TESTAGEM ADAPTATIVA POR COMPUTADOR (CAT): ASPECTOS CONCEITUAIS E UM PANORAMA DA PRODUÇÃO BRASILEIRA

COMPUTERIZED ADAPTIVE TESTING (CAT): CONCEPTUAL ASPECTS AND A PANORAMA OF THE BRAZILIAN PRODUCTION

PRUEBAS ADAPTATIVAS COMPUTARIZADAS (CAT): ASPECTOS CONCEPTUALES Y UN PANORAMA DE LA PRODUCCIÓN BRASILEÑA

Alexandre José de Souza Peres

RESUMO

Este trabalho objetiva introduzir os aspectos conceituais da testagem adaptativa por computador (CAT), bem como oferecer um panorama da pesquisa brasileira na área. A CAT é uma estratégia de administração de testes e estimação de variáveis latentes por meio de algoritmos que, de maneira interativa, selecionam e apresentam itens ao testando de acordo com seu desempenho. Diferentemente da testagem convencional, na CAT diferentes itens poderão ser administrados para diferentes testando, de forma adaptada de acordo com as estimações do traço latente do sujeito realizadas durante a própria testagem. Neste trabalho, apresentamos um breve histórico da testagem adaptativa, além de discutir as vantagens da CAT sobre as estratégias convencionais de testagem. Também apresentamos os elementos constitutivos da CAT a partir da Teoria de Resposta ao Item, bem como elencamos as principais estratégias para seu desenvolvimento apontado na literatura. Por fim, relatamos os resultados de uma revisão da literatura nacional na área.

Palavras-chave: psicometria; testagem adaptativa computadorizada CAT; testagem adaptativa informatizada TAI; avaliação educacional; avaliação psicológica.

ABSTRACT

This paper aimed to introduce the conceptual aspects of computerized adaptive testing (CAT), and to offer a scenario of Brazilian academic research in the area. CAT is a strategy for administering tests and estimating latent variables through algorithms which, in an interactive way, select and present items according to the performance of the individual. CAT differs from conventional testing by administering different sets of items to different test subjects, according to estimates of the subject's latent traits made during the test. In this paper, we seek to introduce a brief history of adaptive testing, in addition to discussing CAT's advantages over conventional testing. We also present CAT's constitutive elements based on Item Response Theory and list the main strategies for CAT construction we found in literature. Finally, we report the results of a literature review concerning the Brazilian research in this area.

Keywords: psychometrics; computerized adaptive testing (CAT); educational assessment; psychological assessment.

RESUMEN

Este trabajo tuvo como objetivo introducir los aspectos conceptuales de las pruebas adaptativas computarizadas (CAT). La CAT es un método para administrar pruebas y estimar rasgos latentes a través de algoritmos interactivos que seleccionan y presentan ítems de acuerdo con el desempeño del examinado. Se puede administrar diferentes conjuntos de ítems a diferentes examinandos, de acuerdo con las estimaciones del rasgo latente realizadas durante la prueba. En este documento, tratamos de introducir los principios generales que subyacen a la CAT y discutir sus ventajas sobre las pruebas convencionales. También presentamos los elementos y las principales estrategias para la construcción de CAT utilizando la Teoría de Respuesta al Ítem. Finalmente, informamos los resultados de una revisión de la literatura sobre la investigación brasileña en esta área.

Palabras clave: psicometría; pruebas adaptativas computadorizadas CAT; pruebas educativas informatizadas; evaluación educacional; pruebas psicológicas.

Introdução

Tradicionalmente, os testes psicométricos são constituídos por um número fixo de itens, ou seja, o teste é formado por um mesmo conjunto de itens administrados a todos os testando, independentemente de suas características individuais (e.g., seu nível de habilidade, em um teste educacional). Esses instrumentos usualmente são apresentados no formato "papel e lápis" ou mesmo informatizados e sem correção automatizada (e.g., formulários eletrônicos). Geralmente, esses testes são construídos a partir dos pressupostos da Teoria Clássica dos Testes (TCT) e apresentam diversas limitações psicométricas e práticas (WEISS, 2004).

Uma das principais limitações desses testes está relacionada à estratégia de selecionar o conjunto de itens que formará o teste com o objetivo estrito de maximizar sua consistência interna e, consequentemente, reduzir o erro padrão da medida. Essa abordagem assume que o erro padrão é uniforme em toda a extensão da escala de medida. De acordo com esse critério, os itens selecionados serão aqueles com dificuldade apropriada a um testando com nível de traço latente (*i.e.*, *teta* ou habilidade) mediano, ou seja, itens com probabilidade de acerto próxima a $p = 0{,}50$, oferecendo uma medida pobre para sujeitos com *teta* acima ou abaixo dessa faixa (WEISS, 2004).

Além de questões relacionadas à qualidade da medida, há ainda outras limitações de ordem prática que são típicas da testagem convencional com número fixo de itens. Por exemplo: a aplicação de itens desnecessários para estimar o *teta* dos indivíduos; o tempo de aplicação maior que o necessário, pois é preciso aplicar o mesmo conjunto de itens a todos os testando; e a fadiga do testando. Na testagem adaptativa, esses problemas podem ser mitigados.

"Um teste adaptativo é aquele em que diferentes conjuntos de questões (itens) são administrados para diferentes indivíduos dependendo do status de cada indivíduo no traço que está sendo medido" (WEISS, 1985, p. 774). Ou seja, em uma testagem adaptativa, o teste é ajustado ao desempenho do indivíduo nas tarefas (i.e., itens) de forma a obter, especificamente, a estimativa mais precisa do *teta* desse indivíduo.

Já testagem adaptativa computadorizada ou por computador — também chamado no Brasil de testagem adaptativa informatizada (TAI) e conhecida internacionalmente pela sigla CAT (*computerized adaptive testing*) — é um método de aplicação de testes e estimação de *teta* por meio de algoritmos e sistemas informatizados interativos. Na CAT, a sequência de apresentação dos itens é definida a partir da estimativa do nível de *teta* próximo ao do testando e a aplicação do teste é encerrada quando uma estimativa precisa desse nível é atingida, conforme alguns critérios psicométrico, ou de acordo com outro critério de parada, como verão na seção deste trabalho dedicada aos elementos constitutivos da CAT. Weiss (2004) e van der Linden e Glas (2000) definem a CAT como um redesenho dos testes psicométricos de maneira que eles sejam administrados por meio de computadores de maneira interativa, eficiente e eficaz.

De fato, a CAT apresenta diversas vantagens potenciais sobre a testagem convencional, como por exemplo, (LINACRE, 2000; THOMPSON; WEISS, 2011): evitar a aplicação de itens irrelevantes (e.g., itens muito fáceis ou muito difíceis para o testando), amenizando problemas como respostas dadas ao acaso (i.e., marcar uma alternativa aleatoriamente ou tentar adivinhar ou chutar a resposta correta); diminuir o tamanho do teste; ser mais rápida no que diz respeito ao seu desenvolvimento, implementação e entrega de resultados; oferecer uma experiência melhor ao testando, uma vez que possibilita diferentes configurações de tamanho e tempo e diminui problemas relacionados ao cansaço e à frustração; evitar erros na coleta de dados, como problemas na digitalização do material do teste (*i.e.*, dados ausentes ou digitados erroneamente) e na atribuição de escore; e coletar dados ao mesmo tempo em que o teste é administrado.

Um breve histórico da testagem adaptativa

Pelo menos desde a década de 1970 vêm sendo desenvolvidas pesquisas com os objetivos de investigar aspectos técnicos relacionados a modelos psicométricos que permitam a testagem adaptativa. No entanto, é possível identificar já em um dos primeiros testes psicológicos a primeira experiência de testagem adaptativa: o teste de inteligência de Binet (VAN DER LINDEN; GLASS, 2000; WEISS, 1985 e 2004). Esse teste adotava uma estratégia de ramificação mecânica — em inglês, *mechanical branching*.

A testagem inicia-se com base em alguma informação sobre o testando (e.g., a idade cronológica do testando ou uma estimativa da idade mental). Então, administra-se um primeiro conjunto de itens correspondentes a estimativa *a priori* de *teta*, chamada nível inicial. A testagem prosseguirá até a identificação dos níveis inferior (i.e., idade basal) e superior (i.e., idade teto) de idade mental para o testando, nessa sequência. O nível inferior será aquele abaixo do nível inicial em que o testando acerta todos os itens, enquanto o superior é aquele acima do nível inicial em que o testando erra todos os itens. Todos esses procedimentos são conduzidos mecanicamente pelo aplicador do teste.

O teste adaptativo de Binet tem, portanto (WEISS, 1985): um método para início da testagem; um método para atribuição de escores (i.e., proporção de acertos); um método para seleção de itens em um banco (itens organizados em conjuntos, de acordo com nível de idade mental, e selecionado a partir do critério do não atingimento da regra de 100% de acertos ou erros, dependendo da direção da ramificação que está sendo seguida); e um método para encerramento da testagem (i.e., encontrar os níveis inferior e superior). Destaca-se que, seguindo as regras de ramificação, o tamanho do teste é reduzido ao evitar a administração de itens mais fáceis que os do nível inferior ou mais difíceis que os do nível superior.

Posteriormente, foram propostas outras estratégias que adotavam critérios fixos de início e término de testagem, chamadas de ramificação fixa (WEISS, 1985). Por exemplo, o teste adaptativo de duas etapas (WEISS, 1974), cujo critério de início fixo consiste na aplicação de um conjunto de itens de dificuldade média (i.e., *routing test*) com o objetivo de decidir sobre qual ramificação será seguida. Se o testando tiver uma proporção alta de acertos, ele será submetido a um conjunto de itens mais difíceis que os iniciais e, ao contrário, ele responderá a um conjunto de itens mais fáceis. Com essa mesma estratégia de ramificação fixa foi também propostos testes adaptativos com três ou quatro etapas (WEISS, 1985).

Já as estratégias de ramificação fixa em pirâmide (*pyramidal*), de níveis flexíveis (*flexilevel*) e estratificados (stratified adaptive ou stradaptive) adotaram um método de seleção item a item, ao invés de administrar conjuntos inteiros de itens a cada etapa (WEISS, 1974). No teste adaptativo de níveis flexíveis é aplicado apenas um item por nível de dificuldade,

enquanto no piramidal pode haver repetição de itens correspondentes a um mesmo nível de dificuldade. Assim, o esquema de níveis flexíveis exige menos itens, uma vez que o aplicador selecionará como próximo item um que ainda não tenha sido administrado. Na proposta estratificada, o próximo item a ser aplicado é aquele mais discriminativo ainda não administrado de um nível de dificuldade mais baixo ou mais alto. Assim, o teste continua seguindo uma ramificação item a item, até que se identifique um nível superior.

De acordo com Linacre (2000) e Weiss (2004), o campo da testagem adaptativa com o uso do computador somente começou a tomar corpo com o desenvolvimento de novos modelos psicométricos, como os da Teoria de Resposta ao Item - TRI (AYALA, 2009; PASQUALI, 2007), e com o desenvolvimento computacional. Um exemplo pioneiro é a metodologia proposta por Reckase (1974), baseada no modelo logístico de um parâmetro de Rasch. Reckase elaborou um *software* que programava uma testagem personalizada (em inglês, *tailored testing*) com um banco de 150 itens relacionados à aprendizagem de métodos de medida. O primeiro item a ser aplicado deveria ser escolhido a partir de estimativas anteriores da habilidade do testando em um banco de dados. Caso não houvesse no banco de dados informações anteriores sobre a habilidade do testando, o primeiro item deveria então ter facilidade igual a 100%. Em caso de acerto, são apresentados itens com metade da facilidade até que haja uma resposta incorreta. Em caso de erro, são apresentados itens com o dobro de facilidade até que uma resposta correta seja obtida. Para que a estimativa de habilidade pudesse ser feita, era necessário que o testando tivesse dado pelo menos uma resposta correta e uma incorreta. A habilidade do sujeito era estimada após a administração de cada item. O critério psicométrico para encerrar a testagem era um ponto-de-corte relacionado ao limite inferior da estimativa de habilidade.

Os elementos da CAT

Estruturalmente, um algoritmo de CAT é composto por cinco elementos: (i) um banco de itens calibrados; (ii) um método para iniciar a testagem (i.e., escolha do primeiro item); (iii) um método para seleção de itens (após o início); (iv) um método para estimação de escore; e (v) um método para

encerrar a testagem. A seguir, introduzimos os princípios relacionados a cada um desses elementos, buscando elencar os principais desafios relacionados ao desenvolvimento de cada um deles.

Banco de itens calibrados

O banco de itens de uma CAT contém um conjunto de itens calibrados, ou seja, cujos parâmetros psicométricos foram previamente estimados. Usualmente, esses parâmetros são estimados por meio de algum modelo da TRI — embora seja possível desenvolver uma CAT a partir da TCT. De acordo com Pasquali (2007), o banco de itens será formado de acordo com os objetivos da testagem. Por um lado, um banco pode ser construído para avaliar toda a extensão de um traço latente (e.g., um teste de personalidade). Por outro, pode ser construído com a finalidade de avaliar apenas uma faixa restrita da escala do traço latente (e.g., seleção de pessoas, certificação de um dado nível de proficiência, diagnóstico etc.).

A primeira etapa para construção de um banco é o desenvolvimento de um *framework* para o teste para o qual se construirá o banco de itens, ou seja, a construção de um modelo conceitual para o teste - no contexto da avaliação educacional frequentemente chamado de matriz de referência. Pasquali (2010) divide a elaboração de testes em três grandes polos: teórico, empírico e analítico. O primeiro polo envolve a delimitação do modelo conceitual do teste. Em educação esse modelo costuma ser chamado de matriz de referência ou tabela de especificação. Além de constar as definições constitutiva e operacional dos traços latentes que se pretende mensurar, o modelo conceitual precisa caracterizar a dimensionalidade do teste a ser construído: uni ou multidimensional. No caso de modelos multidimensionais, é necessário especificar se é o caso de um modelo ortogonal ou oblíquo, hierárquico, bifactor, com fator geral etc. A partir dessas definições, serão então elaborados os itens, responsáveis pela operacionalização do modelo conceitual em comportamentos observáveis (i.e., as respostas dos testando às tarefas que compõem os itens). A escolha de qual tipo de item utilizar dependerá, obviamente, do modelo conceitual e das finalidades da testagem (PASQUALI, 2010; RABELO 2013).

Então, seguem-se os procedimentos empíricos (PASQUALI, 2010), como o planejamento da testagem (e.g., amostragem) e a coleta de dados

(pré-teste). Após a coleta, iniciam-se os procedimentos referentes ao polo analítico. O primeiro passo é a análise da dimensionalidade do conjunto de itens (LAROS, 2012), avaliada à luz do modelo conceitual inicialmente elaborado. Então, realiza-se a calibração dos itens propriamente dita por meio de algum modelo da TRI (PASQUALI, 2007).

Sobre o tamanho da amostra para calibração de itens por meio da TRI, Sahin e Weiss (2015) apontaram que, desde o estudo de Lord (1968), costuma-se seguir a diretriz de uma amostra com pelo menos 1.000 sujeitos para estimativas utilizando o modelo logístico de três parâmetros. No entanto, Sahin e Weiss (2015) ponderaram que, após a adoção de novos métodos de estimação, a partir da década de 1990, como o da máxima verossimilhança marginal, é esperado que amostras menores (e.g., 500 sujeitos) produzam estimativas precisas. A partir de uma série de simulações, Sahin e Weiss concluíram que amostras pequenas podem ser suficientes em alguns cenários com CAT, como uma amostra de 150 testandos para um banco de 200 itens.

Quanto ao tamanho do banco de itens, costuma-se apontar que esse banco deve ser grande (e.g., 1.000 itens ou mais), mas essa orientação parece ser diferente de acordo com as finalidades, os riscos e as consequências de cada processo de testagem (WISE; KINGSBURY, 2000). O tamanho do banco será determinado, pelo menos, por aspectos como a qualidade da estimação dos parâmetros psicométricos dos itens (c.f., SAHIN; WEISS, 2015) e a necessidade de controlar a superexposição de itens (c.f., OZTURK; DOGAN, 2015) e de realizar um balanceamento entre conteúdos medidos pelo teste (c.f., SAHIN; OZBASI, 2017). Testes educacionais padronizados administrados em larga-escala e *high stake* (i.e., utilizados para tomar decisões importantes e de grande impacto social na vida de indivíduos ou organizações) enfrentam esse desafio ao adotar a CAT. Um grande banco de itens será necessário para que não haja superexposição de itens e os testandos simplesmente memorizarem e divulguem os gabaritos dos itens, comprometendo a segurança e a isonomia da testagem. Esse é o caso dos grandes testes nacionais adotados no Brasil, como o Exame Nacional do Ensino Médio (ENEM) e a Prova Brasil, cujos resultados impactam diretamente a vida de estudantes, famílias, professores, escolas, sistemas educacionais da Educação Básica e as Instituições de Ensino Superior.

Segundo Wise e Kingsbury (2000), inicialmente sugeria-se bancos com pelo menos 100 itens. De fato, as simulações de Sahin e Weiss (2015) apontaram que, para determinadas áreas da escala de traço latente, um banco com 100 itens poderia funcionar tão bem quanto bancos com 500 itens, desde que a informação dos itens (AYALA, 2009; PASQUALI, 2007) concentre-se uma mesma área. No entanto, Sahin e Weiss (2015) destacaram que um banco de 200 itens ou mais diminuiria o risco de não possuir informação em determinadas áreas da escala, evitando produzir estimativas imprecisas na maior parte das situações simuladas. Sahin e Weiss concluíram que 300 itens era um tamanho a ser inicialmente considerado inicialmente no planejamento de um grande banco de itens. No entanto, com as restrições de realizar um balanceamento de conteúdos e controlar a superexposição de itens, o tamanho exigido poderá ser maior. De acordo com Wise e Kingsbury (2000), os bancos de CAT em funcionamento à época usualmente possuíam mais de 1.000 itens. Para testes com itens politômicos, segundo Boyd, Dodd e Choi (2010), bancos pequenos (e.g., 30 itens) podem levar a estimativas precisas. Com a restrição de controlar a superexposição de itens, 100 a 120 itens seriam suficientes.

Por fim, o banco precisará passar por manutenção. Thompson e Weiss (2011) alertam que o banco precisará ser renovado. Por um lado, itens podem ser excluídos, de forma provisória ou permanente, devido à baixa qualidade ou à superexposição. Por outro, novos itens podem ser adicionados ao banco já calibrado. Para tanto, utiliza-se técnicas de ligação (*linking)* e equiparação (*equating*) de escores (c.f., VON DAVIER, 2011), geralmente utilizando uma abordagem com ancoragem de itens. Há também uma diversidade de estratégias para calibração *online*, ou seja, calibrar novos itens durante a própria testagem sem a necessidade de realizar pré-testes (c.f., ZHENG, 2014 e 2016). Um aspecto final relacionado à manutenção do banco é a consistência da escala de mensuração com o passar do tempo e com alterações de itens, como a questão do DRIFT (DUNYA, 2018), ou seja, da mudança dos valores originais dos parâmetros dos itens com o passar do tempo (e.g., itens que ficam mais fáceis ou mais difíceis).

Método para iniciar a testagem

Um critério para iniciar a testagem é considerar uma estimativa inicial do *teta* do testando, que pode ser baseada em alguma informação *a priori*

sobre o testando, como o desempenho em uma testagem anterior ou algum outro fator externo à testagem (THOMPSON; WEISS, 2011). Por exemplo, em uma testagem educacional em larga-escala, é possível utilizar o escore médio dos estudantes oriundos da mesma escola, rede ou município do testando. Ou, em uma avaliação com delineamento longitudinal, é possível utilizar uma estimativa de *teta* feita anteriormente.

Assim, considerando que geralmente não se tem uma estimativa anterior do *teta* do testando, uma opção é estabelecer um valor fixo como estimativa de *teta* para todos os testandos, por exemplo, um escore médio. Na TRI esse valor seria 0,0 ou percentil 50%, considerando uma distribuição normal. O problema com esse método é que, caso haja apenas um ou poucos itens com esse nível de dificuldade no banco, o sistema da CAT poderá sempre recorrer ao mesmo item (ou aos mesmos poucos itens) para iniciar a testagem, pois todos os usuários terão o mesmo *teta* inicial. Para contornar esse problema, é possível adotar um critério flexível, selecionando randomicamente itens em torno do escore médio, por exemplo, entre -0,5 e 0,5 (THOMPSON; WEISS, 2011).

Método para seleção de itens

A seleção dos próximos itens a serem apresentados dá-se com base na estimativa do *teta* do testando feita após a resposta a cada item. A partir da ideia de informação do item, o próximo a ser selecionado será aquele que produzirá maior informação com maior precisão na faixa da escala do traço latente na qual se encontra o *teta* do testando de acordo com a estimativa anterior. Thompson e Weiss (2011) alertam que, caso os objetivos da testagem estejam relacionados à classificação dos testandos de acordo com algum ponto-de-corte, deve-se optar por selecionar itens que produzam mais informação próxima a esse ponto.

A superexposição de itens e o balanceamento dos conteúdos do teste mencionados anteriormente são duas questões práticas que influenciam a definição de critérios para a seleção de itens (OZTURK; DOGAN, 2015; SAHIN; OZBASI, 2017). Assim, o algoritmo de seleção de itens pode incluir alguma estratégia de randomização para evitar que os mesmos itens sejam sempre escolhidos por apresentarem as melhores propriedades psicométricas, como os itens com parâmetro de discriminação maior. Da mesma forma,

pode ser necessário que o algoritmo de seleção de itens garanta que todos os conteúdos relevantes sejam cobertos durante a testagem, contribuindo assim para a validade de conteúdo da CAT.

Método para encerrar a testagem

A CAT pode ser desenhada para ter um tamanho fixo ou variável. No caso de CAT com tamanho fixo, a testagem será encerrada após o testando tenha respondido a uma quantidade pré-estabelecida de itens. No entanto, como uma das principais vantagens da CAT é reduzir o tamanho do teste, evitando apresentar itens desnecessários, geralmente objetiva-se construir uma CAT com tamanho variável. Para tanto, há vários métodos possíveis para estabelecer um critério flexível de encerramento da testagem. Thompson e Weiss (2011) apontam que o critério mais comum de encerramento nesses casos é o de erro-padrão mínimo. Nessa estratégia, a testagem será encerrada quando a estimativa tenha atingido um determinado nível de precisão. Outra possibilidade é encerrar a testagem quando a estimativa de *teta* torna-se estável (i.e., sofre pequenas variações) após a apresentação e correção de novos itens. Ainda, há o critério do mínimo de informação do banco, em que a testagem é interrompida quando não há mais itens no banco que possam fornecer um nível mínimo de informação adicional.

Simulações como estratégia para o desenvolvimento de uma CAT

Recomenda-se que se a inicie a construção de uma CAT a partir de uma agenda de simulações com diferentes configurações do algoritmo de CAT que permita avaliar sua viabilidade e planejar seu desenvolvimento (THOMPSON; WEISS, 2011). Em muitas situações, começa-se o desenvolvimento a partir de um teste convencional já operacional, que já conta com bancos de itens e de dados reais. Neste contexto, é possível adotar simulações do tipo Post-Hoc (NYDICK; WEISS, 2009), também chamadas de simulações com dados reais.

No entanto, simulações Post-Hoc enfrentam algumas limitações objetivas. Por exemplo, há casos em que o banco de itens disponível é reduzido ou há poucos dados coletados. Se o teste irá começar do zero, a coleta de dados

pode ser muito dispendiosa. Há também os casos em que não há respostas de todos os sujeitos a todos os itens, como ocorre frequentemente em testes educacionais em larga-escala, nos quais são adotados desenhos como o de Blocos Incompletos Balanceados. Por exemplo, um banco é composto por 500 itens, mas o pré-teste é realizado com cadernos de prova com 100 itens, sendo que 80 são novos e 20 já parametrizados. Assim, os sujeitos não teriam respostas para 400 itens.

Para contornar esses problemas, é possível realizar simulações do tipo Monte Carlo ou do tipo híbrido (NYDICK; WEISS, 2009). Simulações do tipo Monte Carlo são realizadas a partir de dados inteiramente simulados — gerados de maneira aleatória a partir de algum modelo de TRI ou especificando-se, de acordo com o objetivo da simulação, a distribuição (e.g., normal, log-normal, uniforme etc.) dos parâmetros psicométricos dos itens e sujeitos. Nas simulações do tipo híbrido, por sua vez, os dados reais são usados quando possível. Assim, itens e respostas são gerados usando métodos Monte Carlo com base na distribuição dos parâmetros de itens pré-testados e no *nível* de teta de sujeitos reais.

De acordo com Thompson e Weiss (2011), as simulações do tipo Post-Hoc e do tipo híbridas permitem simular com maior eficiência sistemas de CAT. Elas são essenciais para comparar e avaliar diferentes métodos e especificações para os algoritmos que estruturam a CAT. De fato, a simulação do tipo híbrida é uma boa solução, pois pode representar economia financeira e logística ao permitir utilizar bancos de itens e matrizes de respostas já existentes e complementá-los, se necessário, com dados simulados a partir de informações reais.

Software para o desenvolvimento de CAT

Há uma quantidade significativa de softwares disponíveis para o desenvolvimento de CATs. A International Association for Computerized Adaptive Testing (2018) possui uma lista de *softwares* comerciais e de código aberto dedicado à CAT. O repositório The Comprehensive R Archive Network (2019) também possui uma lista com os pacotes disponíveis para TRI e CAT no R. A seguir, destacamos algumas opções livres e de código aberto. Os pacotes catR (MAGIS; BARRADA, 2017) e MAT (CHOI; KING, 2015) do R e o SimulCAT (HAN, 2012) permitem gerar padrões de resposta, bem como possuem

diferentes opções para estimação de habilidade, seleção do primeiro e dos próximos itens, regras de encerramento, controle da exposição dos itens e balanceamento do conteúdo do teste. Já o pacote mirtCAT (CHALMERS, 2016) e a plataforma Concerto (SCALISE; ALLEN, 2015) permitem gerar interfaces HTML para aplicação de testes adaptativos e não-adaptativos, uni e multidimensionais.

A pesquisa com CAT no Brasil

No Brasil, a pesquisa com CAT ainda é incipiente, com poucos trabalhos publicados. Para obter um panorama da produção nacional nessa área, realizamos uma busca na BVS-PSI (Biblioteca Virtual em Saúde - Psicologia Brasil), no Scielo (Scientific Eletronic Library Online) e na BDTD (Biblioteca Digital Brasileira de Teses e Dissertações). A busca foi realizada em janeiro de 2019 e utilizou os termos "testagem adaptativa" e "teste adaptativo". As buscas resultaram em oito trabalhos (i.e., dissertações, teses e artigos). Para complementar os resultados, realizamos novas buscas utilizando outras ferramentas de busca (e.g., Google).

Foram encontradas 11 dissertações de mestrado. Piton-Gonçalves (2004) desenvolveu um ambiente para avaliação de proficiência em língua inglesa utilizando CAT. Costa (2009) explorou métodos estatísticos para construção de CATs, bem como avaliou a adequabilidade para o desenvolvimento de um sistema a partir de um teste de proficiência em língua inglesa. Abreu (2012) realizou simulações visando a criação de uma CAT para ser utilizado como ferramenta para diagnóstico e *feedback* da aprendizagem em matemática na Educação Básica. Sassi (2012) comparou algoritmos de seleção de itens e desenvolveu uma aplicação em VBA-Excel integrado ao R. Galvão (2013) comparou métodos de seleção de itens com base na estimativa do erro padrão da medida. Ricarte (2013) simulou implementações de CAT a partir de um exame de proficiência em língua inglesa e de um inventário de depressão. Araújo (2014) teve como objetivo a criação de CAT como provas simuladas para preparação para concursos públicos. Maia Júnior (2015) explorou o uso da covariável tempo de resposta no algoritmo de CAT com o objetivo de encurtar a testagem. Meneghetti (2015) explorou estratégias de seleção de itens com base em agrupamento por similaridade. Silva (2015) implementou uma CAT a partir de um teste convencional de proficiência em língua inglesa. Santos

(2017) adaptou uma medida psicométrica para rastreio de dislexia para um jogo educacional para dispositivos móveis utilizando um sistema de CAT.

A busca também identificou quatro teses de doutorado. Moreira Junior (2011) sistematizou diretrizes para implantação e manutenção de sistemas de CAT. Piton-Gonçalves (2012) dedicou-se ao estudo e desenvolvimento de CAT baseada na Teoria de Resposta ao Item Multidimensional. Oliveira (2017) desenvolveu um banco de itens para a CAT voltada à avaliação dos cinco grandes fatores da personalidade, utilizando o modelo de resposta gradual da TRI. Por fim, Spenassato (2017) explorou estratégias para manutenção de banco de itens para CAT, como o acréscimo de itens e a verificação de DRIFT.

Além das dissertações e teses, identificamos cinco artigos. Labarrère, Da-Silva e Costa (2011) analisaram o erro nas estimativas de teta a partir de dados reais e simulados. Moreira Junior, Tezza, Andrade e Bornia (2012) propuseram um algoritmo de CAT para estimação de usabilidade de sites de *e-commerce*. Piton-Gonçalves e Aluísio (2015) apresentaram os princípios e métodos para elaboração de uma CAT multidimensional. Santana et al. (2017) compararam as formas convencional e adaptativa de um teste de desempenho acadêmico de estudantes do nível superior do curso de direito. Alavarse, Catalani, Menheghetti e Travitzki (2018) desenvolveram uma versão CAT para o teste padronizado Provinha Brasil, voltada à proficiência em leitura (este foi artigo publicado em revista estrangeira).

Certamente a pesquisa nacional com CAT não se restringe a esses trabalhos. Não incluímos nos resultados, por exemplo, os trabalhos publicados em anais de congressos, como o CONBRATRI (Congresso Brasileiro de Teoria de Resposta ao Item) e os encontros da ABAVE (Associação Brasileira de Avaliação Educacional) e do IBAP (Instituto Brasileiro de Avaliação Psicológica) que costumam reunir pesquisadores da área de psicometria. Também não buscamos produções de pesquisadores brasileiros em periódicos estrangeiros. Além disso, é preciso considerar que muitos pesquisadores e técnicos envolvidos com a construção de CATs trabalham em organizações públicas ou privadas que desenvolvem seus sistemas sem necessariamente incluir em sua agenda a publicação de trabalhos acadêmicos. Por fim, os procedimentos adotados para a busca podem ter falhado na identificação de outros trabalhos publicados (e.g., trabalhos publicados em periódicos ou repositórios não indexados na BVS-PSI, Scielo e BDTD).

Discussão

Apesar de ainda mostrar-se um pouco tímida, com poucos artigos publicados em periódicos, verifica-se que a produção brasileira é crescente. Identificou-se um importante grupo de pesquisadores oriundos de diversas áreas (e.g., computação, estatística, engenharias, psicologia e educação) que investigaram diferentes aspectos da CAT, como a construção e manutenção de bancos de itens, comparação de algoritmos para seleção de itens e métodos de estimação, além da adaptação de testes convencionais em sistemas de CAT em educação, psicologia e tecnologia da informação.

A testagem adaptativa por computador tem o potencial de contornar uma série de problemas presentes nas estratégias de testagem convencionais. A CAT não apenas personaliza a testagem, oferecendo uma melhor experiência ao testando, que não precisará responder a itens desnecessariamente, como também aperfeiçoa a testagem e aumenta a qualidade da medida ao realizar estimativas do traço latente com maior precisão para cada testando. No entanto, como também se buscou demonstrar, o desenvolvimento de sistemas de CAT não é uma tarefa trivial.

O desenvolvimento de cada elemento da CAT envolve uma série de decisões que impactará seu funcionamento. Além disso, o desenvolvimento sempre estará limitado pelas características da medida (e.g., modelo conceitual, modelo psicométrico, qualidade dos itens, qualidade dos dados, distribuição dos parâmetros dos itens e dos sujeitos etc.) e pelas restrições práticas da testagem (e.g., segurança, custos, tempo disponível para aplicação etc.). Assim, o desenvolvimento de uma CAT irá exigir conhecimento e experiência em psicometria de maneira ampla, incluindo tanto o domínio dos modelos estatísticos, como dos procedimentos teóricos relacionados à construção de medidas.

Referências

ABREU, R. C. P. *Ensaio da ferramenta DIA Diagnóstico e Informação do Aluno. 2012*. Dissertação (Mestrado em Ciências Computacionais) - Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2012.

ALAVARSE, O. C., CATALANI, E. M. T., MENEGHETTI, D. R., TRAVITZKI, R. Teste adaptativo informatizado como recurso tecnológico para alfabetização inicial. *Sistemas, Cibernética e Informática*, v. 15, n. 3, p. 68-78, 2018.

ARAÚJO, J. V. M. *Teoria da resposta ao item em processo de decisão. 2014*. Dissertação (Mestrado em Estatística) — Universidade de Brasília, Brasília, 2014.

AYALA, R. J. *The theory and practice of item response theory*. Nova Iorque, Estados Unidos: The Guilfordd Press, 2009.

BOYD, A., DODD, B., CHOI, S. Polytomous models in computerized adaptive testing. In: Nering, M. L., Ostini, R. (Org.) *Handbook of Polytomous Item Response Theory*. Nova Iorque, Estados Unidos: Routledge, 2010, p. 229-256.

CHALMERS, P. mirtCAT: Computerized Adaptive Testing with Multidimensional Item Response Theory. *Journal of Statistical Software*, v. 71, n. 6, p. 1-38, 2016.

CHOI, S. W., KING, D. R. R Package MAT: Simulation of Multidimensional Adaptive Testing for Dichotomous IRT Models. *Applied Psychological Measurement*, v. 39, n. 3, p. 239-240, 2015.

COSTA, D. R. *Métodos estatísticos em testes adaptativos informatizados. 2009*. Dissertação (Mestrado em Estatística). Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

DUNYA, B. A. (2018). Item parameter DRIFT in computer adaptive testing due to lack of content knowledge. *International Journal of Testing*, v. 18, n. 4, p. 346-365, 2018.

GALVÃO, A. F. *Um modelo inteligente para seleção de itens em testes adaptativos computadorizados*. 2013. Dissertação (Mestrado em Ciência da Computação). Universidade Federal de Juiz de Fora, Juiz de For a, 2013.

HAN, K. T. SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement, v. 36, n.* 1, p. 64-66, 2012.

INTERNATIONAL ASSOCIATION FOR COMPUTERIZED ADAPTIVE TESTING – IACAT. CAT Software, 2019. Disponível em: <http://www.iacat.org/content/cat-software>. Acesso em: 31 de jan. de 2019.

LABARRÈRE, J. G., DA-SILVA, C. Q., COSTA, D. R. Testes Adaptativos Computadorizados. *Revista Brasileira de Biometria*, v. 9, n. 2, p. 229-261, 2011.

LAROS, J. A. O uso da análise fatorial: algumas diretrizes para pesquisadores. In: Pasquali, L. (Org.) *Análise fatorial para pesquisadores*. Brasília: LabPAM Saber e Tecnologia, p. 163-193, 2012.

LINACRE, J. M. Computer-adaptive testing: A methodology whose time has come. In: Chae, S., Kang, U., Jeon, E., Linacre, J. M. (Orgs.*) Development of Computerized Middle School Achievement Test* [in Korean]. Seoul: Komesa Press, p. 1-58, 2000.

LORD, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, v. 28, n. 4, p. 989–1020, 1968.

MAGIS, D., BARRADA, J. R. Computerized Adaptive Testing with R: Recent Updates of the Package catR. *Journal of Statistical Software*, 2017, v. 16, n. 1, p. 1-192017.

MAIA JUNIOR, A. G. P. *Uso do tempo de resposta para melhorar a convergência do algoritmo de testes adaptativos informatizados. 2015*. Dissertação (Mestrado em Estatística) — Universidade de Brasília, Brasília, 2015.

MENEGHETTI, D. R. *Metodologia de seleção de itens em testes adaptativos informatizados baseada em agrupamento por similaridade. 2015*.

Dissertação (Mestrado em Engenharia Elétrica) — Centro Universitário da FEI, São Bernardo do Campo, 2015.

MOREIRA JUNIOR, F. J. *Sistemática para implantação de testes adaptativos informatizados baseados na Teoria de Resposta ao Item. 2011*. Tese (Doutorado em Engenharia de Produção) — Universidade Federal de Santa Catarina, Florianópolis, 2011.

MOREIRA JUNIOR, F. J., TEZZA, R., ANDRADE, D. F., BORNIA, A. C. Algoritmo de um teste adaptativo informatizado com base na Teoria da Resposta ao Item para a estimação da usabilidade de sites de e-commerce. *Production*, v. 23, n. 3, p. 525-536, 2013.

NYDICK, S. W., WEISS, D. J. A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Org.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, 2009.

OLIVEIRA, C. M. *Construção e busca de evidências de validade de um banco de itens de personalidade para testagem adaptativa desenvolvido a partir dos princípios do desenho universal. 2017*. Tese — (Doutorado em Psicologia). Universidade Federal de Santa Catarina, Florianópolis, 2017.

OZTURK, N. B., DOGAN, N. Investigating item exposure control methods in computerized adaptive testing. *Educational Sciences: Theory Practice*, v. 15, n. 1, 86-98, 2015.

PASQUALI, L. Testes referentes a construto: teoria e modelo de construção. In: L. Pasquali, L. (Org.) *Instrumentação psicológica*. Fundamentos e práticas. Porto Alegre: Artmed, p. 165-198, 2010.

PASQUALI, L. TRI - Teoria de Resposta ao Item: Teoria, procedimentos e aplicações. Brasília: Laboratório de Pesquisa em Avaliação e Medida – LabPAM, 2007.

PITON-GONÇALVES, J. *A integração de testes adaptativos informatizados e ambientes computacionais de tarefas para o aprendizado do inglês instrumental. 2004*. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo, São Carlos, 2004.

PITON-GONÇALVES, J. *Desafio e perspectivas da implentação computacional de testes adaptativos multidimensionais para avaliações educacionais. 2012*. Tese (Doutorado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo, São Carlos, 2012.

PITON-GONÇALVES, J., ALUÍSIO, S. M. Teste adaptativo computadorizado multidimensional com propósitos educacionais: Princípios e métodos. *Ensaio*: Avaliação e Políticas Públicas em Educação, v. 23, n. 87, p. 389-414, 2015.

RABELO, M. *Avaliação educacional: fundamentos, metodologia e aplicações no contexto brasileiro*. Rio de Janeiro: Sociedade Brasileira de Matemática, 2013.

RECKASE, M.D. An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods and Instrumentation*, v. 6, n. 2, p. 208-212, 1974.

RICARTE, T. A. M. *Teste adaptativo computadorizado nas avaliações educacionais e psicológicas. 2013*. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo, São Carlos, 2013.

SAHIN, A., WEISS, D. J. Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. Educational Sciences: Theory Practice, v. 15, n. 6, p. 1585-1595, 2015.

SAHIN, A.; OZABASI, D. Effects of content balancing and item selection method on ability estimation in computerized adaptive tests. *Eurasian Journal of Educational Research*, v. 69, p. 21-36, 2017

SALGADO, A. M. *Uso de regressão isotônica na escolha de itens em testes adaptativos computadorizados. 2018*. Dissertação (Mestrado em Estatística) - Universidade de Brasília, Brasília, 2018.

SANTANA, L. F., BARTHOLOMEU, D., MONTIEL, J. M., COUTO, G., BERBERIAN, A. A., PESSOTO, F. Avaliação informatiza adaptativa do ENADE pelo MOODLE:

evidências de validade. *Informática na educação: teoria prática*, v. 20, n. 2, p. 222-238, 2017.

SANTOS, J. S. *Mensuração de habilidades cognitivas preditoras do desenvolvimento de leitura em crianças através de jogos educacionais para dispositivos móveis. 2017*. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Campina Grande, Campina Grande, 2017.

SASSI, G. P. *Teoria e prática de um teste adaptativo informatizado. 2012*. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo, São Carlos, 2012.

SCALISE, K. ALLEN, D. D. Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *British Journal of Mathematical and Statistical Psychology*, v. 68, n. 3, p. 478–496, 2015.

SILVA, V. R. *Avaliação da proficiência em inglês acadêmico através de um teste adaptativo informatizado. 2015*. Dissertação (Mestrado em Estatística). Universidade de São Paulo e Universidade Federal de São Carlos, São Carlos, 2015.

SPENASSATO, D. *Manutenção do banco de itens para testes adaptativos computadorizados aplicados em avaliações de alto impacto. 2017*. Tese (Doutorado em Engenharia de Produção). Universidade Federal de Santa Catarina, Florianópolis, 2017.

THE COMPREHENSIVE R ARCHIVE NETWORK. *CRAN Task View: Psychometric Models and Methods, 2019*. Disponível em: < https://cran.r-project.org/web/views/Psychometrics.html>. Acesso em: 31 de jan. de 2019.

THOMPSON, N. A., WEISS, D. J. A framework for the development of computerized adaptive tests. *Practical Assessment, Research Evaluation*, v. 16, n. 1, p. 1-9, 2011.

VAN DER LINDEN, W. J., GLAS, C. W. (ORGS.). *Computerized adaptive testing: theory and practice*. Norwell, Estados Unidos: Kluwer Academic Publishers, 2000.

VON DAVIER, A. A. (Org.). *Statistical models for test equating, scaling, and linking*. Nova Iorque, Estados Unidos: Springer, 2011.

WEISS, D. J. Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, v. 53, n. 6, p. 774-789, 1985.

WEISS, D. J. Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Education*, v. 37, n. 2, p. 70-84, 2004.

WEISS, D. J. *Strategies of adaptive ability measurement*. Minneapolis: University of Minnesota, 1974.

WISE, S. L., KINGSBURY, G. Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, v. 21, p. 135-155, 2000.

ZHENG, Y. *New methods of online calibration for item bank replenishment. 2014*. Tese (Doutorado em Psicologia Educacional) - University of Illinois, Urbana-Champaign, Estados Unidos, 2014.

ZHENG, Y. Online calibration of polytomous items under the Generalized Partial Credit Model. *Applied Psychological Measurement*, v. 40, n. 6, p. 434-450, 2016.

Alexandre José de Souza Peres

Doutor em Psicologia Social, do Trabalho e das Organizações pela Universidade de Brasília (UnB). Professor da Universidade Federal de Mato Grosso do Sul (UFMS), Campus de Paranaíba. Paranaíba (MS), Brasil;
E-mail: alexandre.peres@gmail.com e alexandre.peres@ufms.br

Artigo submetido em 04/02/2019
Aprovado em 24/06/2019