Teste Adaptativo Computadorizado Multidimensional com propósitos educacionais: princípios e métodos[1]

Jean Piton-Gonçalves [a]
Sandra Maria Aluísio [b]

Resumo

O *Teste Adaptativo Computadorizado Multidimensional* (MCAT), baseado na *Teoria de Resposta ao Item Multidimensional* (MIRT), considera as múltiplas habilidades do examinado em testes educacionais. Nesse contexto, o presente artigo: (i) traz subsídios teóricos e metodológicos sobre MCATs com propósitos educacionais e (ii) aborda vantagens e desvantagens da aplicação em cenários educacionais. A revisão da literatura aponta que o MCAT é adequado para testes computadorizados com múltiplas habilidades, administrando um número menor de itens do que os testes tradicionais.

Palavras-chave: Teste Adaptativo Multidimensional. Teoria de Resposta ao Item Multidimensional. Teste Adaptativo. Avaliação educacional.

1 Introdução

A preocupação com a qualidade do ensino e da aprendizagem em diversos níveis de ensino vem crescendo ao longo das últimas décadas. Uma avaliação externa para seleção e certificação tem o propósito de coletar, interpretar e aferir resultados de grandes grupos populacionais, visando à certificação educacional ou ao ingresso em determinadas instituições, sejam estas públicas, privadas, educacionais, de treinamento, militares ou ainda outras (PITON-GONÇALVES, 2012).

[a] Universidade Federal de São Carlos - UFSCar - Centro de Ciências Exatas e de Tecnologia - CCET - Departamento de Matemática. São Carlos, São Paulo, Brasil.

[b] Universidade de São Paulo - USP - Instituto de Ciências Matemáticas e de Computação - ICMC - Departamento de Ciências da Computação. São Carlos, São Paulo, Brasil.

Recebido em: 20 Set., 2013
Aceito em: 06 Nov., 2014

[1] Este artigo teve origem na Tese de Doutorado de Jean Piton-Gonçalves, intitulada "Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais", defendida no ICMC/USP, em 2012.

Existem diversas iniciativas públicas de avaliação externa no Brasil, destacando-se aquelas do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), que realiza, por exemplo, o Exame Nacional do Ensino Médio (ENEM). Observa-se que a maior parte dessas avaliações é composta por testes realizados via lápis e papel, com itens objetivos de múltipla escolha e correção mediante folha de leitura ótica.

Os Testes Computadorizados (do inglês *Computer-Based Testing* – CBT) são uma possibilidade inovadora para as avaliações educacionais, possibilitando correções automatizadas, com a produção de estatísticas. Visto como um avanço na área de CBTs, o Teste Adaptativo Computadorizado (CAT) seleciona os itens do teste adaptativamente, seguindo um determinado critério de seleção e considerando as respostas anteriores do examinado durante a aplicação do teste.

As principais vantagens de um CAT são a aplicação de testes com maior flexibilidade e adaptabilidade, com significativa redução do tempo e maior precisão em relação a testes que apresentam um número fixo de itens (OLEA; PONSODA; PRIETO, 1999). A metodologia de um CAT pode seguir a análise de agrupamentos, redes neurais e, especialmente, a Teoria de Resposta ao Item (TRI).

A TRI está ligada a aspectos psicométricos do teste e propõe uma modelagem estatístico-matemática para as características latentes do indivíduo e para os parâmetros associados aos itens. O CAT baseado na TRI (unidimensional) pode ser denominado Teste Adaptativo Computadorizado Unidimensional (UCAT). Geralmente, a literatura utiliza o termo CAT quando se refere a um UCAT.

Atualmente, cada vez mais os testes educacionais têm apontado para avaliações de desempenho do examinado em múltiplas competências e habilidades, premissas estas que o UCAT não considera. Zukowsky-Tavares (2013) fomenta uma discussão entre as abordagens uni e multidimensionais da TRI, com forte crítica sobre as abordagens unidimensionais, contrapondo que o conhecimento humano é plural e multifacetado.

Enquanto resposta a uma avaliação que considera simultaneamente os múltiplos traços latentes, o Teste Adaptativo Computadorizado Multidimensional (MCAT) supõe que o examinado possua mais de uma habilidade estimada pela Teoria de Resposta ao Item Multidimensional (MIRT) (RECKASE, 1985; SEGALL, 1996).

Na literatura, o termo MCAT geralmente está associado a um conjunto de teorias, métodos e modelos psicométricos da área de estatística e estatístico-matemática multivariada, e não necessariamente a um sistema computacional que permite a utilização com usuários reais. Para suprir esta lacuna, Piton-Gonçalves e Aluísio (2012) desenvolveram a abordagem *Computer-based Multidimensional Adaptive Testing* (CBMAT), que norteia o projeto e o desenvolvimento computacional de testes multidimensionais. Dessa forma, um CBT poderá contemplar um MCAT, realizando automaticamente a montagem, a aplicação e a correção de testes adaptativos.

Em termos de MCATs operacionais, ou seja, testes educacionais que podem ser aplicados com usuários reais, a literatura apresenta (i) o *Multidimensional Adaptive Test for Educational Purposes* (MADEPT) (PITON-GONÇALVES, 2012), (ii) o sistema computacional de Senarat et al. (2013) e (iii) o sistema de Wang, Kuo e Chao (2010).

Como a pesquisa em MCAT é incipiente no Brasil, este artigo tem como objetivos: (i) trazer subsídios teóricos e metodológicos para um debate no âmbito nacional sobre MCATs com propósitos educacionais e (ii) preencher a lacuna da literatura nacional sobre os princípios, os conceitos, os métodos, os modelos e os sistemas computacionais operacionais inerentes ao tema em questão. As próximas seções foram elaboradas com base na Tese de Doutorado de Piton-Gonçalves (2012).

2 Teste computadorizado

Todas as avaliações, de qualquer tipo, precisam ser justas e bem analisadas, podendo ser compostas por itens de múltipla escolha e itens dissertativos. Quando um teste possui itens dissertativos, há o aumento do custo e do treinamento de corretores de provas, principalmente em cenários de larga escala. Como alternativa, os testes objetivos buscam cada vez mais avaliações justas, precisas e rápidas, de acordo com métricas e critérios educacionais preestabelecidos e consolidados. Os testes objetivos podem ser classificados em:

- *Teste Objetivo Tradicional*: é um teste realizado via lápis e papel, com aplicação e correção realizadas por pessoas, manualmente (OLEA; PONSODA; PRIETO, 1999).

- *Teste Tipo Fichas*: é um Teste Objetivo Tradicional, diferenciando-se pela correção automatizada. Os leitores ópticos são um exemplo (OLEA; PONSODA; PRIETO, 1999).

- *Teste Computadorizado ou Informatizado* (do inglês *Computer-based Testing*– CBT): o processo de aplicação e correção é automatizado. O computador fornece as questões e os resultados do teste. O processo de análise e apresentação dos resultados pode envolver um conjunto de métricas e critérios.

Com foco no CBT, o Teste Adaptativo surge como uma metodologia que possibilita o aumento da precisão da estimativa das habilidades medidas, administrando um número menor de itens em testes educacionais.

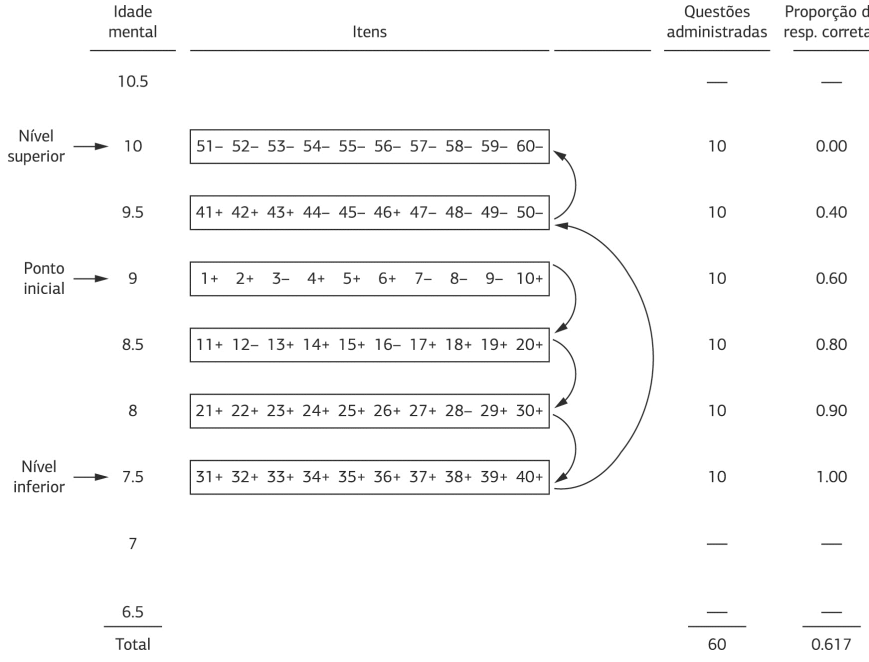
3 Teste Adaptativo

As primeiras ideias sobre testes de Quociente de Inteligência (QI) datam de 1905, com as pesquisas do psicólogo e educador francês Alfred Binet (1857-1911). Ele percebeu que poderia individualizar um teste, classificando os itens em termos de dificuldade. O teste de Binet (*Binet-type adaptive test*) é um teste de inteligência baseado em níveis de dificuldade, criado inicialmente para o diagnóstico do nível de inteligência de uma criança em comparação com sua idade cronológica, analisando a idade mental (WEISS, 1985). Os itens do teste são classificados segundo níveis. Se todos os itens de um determinado nível forem respondidos corretamente, são fornecidos itens de um nível mais alto, até que todos estes sejam respondidos incorretamente (Nível Superior); caso contrário, se todos os itens de certo nível forem respondidos incorretamente, são disponibilizados itens de um nível mais baixo, até que todos estes sejam respondidos corretamente (Nível Inferior). Quando o Nível Superior e o Nível Inferior são identificados, termina-se o teste (WEISS, 1985).

A Figura 1 mostra o procedimento da seleção de itens no teste, no qual os símbolos + e – significam, respectivamente, item respondido correta e incorretamente por um examinado. Ele, então, inicia o teste no primeiro item 1+ e segue até o décimo item 10+. Depois de verificado seu desempenho, ele é levado a um novo conjunto de dez itens, em um segundo nível (idade mental 8.5), seguindo até o nível inferior (7.5), em que responde todos corretamente.

Após responder todos os itens corretamente, será levado ao nível superior no ponto inicial, para se determinar qual é o conjunto de 10 itens em que o examinado responde todos incorretamente. Na Figura 1, este é o nível 10.

Figura 1 - Registro de resposta de um examinado no Teste de Binet. Os símbolos + e – significam, respectivamente, item respondido correta e incorretamente



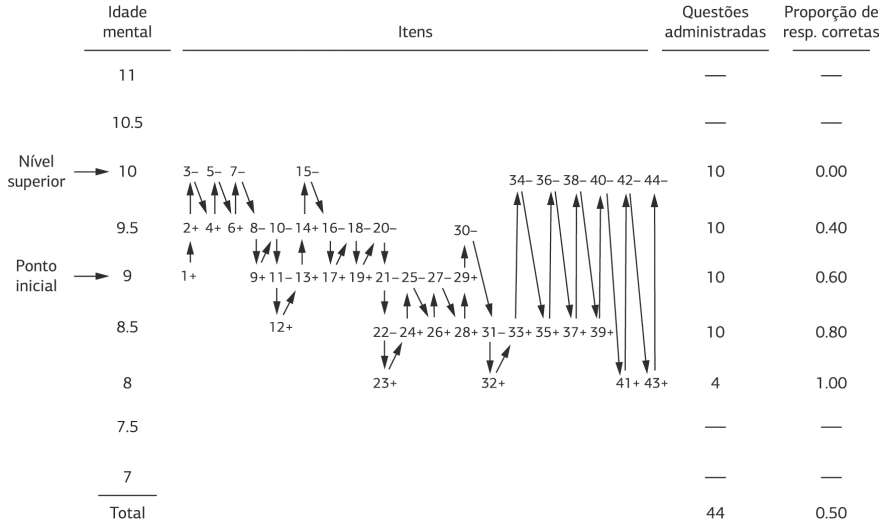
Fonte: Elaborada pelos autores a partir de Weiss (1985).

Após a introdução do teste de Binet, em 1905, na década de 1950, foi desenvolvido o Teste Adaptativo de Dois Estágios (*Two-Stage Adaptive Testing*). Este se divide em dois subtestes de menor dificuldade (*Routing Test*) e maior dificuldade (*Measurement Test*). Segundo as respostas corretas e incorretas obtidas no *Routing Test*, selecionam-se os itens do *Measurement Test* (WEISS, 1985).

Como uma variação do teste de Binet, o Teste Adaptativo Estratificado (*Stratified Adaptive Test*) é diferenciado pela eleição de um próximo item, logo após cada um ser respondido (WEISS, 1985). A Figura 2 ilustra o procedimento de seleção dos itens.

Quando o examinado responde corretamente a um item, o próximo apresenta maior dificuldade. Por outro lado, quando o examinado responder incorretamente, o próximo é de menor dificuldade. O examinado inicia no item 1+, respondendo-o corretamente. Conforme ocorrem acertos (+), o nível de dificuldade dos itens vai se elevando. No caso de errar um item (–), o examinado será levado a um item de

Figura 2 - Registro de resposta de um examinado no Teste Adaptativo Estratificado. Os símbolos + e – significam, respectivamente, item respondido correta e incorretamente



Fonte: Elaborada pelos autores a partir de Weiss (1985).

um nível menor de dificuldade. O teste termina quando for identificado o Nível Superior, nível de dificuldade no qual nenhum item foi respondido corretamente.

Com os avanços de *software*, de *hardware* e da Matemática Computacional, na década de 1960, as áreas de metodologia, análise e aplicação de testes também evoluíram. Como um exemplo, tem-se o trabalho de Reckase (1974), que propôs uma implementação computacional para um teste adaptativo. A informatização do teste adaptativo passou a ser denominado Teste Adaptativo Informatizado ou Teste Adaptativo Computadorizado (do inglês, *Computer* (*ou Computerized*) *Adaptive Test* (*ou Testing*) – CAT).

4 Teste Adaptativo Computadorizado

Um CAT busca maximizar a acurácia do teste, baseando-se no conhecimento do examinado a partir do histórico de itens anteriormente respondidos (WEISS; KINGSBURY, 1984). Desta forma, podem ser aplicados testes com maior flexibilidade e adaptabilidade, com significativa redução do tempo. Além disso, há correção imediata do teste e maior precisão em relação a testes que apresentam um número fixo de itens (OLEA; PONSODA; PRIETO, 1999).

São diversas as aplicações de CATs e, como exemplos, há testes na medicina, na educação, na administração de empresas, na psicologia e na indústria. Cada contexto tem seu objetivo bem definido. Por exemplo, um CAT na área de psicologia pode determinar se um indivíduo possui traços de ansiedade ou uma tendência para algum tipo de disfunção psicológica. No campo empresarial, pode ajudar na gestão do conhecimento da empresa, aplicando-se questionários para um grupo de coordenadores, por exemplo.

Tradicionalmente, os resultados obtidos em avaliações educacionais caracterizam-se por um escore padronizado, o que se mostra útil na comparação da posição relativa do desempenho de um indivíduo no grupo ao qual pertence. Porém, com o modelo de escore padronizado, não é possível realizar comparações de resultados entre diferentes provas aplicadas a diferentes grupos, aspecto importante em cenários de avaliação educacional em larga escala. Para isso, a TRI permite, por exemplo, a comparação entre grupos em instantes diferentes.

A TRI propõe uma modelagem estatístico-matemática para as características latentes do indivíduo, também chamadas de traços latentes, proficiências ou habilidades[2], e para os parâmetros associados aos itens. No contexto educacional, a teoria busca relacionar a probabilidade de um examinado responder corretamente a um item, dada sua habilidade.

Um item pode conter opções de resposta dicotômica (em que apenas uma opção de resposta é a correta e as demais, incorretas), politômicas (por exemplo, quando são consideradas mais de duas opções de resposta como corretas) e dissertativas. Neste caso, itens que apresentam resposta dissertativa devem ser corrigidos de acordo com alguma graduação, o que implica em uma resposta dicotômica ou politômica.

Um requisito para a aplicação da TRI em CATs é a existência de um banco de itens calibrado. A calibração é a estimação dos parâmetros dos itens e ocorre antes da aplicação de um teste formal. Os dados para a calibração são obtidos a partir da aplicação de um pré-teste em uma amostra de examinados da população em questão. Se o teste for definido para respostas dicotômicas, então, a partir dos acertos e dos erros dos examinados, determinam-se os parâmetros de cada item. Além da calibração, dois pontos fundamentais da TRI são a estimação da habilidade θ do examinado e o modelo de resposta a ser considerado. Klein

[2] Termo mais empregado na área de testes adaptativos educacionais.

(2013) faz um amplo estudo sobre a influência que determinados modelos têm sobre a estimação das habilidades.

A TRI assume modelos de resposta ao item, que estabelecem uma relação entre a habilidade do examinado e a sua probabilidade de acerto. Os modelos são compostos, essencialmente, por parâmetros individuais (ou do examinado), parâmetros de itens e uma função que relaciona esses parâmetros com a probabilidade das possíveis respostas aos itens. Procedimentos de inferência estatística são utilizados para obtenção de estimadores desses parâmetros a partir das respostas obtidas aos itens.

A escolha dos modelos e dos métodos da TRI depende diretamente do objetivo do teste, do tamanho do banco de itens e do desempenho computacional, por envolver métodos numéricos. Do ponto de vista do número de habilidades consideradas, a TRI pode ser dividida em duas classes (PITON-GONÇALVES, 2012):

1. Teoria de Resposta ao Item Unidimensional (ou simplesmente TRI): define uma única habilidade associada ao examinado e, consequentemente, os modelos e os métodos envolvidos também são unidimensionais (ou univariados). Do ponto de vista educacional, a habilidade θ pode representar, por exemplo, a proficiência em Matemática.

2. Teoria de Resposta ao Item Multidimensional (**MIRT**): define um vetor $\boldsymbol{\theta}$ de habilidades associado ao examinado. Do ponto de vista educacional, pode-se, por exemplo, interpretar um vetor habilidade $\boldsymbol{\theta}$ tridimensional, como a proficiência em álgebra, aritmética e geometria, no espaço "conhecimento de Matemática".

Na concepção moderna de CAT baseado na TRI, este caracteriza-se por selecionar os itens segundo os níveis de habilidade do examinado, "individualizando" um teste. Nesse contexto, o CAT baseado na TRI divide-se essencialmente em duas classes (PITON-GONÇALVES, 2012):

1. O Teste Adaptativo Computadorizado Unidimensional (em inglês, *Unidimensional Computer Adaptive Test* – UCAT) é um CAT baseado na Teoria de Resposta ao Item Unidimensional.

2. O Teste Adaptativo Computadorizado Multidimensional (MCAT) é um CAT baseado na Teoria de Resposta ao Item Multidimensional.

A partir da relação de teses defendidas[3] no Brasil e de nossas buscas sistemáticas no Google, foi encontrada somente uma tese de doutorado em MCAT (PITON-GONÇALVES, 2012). As próximas seções detalham os princípios e os métodos relacionados ao MCAT, tema central deste artigo.

5 Teste Adaptativo Computadorizado Multidimensional

O trabalho de Zukowsky-Tavares (2013) fomenta uma discussão entre as abordagens uni e multidimensionais da TRI, com forte crítica sobre a premissa da unidimensionalidade em seus modelos/critérios, em avaliações educacionais. Para a autora, "[...] os modelos multidimensionais da Teoria da Resposta ao Item ou a Teoria da Resposta ao Item Multidimensional (Trim) poderiam ser uma resposta para esse desafio (ZUKOWSKY-TAVARES, 2013 p. 74)".

Porém, ao refletir sobre questões de aplicação da multidimensionalidade, a mesma autora frisa que "[...] isso ocorrerá apenas quando os modelos multidimensionais deixarem o estatuto de 'teoria' e forem aplicáveis na prática (ZUKOWSKY-TAVARES, 2013 p. 74)". Nesse sentido, os trabalhos de Piton-Gonçalves e Aluísio (2012) e Piton-Gonçalves (2012) buscam preencher esta lacuna da literatura apontada por Zukowsky-Tavares (2013). O MCAT supõe que o examinado possua um construto mental, representado por vários traços latentes estimados pela MIRT (RECKASE, 1985). No caso de múltiplas habilidades, o domínio de conhecimento Física pode ser interpretado, por exemplo, como a composição das habilidades em Cinemática, Ótica e Eletromagnetismo. Neste exemplo, diz-se que o espaço das habilidades possui dimensão três, ou seja, a habilidade multidimensional é representada pelo vetor $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$.

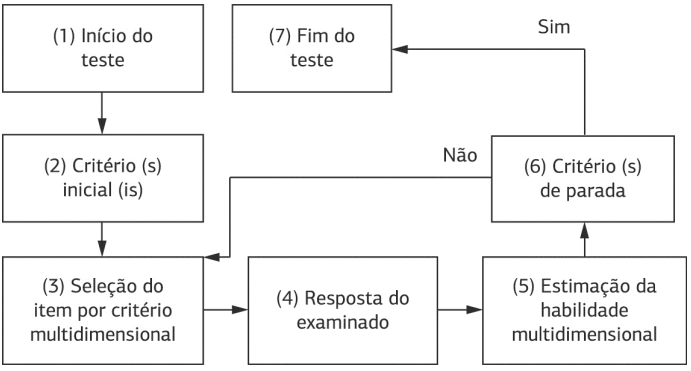
Essencialmente, um MCAT requer (i) um modelo de resposta multidimensional, (ii) um banco de itens calibrado a partir de um método multidimensional, (iii) um método para estimar as habilidades dos examinados, (iv) um critério de seleção de itens, (v) um critério inicial e (vi) um critério de parada do teste. O fluxograma da Figura 3 mostra esquema geral de funcionamento de um MCAT.

Os passos para a execução de um MCAT são:

I. Inicia-se o teste (1) e aplica-se um ou mais critérios iniciais (2).

[3] Tese e Dissertações catalogadas no Banco de Teses CAPES – http://capesdw.capes.gov.br

Figura 3 - Fluxograma geral de um MCAT.



Fonte: Piton-Gonçalves (2012).

II. Aplica-se um critério multidimensional de seleção de item (3). Exibe-se o item e o examinado responde ao item (4).

III. Estima-se multidimensionalmente a habilidade do examinado (5) e decide-se de acordo com a satisfação do(s) critério(s) de parada (6). Se satisfeito, então fim de teste (7). Se não, vá ao passo (II).

6 Modelo Logístico Multidimensional de Três Parâmetros

O Modelo Logístico Multidimensional de Três Parâmetros (MLM3P) é o modelo de resposta ao item mais adequado em contextos educacionais, pois considera o "chute" do examinado, ação muito comum em avaliações educacionais. Seja o vetor $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p) \in \Re^p$ e p a dimensão no espaço das habilidades, a função de probabilidade de acerto para um i-ésimo item aplicado é (HATTIE, 1981, p. 68), conforme a Equação 1:

$$P_i(\boldsymbol{\theta}) = P(U_i = 1 \mid \boldsymbol{\theta}, \mathbf{a}_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + e^{-D\,\mathbf{a}_i(\boldsymbol{\theta}^t - b_i\mathbf{1})}}, \quad (1)$$

em que:

- U_i é a variável randômica que corresponde à resposta do *i*-ésimo item. Para o caso dicotômico (em que apenas uma opção de resposta é a correta e as demais incorretas), $U_i = 1$ para resposta correta e $U_i = 0$, caso contrário.

- $\mathbf{a}_i = (a_1,...,a_p)$ é o vetor $1 \times p$ parâmetro discriminação do *i*-ésimo item e relaciona-se com a inclinação da *Superfície de Resposta ao Item* gerada pela Equação 1, representada bidimensionalmente nas Figuras 3, 4 e 5. De acordo com Reckase e McKinley (1991), o *Poder de Discriminação do Item* (MDISC), que resume a interpretação do vetor $\mathbf{a}_i$, é definido pela Equação 2:

$$MDISC_i = \| \mathbf{a}_i \| = \sqrt{\sum_{j=1}^{p} a_{i_j}^2}, \quad (2)$$

em que:

- O MDISC mede a intensidade da variação da região de baixa probabilidade de acerto para a alta probabilidade do modelo de resposta. Dessa maneira, um item com poder de discriminação ideal é aquele que todos os examinados de baixa habilidade erram e todos de alta habilidade acertam.

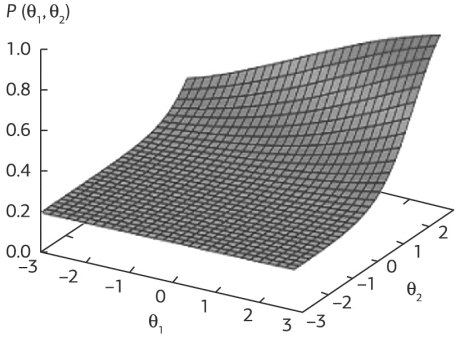
- $b_i\mathbf{1}$ é o vetor $p \times 1$ do parâmetro dificuldade b_i do i-ésimo item e **1** é o vetor coluna de 1's.

- c_i representa a probabilidade (de acerto casual) de obter-se uma resposta correta mediante um examinado com baixa habilidade. No cenário educacional é chamado de "chute" ou adivinhação.

- *D* é um fator igual a 1.7 quando a função de probabilidade é, aproximadamente, a da ogiva normal ou 1.0 para a logística.

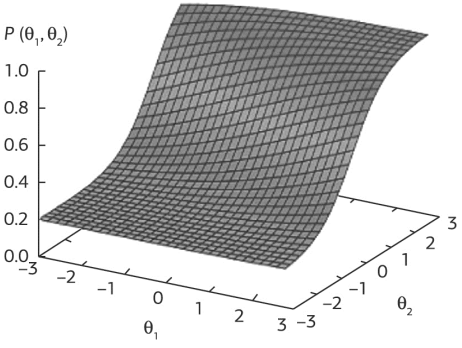
As Figuras 4, 5 e 6 representam a superfície de resposta para três itens que se diferenciam pelo parâmetro dificuldade. Por exemplo, se a habilidade do examinado for o vetor $\boldsymbol{\theta}_e = (1.2,0.2)$, então a probabilidade de acerto é de aproximadamente 24% para a superfície de resposta ao item descrita na Figura 4. Considerando a superfície de resposta descrita na Figura 5, a probabilidade de acerto é de aproximadamente 29% em $\boldsymbol{\theta}_e$. Quando se diminui a dificuldade do item para $b = -1.5$ (Figura 6), a probabilidade de acerto aumenta para aproximadamente 98% em $\boldsymbol{\theta}_e$.

Figura 4 - Superfície de resposta para um item com parâmetros **a** = (0.5,1.5), *b* = 2.0 e *c* = 0.2



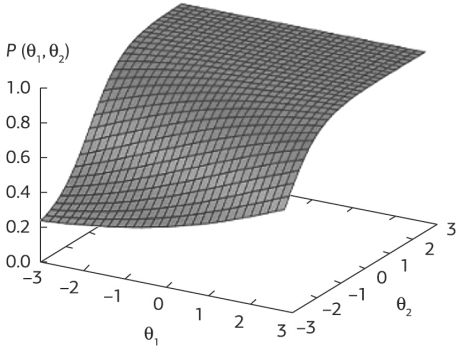
Fonte: Piton-Gonçalves (2012).

Figura 5 - Superfície de resposta para um item com parâmetros **a** = (0.5,1.5), *b* = 1.5 e *c* = 0.2



Fonte: Piton-Gonçalves (2012).

Figura 6 - Superfície de resposta para um item com parâmetros **a** = (0.5,1.5), *b* = 1.5 e *c* = 0.2



Fonte: Piton-Gonçalves (2012).

7 Banco de itens

Com o objetivo de organizar os itens de teste, o banco de itens é um banco de dados que contém itens e outros dados associados, tais como as respostas corretas (para o caso objetivo), dados estatísticos (por exemplo: número de estudantes que acertaram/erraram o item, nota média dos examinados que responderam ao item), *links* associados ao item (por exemplo, um texto, vídeo ou áudio), identificadores (ID) do item, parâmetros psicométricos (por exemplo, o nível de dificuldade do item), marcadores (*tags*) dos diferentes conteúdos/temas do item, entre outros (PITON-GONÇALVES, 2012).

A calibração de itens é um procedimento caro e computacionalmente lento, que deve ser realizado e analisado com antecedência à aplicação do MCAT, incluindo a análise de dimensionalidade do banco. Um *software* (comercial) para calibração multidimensional é o TESTFACT[4].

8 Estimação da habilidade

Estimar a habilidade, após cada resposta do examinado em um teste, possibilita a sua continuidade, provendo um $\boldsymbol{\theta}$ provisório até o término do teste. $\boldsymbol{\theta}$ pode ser estimado pelos três métodos mais conhecidos na literatura, que são a Máxima Verossimilhança (em inglês, *Maximum Likelihood* – ML), o Máximo *a Posteriori* (em inglês, *Bayesian Maximum a Posteriori* – MAP) e a Esperança *a Posteriori* (em inglês, *Bayesian Expected a Posteriori* – EAP).

Klein (2013) afirma que o método EAP é o mais utilizado na TRI. Porém, quando se trata de MCAT, deve-se considerar o tempo computacional de processamento, uma vez que a habilidade é estimada após cada item administrado. Nesse sentido, Chen (2009) afirma que a EAP é impraticável em dimensões mais altas devido ao alto tempo computacional de processamento, o que não ocorre com o método MAP.

O método MAP aplicado em MCAT é amplamente discutido no trabalho de Segall (1996). A função de verossimilhança[5] (SEGALL, 1996) é dada pela Equação 3:

[4] http://www.ssicentral.com

[5] Adota-se aqui a mesma notação de Segall (1996, 2000), em que $L(\boldsymbol{u} \mid \boldsymbol{\theta})$ é equivalente a $L(\boldsymbol{\theta} \mid \boldsymbol{u})$ na notação usual.

$$L(\mathbf{u} \mid \boldsymbol{\theta}) = \prod_{i \in \mathbf{v}} P_i(\boldsymbol{\theta})^{u_i} Q_i(\boldsymbol{\theta})^{1-u_i}, \quad (3)$$

em que parte do vetor respostas observadas **u**. $P_i(\boldsymbol{\theta})$ é definida pela escolha de um modelo de resposta ao item, como, por exemplo, a Equação 1. $Q_i(\boldsymbol{\theta}) = 1 - P_i(\boldsymbol{\theta})$ e **v** é o vetor que contém os identificadores dos itens administrados. A igualdade provém da suposição de independência entre as respostas condicionada às habilidades do indivíduo. Assim, $\boldsymbol{\theta}$ estimado será o valor que maximizará $L(\mathbf{u} \mid \boldsymbol{\theta})$, ou seja, a máxima verossimilhança é dada pela solução da Equação 4:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\mathbf{u} \mid \boldsymbol{\theta}) = \mathbf{0} \quad (4)$$

para *p* habilidades.

A estimação Bayesiana (modal) de $\hat{\boldsymbol{\theta}}$ proporciona o valor que maximiza a função densidade *a posteriori* $f(\mathbf{u} \mid \boldsymbol{\theta})$, em que **u** é o vetor de respostas do examinado. A estimação bayesiana utilizada por Segall (1996) é consonante com o Teorema de Bayes. Assim, a função densidade *a posteriori* de θ é dada pela Equação 5:

$$f(\boldsymbol{\theta} \mid \mathbf{u}) = L(\mathbf{u} \mid \boldsymbol{\theta}) \frac{f(\boldsymbol{\theta})}{f(\mathbf{u})}. \quad (5)$$

$f(\boldsymbol{u})$ é a probabilidade marginal de $\boldsymbol{u}$. $f(\boldsymbol{\theta})$ é a distribuição *a priori* de $\boldsymbol{\theta}$ e, no campo educacional, é adotada por Segall (1996) como a distribuição Normal Multivariada $N(\boldsymbol{\mu}, \boldsymbol{\Phi})$. A estimação pontual de $\boldsymbol{\theta}$ pode ser obtida a partir de medidas resumo de $f(\boldsymbol{\theta} \mid \boldsymbol{u})$, como, por exemplo, o máximo ou a esperança. $f(\boldsymbol{\theta} \mid \boldsymbol{u})$ é maximizada de acordo com a Equação 6:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{\theta} \mid \mathbf{u}) = \mathbf{0}, \quad (6)$$

para *p* habilidades. Utilizando o Método de Newton em várias variáveis, $\boldsymbol{\theta}^{(j+1)}$ é calculado iterativamente segundo a Equação 7:

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - [\mathbf{H}(\boldsymbol{\theta}^{(j)})]^{-1} \times \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{\theta}^{(j)} \mid \mathbf{u}), \quad (7)$$

em que $\boldsymbol{H}(\boldsymbol{\theta}^{(j)})$ é a matriz Hessiana $p \times p$ avaliada na *j*-ésima instância. Segall (2000) sugere que a aproximação inicial do Método de Newton para $\boldsymbol{\theta}^{(j)}$ seja a habilidade provisória estimada na instância k – 1 do teste, diminuindo assim os problemas de convergência da solução do sistema não linear.

9 Seleção de itens

Desde a última década, a literatura vem apontando para critérios de seleção de itens que interpretem o nível da informação do teste ou do item. O objetivo é selecionar o item que maximiza a informação de acordo com a habilidade estimada, sendo que a seleção pode ser por Informação Local (SEGALL, 1996) ou por Informação Global (CHANG; YING, 1996; WANG; CHANG; BOUGHTON, 2011).

9.1 Seleção por Informação Local

Para a Informação Local, o destaque é o critério baseado nos princípios da abordagem Bayesiana. O *Maior Decremento no Volume do Elipsoide de Confiança Bayesiano* (em inglês, *Largest Decrement in the Volume of the Bayesian Credibility Ellipsoi*d – B^{γ}) é um critério de seleção de itens proposto por Segall (1996). Neste, seleciona-se o *k*-ésimo item que satisfaz a Equação 8:

$$\arg\max_{i_k \in R_k}[V_{i_k}] = \arg\max_{i_k \in R_k}[I(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k) + I(\boldsymbol{\theta}, u_{k+1}) + \boldsymbol{\Phi}^{-1}] \tag{8}$$

para o *i*-ésimo item do banco no *k*-ésimo item administrado pelo teste. $I(\cdot,\cdot)$ é a matriz de Informação de Fisher, $\boldsymbol{\Phi}^{-1}$ é a inversa da matriz de covariância de $N(\boldsymbol{\mu},\boldsymbol{\Phi})$ e R_k é um conjunto de itens do banco ainda não administrados no teste até o passo *k*. A distribuição condicional do estimador $\boldsymbol{\theta}_k$ é assintoticamente normal, relacionada com a matriz de Informação de Fisher (RECKASE, 2009).

9.2 Seleção por Informação Global

A Informação Global em critérios de seleção de itens é baseada nos princípios da medida de informação de *Kullback-Leibler* (ou Entropia Relativa) (KULLBACK; LEIBLER, 1951). A Informação de Kullback-Leibler entre duas distribuições de probabilidade $f(X)$ e $g(X)$, em notação moderna, é definida pela Equação 9:

$$K(f,g) = E_f\left[\log\frac{f(X)}{g(X)}\right]. \tag{9}$$

Os trabalhos de Mulder e van der Linden (2010), Wang, Chang e Boughton (2011), Wang e Chang (2011) e Chang e Ying (1996) discutem e propõem critérios bayesianos para a seleção de itens baseados em Kullback-Leibler para

espaços uni e multidimensionais, com estudos categorizados em habilidades intencionais (predeterminadas) ou ruidosas (não previstas). As próximas seções detalham alguns destes critérios de seleção.

9.3 Índice de Kullback-Leibler

Chang e Ying (1996) aplicam a Equação 9 na seleção de itens em UCAT e, mais tarde, Veldkamp e van der Linden (2002) propuseram uma versão multidimensional. O Índice de Kullback-Leibler é definido pela Equação 10:

$$KI_{i_k}(\hat{\boldsymbol{\theta}}_{k-1}) = \int K_{i_k}(\hat{\boldsymbol{\theta}}_{k-1}, \boldsymbol{\theta})\, d\boldsymbol{\theta}. \quad (10)$$

Em aplicações reais, θ é desconhecido e, neste caso, utilizam-se os métodos de estimação usuais (WANG; CHANG; BOUGHTON, 2011). O *k*-ésimo item selecionado é aquele em que $\arg\max_{i_k \in R_k}[KI_{i_k}(\hat{\boldsymbol{\theta}}_{k-1})]$.

9.4 Informação Esperada de Kullback-Leibler

Veldkamp e van der Linden (2002) propuseram uma versão bayesiana para o Índice de Kullback-Leibler. A partir da Equação 9, K^B é expresso pela Equação 11:

$$K^B_{i_k}(\hat{\boldsymbol{\theta}}_{k-1}) = \int K_{i_k}(\hat{\boldsymbol{\theta}}_{k-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1})\, d\boldsymbol{\theta}, \quad (11)$$

em que $\mathbf{u}_{k-1}$ é o vetor de respostas dos *k*–1 itens anteriormente administrados no teste, $\hat{\boldsymbol{\theta}}_{k-1}$ é a estimativa de θ *a posteriori* após *k*–1 itens respondidos pelo examinado e $f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1})$ é a distribuição *a posteriori* para $\boldsymbol{\theta}$ após *k*–1 itens respondidos. Como θ é desconhecido, a seleção do próximo item terá como base a informação esperada *a posteriori* para a resposta do *i*-ésimo item do banco após *k*–1 itens do teste administrados. O *k*-ésimo item selecionado é aquele em que $\arg\max_{i_k \in R_k}[K^B_{i_k}(\hat{\boldsymbol{\theta}}_{k-1})]$.

9.5 Informação Mútua

A *Informação Mútua* pode ser interpretada como a redução da incerteza da distribuição *a posteriori* de $\boldsymbol{\theta}$ por conhecimento da distribuição da resposta ao item em questão. A informação mútua é definida (MULDER; VAN DER LINDEN, 2010) pela Equação 12:

$$I_M(\boldsymbol{\theta}, u_{i_k}) = \sum_{u_{i_k}=0}^{1} \int f(\boldsymbol{\theta}, u_{i_k} \mid \mathbf{u}_{k-1}) \log \frac{f(\boldsymbol{\theta}, u_{i_k} \mid \mathbf{u}_{k-1})}{f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1}) f(u_{i_k} \mid \mathbf{u}_{k-1})} \, d\boldsymbol{\theta}. \quad (12)$$

O *k*-ésimo item selecionado na Informação Mútua é aquele em que $\arg\max_{i_k \in R_k} I_M(\boldsymbol{\theta}, u_{i_k})$. Detalhes dos critérios K^B e Informação Mútua são encontrados nos trabalhos de Wang, Chang e Boughton (2011) e Wang e Chang (2011).

9.6 Kullback Leibler entre Posteriores Subsequentes

Se uma distribuição *a posteriori* de $\boldsymbol{\theta}$ não se altera significativamente após um item administrado, é razoável que deva ser evitado selecionar um item difícil para um examinado com baixa habilidade (MULDER; VAN DER LINDEN, 2010). Os autores formalizam a informação de *Kullback Leibler entre Posteriores Subsequentes* (em inglês, *Kullback-Leibler between Subsequent Posteriors* -K^P) de acordo com a Equação 13:

$$K^P_{i_k}[f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1})] = \sum_{u_{i_k}=0}^{1} f(u_{i_k} \mid \mathbf{u}_{k-1}) K(f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1}), f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1}, u_{i_k})). \quad (13)$$

$K(f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1}), f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1}, u_{i_k}))$ é a informação de Kullback-Leibler entre duas densidades *a posteriori* para o *k*-ésimo item administrado e $\mathbf{u}_{k-1}$ é o vetor resposta aos *k*–1 itens administrados anteriormente. $f(u_{i_k} \mid \mathbf{u}_{k-1})$ é a função de probabilidade preditiva *a posteriori*, com ($u_{i_k} = 0,1$) para itens dicotômicos. O item selecionado dentre os *i*-ésimos itens disponibilizados para seleção é aquele em que $\arg\max_{i_k \in R_k} K^P_{i_k}[f(\boldsymbol{\theta} \mid \mathbf{u}_{k-1})]$.

10 Critérios iniciais

O primeiro item selecionado ao examinado em um MCAT pode seguir dois cenários (PITON-GONÇALVES, 2012). O primeiro, *sem informação prévia*, é o cenário em que todos os examinados são considerados do mesmo nível inicial, uma vez que este nível é desconhecido. Neste caso, alguns critérios podem ser utilizados: (i) fixar um nível inicial do examinado e selecionar o item, (ii) separar um conjunto de itens que possuam um mesmo nível de dificuldade e escolher o primeiro item aleatoriamente (PARSHALL et al., 2002) e (iii) selecionar alguns itens de nível fácil, com o propósito do examinado se "aquecer" (PARSHALL et al., 2002) e depois partir para os itens formais.

O segundo cenário é com *informação prévia,* em que se utiliza alguma informação *a priori* sobre o examinado ou o grupo de examinados. No campo unidimensional, o método Dados Colaterais (PARSHALL et al., 2002) baseia-se no conhecimento de escores ou informações anteriores do examinado, tais como o currículo escolar ou o escore obtido em um exame anterior. Na literatura, não foram encontrados critérios multidimensionais que contenham a *informação prévia*.

11 Critérios de parada

Um elemento fundamental em um MCAT é o critério de parada, que dependerá dos objetivos do teste, dos modelos e métodos estatísticos adotados, do estresse do examinado, dentre outros fatores. Um teste adaptativo poderá finalizar quando (LINACRE, 2000) (i) há o esgotamento de itens do banco, o que geralmente ocorre com bancos de itens pequenos para testes longos, (ii) o número pré-determinado de itens ao examinado é atingido (recomenda-se que o número máximo seja igual a um teste equivalente via lápis e papel), (iii) o examinado apresentar algum comportamento anormal, como, por exemplo, responder a um item muito rapidamente ou muito lentamente, (iv) o erro padrão da estimação estar abaixo de um valor pré-definido, (v) quando a medida de informação do item deixar de existir para um examinado (SIMMS; CLARK, 2005) e (vi) quando a Equação 8 for satisfeita (para o caso de seleção via abordagem bayesiana). No último caso, após cada item administrado, o volume é computado e, se for menor que um valor pré-fixado, o teste termina (RECKASE, 2009).

Um teste adaptativo deve ser íntegro quanto a sua aplicação e seus dados e, por isso, não poderá terminar antes (LINACRE, 2000):

- do examinado responder a um número mínimo de itens. Em algumas situações, se um examinado responder poucos itens, poderá causar um sentimento de injustiça e acarretar argumentos como "eu apenas tive falta de sorte no início do teste, se me dessem mais questões, meu resultado seria diferente";
- de todos os tópicos do teste serem cobertos. Isso ocorre em testes que possuem mais de um tópico, em que é necessário que o examinado seja avaliado em todos os tópicos;
- que um número suficiente de itens tenha sido administrado, mantendo a validade estatística do teste.

A próxima seção abordará MCATs operacionais que podem ser aplicados com usuários reais em testes educacionais.

12 Sistemas computadorizados

Em termos de MCATs operacionais, ou seja, aqueles testes que podem ser aplicados com usuários reais em cenários educacionais, a revisão da literatura aponta para três trabalhos:

- O *Multidimensional Adaptive Test for Educational Purposes* (MADEPT) (PITON-GONÇALVES, 2012) é um sistema Web desenvolvido em PHP, MySQL, Linguagem R e Mathjax. O MADEPT segue a abordagem *Computer-based Multidimensional Adaptive Testing* (CBMAT) (PITON-GONÇALVES; ALUÍSIO, 2012), que é uma arquitetura computacional distribuída em seis módulos: Administrador, MCAT, Corretude, Base de Dados, Integração e Item. Segundo os autores, estes módulos são os componentes necessários para o êxito no projeto e no desenvolvimento computacional de um MCAT operacional. O MADEPT contempla essencialmente o Modelo Logístico Multidimensional de Três Parâmetros em qualquer dimensão (incluindo a unidimensional), a estimação bayesiana (modal) de Segall (1996) e a seleção de itens é por critérios do Maior Decremento no Volume do Elipsoide de Confiança Bayesiano e Kullback Leibler entre Posteriores Subsequentes. O banco de itens é composto por 25 itens da 5.ª série do Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP) do ano de 2005, da área de Matemática, calibrados no *software* TESTFACT, culminando em uma dimensão p=2.

- Senarat et al. (2013) desenvolveram um MCAT para diagnosticar o processo cognitivo na aprendizagem da álgebra elementar para estudantes escolares. Desenvolvido em *Visual Basic*, este MCAT utiliza o modelo ogiva normal com estimação bayesiana e da seleção do próximo item por meio do *Maior Decremento no Volume do Elipsoide de Confiança Bayesiano*, em um banco com 163 itens.

- O trabalho de Wang, Kuo e Chao (2010) trata da implementação, da modelagem e da validação (em duas dimensões) de um MCAT que avalia quanto à língua chinesa. A seleção de itens é por meio do método da máxima verossimilhança, o modelo de resposta é do modelo logístico de dois parâmetros e o teste termina quando atingir um determinado número (fixo) de itens administrados.

13 Considerações finais

O presente artigo trouxe subsídios teóricos e metodológicos para fomentar um debate, no âmbito nacional, sobre testes computadorizados para avaliações educacionais. Estas, cada vez mais, priorizam avaliações de desempenho do examinado em múltiplas competências e habilidades, premissas estas que o UCAT não considera. Neste caso, o MCAT mostra-se como uma possível solução.

Tanto a pesquisa quanto o desenvolvimento de MCATs que possam ser aplicados no cotidiano ainda são incipientes nos cenários nacional e internacional, conforme a revisão da literatura. Além dos benefícios de aplicação e correção automatizadas que um teste computadorizado possui, os testes adaptativos apresentam as seguintes vantagens:

1. O desempenho do examinado determina o seu próprio elenco de itens administrados, quando responde correta ou incorretamente a um item.

2. Como o número de itens administrados é menor do que os testes tradicionais, então o tempo de teste é reduzido. Isso implica na redução da fadiga do examinado em testes longos, sem a necessidade de que todos os examinados realizem simultaneamente o exame (COSTA, 2009), caso o teste seja baseado na TRI.

3. Quando o teste é baseado na TRI ou na MIRT, possibilitam-se a análise e a comparação dos resultados de diferentes grupos de examinados, a partir de diferentes itens.

4. Um UCAT ajusta adequadamente o nível de dificuldade dos itens, sem prejudicar a acurácia das estimativas (COSTA, 2009).

Algumas características do MCAT são destacadas (PITON-GONÇALVES, 2012; PITON-GONÇALVES; ALUÍSIO, 2012):

1. A elaboração e a calibração multidimensional do banco de itens requerem conhecimento especialista do conteúdo específico, do educacional e do estatístico-matemático da MIRT.

2. A implementação de um sistema computacional para usuários reais é muito custosa e trabalhosa, e exige o detalhamento e a compreensão das teorias,

dos modelos e dos métodos numéricos no campo multidimensional. A análise e a interpretação pedagógica dos resultados e dos parâmetros de itens em um MCAT com propósitos educacionais possuem as mesmas dificuldades que os apresentados pela MIRT.

3. O desenvolvimento e a manutenção de um UCAT são caros (VELDKAMP; MATTEUCCI, 2013), e no MCAT, é ainda mais caro.

4. Os algoritmos que calculam numericamente equações em espaços multidimensionais (ou multivariados) de forma otimizada devem ser projetados de forma que o processamento seja correto, seguro e rápido, assegurando que todos os resultados sejam confiáveis e fidedignos, incluindo estudos de convergência numérica das soluções das equações no campo multidimensional. Ressalta-se que todos os cálculos envolvidos nos critérios de estimação e seleção são realizados em tempo de execução, fato que não permite falhas ou erros numéricos do sistema. O trabalho de Piton-Gonçalves e Aluísio (2012) aborda essa questão na ótica matemático-computacional.

5. Os estudos de Piton-Gonçalves e Aluísio (2012) mostram que, devido à abordagem multidimensional, o tempo computacional de processamento numérico passa a ser considerado, uma vez que o examinado não deve aguardar um longo período para receber o próximo item.

6. Quando se tem um MCAT composto por um banco de itens grande e que mede a habilidade em altas dimensões para testes longos, o tempo computacional de processamento poderá ser muito grande. Neste caso, *clusters* e processamento paralelizado são indicados, com o cuidado na implementação dos algoritmos paralelizados, que podem aumentar substancialmente os erros numéricos.

7. Dependendo da metodologia adotada para o teste, a inserção e a remoção de itens do banco são procedimentos caros, pois envolvem uma análise criteriosa dos itens. Esse processo envolve especialistas de conteúdo e métodos estatísticos multivariados para avaliação dos itens.

Apesar das limitações e dificuldades inerentes ao MCAT, esse se mostra como um caminho para a aplicação de testes computadorizados que mais se aproximam das múltiplas habilidades e conhecimentos que o examinado possui, administrando um teste com acurácia e mais curto do que os tradicionais.

Referências

CHANG, H.-H.; YING, Z. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, St. Paul, v. 20, n. 3 , p. 213-229, 1996. http://dx.doi.org/10.1177/014662169602000303

CHEN, P. H. Comparison of adaptive bayesian estimation and weighted bayesian estimation in multidimensional computerized adaptive testing. In: CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING (GMAC), 2009. *Proceedings*...

COSTA, D. R. *Métodos Estatísticos em Testes Adaptativos Informatizados*. 2009. 107 f. Dissertação (Mestrado)-Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

HATTIE, J. *Decision criteria for determining unidimensionality*. Tese (Doutorado)-University of Toronto, Canada, 1981.

KLEIN, R. Alguns Aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências. *Estudos em Avaliação Educacional*, São Paulo, v. 21, n. 78, p. 35-56, jan./mar. 2013.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, [S.l.], v. 22, n. 1, p. 79-86, 1951. http://dx.doi.org/10.1214/aoms/1177729694

LINACRE, J. M. Computer-Adaptive Testing: a methodology whose time has come. In: CHAE, S. et al. *Development of computerized middle school achievement test*. Seoul: Komesa Press, 2000.

MULDER, J.; VAN DER LINDEN, W. Elements of adaptive testing. In: VAN DER LINDEN, W. J.; GLAS, C. A. W. *Multidimensional Adaptive Testing with Kullback-Leibler information item selection*. New York: Springer, 2010. p. 77-101.

OLEA, J.; PONSODA, V.; PRIETO, G. *Tests informatizados fundamentos y aplicaciones*. [S.l.]: Ediciones Pirámede, 1999.

PARSHALL, C. G. et al. *Practical considerations in computer-based testing*. New York: Springer-Verlag, 2002.

PITON-GONÇALVES, J. *Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais*. 2012. 153 f. Tese (Doutorado)-Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012.

PITON-GONÇALVES, J.; ALUÍSIO, S. M. An architecture for multidimensional computer adaptive test with educational purposes. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB (WebMedia '12), 18., 2012, New York, NY. *Proceedings*... New York: ACM, 2012. p. 17-24.

RECKASE, M. D. An interactive computer program for tailored testing based on the oneparameter logistic model. *Behaviour Reserach Methods and Instrumentation*, Austin, v. 6, n. 2, p. 208-212, 1974. http://dx.doi.org/10.3758/BF03200330

RECKASE, M. D. *Multidimensional item response theory*. New York: Springer, 2009.

RECKASE, M. D. The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, St. Paul, v. 9, n. 4, p. 401-412, 1985. http://dx.doi.org/10.1177/014662168500900409

RECKASEM, D.; MCKINLEY, L. The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, St. Paul, v. 15, n. 4, p. 361-373, 1991. http://dx.doi.org/10.1177/014662169101500407

SEGALL, D. O. Multidimensional adaptive testing. *Psychometrika*, United States, v. 61, n. 2, p. 331-354, 1996. http://dx.doi.org/10.1007/BF02294343

SEGALL, D. O. Computerized adaptive testing: theory and practice. In: VAN DER LINDEN, W. J.; GLAS, C. A. W. (Ed.). *Principles fo multidimensional adaptive testing*. New York: Kluwer Academic Publishers, 2000. p. 53-73.

SENARAT, S. et al. Development of a computerized adaptive testing for diagnosing the cognitive process of grade 7 students in learning algebra, using

multidimensional item response theory. *Educational Research and Reviews*, Kenya, v. 8, n. 13, p. 1009-1021, jul. 2013.

SIMMS, L. J.; CLARK, L. A. Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment*, Arlington, v. 17, n. 1, p. 28-43, 2005. http://dx.doi.org/10.1037/1040-3590.17.1.28 PMid:15769226

VELDKAMP, B. P.; MATTEUCCI, M. Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Publicas em Educação*, Rio de Janeiro, v. 21, n. 78, p. 57-82, 2013.

VELDKAMP, B. P.; VAN DER LINDEN, W. J. Multidimensional adaptive testing with constraints on the test content. *Psychometrika*, United States, v. 67, n. 4, p. 575-588, 2002. http://dx.doi.org/10.1007/BF02295132

WANG, C.; CHANG, H.-H. Item selection in multidimensional computerized adaptive testing-gaining information from different angles. *Psychometrika*, United States, v. 76, n. 3, p. 363-384, 2011. http://dx.doi.org/10.1007/s11336-011-9215-7

WANG, C.; CHANG, H.-H.; BOUGHTON, K. A. Kullback-leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, United States, v. 76, n. 1, p. 13-19, 2011. http://dx.doi.org/10.1007/s11336-010-9186-0

WANG, H.-P.; KUO, B.-C.; CHAO, R.-C. A multidimensional computerized adaptive testing system for enhancing the chinese as second language proficiency test. In: WSEAS INTERNATIONAL CONFERENCE ON EDUCATION AND EDUCATION TECHNOLOGY (EDU'10), 9., 2010. *Proceedings*...

WEISS, D. J. Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, Washington, v. 53, n. 6, p. 774-789, 1985. http://dx.doi.org/10.1037/0022-006X.53.6.774 PMid:3841355

WEISS, D. J.; KINGSBURY, G. G. Application of computerized adaptive testing to educational problems. *Journal of Education Measurement*, United Kingdom, v. 21, n. 4, p. 361-375, 1984. http://dx.doi.org/10.1111/j.1745-3984.1984.tb01040.x

ZUKOWSKY-TAVARES, C. Teoria de resposta ao item: Uma análise crítica dos pressupostos epistemológicos. *Estudos em Avaliação Educacional*, São Paulo, v. 24, n. 54, p. 56-76, jan./abr. 2013.

Multidimensional Computer Adaptive test with educational purposes: principles and methods

Abstract

The Multidimensional Computer Adaptive Test (MCAT) based on the Multidimensional Item Response Theory (MIRT) considers multiple skills in educational tests. The purpose of this paper is (i) to present theoretical and methodological support on the educational purposes of MCATs, and (ii) address the advantages and disadvantages of their application in educational tests. The literature review indicates that the MCAT is suitable for computerized tests with multiple skills, administering a smaller number of items than the traditional tests.

Keywords: *Multidimensional Computer Adaptive Test. Multidimensional Item Response Theory. Adaptive test. Educational assessment.*

Test Adaptativo Computarizado Multidimensional con finalidad educativa: principios y métodos

Resumen

El Test Adaptativo Computarizado Multidimensional (MCAT) basado en la Teoría de la Respuesta al Ítem Multidimensional (MIRT) considera las múltiples habilidades en las evaluaciones educativas. En este contexto, el presente artículo (i) presenta base teórica y metodológica sobre los MCATS educativos y (ii) se refiere a las ventajas y desventajas de la aplicación en escenarios educativos. Los resultados presentados en la literatura muestran que el MCAT es adecuado para pruebas computarizadas con múltiples habilidades, con un número menor de ítems del que tienen las pruebas tradicionales.

Palabras-clave: *Test Adaptativo Computarizado Multidimensional. Teoría de la Respuesta al Ítem Multidimensional. Test adaptativo. Evaluación educacional.*

Informações dos autores

Jean Piton-Gonçalves: Professor Adjunto da Universidade Federal de São Carlos - UFSCar. Doutor em Ciências da Computação e Matemática Computacional pela Universidade de São Paulo - USP. Contato: jpiton@ufscar.br

Sandra Maria Aluísio: Professora Efetiva do ICMC-USP. Doutora em Inteligência Artificial, mais especificamente, Processamento de Língua Natural, pela Universidade de São Paulo. Contato: sandra@icmc.usp.br