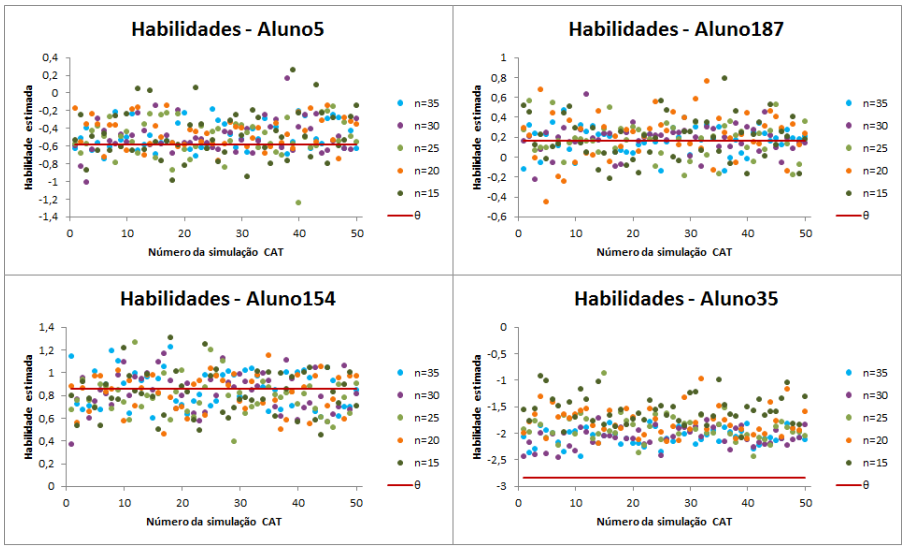

Figura 6 - Habilidades estimadas a cada repetição da simulação CAT por quantidade de itens e habilidade real.

5.6.1 Erros dos estudos de simulação

Nas Tabelas que seguem, pode-se analisar os erros provenientes dos estudos realizados. A Tabela 4 é referente ao Estudo II, a Tabela 5, ao Estudo III, e a Tabela 6 resulta do Estudo IV. As Equações 10, 11 e 12 explicitam como os erros, por meio do vício (B), erro quadrático médio (EQM) e variância (Var), respectivamente, foram calculados.

$$B = \sum_{j=1}^{n} \frac{(\hat{\theta}_j - \theta_j)}{n} = \sum_{i=1}^{50} \frac{\hat{\theta}_j}{50} - \theta_j. \tag{10}$$

$$EQM = \sum_{j=1}^{n} \frac{(\hat{\theta}_j - \theta_j)^2}{n}. \tag{11}$$

$$Var = EQM - B^2. \tag{12}$$

Tabela 4 - Vício, variância e EQM das simulações CAT com erros preestabelecidos definidos como critério de parada - Estudo II

Erro estipulado	$\overline{B(\widehat{\theta})}$	$\overline{Var(\widehat{\theta})}$	$\overline{EQM(\widehat{\theta})}$
0,30	0,10080	0,06850	0,17914
0,25	0,06696	0,05428	0,10376
0,20	0,03820	0,03625	0,05514

Tabela 5 - Vício, variância e EQM das simulações CAT com quantidade fixa de itens - Estudo III

Nº de itens	$\overline{B(\widehat{\theta})}$	$\overline{Var(\widehat{\theta})}$	$\overline{EQM(\widehat{\theta})}$
n=35	0,02435	0,02730	0,04333
n=30	0,03047	0,02841	0,04766
n=25	0,03591	0,03208	0,05737
n=20	0,04517	0,03638	0,06977
n=15	0,05825	0,04398	0,09577

Tabela 6 - Vício, variância e EQM das simulações dos testes tradicionais - Estudo IV. A coluna 1 da tabela refere-se à prova a qual o item originalmente pertence

Id da prova	$\overline{B(\widehat{\theta})}$	$\overline{Var(\widehat{\theta})}$	$\overline{EQM(\widehat{\theta})}$
105	-0,00811	0,10061	0,10061
205	-0,00608	0,10758	0,10758
106	-0,03283	0,10180	0,10180
206	-0,03155	0,08510	0,08510
207	0,02698	0,10904	0,10904
108	-0,04890	0,14469	0,14469
208	-0,03319	0,11890	0,11890
109	-0,01937	0,09336	0,09336
209	-0,00385	0,04863	0,04863

Ao analisar o vício médio absoluto das simulações, verifica-se que a prova de menor vício é a tradicional 209, e a de maior vício, o CAT com critério de parada definido como sendo um erro inferior a $0,30$. Esse resultado é esperado se for recordada a quantidade de itens em cada um destes testes: 87 itens no teste tradicional 209 e $7,7$ itens em média no CAT de erro 30%. Com exceção das provas tradicionais 205, 109 e 209, o CAT com critério de parada estipulado em 35 itens teve menor vício médio dos que as demais provas tradicionais, cuja quantidade de itens varia entre 39 a 44.

Observando a variância, por sua vez, todas as simulações CAT com quantidade fixa de itens resultaram em menor variância média das proficiências estimadas, se comparadas com as variâncias resultantes da provas tradicionais. O mesmo ocorreu com o CAT de erro $0,20$. As variâncias dos testes adaptativos de erros estipulados em 25% e 30% só foram superiores ao teste tradicional 209, que é o de maior quantidade de itens.

Em relação ao EQM, as simulações CAT de quantidade de itens fixados em 35 e 30 obtiveram melhores resultados do que todas as provas tradicionais. Os de quantidade de itens 25 e 15 somente não resultaram em EQM inferiores ao do teste tradicional 209. Já o CAT de 15 itens, teve menor EQM do que 6 das 9 provas tradicionais. O CAT com erro de 20% teve EQM inferior aos das provas tradicionais, exceto a 209, e a de erro de 25% teve menor EQM em relação a 4

provas tradicionais e bem próximo a outras 2 provas tradicionais, com diferenças inferiores a 0,4%. O CAT de erro 0,30 teve o maior EQM de todas as simulações feitas.

Considerando tanto o significativo encurtamento do teste quanto a precisão das estimativas, pelos resultados obtidos, pôde-se conferir a eficácia de um CAT e evidenciar seus benefícios.

Conclusões

É bastante atraente a flexibilidade proporcionada pela possibilidade de se formularem testes que se vão adaptando ao nível de capacidade do examinando. Dessa forma, apenas itens importantes para o indivíduo são necessários para um preciso resultado na captação do conhecimento. A eficácia do CAT foi demonstrada, empiricamente, através de nossos estudos de simulação, e, seus benefícios, foram evidenciados. Um fator importante nesse argumento é a redução da quantidade de itens da prova e, consequentemente, do tempo de execução, sem que se perca qualidade na mensuração da proficiência.

Como sugestão de trabalhos futuros, fica a melhoria estatística do algoritmo de seleção dos itens, com simulações que testem o Critério de Máxima Informação Global e o Critério de Máxima Informação Esperada. Também fica como sugestão o incremento do algoritmo com a inclusão do controle da frequência de exposição de itens e o balanceamento do conteúdo do teste. Além da inclusão de novos itens, que precisem ter seus parâmetros estimados durante o teste.

Agradecimentos

Os autores agradecem ao CESPE (Centro de Seleção e Promoção de Eventos), por disponibilizar bancos de dados necessários a este trabalho.

LABARRÈRE, J. G.; DA-SILVA, C. Q.; COSTA, D. R. Computerized adaptive testing. *Rev. Bras. Biom.*, São Paulo, v.29, n.2, p.229-261, 2011.

- *ABSTRACT: The Computerized Adaptive Tests (CAT) describe a class of eletronic exams in which the items are selected according to the test taker ability. In that respect his or her proficiency is interactively measured. In order to implement the CAT it is necessary to set up a sound collection of items (Item Bank). Such items must possess both pedagogic and psichometric quality. For the CAT, Item Response Models (IRM) are very important for calibrating the items. In the IRM the probability of answering an item correctly grows with the test taker ability. Once the items in the Item Bank were calibrated it is possible to release an online estimate of the test taker ability. The methodology also makes possible the comparision of the abilities of different examinees. Using simulated data and also real data we analyze the errors in the estimates of examinees abilities submitted to both CAT and traditional exams.*

- *KEYWORDS: Computerized adaptive testing; item response theory; computerized evaluation.*

Referências

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da resposta ao item:* conceitos e aplicações. São Paulo: Associação Brasileira de Estatística. Disponível em: <http://www.inf.ufsc.br/∼dandrade/TRI/>, 2000. Acesso em: 22 mar. 2010.

BOCK, R. D.; LIEBERMAN, M. Fitting a response model for n dichotomously scored items. *Psychometrika*, New York, v.35, p.179-197, 1970.

BOCK, R. D.; MISLEVY, R. J. Adaptive EAP estimation of ability in a microcomputer environment. *Appl. Psychol. Meas.*, Thousand Oaks, v.6, n.4, p.431-444, 1982.

COSTA, D. R. *Métodos estatísticos em testes adaptativos informatizados.* 2009. Dissertação (Mestrado em Estatística) – Universidade Federal do Rio de Janeiro, Rio Janeiro, 2009. Disponível em: <http://www.dme.ufrj.br/teses.htm>. Acesso em: 22 mar. 2010.

CHANG, H. H.; YING, Z. A global information approach to computerized adaptive testing. *Appl.Psychol. Meas.*, Thousand Oaks, v.20, p.213-229, 1996.

CHANG, H. H.; QIAN, J.; YING, Z. A-stratified multistage computerized adaptive testing with b-blocking. *Appl. Psychol. Meas.*, Thousand Oaks, v.25, p.333-341, 2001.

FLIEGE, H.; BECKER, J.; WALTER, O. B.; BJORNER, J. B.; KLAPP, B. F.; ROSE, M. Development of a computer-adaptive test for depression (D-CAT). *Qual. Life Res.*, Dordrecht, v.14, p.2277-2291, 2005.

GEORGIADOU, E.; TRIANTAFILLOU, E.; ECONOMIDES, A. A. A review of item exposure control strategies for computerised adaptive testing developed from 1983 to 2005. *J. Technol. Learn. Assessment*, Chestnut Hill, v.5, n.8, 2007.

GLAS, C. A. W.; VAN DER LINDEN, W. J. Computerized adaptive testing with item cloning. *Appl. Psychol. Meas.*, Thousand Oaks, v.27, n.4, p.247-261, 2003.

GRAY, R. *BIO 248 cd*: advanced statistical computing. 2001. 342p. (Course Notes).

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. *Fundamentals of item response theory.* Newbury Park: Sage Publications, 1991. 174p.

HARTA, D. L.; COOK, K. F.; MIODUSKID, J. E.; TEAL, C. R.; CRANEG, P. K. Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. *J. Clin. Epidemiol.*, Philadelphia, v.59, p.290-298, 2006.

HETTER, R. D.; SYMPSON, B. Item exposure control in CAT-ASBAV. In: SANDS, W. A.; WATERS, B. K.; MCBRIDE J. R. (Ed.). *Computerized adaptive testing:* from inquiry to operation. Washington: American Psychologistical Association, 1997. p.141-144.

KINGSBURY, G. G.; ZARA, A. R. Procedures for selecting items for computerized adaptive tests. *Appl. Meas. Educ.*, Philadelphia, n.4, p.359-375, 1989.

LORD, M. F. *Applications of item response theory to practical testing problems.* Hillsdale: Erlbaum, 1980. 274p.

VAN DER LINDEN, W. J.; PASHLEY, P. J. Item selection and ability estimation in adaptive testing. In: VAN DER LINDEN, W. J.; GLAS, C. A. W. (Ed.). *Computerized adaptive testing*: Theory and practice. Boston: Kluwer, 2000. p.1–25.

VAN DER LINDEN, W. J.; REESE, L. M. A model for optimal constrained adaptive testing. *Appl. Psychol. Meas.*, Thousand Oaks, v.22, p.259–270, 1998.

VAN DER LINDEN, W. J.; SCRAMS, D. J.; SCHNIPKE, D. L. Using response-time constraints to control for differential speededness in computerized adaptive testing. *Appl. Psychol. Meas.*, Thousand Oaks, v.23, n.3, p.195-210, 1999.

TRIANTAFILLOU, E.; GEORGIADOU, E.; ECONOMIDES, A. A. The design and evaluation of a computerized adaptive test on mobile devices. *Comput. Educ.*, Oxford, v.50, p,1319-1330, 2008.

WARE JR., J. E.; KOSINSKI, M.; BJORNER, J. B.; BAYLISS, M. S.; BATENHORST, A.; DAHLÖF, C. G. H.; TEPPER, S.; DOWSON, A. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Qual. Life Res.*, Dordrecht, v.12, p.935-952, 2003.

Recebido em 18.02.2011.

Aprovado após revisão em 26.07.2011.