

TESTES ADAPTATIVOS COMPUTADORIZADOS - CAT

Carlos Henrique Sancinetto da Silva Nunes

Débora Spenassato

Cassandra Melo Oliveira

Antonio Cezar Bornia

Ricardo Primi

1. INTRODUÇÃO

O uso de testes para avaliação educacional, psicológica, no contexto ocupacional e em outras áreas alcançou reconhecimento principalmente a partir do século XX. Apesar de variados formatos terem sido explorados desde o início das pesquisas de desenvolvimento dos testes, o formato que envolve o uso de “papel e lápis” (P&P - do termo inglês *paper-and-pencil*) ganhou grande destaque (Anastasi & Urbina, 2000). No entanto, com a evolução da tecnologia da informática, surgiram os Testes Baseados em Computador e, a partir de 1960, aplicações de testes adaptativos computadorizados (CAT - do inglês *Computerized Adaptive Testing*) foram sendo implementadas por pesquisadores, aumentando a popularidade e as aplicações desses instrumentos (Chang & Ying, 2009; Chen, Ankenmann, & Chang, 2000; Leroux, Lopez, Hembry, & Dodd, 2013). Um mapeamento do histórico dos testes CATs indica um incremento de centenas de testes desenvolvidos até o início da década de 1990 para mais de um milhão até o ano de 1999 (Wainer, 2000).

Ao longo das últimas décadas, os CATs têm sido amplamente utilizados para testes de nivelamento ou avaliações de desempenho, nas áreas da educação, da certificação, em testes psicológicos e na área da saúde. Pesquisas e aplicações deste método de avaliação ainda são recentes no Brasil, se comparadas a outros países, como nos Estados Unidos, onde já existem várias aplicações em larga escala.

No cenário internacional, inúmeras instituições utilizam esta forma de avaliação. No site da *International Association for Computerized Adaptive Testing* (<http://www.iacat.org/>), é possível encontrar uma lista desses testes. Alguns deles são: *Graduate Management Admission Test* (GMAT) – testes para admissões na pós-graduação ou escola de negócios, *National Council of State Boards of Nursing* (NCSBN) – oferece testes de certificação para profissionais na área de Enfermagem, *Armed Services Vocational Aptitude Battery* (ASVAB) – teste de aptidão para

recrutamento de pessoal para serviços militares, *Test of English as a Foreign Language* (TOEFL) – teste de proficiência na língua inglesa reconhecido no mundo todo, *Microsoft Certified Professional exams* – testes de certificação que avaliam competências de profissionais na área de tecnologia da informação sobre conhecimentos dos produtos e aplicativos da Microsoft, *National Assessment of Educational Progress* (NAEP) – avalia o progresso acadêmico do estudante do ensino fundamental e médio nos Estados Unidos ao longo do tempo, *Dynamic Health Assessments* (DYNHA) – aplicação de testes para o desenvolvimento e melhoria de pesquisas na área de saúde funcional do indivíduo e bem-estar, entre outros.

Os CAT são testes aplicados de forma adaptativa aos indivíduos; sendo assim, cada respondente recebe um teste personalizado de acordo com sua proficiência (θ) e conforme as regras predefinidas no algoritmo. O algoritmo computacional seleciona os itens a serem aplicados a partir de um banco de itens (BI). Os itens que compõem o BI já passaram por etapas prévias, tais como: desenvolvimento, análise de juízes, busca por evidências de validade e estimação dos parâmetros pela Teoria da Resposta ao Item (TRI).

De acordo com Thissen et al. (2007), a maioria dos desenvolvedores de BI usam modelos da TRI para calibrar os itens, pois a TRI permite que qualquer subconjunto de itens selecionados deste banco seja comparável, desde que a estimativa dos parâmetros estejam na mesma métrica. Desta forma, a proficiência dos indivíduos é comparável, mesmo respondendo a itens distintos (Kopec et al., 2008; Luecht, De Champlain, & Nungester, 1998; Zitny, Halama, Jelinek, & Kveton, 2012).

O algoritmo geral de um CAT é apresentado na Figura 1. Na etapa inicial do teste, um ou mais itens são apresentados ao respondente, proporcionando uma estimativa provisória de sua proficiência. Na etapa adaptativa, um item mais informativo para a proficiência do respondente é aplicado, reestima-se a proficiência e repete-se o ciclo até que uma regra de parada pré-estabelecida seja satisfeita, chegando-se à etapa final, na qual o resultado é divulgado (Magis & Barrada, 2014).

Quando o indivíduo responde corretamente a um item, o próximo item a ser apresentado será mais difícil; caso a resposta seja incorreta, um item mais fácil será apresentado, e assim por diante, até atingir a regra de parada. Itens muito fáceis ou muito difíceis fornecem poucas informações úteis sobre o nível de proficiência dos respondentes (Luecht et al., 1998; Wainer, 2000) e acabam, muitas vezes, por desestimulá-los.

A Figura 2 apresenta uma ilustração da aplicação do CAT a um respondente com estimativa da proficiência igual a 1,2 (linha tracejada) e uma regra de parada de 20 itens. É possível observar o padrão de respostas, em que zero significa resposta incorreta e um significa resposta correta.

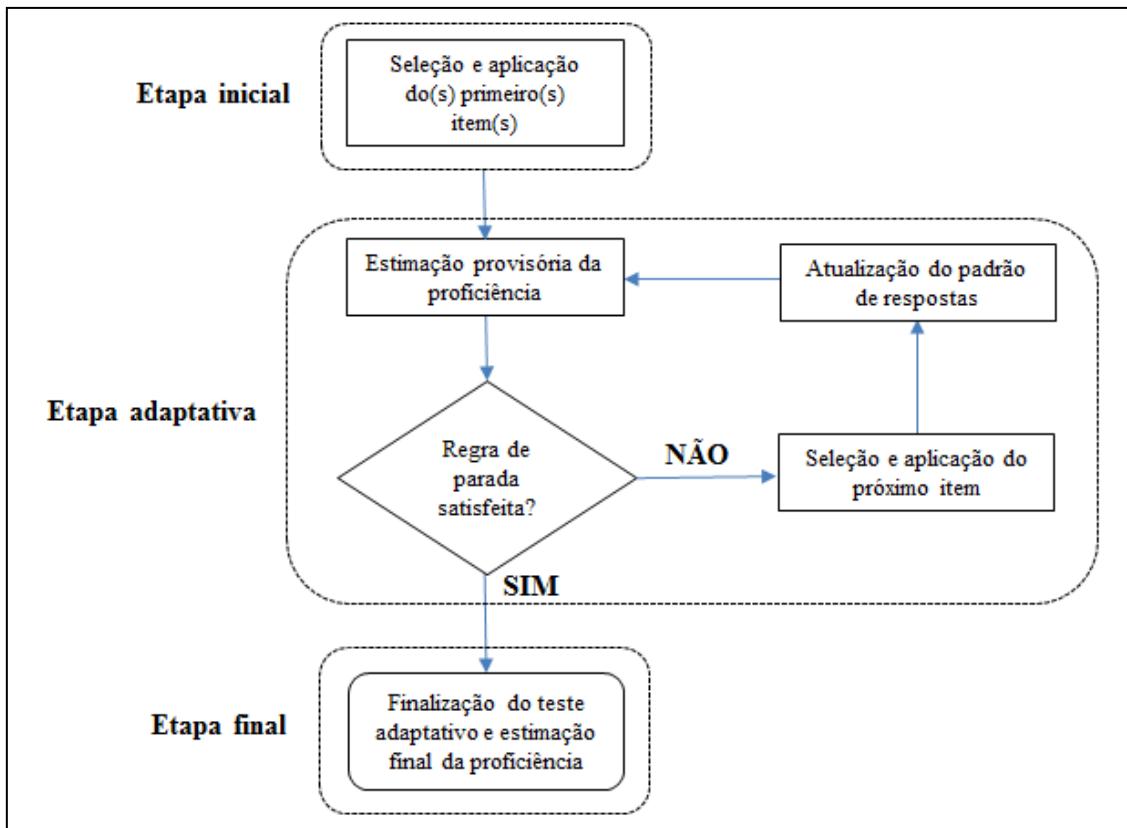


Figura 1 - Algoritmo geral de um CAT.

Fonte: Adaptado de Magis e Barrada (2014).

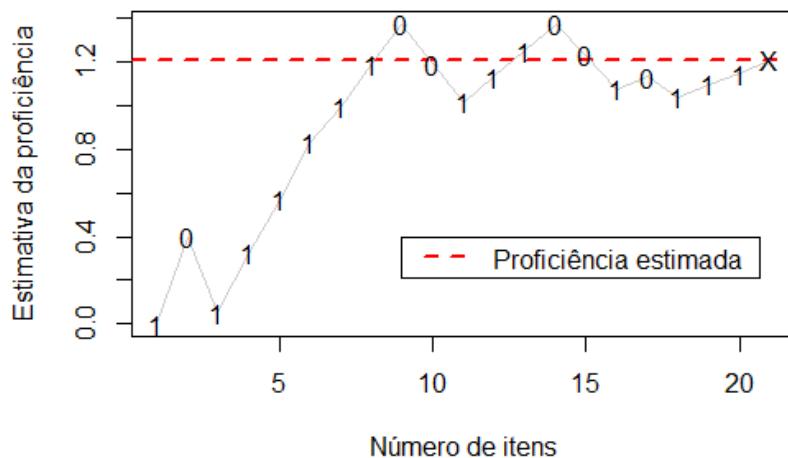


Figura 2 - Padrão de resposta de um indivíduo submetido a um CAT.

Fonte: Elaborada pelos autores.

Os itens são selecionados após a estimação da proficiência a cada rodada e, assim, acabam sendo, quando comparados com testes tradicionais, mais informativos e eficientes (Kopec et al., 2008). Desta forma, CATs tendem a reduzir o número de itens aplicados sem gerar a redução da precisão do teste, ocasionando economia de itens do BI, diminuindo o tempo de teste e possível fadiga dos respondentes, a qual pode comprometer o desempenho (Bjorner, Chang, Thissen, & Reeve, 2007; Luecht et al., 1998; Vispoel, 1998; Wang & Vispoel, 1998).

CATs também oferecem maior segurança quanto à distribuição e aplicação do teste, pois é oriundo de um BI que é apenas parcialmente utilizado a cada aplicação, diferentemente dos testes P&P, em que todos os itens do instrumento são apresentados. Além da redução de custos com materiais para o desenvolvimento dos testes, armazenamento e correção, destaca-se, também, a facilidade para pré-testar itens novos, que podem ser aplicados na sequência do CAT (Wainer, 2000), como será melhor detalhado subsequentemente neste texto.

Esses testes trazem benefícios quando da sua aplicação em avaliações na área da educação à distância (EaD) (Huang, Lin, & Cheng, 2009; Kaya & Tan, 2014; Ozyurt, Ozyurt, Baki, & Guven, 2012b; Ozyurt, Ozyurt, Baki, & Guven, 2012b; Salcedo, Pinninghoff, & Contreras, 2005) e no processo de aprendizagem dos alunos em ambiente de *e-learning*, em que o material didático tem a possibilidade de ser adaptado ao nível do aluno (Huang et al., 2009; Jeong & Hong, 2013; Ozyurt et al., 2012b; Ozyurt et al., 2012b).

De acordo com Huff e Sireci (2001) e Parshall et al. (2010), formatos inovadores de itens fazem uso da capacidade do computador para melhor mensurar um domínio do construto. Esta forma de aplicação de testes proporciona grande variedade no modo de apresentação dos itens que não são possíveis em testes P&P, como anexar várias multimídias - gráficos, áudio, vídeo e animação no enunciado do item e/ou nas opções de respostas (Huang et al., 2009; Parshall, Harmes, Davey, & Pashley, 2010; Wainer, 2000). Inclusive, há para os indivíduos a opção de responder a um item realçando o texto, clicando em gráficos, arrastando ou movendo objetos ao redor da tela, ou, até mesmo reordenando uma série de declarações ou imagens (Parshall et al., 2010).

Parshall et al. (2010) classificam a taxonomia dos itens inovadores em sete aspectos: (1) estrutura de avaliação - define a estrutura da apresentação do item e o tipo de resposta coletada; (2) ação de resposta - meios pelos quais os respondentes fornecem suas respostas, por exemplo, teclado e mouse; (3) inclusão de mídia - uso de elementos

como gráficos, som ou vídeo em um item; (4) interatividade - descreve a extensão em que um item reage ou responde ao *input* do respondente; (5) complexidade - número e variedade de elementos que o respondente necessita interpretar e utilizar, a fim de responder a um item; (6) fidelidade - grau em que um item fornece uma representação realista e precisa dos objetos reais, situações ou tarefas que fazem parte do construto a ser medido; e (7) método de escore - como as respostas são traduzidas em escores quantitativos. O uso de itens inovadores em CAT favorece a sua utilização de forma conjunta com o Desenho Universal (DU).

O DU é um conceito que surgiu no campo da arquitetura e busca, por meio da aplicação dos seus princípios, a acessibilidade plena para pessoas com e sem deficiência. Quando aplicado na elaboração de instrumentos, prima pela acessibilidade do mesmo, sendo denominado de Testagem Universal (TU). A utilização do CAT nos instrumentos de TU permite que a acessibilidade seja potencializada, já que se aliam a todas as vantagens dos testes informatizados (por exemplo: a já citada utilização de itens inovadores) a um número menor de respostas necessárias para estimativas mais eficientes das proficiências dos indivíduos, incrementando, assim, o DU dos instrumentos.

2. TESTAGEM UNIVERSAL E CAT

O desenvolvimento e a adaptação de testes psicológicos e educacionais, bem como todos os aspectos que os tangem tornam-se um problema científico de interesse da TU. A TU respeita todos os pressupostos teóricos instituídos no tocante aos testes para os diferentes campos, preocupando-se de forma mais específica com o formato dos testes e sua influência na realização do mesmo. Foca as adaptações e os recursos de tecnologias assistivas quanto proporcionadores de acessibilidade, analisando sua repercussão na qualidade dos instrumentos. Os princípios do DU, na sua aplicação à testagem, foram sistematizados em sete princípios da TU por Thompson, Johnstone e Thurlow (2002) (Thompson, Johnston, & Thurlow, 2002): (1) população de avaliação ampla e inclusiva; (2) definição precisa do construto; (3) itens acessíveis e não tendenciosos; (4) testes flexíveis a acomodações; (5) instruções e procedimentos simples, claros e intuitivos; (6) leitura agradável e de máxima inteligibilidade; e (7) máxima legibilidade. Um teste será considerado de TU se atender a maioria destes sete princípios.

No desenvolvimento de testes psicológicos de TU destinados às pessoas com deficiência, faz-se imprescindível a utilização de tecnologias assistivas, as quais criam um lócus de acessibilidade. A tecnologia assistiva, quando bem empregada, viabiliza a realização dos instrumentos por pessoas com deficiência de forma equitativa às pessoas sem deficiência. Porém, tais recursos exigem que o instrumento seja planejado desde o princípio com vistas a garantir a exequibilidade técnica do mesmo. Os testes computadorizados se tornam grandes aliados na construção de testes de TU, por permitirem que os instrumentos possuam maior flexibilidade, agregando várias possibilidades de usos e recursos adicionais para as pessoas com deficiência.

Na TU, almejam-se sistemas realmente inclusivos desde a sua forma de acesso, seus formatos, seus itens, a utilização de tecnologias assistivas variadas até as estimativas das proficiências dos indivíduos; nestas últimas, os testes adaptativos contribuem enormemente. A união da CAT com a TU permite obter estimativas da proficiência mais eficientes, evitando a fadiga e, portanto, facilitando a realização pelos indivíduos com deficiências, sobretudo quando da presença de vários comprometimentos físicos e/ou cognitivos(Almond et al., 2010; Ketterlin-Geller, 2005).

Sistemas computadorizados adaptativos são utilizados nos instrumentos de TU para o incremento da precisão e da eficiência. Medir os resultados dos respondentes que se encontram na extremidade inferior ou superior da escala de proficiência, tanto de indivíduos com deficiência quanto sem deficiência, é bastante relevante em certos construtos no campo da saúde mental (por exemplo, depressão e ansiedade). Tais estimativas potencializam o objetivo da TU de maior inclusão. Do mesmo modo que aos demais testes computadorizados, todos os princípios da TU são aplicáveis aos testes de CAT. Apesar de não serem imprescindíveis à elaboração de instrumentos de TU, os recursos do CAT se tornam ideais para a criação de tais instrumentos (Almond et al., 2010; Ketterlin-Geller, 2005). Assim, o CAT é um dos recursos mais potentes utilizados na TU.

3. COMPONENTES DE UM CAT

Os testes adaptativos computadorizados apresentam, basicamente, cinco componentes: (1) conjunto de itens calibrados pela TRI (banco de itens); (2) método para iniciar o teste; (3) método de seleção dos itens; (4) método de estimação da proficiência; e (5) regra de parada do teste. Testes que exigem maior controle utilizam

algumas restrições, as principais são balanceamento de conteúdo, controle da exposição do item e tempo de resposta.

3.1 Banco de Itens

O desenvolvimento de um BI adequado é importante para o sucesso de um CAT. É preciso ter um BI com boas qualidades psicométricas para mensurar com precisão e acurácia a proficiência dos respondentes em todos os níveis da escala. Deve haver validade de conteúdo, pois se almeja a cobertura de todos os aspectos do construto a ser mensurado (Bjorner et al., 2007). Além disso, é de grande relevância o estudo cuidadoso das questões levantadas anteriormente sobre o uso do TU/DU, para a especificação do formato do teste, de tal forma que maximize a sua acessibilidade para pessoas com perfis variados.

Um BI tem a função de armazenar um conjunto de itens de teste acompanhado de suas classificações e estatísticas (Bergstrom & Gershon, 1995), fornecendo a base para a seleção dos itens mais informativos para cada respondente durante o teste (Bjorner et al., 2007; Walker, Böhnke, Cerny, & Strasser, 2010). Este BI pode ser continuamente ampliado com a inserção de novos itens, assim como devem ser excluídos itens que são expostos com muita frequência ou que seu conteúdo se tornou obsoleto (Veldkamp & Matteucci, 2013).

3.1.1 Pré-testagem de itens

Uma das etapas essenciais à construção de um BI envolve o processo de pré-testagem dos itens, processo este que tem como objetivo a coleta de dados para a etapa de calibração dos mesmos. O processo de pré-testagem é realizado em momentos diferentes com estratégias muito variadas. Segue uma descrição geral de tal processo considerando-se dois momentos possíveis para a sua realização: *Pré-testagem de itens para a construção do BI inicial* e *Pré-testagem de itens ao longo do uso de testes CAT*.

Pré-testagem de itens para a construção do BI inicial

Quando um pesquisador está construindo um teste adaptativo computadorizado deve, entre outras coisas, dedicar-se ao desenvolvimento de itens e a sua pré-testagem para permitir a verificação de suas propriedades psicométricas que, se for favorável, resultará na sua inclusão no BI. Quando o pesquisador não dispõe de um conjunto suficiente de itens para compor o BI, deve utilizar um planejamento de coleta de dados

que permita atender aos requisitos para a futura calibração dos itens. Para tanto, cita-se dois cenários principais: (a) o pesquisador já tem um grande número de itens a serem pré-testados e calibrados e (b) o pesquisador tem um conjunto inicial de itens desenvolvidos e pretende continuar a construção dos itens ao longo do tempo para posteriormente calibrá-los.

Na primeira situação, o pesquisador precisa buscar uma forma para a coleta de dados que viabilize a posterior calibração dos itens para o construto ou construtos de tal forma que sejam colocados na mesma métrica. Uma das possíveis alternativas para tanto, sem que seja solicitado aos participantes que respondam a um número muito elevado de itens, envolve o uso de *Blocos Incompletos Balanceados (BIB)*. O uso de BIB é adequado quando: a. precisa-se realizar a coleta de dados a partir de um grande conjunto de itens cuja aplicação não seria viável aos participantes em função de sua excessiva extensão e b. Quando as análises a serem realizadas requerem uma matriz de correlações completa entre os itens. O uso de desenhos metodológicos de BIB é relativamente antigo e tal método propõe uma resolução para a situação apontada por meio da divisão do banco de itens em conjuntos menores (subtestes), os quais são combinados em cadernos de aplicação (blocos) (Bose, 1939; Bose & Nair, 1939).

Para ilustrar o uso de BIB na pré-testagem de itens pode-se indicar uma situação em que dispõe-se de aproximadamente 500 itens construídos para a avaliação de um dado construto. Diante de uma situação na qual não seja viável a pré-testagem de todos os itens conjuntamente, seja pelo provável efeito excessivo da fadiga, seja por questões práticas ou logísticas, é possível fazer o uso de BIB para que os dados coletados permitam a realização de procedimentos estatísticos que requeiram respostas válidas entre todos os pares de itens. Um desenho possível seria dividir os itens em 15 subtestes compostos por 35 itens, os quais deveriam ser combinados em cadernos de aplicação compostos por três subtestes gerando, assim, cadernos (ou blocos) compostos por 105 itens. Seriam necessários 35 cadernos diferentes para que seja obtido um desenho balanceado, em que os subtestes são apresentados em igual número e todos os pares de subtestes sejam contemplados. Vale ressaltar que o delineamento de BIB pode ser feito com relativa facilidade com o uso de pacotes específicos para tal propósito, como o *crossdes*, que roda sob o ambiente R (Sailer, 2005). A Figura 3 ilustra o desenho de coleta para um banco de itens com uso BIB.

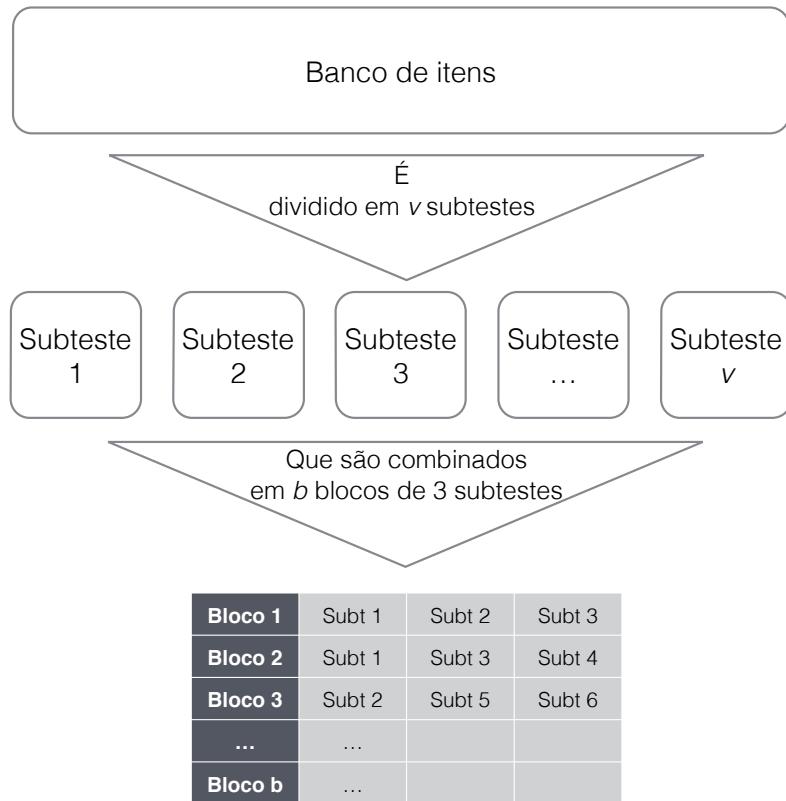


Figura 3. Principais elementos de um desenho de BIB.

Fonte: Elaborada pelos autores.

O uso da pré-testagem de itens com o desenho de BIB permite estudos posteriores sobre a dimensionalidade dos itens com uso dos modelos tradicionais de análise fatorial, uma vez que é viabilizada a estimativa da matriz completa de correlações entre todos os itens. Por fim, vale citar que a calibração dos itens tipicamente é feita mantendo-se livres todos os parâmetros de tal forma que o BI é calibrado conjuntamente.

Outro método que pode ser utilizado para a realização da pré-testagem de itens é o uso de cadernos de aplicação. Tais cadernos apresentam um conjunto de itens comuns, já calibrados e com propriedades psicométricas favoráveis, os quais são utilizados como itens-âncora. A ocorrência dos itens comuns entre os cadernos de aplicação viabiliza que os demais itens, sobre os quais não se tem informação psicométrica prévia, sejam equalizados entre si. Este método é referido na literatura como *grupos não equivalentes com uma âncora - NEAT* e é realizado pela manutenção dos parâmetros pré-estimados dos itens âncora e a livre calibração dos itens novos, que são escalonados para a mesma

métrica (Stocking & Lord, 1983; von, A. A. & Wilson, 2007), processo este chamado de equalização.

A equalização de vários conjuntos de itens com o uso da TRI é possibilitado pela relativa invariância de seus parâmetros com amostras independentes. Para a sua realização, é necessário que as amostras usadas para a calibração dos itens tenham suficiente variância em relação à magnitude do traço latente mensurado, que o construto avaliado pelos diferentes conjuntos de itens seja o mesmo e que haja itens comuns entre os blocos de itens a serem equalizados (Andrade, Tavares, & Valle, 2000; Pasquali, 2007a). Uma das vantagens desse método sobre o primeiro apresentado, com uso de BIB, é que o desenvolvimento dos itens e a sua aplicação nos variados cadernos podem ocorrer de forma sequencial, ou seja, não exige que todos os itens estejam prontos para o momento da pré-testagem. Como uma importante desvantagem pode ser citado que, pelo planejamento adotado, não é possível a obtenção de uma matriz de correlações completa entre todos os itens pré-testados e isso limita grandemente a realização de estudos sobre a sua dimensionalidade a partir de métodos tradicionalmente utilizados para tanto, como a análise fatorial exploratória (Hair, Anderson, Tatham, & Black, 2005; Hurley et al., 1997). Existem algumas possibilidades para superar tal dificuldade, como a realização de análises fatoriais para cada caderno – o que requer uma amostra bastante numerosa para cada um deles – ou o uso de procedimentos de *factor extension*, os quais foram desenvolvidos para desenhos metodológicos como o apresentado (Gorsuch, 1997).

Pré-testagem de itens ao longo do uso de testes CAT

Uma característica importante de sistemas adaptativos computadorizados é que permitem, com grande facilidade, a realização da pré-testagem de itens. Tal processo tipicamente ocorre pela inclusão de itens de pré-teste ao longo da aplicação CAT, seja com a apresentação de itens operacionais e de pré-testes de forma mista ou pela apresentação dos itens novos em bloco, tipicamente após a aplicação dos itens operacionais do teste.

Além das variações na ordem de apresentação dos itens de pré-teste, também há propostas variadas sobre a forma da escolha dos itens que serão apresentados. O formato mais simples para lidar com essa questão é a apresentação de um conjunto fixo de itens também em uma ordem fixa. As principais vantagens desse método é a facilidade para a programação do sistema de aplicação dos itens e a maior facilidade

para a montagem do banco de dados para a sua calibração. Uma das desvantagens mais evidentes desse método é o possível efeito da ordem da apresentação dos itens em relação às suas propriedades psicométricas. Assim, por exemplo, os últimos itens potencialmente sofrem um maior efeito da fadiga dos respondentes do que os primeiros apresentados. Uma solução relativamente simples para este problema envolve o uso de itens fixos mas com ordem de apresentação aleatória.

Outra abordagem viável relacionada à escolha dos itens de pré-teste envolve o uso dos resultados parciais obtidos pela aplicação dos itens operacionais. Assim, por exemplo, quando é estimado que um respondente apresenta proficiência média, podem ser escolhidos itens novos cuja dificuldade prevista seja compatível com tal nível. Esta abordagem é consideravelmente mais sofisticada e complexa para ser implementada, uma vez que não se tem de antemão uma expectativa precisa do nível de dificuldade/severidade dos itens novos. Uma forma de minimizar tal questão é construir um algoritmo que faça a calibração dos itens conforme este vai sendo aplicado. Outra questão relevante relativa a este método é a necessidade da apresentação dos itens novos para respondentes com uma razoável variabilidade no seu nível de proficiência; caso contrário, a calibração destes não pode ser feita de forma adequada. Por fim, ressalta-se que este método requer ainda um mecanismo de controle de exposição dos itens de pré-teste, uma vez que, se aplicados em uma amostra que apresente uma distribuição normal de proficiência, os itens com dificuldades medianas terão maior chance de serem escolhidos pelo algoritmo, fazendo com que os itens de dificuldade extrema sejam bem mais lentamente aplicados.

3.1.2 Dimensionalidade do banco de itens

A avaliação da estrutura fatorial dos itens pré-testados é um processo muito importante na construção de um BI. A avaliação da dimensionalidade faz parte da busca de evidências de validade do teste baseada na sua estrutura interna (American Educational Research Association, American Psychological Association, National, Council on Measurement in Education, 1999), a qual pode ser realizada por uma variedade de métodos tais como por análise fatorial e modelagem de equações estruturais (De Ayala, 2009), sendo que este último também pode ser desdobrado em uma forma específica de análise fatorial, a saber, a confirmatória (Borkenau & Ostendorf, 1990).

Os estudos da dimensionalidade do BI visam identificar como os itens podem ser agrupados de tal forma que formem dimensões interpretáveis mais amplas, de tal forma que individualmente estas possam ser entendidas como unidimensionais. O pressuposto da unidimensionalidade afirma que existe um traço latente dominante responsável pela realização do conjunto de itens e é requisito para o uso dos modelos da Teoria de Resposta ao Item mais amplamente utilizados até o momento (Andrade et al., 2000). Violar este pressuposto pode levar a um desajuste na estimativa dos parâmetros ou erros padrão elevados (DeMars, 2010).

Não existem regras definitivas para decidir quando multidimensionalidade e dependência local possuem magnitude suficiente para causar problemas na mensuração do traço latente, ficando sob responsabilidade do pesquisador e dos profissionais especializados no assunto, definir se os efeitos são significativos para o escore e se é razoável supor que existe um fator dominante responsável pelo conjunto de itens (Thissen et al., 2007).

3.1.3 Modelos da Teoria da Resposta ao Item para calibração do BI

A TRI surgiu como uma alternativa para resolução de vários problemas da Teoria Clássica dos Testes (TCT). A Teoria da Resposta ao Item se insere no âmbito das teorias do traço latente que consistem em assumir que os comportamentos medidos por meio dos itens são uma representação destes mesmos traços. Os traços latentes são variáveis hipotéticas não observáveis as quais podem ser inferidas pelos comportamentos manifestos, que são mensurados pelos itens de um teste (Embretson & Reise, 2000; Nunes & Primi, 2009; Pasquali, 2007b; Urbina, 2007).

A utilização da TRI em estudos na psicologia e educação agrega vantagens em relação à aplicação da TCT. Esta última possui limitações bastante conhecidas e discutidas na literatura da área, tais como: os resultados dos instrumentos desenvolvidos por esse modelo são dependentes dos itens que os compõem (*test-dependent*), os parâmetros dos itens de um teste dependem da amostra de sujeitos em que eles foram calculados (*subject-dependent*) e, ainda, tal teoria parte da suposição de que a variância dos erros de medida é igual para todos os respondentes (Embretson & Reise, 2000; Nunes & Primi, 2009; Pasquali, 2007b; Urbina, 2007).

Na aplicação da TRI, muitas das limitações da TCT são superadas. Algumas destas são que: (1) na TRI, há independência em relação ao conjunto de itens, uma vez

que a estimativa da proficiência do sujeito independe da amostra de itens utilizados; (2) a estimativa dos parâmetros dos itens independe da amostra de sujeitos - desde que seja atendido o pressuposto de variância da amostra e o erro de medida associado não seja muito elevado; (3) a TRI permite selecionar itens em consonância com a proficiência do sujeito; (4) não se precisa supor que os erros de medida são iguais para todos os respondentes e para os diferentes níveis de proficiência; e (5) acresce-se, ainda, que não é necessário trabalhar com testes estritamente paralelos na realização da equalização de testes, como exige a Psicometria Clássica (Nunes & Primi, 2009; Pasquali, 2007b).

Os modelos da TRI se baseiam no pressuposto que o desempenho do indivíduo em qualquer item de um teste pode ser predito a partir da estimativa da sua proficiência ou traço latente – sendo uma função deste. Assim, é possível estimar a probabilidade que o indivíduo terá de apresentar certos padrões de respostas em um dado item, considerando seu nível estimado no traço latente (Pasquali, 2007b).

No que se refere ao traço latente medido, os modelos de TRI apresentam-se como unidimensionais e multidimensionais. Os modelos unidimensionais são aqueles que consideram que os itens que compõem um teste representam medidas dos diferentes níveis de um mesmo traço latente. Nos modelos multidimensionais, mais de um traço latente traduz as diferenças entre os indivíduos, ou seja, os itens serão representativos de vários construtos. Neste capítulo, são abordados apenas modelos unidimensionais, uma vez que são os mais utilizados atualmente (Embretson & Reise, 2000; Urbina, 2007).

Existem modelos da TRI para itens dicotômicos (respostas classificadas como correta ou incorreta) e politônicos (categorias de resposta ordinais ou não). Os modelos mais utilizados da TRI baseiam-se na distribuição logística, podendo ser de 1, 2 ou 3 parâmetros. Considerando um modelo logístico de três parâmetros (3PL) e J indivíduos submetidos a um teste, o qual possui I itens. Se um indivíduo j responde corretamente ao item i , tem-se $y_{ij} = 1$; caso a resposta esteja incorreta, tem-se $y_{ij} = 0$. A probabilidade p_{ij} de que o indivíduo j responda corretamente ao item i , é dada pela função de resposta ao item (Andrade et al., 2000):

$$p_{ij} = P(Y_{ij} = 1 | \theta_j) = c_i + (1 - c_i) \frac{1}{1 + e^{-a_i(\theta_j - b_i)}},$$

onde $i = 1, \dots, I$; $j = 1, \dots, J$ e θ_j está associado à proficiência do indivíduo j .

- $P(Y_{ij} = 1 | \theta_j)$ é a probabilidade de um indivíduo j com habilidade θ_j responder corretamente o item i e é representada pela curva característica do item (CCI);
- b_i é o parâmetro que representa a dificuldade do item i , medido na mesma escala da habilidade, $b_i \in (-\infty, +\infty)$. Quanto maior o seu valor, maior é a proficiência exigida do indivíduo para responder corretamente determinado item;
- $a_i > 0$ é proporcional à inclinação da CCI no ponto b_i e corresponde a discriminação do item i . Quanto maior o valor deste parâmetro, maior será o poder do item em discriminar os respondentes;
- c_i é o parâmetro do item que representa a probabilidade de acerto ao acaso, ou seja, de indivíduos com baixa proficiência responderem corretamente o item i ; como é uma probabilidade, $c_i \in [0,1]$

O modelo logístico de dois parâmetros da TRI (2PL) é utilizado quando não há a possibilidade de resposta ao acaso, ou seja, $c = 0$. De forma semelhante, têm-se os modelos de Rasch e de 1PL. Quando se assume $c = 0$ e a constante para todos os itens, tem-se o modelo logístico de um parâmetro (1PL) e quando a é igual a um ($a = 1$) para todos os itens, tem-se o modelo de Rasch (1960). O processo de estimação dos parâmetros do modelo é conhecido como calibração. Esta etapa é muito importante, pois será definida a escala que servirá de referência para a interpretação dos resultados do teste (Baker, 2001).

Muitos dos construtos estudados em psicologia necessitam de mais informação do que apenas as respostas em um formato dicotômico, como certo/errado ou verdadeiro/falso. Neste caso, informações relevantes são perdidas ao utilizarem-se itens dicotômicos quando o traço latente medido adequa-se melhor a um modelo politômico da TRI. Nos modelos politônicos, estimam-se as probabilidades de um indivíduo endossar ou marcar uma resposta “x” a cada uma das categorias ou alternativas de um dado item (Embretson & Reise, 2000; Valentini & Laros, 2011).

Existem vários modelos para itens de respostas politômicas, dentre os quais o de resposta gradual (GRM), proposto por Samejima (1969a). Este modelo é muito utilizado em sistemas adaptativos computadorizados e consiste em uma extensão do modelo de 2PL da TRI para uma variável ordinal. A aplicação da TRI para itens com mais de duas categorias potencializa o acesso a informação, porém exigem que a amostra de calibração do instrumento seja maior, pois há mais parâmetros a serem estimados

(Andrade et al., 2000; Assessment Systems Corporation, 2012; Embretson & Reise, 2000; Valentini & Laros, 2011).

No GRM, as categorias de um item i são ordenadas da menor para a maior e representadas por k . A probabilidade de um indivíduo j escolher uma categoria particular ou outra mais alta do item i pode ser dada por (Andrade, Tavares & Valle, 2000):

$$P_{i,k}^+(\theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_{i,k})}},$$

onde $b_{i,k}$ é o parâmetro de dificuldade da k -ésima categoria do item i e $k = 0, 1, \dots, m_i$, ($m_i + 1$) é o número de categorias do i -ésimo item. A probabilidade de um indivíduo j responder a categoria k no item i é dada pela expressão:

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

Por definição, $P_{i,0}^+(\theta_j) = 1$ e $P_{i,m_i+1}^+(\theta_j) = 0$. Assim, é necessário estimar, além do parâmetro de discriminação do item, m_i valores de dificuldade.

As categorias de cada item, neste modelo, têm valores de dificuldades graduais, por seu caráter ordinal, ou seja, $b_{i,1} \leq b_{i,2} \leq \dots \leq b_{i,m_i}$. Deste modo, para um item com cinco possibilidades de resposta, obtém-se quatro *thresholds*, que representam o nível de habilidade em que uma categoria apresenta uma probabilidade de resposta igual à sua sucessora. Nesse modelo, o *threshold* de cada categoria será sempre maior do que o de sua antecessora e o parâmetro de discriminação é constante para todas as categorias de resposta em um mesmo item, possibilitando sua variação apenas entre os itens (Assessment Systems Corporation, 2012; Embretson & Reise, 2000).

Existem ainda outros modelos para uso de itens com respostas politômicas dentro da abordagem da Teoria de Resposta ao Item. A tabela 1 apresenta, de forma resumida, suas principais diferenças e referências que apresentam maiores detalhes sobre os modelos.

Tabela 1. Especificidades dos modelos de TRI para itens politômicos.

Modelo	Discriminação	Dificuldade	Indicação de uso	Referência

<i>Rating Scale</i>	Igual a 1	<i>thresholds</i> iguais para os itens. É somado um valor (b_i) específico por item		(Wright & Masters, 1982)
<i>Partial Credit Model</i>	Igual a 1	<i>thresholds</i> calculados por item	Itens de desempenho	(Masters, 1982)
<i>Graded Response Model</i>	Calculada por item	<i>boundary</i> calculados para cada item	Uso em escalas tipo Likert	(Samejima, 1997)
<i>General rating scale model</i>	Calculada por item	b calculado por itens; um conjunto comum de <i>boundary</i> é calculado para o teste	Uso em escalas tipo Likert	(Assessment Systems Corporation, 2012)
<i>General Partial Credit Model</i>	Calculada por item	<i>boundary</i> calculados por itens	Itens de desempenho	(Muraki, 1992)

3.1.4 A verificação da ocorrência de *funcionamento diferencial do item - DIF*

Para um teste ser adequado para mensurar a proficiência de indivíduos, entre outros fatores, é necessário que os itens que o compõem funcionem da mesma forma para os diferentes subgrupos, ou seja, deve-se verificar o Funcionamento Diferencial do Item (DIF - do inglês *Differential Item Functioning*), o qual ocorre quando indivíduos com a mesma proficiência não tem a mesma probabilidade de responder a um item específico por pertencerem a grupos diferentes (culturalmente/socialmente) (Embretson & Reise, 2000), desfavorecendo um grupo em detrimento de outro.

Itens com DIF também são problemáticos em CAT, uma vez que diferentes itens são aplicados aos respondentes e não há como saber previamente se muitos itens com DIF serão destinados à mesma pessoa ou a um grupo, podendo prejudicar seu

desempenho ou produzir uma estimativa tendenciosa do traço latente (Hart, Deutscher, Crane, & Wang, 2009; Makransky & Glas, 2013).

Existem vários métodos passíveis de serem utilizados para verificar a existência de DIF, por exemplo, Mantel-Haenszel (MH) (Holland & Thayer, 1988) e suas variações, Regressão Logística (RL) (Rogers & Swaminathan, 1989), métodos baseados na TRI como testes de razão de verossimilhança da TRI (LR-IRT) (Thissen, Steinberg, & Wainer, 1988), teste de Lord (1980), teste de Raju (1988), entre outros. No entanto, diversos autores (Camilli & Shepard, 1994; O'Neill & McPeek, 1993; Zumbo, 2007) ressaltam que o foco nos resultados matemático-estatístico não deve ser o único enfoque a ser utilizado pelos pesquisadores. Busca-se não só constatar a presença de valores de DIF que prejudicam a medida, mas, também as causas para que o item ou teste apresente esses valores. Quando valores prejudiciais são encontrados é recomendada a retirada do item ou mesmo a reformulação do instrumento (Embretson & Reise, 2000; Zumbo, 2007).

3.2 Desenvolvimento do CAT

Vários estudos abordam as etapas envolvidas no desenvolvimento de um CAT (Cella, Gershon, Lai, & Choi, 2007; Moreira Junior, 2011; Thompson & Weiss, 2011; Veldkamp & Matteucci, 2013; Walker et al., 2010; Wise & Kingsbury, 2000). Dentre elas, destacam-se: (1) banco de itens: questões relacionadas ao tamanho do BI, verificação da dimensionalidade de um conjunto de itens, independência local e avaliação da equivalência dos parâmetros dos itens entre os subgrupos (análise do funcionamento diferencial dos itens – DIF), modelos de resposta, remoção e revisão do item, adição de itens ao banco, consistência da escala ao longo do tempo; (2) desenvolvimento do CAT: efetuar simulações para comparar e avaliar diferentes métodos e especificações incorporadas no algoritmo; (3) segurança do teste: preocupações com a divulgação e roubo de itens e (4) questões relacionadas aos respondentes: tempo de teste, apresentação dos itens e comparabilidade.

3.2.1 Métodos para iniciar o teste

Nesta etapa, estabelece-se uma regra para inicialização do teste. Usualmente, uma estimativa inicial do nível de proficiência do respondente é fornecida; por exemplo, a média da distribuição de proficiência da população, ou a seleção de forma aleatória a partir da distribuição da proficiência (Veldkamp & Matteucci, 2013). Em alguns casos, é

possível utilizar informações conhecidas a priori sobre o respondente, como o escore de um teste feito anteriormente, para que a estimativa inicial fique mais próxima o possível da estimativa da proficiência após a realização dos primeiros itens (Van der Linden & Pashley, 2010). Uma vez que, quanto mais acurada é a estimativa inicial, mais apropriada será a seleção dos próximos itens (Chen et al., 2000).

Duas estratégias bastante utilizadas para lidar com a seleção do primeiro item, cuja efetividade deve ser avaliada por meio de simulações, envolvem: (a) a seleção de itens considerando que a proficiência dos participantes é igual a zero e (b) o sorteio da proficiência inicial em uma faixa de -1 a +1 (Weiss & Guyer, 2010).

Uma alternativa para tentar minimizar o problema de imprecisão das estimativas iniciais da proficiência é aplicar vários itens - por exemplo, seis itens – e estimar a proficiência (van Krimpen-Stoop & Meijer, 1999; Wouters, Zwinderman, Van Gool, Schmand, & Lindeboom, 2009). Aplicações de CAT na área da saúde geralmente utilizam o mesmo item inicial para todos os respondentes, o qual apresenta uma dificuldade mediana e, na maioria das vezes, define o construto. Por exemplo, para medir fadiga, o item administrado a todos é “eu tenho falta de energia?” (Cella et al., 2007).

Chang e Ying (2009, 1999) sugerem a utilização de uma estratégia de seleção de itens que estratifica o BI e empregam itens menos discriminativos no início do teste, quando não se tem informações sobre o nível de proficiência do respondente, garantindo que itens com alta discriminação sejam aplicados em fases posteriores, permitindo melhor estimação da proficiência final.

3.2.2 Métodos para seleção dos itens

Entre as estratégias para seleção de itens, tipicamente comparadas por simulação para a construção do sistema CAT, cita-se algumas relativamente simples, como a que os itens são escolhidos por serem os mais informativos para o resultado parcial estimado (Máxima Informação de Fisher - MFI). A principal desvantagem deste método é que, se a qualidade psicométrica do banco de itens for muito variada, apenas uma parcela destes será efetivamente utilizada em aplicações CAT (Georgiadou, Triantafillou, & Economides, 2007b).

Uma técnica derivada da anterior envolve o sorteio de um item dentre os “ n ” melhores para a mensuração da proficiência estimada até a rodada realizada. A seleção do conjunto de itens que são usados para o sorteio é feita pelos n itens que apresentam

maior informação para a proficiência avaliada, o que ocorre na região em que se localiza a dificuldade dos itens, quando estes são dicotômicos. Para itens politônicos, a lógica é a mesma, mas é utilizada a função de informação que agrupa as características dos pontos da escala adotada. Em ambos os casos, após a administração de cada item, é estimada a proficiência e é verificado quais seriam os próximos itens disponíveis que maximizariam a função de informação, sendo muito comum o uso da função de informação esperada de Fisher (Simms & Clark, 2005). O método descrito neste parágrafo é considerado o mais eficiente para a realização de CAT, pois propicia a mais rápida redução do erro padrão de medida (Weiss & Guyer, 2010).

Apesar do método da máxima informação ser considerado referência para a seleção de itens em sistemas CAT, é importante destacar que outros métodos buscam, por estratégias variadas, alcançar maior eficiência (Choi & Swartz, 2009). Dentre os métodos propostos, Verkamp e Berger (1997) indicaram que o *maximum likelihood weighted information* (MLWI), no qual é aplicado um peso à função de informação baseado na probabilidade do *theta*, tende a apresentar maior eficiência do que o método tradicional (MFI) para itens dicotômicos. Os autores também indicaram que a seleção de itens baseada no método MFI avaliada pelo método *expected a posteriori* (EAP) apresenta, para itens dicotômicos, vantagens semelhantes às obtidas com o método MLWI.

Neste mesmo tocante, Van der Linden e Pashey (2000) compararam métodos para seleção aplicados em itens dicotômicos, a saber MFI, MEI (*maximum expected information*), MEPV (*minimum expected posterior variance*), MEPWI (*maximum expected posterior weighted information*) e MPWI (*maximum posterior weighted information*), sendo que os quatro últimos usam conceitos Bayesianos para gerar informações globais. Os autores encontraram resultados que indicaram que os métodos MPWI e MFI apresentaram níveis de eficiência semelhantes em testes com 10 ou mais itens. Para testes mais curtos, os métodos MEPV, MEPWI e MPWI foram mais eficientes do que o MFI.

Choi e Swartz (2009) realizaram um estudo no qual compararam a eficiência de seis métodos para seleção de itens politônicos calibrados no modelo GRM (Samejima, 1969b). Os procedimentos adotados pelos autores foram os mesmos avaliados por van der Linden e Pashey (2010), a saber, MFI, MLWI, MPWI, MEI, MEPV, MEPWI, sem que fossem utilizados processos para controle de exposição e balanceamento de conteúdo. Os autores utilizaram tanto dados reais quanto simulados para a realização

das análises, fazendo a combinação de bancos reais e simulados com itens reais e simulados. Nos estudos realizados, verificaram como os procedimentos para seleção de itens funcionaram considerando-se três tamanhos fixos de testes, de 5, 19 e 20 itens. Como resultados, os autores evidenciaram que há equivalência matemática entre os métodos MEPWI e MPWI e os resultados verificados por tais modelos foram idênticos.

Ainda no mesmo estudo (Choi & Swartz, 2009), em se referindo as análises que envolveram bancos reais com itens reais não foram evidenciadas diferenças pronunciadas entre os métodos avaliados. Pode-se ilustrar com a comparação do erro padrão quadrático médio dos métodos MFI, o qual obteve pior resultado, e do MEI, com o melhor resultado. A diferença entre os dois resultados foi de apenas 0,0075 para um teste com cinco itens. Uma avaliação detalhada dos resultados obtidos nos outros estudos realizados pelos autores indica que as diferenças verificadas entre os indicadores de eficiência do método MFI em relação aos métodos Bayesianos testados foram mínimas em todas as condições avaliadas. Concluíram os autores que, ao menos para itens politônicos, o uso do método MFI é facilmente defensável face a sua relativa simplicidade e ampla implementação nos sistemas disponíveis para a avaliação computadorizada adaptativa.

Os procedimentos de seleção de itens para a maioria dos programas CAT são baseados em critérios estatísticos e em restrições, as quais podem ser estabelecidas diretamente no BI ou no algoritmo e servem para controlar atributos ou estrutura do BI (Van der Linden & Pashley, 2010). Em van der Linden e Reese (1998), é possível encontrar uma lista de possíveis restrições em testes. No entanto, deve-se evitar a inserção de restrições desnecessárias, uma vez que podem reduzir a precisão do teste quando o BI é pequeno e os testes são curtos, e sua implantação gera custos (Luecht et al., 1998; Van Der Linden, 1999; Van Der Linden, Scrams, & Schnipke, 1999).

A restrição de balanceamento de conteúdo garante que itens de todos os conteúdos ou domínios sejam aplicados aos respondentes, pois, se nenhuma restrição de conteúdo é imposta, não haverá garantias de que todos os domínios do construto serão representados, principalmente se a seleção de itens busca apenas maximizar a informação da proficiência estimada (Van der Linden & Pashley, 2010).

Esta estratégia controla o número de itens selecionados de cada domínio, além de maximizar a informação dentro dos domínios para selecionar o próximo item a ser aplicado (Luecht et al., 1998). Alguns estudos visam avaliar os impactos causados pela falta dessa restrição para avaliação adequada do traço latente investigado, como os de

Fliege et al. (2009), Luecht, de Champlain e Nungester (1998) e Zheng, Chang e Chang (2013).

Alguns métodos para impor esta restrição são: estratificação do BI por domínios (Chang & Ying, 2009; Van der Linden & Pashley, 2010); Multi-estágios (Van Der Linden & Reese, 1998; Van der Linden & Pashley, 2010); *Shadow test* (Belov & Armstrong, 2009; Van der Linden & Pashley, 2010) e modelo de programação linear LP 0-1 (Van Der Linden & Reese, 1998).

Outro tipo de restrição que tipicamente é realizada em testes adaptativos é o controle de exposição dos itens e, assim como os demais parâmetros técnicos para o funcionamento do sistema de aplicação, seu efeito na aplicação do teste geralmente é estudado por meio de simulações. O controle de exposição dos itens envolve a restrição da apresentação de itens populares¹ para evitar que estes se tornem conhecidos em função da sua utilização muito frequente. O problema de o item tornar-se conhecido é que os respondentes podem se preparar para respondê-los, o que geraria uma mudança em seus parâmetros psicométricos e decorrente viés na estimativa das proficiências dos respondentes. O controle de exposição dos itens tem sido adotado com diferentes objetivos, sendo que um deles envolve aspectos de segurança do banco de itens, especialmente relevante quando o teste é utilizado em contextos de alto impacto no qual há a preocupação que os itens sejam divulgados para grupos específicos de pessoas (Georgiadou et al., 2007b).

Outro motivo que leva ao uso do controle de exposição de itens é garantir que mesmo itens associados com níveis extremos de dificuldade tenham uma utilização assegurada no sistema adaptativo. Isto decorre do fato de que estratégias de seleção de itens baseadas na função de informação tendem a selecionar itens mais discriminativos para regiões de proficiência mais comuns (níveis medianos), reduzindo grandemente a chance de um item mais extremo ser selecionado.

De acordo com Veldkamp e Matteucci (2013), quando a seleção de itens é realizada por métodos que maximizem algum tipo de critério de informação, tipicamente 20% dos itens do BI são selecionados para a aplicação, e esta superexposição de alguns itens causa a subutilização de outros itens. Quanto mais baixa

¹ São considerados itens populares aqueles que, nos estudos de simulação ou no mapeamento empírico do uso do BI, são identificados como tendo elevadas taxas de exposição (Pastor, Dodd, & Chang, 2002)

for a taxa de exposição, menor é a quantidade de sobreposição de testes, ou seja, número de itens comuns entre os respondentes (Stocking, 1994).

Conforme Georgiadou, Triantafillou e Economides (1994), os métodos de controle de exposição são classificados em cinco grupos de estratégias: procedimentos de seleção condicionais, aleatórios, estratificados, métodos combinados e *design* de testes adaptativos em multi-estágios. De acordo Leroux et al. (Leroux et al., 2013), procedimentos condicionais tentam controlar as taxas de exposição com base em determinados critérios, por exemplo, a frequência de uso (método de Sympson e Hetter, (1985)). Os procedimentos aleatórios introduzem alguma randomização no processo de seleção de itens para um determinado subconjunto de itens, por exemplo, os métodos 5-4-3-2-1 de McBride e Martin (1983) e *randomesque* de Kingsbury e Zara (1989). Métodos estratificados buscam estratificar o conjunto de itens de acordo com propriedades estatísticas e são aplicados a partir de um determinado estrato, por exemplo, métodos *a*-estratificado de Chang e Ying (1999) e Chang, Qian e Ying (2001). Os procedimentos combinados ocorrem quando dois ou mais métodos de controle de exposição são combinados, por exemplo, método progressivo restrito de Revuelta e Ponsoda (Revuelta & Ponsoda, 1998) e erro padrão progressivo restrito de McClarty et al. (2006). O método proposto por Chang e Ying (1999) sugere estratificar o BI em relação à discriminação dos itens para tentar utilizar de forma mais homogênea o BI e não superexpor apenas os itens com valores altos de discriminação.

Vários estudos têm comparado métodos de controle da exposição de itens para respostas dicotômicas em busca do melhor método (Barrada, Abad, & Veldkamp, 2009; Chang & Ansley, 2003; Leroux et al., 2013; Parshall, Davey, & Nering, 1998; Revuelta & Ponsoda, 1998). Porém, Revuelta e Ponsoda (1998), Georgiadou, Triantafillou e Economides (Georgiadou, Triantafillou, & Economides, 2007a) e Leroux et al. (2013) destacam que cada um deles tem suas particularidades, vantagens e desvantagens, não sendo generalizável para todas as situações, por isso, métodos devem ser testados por simulação antes da aplicação de CATs.

Por fim, a proporção de vezes que um item será utilizado em aplicações CAT dependerá de suas propriedades psicométricas, dos outros itens que estão disponíveis no banco e da distribuição de proficiência dos respondentes (Revuelta & Ponsoda, 1998). Portanto, não há uma regra fixa para uma taxa máxima de exposição dos itens, pois, conforme Stocking (1994), ela é influenciada pela quantidade de respondentes.

3.2.3 Estimação da proficiência

Selecionar um método adequado para estimação da proficiência dos respondentes é especialmente importante. A estimativa da habilidade afeta não apenas o resultado final do teste, mas também quais itens são aplicados ao longo do teste (Wang & Vispoel, 1998).

Os principais métodos de estimação são: Esperança a posteriori (EAP) (Bock, Gibbons, & Muraki, 1988), Máxima a posteriori ou moda a posteriori (MAP) (Wang, Hanson, & Lau, 1999; Wang & Vispoel, 1998), estimação pela máxima verossimilhança (MV) (Cheng & Liou, 2000; van der Linden & Pashley, 2000; Wang & Vispoel, 1998), e estimação pela verossimilhança ponderada (WLE) (Cheng & Liou, 2000; van der Linden & Pashley, 2000; Warm, 1989).

As avaliações dos métodos são normalmente baseadas em índices como RMSE, viés e erro padrão (SE). De acordo com Wang e Vispoel (1998), a escolha entre MV e uma abordagem bayesiana dependerá do papel que o SE e o viés exercem na tomada de decisões a partir dos resultados do CAT.

3.2.4 Regras para a finalização de testes adaptativos

A conclusão da aplicação de um teste computadorizado adaptativo é determinada por variados critérios que são utilizados unicamente ou de forma combinada. A escolha de regras de finalização do teste é afetada por vários fatores, por exemplo, a comparabilidade entre os testes P&P e CATs, tempo de teste e eficiência da mensuração (Yi, Wang, & Ban, 2001). Dada a grande combinação de regras para a finalização de um teste CAT, tipicamente a decisão sobre o seu uso é embasada em resultados oriundos de simulações, as quais permitem balancear o impacto da adoção de diferentes regras e parâmetros para a conclusão sob diferentes características do teste, como a precisão alcançada, sua extensão, o uso repetido de conjuntos de itens entre os respondentes, seu impacto na taxa de exposição dos itens, entre outros. Apesar da grande variedade de regras para conclusão de testes adaptativos, enumeram-se algumas referidas na literatura e implementadas em pacotes estatísticos desenvolvidos para este propósito, como o CATSIM (Weiss & Guyer, 2010). Tais regras propõem que o teste seja encerrado quando:

1. O erro padrão de medida da proficiência é igual ou menor do que um valor estipulado. Esta é uma regra clássica em testes adaptativos, uma vez que operacionaliza a tentativa de alcançar uma medida com um determinado nível de

- precisão para todos os respondentes com o menor número de itens;
2. A diminuição dos erros padrão de medida (SE) sucessivos é menor ou igual do que um valor definido. A lógica deste critério é que com o avanço das rodadas de aplicação dos itens em um teste CAT, o SE deve ser gradualmente reduzido. A desaceleração dessa redução tipicamente significa que os itens que restam no banco são pouco informativos para o nível de proficiência do respondente. Este critério deve ser usado com cautela em sistemas que apresentem restrições para a seleção de itens com critérios muito exigentes, uma vez que, em certas rodadas de aplicação, podem não ser selecionados itens muito informativos em função dos mecanismos de balanceamento de conteúdo ou dos controles de exposição;
 3. O valor absoluto da mudança em sucessivas estimativas da proficiência é igual ou menor que um determinado valor. Este também é um critério muito usado em testes adaptativos e o argumento principal para seu uso é que, conforme o teste adaptativo vai se aproximando de uma estimativa “real” do traço latente, este tende a se estabilizar;
 4. O erro padrão de medida da proficiência aumenta em um valor estipulado. Este critério está bastante relacionado com o segundo. A principal diferença é que aqui supõe-se que o SE em uma medida por CAT não deve aumentar ao longo da aplicação;
 5. A informação obtida no último item usado é menor ou igual a um determinado valor. Em uma aplicação CAT, é esperado que todo item administrado seja capaz de fornecer um nível razoavelmente elevado de informação, colaborando assim para um incremento na precisão do teste. O baixo nível de informação nos últimos itens selecionados é um indicativo de exaustão do BI para a proficiência estimada, salvo quando regras muito restritas para exposição de itens e balanceamento de conteúdo são adotadas no teste;
 6. Um determinado número de itens foi administrado. Apesar de todas as vantagens já reconhecidas de aplicações CAT, há situações nas quais não é possível alcançar uma medida precisa com um baixo número de itens. Quando é assim, este critério pode ser acionado, evitando que um número excessivo de itens seja utilizado. Os argumentos para a definição do tamanho máximo de um teste adaptativo são para evitar o efeito da fadiga entre os respondentes e para

- evitar uma exposição exagerada do BI;
7. Não há itens disponíveis para a aplicação. Este é um critério obrigatório em uma aplicação CAT, apesar de seu uso não ser frequente quando o critério anterior é utilizado. Este, visa controlar situações nas quais o BI foi exaurido em função das restrições para seleção de itens. É mais provável a ocorrência desta situação quando as condições para aplicação são muito restritas como, por exemplo, o controle de exposição é excessivo. Nestas situações, mesmo com bancos de itens extensos, pode ocorrer que muitos itens sejam eliminados por condições aplicadas paralelamente.

4. EXEMPLO DA CRIAÇÃO DE UM TESTE ADAPTATIVO PARA AVALIAÇÃO DA PERSONALIDADE

Como uma forma de ilustrar a aplicação dos tópicos abordados neste capítulo, é feita a descrição do processo de construção de uma bateria computadorizada adaptativa para a avaliação da personalidade a ser utilizada no Brasil, a qual foi chamada de Bateria Adaptativa de Personalidade – BAP. A BAP é um instrumento objetivo para avaliação da personalidade embasada no modelo dos Cinco Grandes Fatores de Personalidade (CGF). Nas últimas décadas, foi evidenciada a convergência das pesquisas em âmbito internacional envolvendo modelos explicativos da Personalidade, tendo como referência principal o modelo dos Cinco Grandes Fatores (CGF). O modelo dos CGF representa uma versão moderna das teorias de traço e tem sido considerado internacionalmente uma das formas mais eficazes para a avaliação da personalidade. Neste modelo, a personalidade é descrita a partir de cinco grandes domínios, denominados no Brasil como Extroversão, Socialização, Realização, Neuroticismo e Abertura (Digman, 2002; Digman, 1996; Nunes, Hutz, & Nunes, 2010; Silva & Nakano, 2011). Nesta seção, são descritas, em linhas gerais, as principais etapas para a construção da BAP, desde a calibração dos itens até os cuidados finais com o algoritmo desenvolvido no ambiente Concerto (<http://www.psychometrics.cam.ac.uk/newconcerto>).

4.1. Calibração do banco de itens

O banco de itens para a composição da BAP foi montado em um projeto

realizado anteriormente em três fases: na primeira, foi selecionado um conjunto de itens da Bateria Fatorial de Personalidade (Nunes et al., 2010), que foram utilizados como âncora nas aplicações subsequentes; na segunda fase, foram construídos 675 itens novos para a avaliação dos cinco fatores, dos quais, foram pré-testados 400. Desses, 313 itens foram considerados com propriedades psicométricas favoráveis e foram equalizados com os itens âncora. O processo de coleta de dados utilizado para a pré-testagem de itens é contínuo e feito em cadernos, os quais contam com 100 itens novos e 40 âncora; na terceira fase, foram calibrados aproximadamente 110 itens para cada um dos cinco fatores, os quais já haviam sido aplicados com os itens âncora, permitindo a sua equalização. Os estudos para a verificação das propriedades psicométricas dos itens foram realizados com amostras variadas, que incluíram a que compôs os estudos de validade da BFP ($N=6.599$), a coletada para a equalização dos itens novos (1.290) e as que relacionaram as escalas individuais para a avaliação dos CGF com os itens âncora (3.953).

Para a construção da versão final da BAP, foi feita uma extensa revisão da literatura sobre os modelos utilizados nas medidas atuais baseadas em CAT que trabalham com itens politômicos. A adoção do modelo de Samejima (Samejima, 1997; Samejima, 1969b), referido como *graded response model*, tem-se mostrado bastante frequente em testes recentes (Choi & Swartz, 2009). Uma das principais características do modelo que o torna relevante no contexto de testes adaptativos é que a discriminação de cada item é estimada, bem como os *thresholds* entre as categorias utilizadas. A inclusão desses parâmetros são muito relevantes em CAT pois tipicamente os algoritmos para seleção de itens e alguns critérios para término utilizam a curva de informação dos itens para a tomada de decisões.

Os itens da BAP foram calibrados com o uso do software XCalibre versão 4.1 (Assessment Systems Corporation, 2012) separadamente por fator para o qual foram construídos. Vale salientar que, nos estudos anteriores sobre suas propriedades psicométricas, foram realizados procedimentos para verificar a hipótese de sua unidimensionalidade referente a cada fator avaliado. As estimativas foram feitas considerando o valor da constante D como 1.0, e o valor inicial do parâmetro "a" também igual a 1.0. Foram consideradas 40 pontos de quadratura durante a calibração dos parâmetros dos itens e o desvio-padrão do *theta* foi padronizado para 1.0. O método adotado para a calibração foi o *marginal maximum-likelihood* (MML), o que permitiu

uma direta correspondência com os resultados estimados com o pacote ltm e catIrt do R, os quais também adotam o valor de D igual a 1.0. Vale notar que apesar dos parâmetros estimados pelo pacote ltm serem idênticos aos gerados pelo Xcalibre, a calibração foi realizada com o último pacote, pelo maior detalhamento das análises e, por consequência, maior facilidade para as tomadas de decisão derivadas dos resultados.

A amostra foi composta por 8357 pessoas, sendo que 67% destas eram mulheres. A idade média dos participantes foi de 24.7 anos ($DP=10.20$). Um grupo de 1290 pessoas dessa amostra respondeu especificamente aos cadernos de pré-testagem para a ampliação do banco de itens e informaram o Estado brasileiro onde residiam. Este grupo foi oriundo de 25 Estados, sendo que os mais frequentes foram Santa Catarina (42.6%), São Paulo (22.25%) e Bahia (18.14%).

A tabela 1 apresenta os parâmetros psicométricos resumidos de um conjunto de itens para a mensuração de Extroversão. A tabela apresenta a posição do item no banco, sua identificação, sua média, correlação item-theta, a discriminação (a) e thresholds (b_k).

Tabela 1. Propriedades psicométricas dos itens de Extroversão

Seq.	Item ID	Item Mean	R	a	b_k
1	E2I3	3.043	0.402	0.812	-1.42, -0.66, 0.79, 1.69
2	E4I8	3.971	0.526	1.298	-2.64, -1.86, -0.61, 0.39
3	E1I17	3.363	0.557	1.369	-1.11, -0.61, 0.20, 0.59
4	E3I26	3.375	0.412	0.842	-2.49, -1.39, 0.37, 1.50
5	E1I38	3.068	0.571	1.389	-0.77, -0.32, 0.50, 0.91
6	E4I50	3.043	0.618	1.568	-0.99, -0.33, 0.57, 1.13

A Figura 1 apresenta a Função de Informação do Teste (TIF) para todos os itens calibrados para o fator Extroversão. A TIF é uma representação gráfica de quanta informação o teste é capaz de prover para cada nível de proficiência (Θ). O valor máximo de informação foi 36.35 para a proficiência de -0.10.

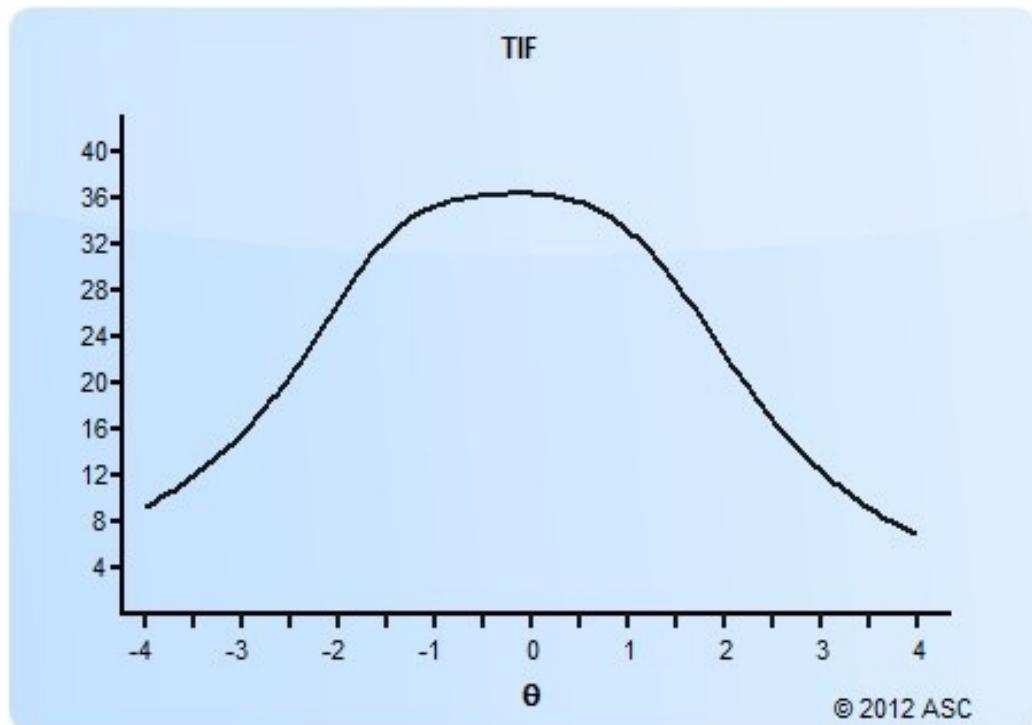


Figura 1: Função de informação do Teste para Extroversão.

4.2. Simulações para a especificação dos parâmetros de funcionamento da BAP

Para a realização das simulações com o banco de itens da BAP, foi utilizado o pacote estatístico CATSim versão 4.0 (Weiss & Guyer, 2010), específico para tais propósitos. O método utilizado para as simulações foi o híbrido, no qual são fornecidos ao programa os parâmetros psicométricos dos itens, bem como o banco de dados utilizado para tal calibração. Para resolver as questões dos *missings* no banco de itens, o CATSim estima a proficiência para cada respondente a partir dos itens respondidos e faz a imputação de dados considerando a resposta mais provável aos demais itens considerando seus parâmetros psicométricos. Na sequência, utilizando a base de dados completa, realiza as simulações de CAT conforme as condições definidas pelo pesquisador.

Os itens selecionados na calibração dos itens para os cinco fatores avaliados pela BAP foram analisados separadamente nas simulações. Ao todo, foram realizadas 75 simulações, nas quais foram manipuladas as seguintes variáveis:

- método para estimativa dos parâmetros, entre MML, Bayes EAP e Bayes MAP;

- proficiência inicial variando entre faixa (de -1 a +1) e zero;
- número de itens selecionados por rodada, entre 2 e 5;
- os itens selecionados foram sorteados de uma lista dos 10 melhores para a proficiência parcial;
- o número mínimo de itens foi mantido constante, definido em 8;
- o número máximo de itens foi definido como “livre” (até todos os itens para o fator), 25 e 20 itens;
- controle de exposição, com valores de 0.1, 0.15, 0.20, 0.30, 0.34 e 0.45.

Algumas combinações dessas variáveis e valores foram usadas para a realização das simulações e foram apurados os seguintes resultados:

- proficiência e erro padrão para o teste completo;
- proficiência e erro padrão para o teste aplicado em CAT;
- correlação entre os *theta*'s nos dois formatos;
- indicadores variados de resíduo: ASD, AAD, MSD, RMSD, SEM_full_mean, SEM_full_sd, SEM_cat_mean, SEM_cat_sd;
- precisão estimada pelo Alfa de Cronbach;
- média e desvio padrão do número de itens;
- número máximo de itens na simulação;
- % de casos em que o teste foi concluído com até 15, 20, 25 e 30 itens;
- número de casos encerrados por nove regras simuladas, que serão detalhadas posteriormente;
- proporção de itens com baixa e elevada taxa de exposição.

Após uma avaliação detalhada dos resultados obtidos, foram definidos os parâmetros de funcionamento do sistema adaptativo da BAP, como os SE para as regras de término do teste, os controles de exposição dos itens, etc. O anexo 1 apresenta, de forma ilustrativa, os resultados obtidos para o fator neuroticismo. Os resultados das simulações com os demais fatores foram convergentes com os apresentados nesta tabela, indicando a viabilidade da especificação de fatores globais de funcionamento da BAP.

A única variável especificada individualmente para os fatores foi o erro padrão de medida (SE) mínimo para gerar o encerramento da aplicação. Como este parâmetro depende diretamente das propriedades psicométricas dos itens, varia de fator para fator. O SE foi estimado de forma individual para cada um dos cinco fatores avaliados pela

BAP, para que fosse obtido um alfa de Cronbach de ao menos 0.85. Os valores resultantes para os fatores extroversão, socialização, realização, neuroticismo e abertura foram de 0.65, 0.65, 0.50, 0.50, 0.40, respectivamente.

O método de *maximum-likelihood* foi selecionado para a estimativa do *theta*, tendo este o valor inicial igual a zero. A forma de seleção dos itens escolhida foi o sorteio entre os 10 melhores, considerando a sua função de informação para a proficiência parcial. O número mínimo de itens a ser aplicado por fator foi fixado em 8 e o máximo em 15. Por fim, a taxa máxima de exposição dos itens foi definida como 0.35.

A especificação selecionada obteve, no conjunto, uma excelente combinação de resultados nas simulações. Foi obtida, por exemplo, uma correlação entre as proficiências estimadas com o teste completo e na versão adaptativa igual a 0.96. A diferença média entre tais proficiências foi de 0.01, o Alfa de Cronbach foi de 0.90 e a média do número de itens aplicados foi de 8.53, o que foi bastante próximo do número mínimo de itens aplicados especificado. Destaca-se ainda que 97.3% dos casos simulados tiveram o teste encerrado com menos de 15 itens e que, em 94.24% dos respondentes, o término ocorreu pelo fato de o SE ter alcançado a linha de corte definida, ou seja, foi obtida uma medida breve e com excelentes níveis de precisão. Apesar dos resultados bastante favoráveis para esta configuração, foi possível evidenciar a ocorrência de baixa taxa de exposição (abaixo de 0.10) em 73% dos itens. Comparando este resultado com o de outras configurações simuladas foi possível identificar que este resultado é decorrente do valor relativamente elevado da taxa de exposição utilizada. No entanto, como já estava prevista a inclusão da equiparação de conteúdo, o que resultaria na minimização deste efeito, o modelo foi escolhido.

4.3. Adaptação da BAP no ambiente Concerto

Após a etapa de simulações e definição dos parâmetros de funcionamento do sistema adaptativo da BAP, foi feita sua adaptação ao ambiente Concerto, o qual utiliza como *framework* o ambiente de programação do R. Assim, a parte do programa relacionado ao tratamento de dados, análises estatísticas e lógica para o fluxo de aplicação da BAP foi desenvolvida diretamente no R, utilizando como interface o RStudio (<http://www.rstudio.com>). Tal estratégia foi adotada pela maior flexibilidade e rapidez oferecida pelo RStudio para o depuramento de programas. Desta forma, o

código apenas foi transferido ao Concerto quando já estava bastante estável. Vale notar, no entanto, que isto só foi possível pela adoção da versão 4 do Concerto, uma vez que as anteriores não permitiam uma fácil transferência entre o código de programação do R para seu ambiente.

As principais mudanças necessárias para o funcionamento da BAP no ambiente concerto envolveram o desenvolvimento dos módulos de leitura e gravação de dados, baseados em mysql. Também foi necessária a mudança da interface com o usuário, a qual originalmente utilizava o sistema nativo do Rstudio, para o html, adotado pelo Concerto. A BAP foi desenvolvida em módulos para facilitar a sua atualização e customização em versões paralelas a serem utilizadas para diferentes demandas. Seus principais módulos são descritos na sequência.

Módulo inicial. Este módulo é responsável pelas etapas de inicialização do sistema, o que envolve a leitura do banco de itens no servidor e a preparação das matrizes e *arrays* de dados usados pelos módulos subsequentes. Neste módulo são também definidas as configurações de funcionamento do sistema adaptativo, derivados da etapa de simulação, e que dizem respeito às regras de início, bem como aos que critérios serão ativados e as linhas de corte que adotarão, entre outras. O módulo inicial é responsável ainda pela apresentação da pesquisa, do TCLE, do questionário sócio-demográfico e das instruções para a realização do teste.

Módulo para seleção de itens. Este é um dos módulos mais complexos da BAP e envolve desde a identificação de que facetas estão disponíveis para seleção de itens até o controle de exposição dos mesmos. O sistema foi produzido para realizar a seleção de cinco itens por rodada, inicialmente um para cada fator do modelo de personalidade adotado (dos CGF). Cada item é selecionado por sorteio entre os 10 melhores para a proficiência parcial do respondente, sendo que nas oito primeiras rodadas os itens são selecionados para os fatores Extroversão, Socialização, Realização, Neuroticismo e Abertura, nesta ordem. A partir da nona rodada de aplicação, são ativados os critérios para término da aplicação, a serem detalhados na sequência, e então a aplicação de um ou mais fatores pode ser concluída. Quando isso ocorre, são abertos “slots” para a aplicação na rodada subsequente e estes são preenchidos com os fatores ativos que apresentam maior erro padrão de medida até o momento. Assim, por exemplo, a seleção dos itens nas rodadas podem ocorrer conforme a tabela 6. É possível notar, no exemplo,

que até a oitava rodada os fatores para os quais os itens são selecionados permanecem inalterados mas, ao término desta, os fatores Extroversão e Abertura alcançaram algum dos critérios para término e então os espaços correspondentes para a sua avaliação foram preenchidos pelos fatores com maior erro de medida até o momento. Neste exemplo, o teste foi concluído na décima segunda rodada, na qual apenas o fator Neuroticismo foi avaliado.

É importante detalhar que o processo de seleção de itens é feito considerando a função da informação dos itens não aplicados no banco de itens, mais especificamente a técnica *unweighted Fisher information at a point*, utilizando como referência o theta estimado até a última rodada realizada.

Tabela 6. Exemplo da distribuição dos fatores avaliados ao longo da aplicação da BAP.

Rodada	Item 1	Item 2	Item 3	Item 4	Item 5
1	Extroversão	Socialização	Realização	Neuroticismo	Abertura
2	Extroversão	Socialização	Realização	Neuroticismo	Abertura
...					
8	Extroversão	Socialização	Realização	Neuroticismo	Abertura
9	Socialização	Socialização	Realização	Realização	Neuroticismo
...					
12	Neuroticismo	Neuroticismo	Neuroticismo	Neuroticismo	Neuroticismo

Além do controle dos fatores cobertos na seleção dos itens a cada rodada, existe um mecanismo para a balanceamento de conteúdo dos itens. Este mecanismo é utilizado com dois objetivos primordiais, a saber: a. garantir que todos os domínios do construto efetivamente sejam avaliados, gerando assim um resultado por fator mais representativo; e b. realizar um controle de exposição de itens, uma vez que conceitualmente pode haver uma faceta específica de um fator que apresente itens com melhores propriedades psicométricas. Caso isso ocorra, estes tenderão a ser mais frequentemente escolhidos do que as demais facetas, efeito este que é minimizado com a equiparação de conteúdo.

O balanceamento de conteúdo é feito pela escolha dos itens para as facetas de forma sequencial. Assim, por exemplo, se há itens em quantidade suficiente para um

fator com três facetas, na primeira rodada o item deverá ser da faceta 1; na próxima rodada, da faceta 2; na seguinte, da faceta 3 e depois o ciclo é reiniciado. Se neste processo o sistema verifica que alguma faceta não tem um número de itens necessário então apenas as demais serão usadas no processo de seleção de itens. O balanceamento de conteúdo é interrompido quando mais de um item deve ser aplicado para o fator em questão, o que só ocorre a partir do momento em que as regras para conclusão são ativadas, ou quando não há itens em quantidade suficiente para todas as facetas.

O módulo de seleção também é responsável pelo controle de exposição dos itens. Para tanto, utiliza o mesmo método adotado pelo CATSim, proposto por Sympson e Hetter (Hetter & Sympson, 1997). Apesar de bastante eficaz, esse método exige um cuidado especial principalmente quando a seleção de itens é feita com o balanceamento de conteúdo pois um fator que inicialmente apresenta um número suficiente de itens pode ser exaurido durante o processo de controle de exposição. Neste caso o sorteio dos itens deverá ser continuado na próxima faceta com itens suficientes ou com o fator geral quando não existe tal faceta. Vale salientar que o controle de exposição e balanceamento de conteúdo não são oferecidos na versão atual do pacote catR, utilizado como base para a construção da BAP. Dessa forma, foi necessária a programação de tais procedimentos com o uso dos recursos oferecidos no ambiente R

Módulo para aplicação de itens. O primeiro processo executado por este módulo é a verificação da ocorrência de exaustão do banco de itens, o que é mais provável em testagens compostas por muitas rodadas. Tal situação representa que não há mais itens disponíveis para os fatores ainda ativos, o que é mais provável quando o participante apresenta um padrão inesperado de respostas, o que dificulta a convergência da estimativa de seu traço latente, ou quando o mesmo apresenta um *theta* em uma região em que há bons itens para a avaliação. Quando é detectada tal situação, o sistema da BAP considera encerrada a aplicação para o indivíduo e grava uma observação no banco de dados indicando que os fatores não puderam ser mensurados adequadamente.

Quando é verificado que a situação de aplicação está avançando regularmente, sem a ocorrência de exaustão do banco de itens, os itens selecionados no módulo anterior são apresentados ao participante sempre de cinco em cinco. Quando algum deles não é respondido é apresentada uma nova tela, apenas com os itens cuja resposta falta ser informada, apenas após a conclusão desta é que o módulo passa para a etapa subsequente - de estimativa do *theta* do participante. Este controle restrito quanto à

ocorrência de *missing* é feito para que a cada rodada se obtenha, efetivamente, as melhores condições à convergência das medidas em estimação.

O cálculo do *theta*, conforme já comentado, é feito a partir do modelo *Graded response model*, proposto por Samejima (Samejima, 1969b) com o uso do método de máxima verossimilhança. Nas rodadas mais avançadas, nas quais são aplicados mais de um item para determinados fatores, o *theta* é estimado apenas uma vez incluindo-os simultaneamente na matriz de respostas e na matriz dos parâmetros dos itens. Nesta etapa também são calculados o erro padrão de medida e a informação gerada pelo conjunto de itens aplicados até então.

Módulo de regras para finalização. As regras para finalização representam uma etapa crucial em um teste adaptativo, uma vez que busca balancear variadas condições que devem tornar o teste o mais breve e preciso quanto for possível. No desenvolvimento da BAP, foram implantadas todas as regras para finalização adotadas no pacote CATSim. O sistema foi projetado de tal forma que a ativação de cada regra é possível mediante a configuração do módulo inicial.

É importante destacar que as regras para término da BAP ficam desativadas até a oitava rodada de aplicação dos itens. Esta especificação é empregada para que ao menos 8 itens sejam aplicados por fator. Além disso, o uso de regras para término nas primeiras rodadas de um teste CAT pode gerar precocemente o término da aplicação em decorrência de flutuações naturais nos parâmetros estimados.

Módulo para estudos de validade por relações com outras variáveis. É muito frequente, principalmente no uso de testes relativamente novos, a realização de estudos que visem a busca de evidências de validade pela relação com outras variáveis. Neste tipo de validade busca-se obter padrões de correlações entre os resultados do instrumento em validação e outros já validados (American Educational Research Association, American Psychological Association, National, Council on Measurement in Education, 1999).

Objetivando viabilizar a fácil realização de estudos de validade baseados em relações com outras variáveis foi desenvolvido um módulo para a coleta de dados de outros instrumentos (baseados em itens fixos e formato informatizado). Este módulo permite a fácil customização de telas de instruções, tanto de escalas diversas para pontuação dos itens quanto de testes com extensões variadas. As respostas dadas são

salvas em uma base própria para cada teste, mas são incluídas as informações necessárias para a realização do pareamento de tais bancos com os da BAP. Até o momento, tal módulo foi usado para a coleta de dados da versão brasileira do *Big Five Inventory* (Andrade, 2008).

Módulo para pré-testagem de itens.

No sistema desenvolvido para aplicação da BAP optou-se pela apresentação dos itens de pré-teste após a aplicação dos itens já calibrados, com o objetivo de priorizar os itens operacionais, minimizando o efeito da fadiga dos respondentes enquanto estes são apresentados. Também é importante destacar que o módulo para pré-testagem de itens foi construído de forma bastante flexível com vistas a viabilizar diferentes cenários de pré-teste. A inclusão de itens no sistema, por exemplo, é bastante simples, envolvendo a realização do *upload* de uma planilha contendo informações básicas sobre os itens para aplicação. A única restrição quanto ao número de itens de pré-teste é que este seja múltiplo de cinco, mas o número total não tem um limite definido. Caso seja incluído um número não múltiplo de cinco no módulo para pré-testagem, o sistema automaticamente ignora os itens que não compuserem a última tela por completo (se forem incluídos 52 itens, por exemplo, apenas 50 serão aplicados). Evidentemente a extensão do bloco de itens de pré-teste deve ser planejada considerando-se o perfil educacional do público alvo, bem como se o número máximo de itens a serem respondidos é relevante para a tal público. Para evitar o efeito da fadiga de uma forma desuniforme aos itens é feita a sua apresentação em ordem aleatória. Também é relevante informar que o sistema de pré-teste adotado foi construído para trabalhar com um bloco fixo de itens, a serem calibrados em conjunto.

Detalhes e algoritmos da BAP

Informações mais detalhadas sobre a Bateria Fatorial de Personalidade podem ser encontradas no site do Laboratório de Pesquisa em Avaliação Psicológica – LPAP (<http://lpap.paginas.ufsc.br>), no qual são disponibilizados exemplos do algoritmo para alguns dos módulos que compõem o sistema adaptativo. Também são disponibilizados no site do LPAP informações para uso da BAP em contextos de pesquisa e os produtos científicos decorrentes do projeto.

5. PLATAFORMAS PARA DESENVOLVIMENTO E APLICAÇÃO DE CAT

Existem algumas plataformas que possibilitam o desenvolvimento e aplicação de testes adaptativos computadorizados, de forma total ou parcial, estando disponíveis na modalidade *free* ou comercial para seu acesso e utilização. A seguir, apresenta-se o Quadro 1 com algumas sugestões disponíveis no mercado e seus *links* para obtenção de mais informações.

Plataforma	Objetivo	Desenvolvimento	Licença	Link
Concerto	É uma plataforma <i>open-source</i> baseada em R e apresentação em HTML. Permite aos usuários criar várias ferramentas de avaliação, desde levantamentos simples até testes adaptativos baseados na TRI. Por ser <i>open-source</i> , é constantemente atualizado e melhorado por uma comunidade de usuários. É possível instalar o Concerto em um servidor próprio ou hospedar o teste gratuitamente no servidor dos desenvolvedores.	Pesquisadores do Centro Psicométrico da Universidade de Cambridge	livre	https://code.google.com/p/concerto-platform/
fastTEST	Completo desenvolvimento e entrega de testes.	Assessment Systems Corporation	pago	http://www.assess.com/xcart/product.php?productid=17
Teste Adaptativo Computadorizado IRT	Tem como objetivo proporcionar um programa de acesso aberto com base na Internet para CAT.		livre	http://irt-cat.sourceforge.net/
OSCATS v 0.6	Sistema <i>open-source</i> - é uma biblioteca para os modelos psicométricos e algoritmos utilizados em CAT. Esta biblioteca pode ser utilizada na realização de simulações para o CAT ou como parte do engine ou back-end para um CAT operacional. OSCATS não pode ser utilizado para aplicar testes, é uma biblioteca a ser incorporada em um grande programa de administração de CAT e para estudos de		livre	https://code.google.com/p/oscats/

	simulação CAT.			
Pearson VUE	<i>Design</i> de conteúdo, acesso e uso mediado pelo fornecedor - oferece um conjunto completo de serviços de desenvolvimento de teste para gerenciamento de dados e entrega de testes.	Pearson VUE	pago	http://www.pearsonvue.com/
Prometric	<i>Design</i> de conteúdo, acesso e uso mediado pelo fornecedor.	Prometric	pago	https://www.prometric.com/en-us/Pages/home.aspx
McCann	<i>Design</i> de conteúdo, acesso e uso mediado pelo fornecedor - desenvolvimento e distribuição de testes para avaliação, certificação, <i>business intelligence</i> e soluções de desenvolvimento pessoal.	McCann Associates	pago	http://www.mccanntesting.com/
Assessment Center SM	É uma ferramenta on-line gratuita que fornece a funcionalidade para criar e entregar os testes adaptativos.	Northwestern University – Department of Medical Social Sciences: David Cella, Richard Gershon, Michael Bass e Nan Rothrock.	livre	http://www.assessmentcenter.net/
Smart Test Technology®	Plataforma para CAT de testes multidimensionais.	Adaptive Assessment Services	pago	http://www.aastest.com/
WebeXaminer FASTCAT	É um sistema para CAT com ferramentas de desenvolvimento de teste e análise pela TRI.	WebExaminer Limited	pago	http://www.webexaminer.com/our_platform.php

Quadro 1 - Plataformas que são utilizadas para desenvolvimento de CAT.

6. REFERÊNCIAS

- Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., ...
 Beddow, P. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. *The Journal of Technology, Learning and Assessment*, 10(5).
- American Educational Research Association, American Psychological Association, National, Council on Measurement in Education. (1999). *Standards for Educational and psychological testing*. Washington, DC: Author.
- Anastasi, A., & Urbina, S. (2000). *Testagem Psicológica* (7a ed.). Porto Alegre: Artes Médicas.
- Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria de Resposta ao Item - Conceitos e Aplicações*. São Paulo, SP: Associação Brasileira de Estatística.
- Andrade, J. M. D. (2008). *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil*. Tese de doutorado. UNB, Brasília.
- Assessment Systems Corporation. (2012). *User's manual for the XCALIBRE - Item Response Theory Calibration Software*. St. Paul, MN: Assessment Systems Corporation.
- Baker, F. (2001). *The Basics of Item Response Theory* (2 ed.). USA: ERIC Clearinghouse on Assessment and Evaluation.
- Barrada, J. R., Abad, F. J., & Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema*, 21(2), 313-320.
- Belov, D. I., & Armstrong, R. D. (2009). Direct and Inverse Problems of Item Pool Design for Computerized Adaptive Testing. *Educational and Psychological Measurement*, 69(4), 533-547. doi:10.1177/0013164409332224
- Bergstrom, B. A., & Gershon, R. C. (1995). *Item Banking* (J. C. I., Trans.). Lincoln, NE: Buros.
- Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research*, 16, 95-108. doi:10.1007/s11136-007-9168-6
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12(3), 261-280.
- Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor

- analisis: A study on the 5-factor model of personality. *Personality and Individual Differences*, 11, 515-524. doi:10.1016/0191-8869(90)90065-Y
- Bose, R. C. (1939). On the construction of balanced incomplete block designs. *Annals of Eugenics*, 9(4), 353-399. Retirado de <http://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1939.tb02219.x/abstract>
- Bose, R. C., & Nair, K. R. (1939). Partially balanced incomplete block designs. *Sankhyā: The Indian Journal of Statistics*, 337-372. Retirado de <http://library.isical.ac.in/jspui/bitstream/10263/270/1/39.05.pdf>
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage publications London.
- Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, 16, 133-141. doi:10.1007/s11136-007-9204-6
- Chang, H. H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Annals of Statistics*, 37(3), 1466-1488. doi:10.1214/08-aos614
- Chang, H.-H., & Ying, Z. (1999). a-Stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement*, 23(3), 211-222. doi:10.1177/01466219922031338
- Chang, H.-H., Qian, J., & Ying, Z. (2001). a-Stratified Multistage Computerized Adaptive Testing with b Blocking. *Applied Psychological Measurement*, 25(4), 333-341.
- Chang, S.-W., & Ansley, T. N. (2003). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, 40(1), 71-103.
- Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, 24(3), 241-255. Retirado de <http://www.scopus.com/inward/record.url?eid=2-s2.0-0034258277&partnerID=40&md5=4ffd7265a74f18bbb5a1bd9b823d764a>
- Cheng, P. E., & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, 24(3), 257-265. Retirado de <http://www.scopus.com/inward/record.url?eid=2-s2.0-0034258277&partnerID=40&md5=4ffd7265a74f18bbb5a1bd9b823d764a>

- 0034258078&partnerID=40&md5=e939cc14236e7daf52cc6c9bb5a8c494
- Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT Item Selection Criteria for Polytomous Items. *Applied Psychological Measurement*. Retirado de <http://apm.sagepub.com>
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.
- DeMars, C. (2010). *Item response theory*. New York: Oxford University Press, Inc.
- Digman, J. M. (2002). Historical Antecedents of the Five-Factor Model. Em P. T. Costa & T. A. Widiger (Orgs.), *Personality Disorders and the Five-Factor Model of Personality* (2 Org, pp. 17-22). Washington, DC: American Psychological Association.
- Digman, J. M. (1996). The Curious History of the Five Factor Model. Em J. S. Wiggins (Org.), *The Five Factor Model of Personality. Theoretical Perspectives*. New York e London: The Guilford Press.
- Embretson, S., & Reise, S. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. B., & Klapp, B. F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research*, 18(1), 23-36. doi:10.1002/mpr.274
- Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007a). A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment*, 5(8).
- Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007b). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, 5(8).
- Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of personality assessment*, 68(3), 532-560. Retirado de http://www.tandfonline.com/doi/abs/10.1207/s15327752jpa6803_5
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2005). *Análise Multivariada de Dados*. Porto Alegre: Bookman.
- Hart, D. L., Deutscher, D., Crane, P. K., & Wang, Y. C. (2009). Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. *Quality of Life Research*, 18(8),

- 1067-1083. doi:10.1007/s11136-009-9517-8
- Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. Em W. A. Sands, B. K. Waters, & J. R. McBride (Orgs.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington DC: American Psychological Association. Retirado em <http://psycnet.apa.org/index.cfm?fa=main.doiLanding&uid=1997-36257-014>
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. Em H. B. R. A. U. N. WAINER, H. (Org.), *Test validity* (pp. 129-145). Hillsdale, NJ:: Lawrence Erlbaum Associates.
- Huang, Y. M., Lin, Y. T., & Cheng, S. C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers and Education*, 52(1), 53-67. doi:10.1016/j.compedu.2008.06.007
- Huff, K. L., & Sireci, S. G. (2001). Validity Issues in Computer based Testing. *Educational Measurement: Issues and Practice*, 16-25.
- Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. R., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: guidelines, issues, and alternatives. *Journal Of Organizational Behavior*, 18, 667-683.
- Jeong, H. Y., & Hong, B. H. (2013). A service component based CAT system with SCORM for advanced learning effects. *Multimedia Tools and Applications*, 63(1), 217-226. doi:10.1007/s11042-012-1027-y
- Kaya, Z., & Tan, S. (2014). New trends of measurement and assessment in distance education. *Turkish Online Journal of Distance Education*, 15(1), 206-217.
- Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *The Journal of Technology, Learning and Assessment*, 4(2). Retirado de <http://napoleon.bc.edu/ojs/index.php/jtla/article/download/1649/1491>
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for Selecting Items for Computerized Adaptive Tests. *Applied Measurement in Education*, 2(4), 359-375. doi:10.1207/s15324818ame0204_6
- Kopec, J. A., Badii, M., McKenna, M., Lima, V. D., Sayre, E. C., & Dvorak, M. (2008). Computerized adaptive testing in back pain - Validation of the CAT-5D-QOL. *Spine*, 33(12), 1384-1390. doi:10.1097/BRS.0b013e3181732a3b
- Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A Comparison of

- Exposure Control Procedures in CATs Using the 3PL Model. *Educational and Psychological Measurement*, 73(5), 857-874. doi:10.1177/0013164413486802
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Luecht, R. M., De Champlain, A., & Nungester, R. J. (1998). Maintaining Content Validity in Computerized Adaptive Testing. *Advances in Health Sciences Education*, 3(1), 29-41. Retirado de <http://www.scopus.com/inward/record.url?eid=2-s2.0-1842621852&partnerID=40&md5=2c931e774db7ac3afc41855a81934669>
- Magis, D., & Barrada, J. R. (2014). Open-source CAT software: R packages and Concerto. Retirado em de
- Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement: Journal of the International Measurement Confederation*, 46(9), 3228-3237. doi:10.1016/j.measurement.2013.06.020
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174. doi:10.1007/BF02296272
- McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. Em D. J. WEISS (Org.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224-236). New York, NY: Academic Press.
- Moreira Junior, F. D. J. (2011). *Sistemática para a implantação de testes adaptativos informatizados baseados na teoria da resposta ao item*. Tese de Doutorado. Universidade Federal de Santa Catarina, Florianópolis - SC.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement*, 16(2), 159-176. Retirado de <http://apm.sagepub.com/content/16/2/159.short>
- Nunes, C. H. S. S., & Primi, R. (2009). Teoria de resposta ao item: conceitos e aplicações na psicologia e na educação. Em C. S. Hutz (Org.), *Avanços e polêmicas em avaliação psicológica* (pp. 25-69). São Paulo: Casa do Psicólogo.
- Nunes, C. H. S. S., Hutz, C. S., & Nunes, M. F. O. (2010). *Bateria Fatorial de Personalidade (BFP): manual técnico*.
- O'Neill, K. A., & McPeek, W. M. (1993). *Item and test characteristics that are associated with differential item functioning*.

- Ozyurt, H., Ozyurt, O., Baki, A., & Guven, B. (2012b). An Application of Individualized Assessment in Educational Hypermedia: Design of Computerized Adaptive Testing System and its Integration Into UZWEBMAT. *Procedia - Social and Behavioral Sciences*, 46(0), 3191-3196.
doi:<http://dx.doi.org/10.1016/j.sbspro.2012.06.035>
- Ozyurt, H., Ozyurt, O., Baki, A., & Guven, B. (2012b). Integrating computerized adaptive testing into UZWEBMAT: Implementation of individualized assessment module in an e-learning system. *Expert Systems with Applications*, 39(10), 9837-9847. doi:[10.1016/j.eswa.2012.02.168](https://doi.org/10.1016/j.eswa.2012.02.168)
- Parshall, C., Davey, T., & Nering, M. L. (1998). *Test development exposure control for adaptive testing*. Programas e resumos do Annual meeting of the National Council of Measurement in Education (NCME), San Diego, CA.
- Parshall, C., Harmes, J. C., Davey, T., & Pashley, P. (2010). Innovative Items for Computerized Testing. Em W. J. van der Linden & C. A. W. Glas (Orgs.), *Elements of Adaptive Testing* (pp. 215-230). Springer New York.
doi:[10.1007/978-0-387-85461-8_11](https://doi.org/10.1007/978-0-387-85461-8_11)
- Pasquali, L. (2007a). Aplicações práticas da TRI: Testes sob medida - CAT. Em *Teoria de Resposta ao Item: Teoria, Procedimentos e Aplicações* (pp. 187-202). Brasília: Laboratório de Pesquisa em Avaliação e Medida - LabPAM/UnB.
- Pasquali, L. (2007b). *Teoria de Resposta ao Item - Teoria, Procedimentos e Aplicações*. Brasília: LabPAM/UnB.
- Pastor, D. A., Dodd, B. G., & Chang, H.-H. (2002). A Comparison of Item Selection Techniques and Exposure Control Mechanisms in CATs Using the Generalized Partial Credit Model. *Applied Psychological Measurement*, 26(2), 147-163.
doi:[10.1177/01421602026002003](https://doi.org/10.1177/01421602026002003)
- Raju, N. (1988). The area between two item characteristic curves. *Psychometrika*, 53(4), 495-502. doi:[10.1007/BF02294403](https://doi.org/10.1007/BF02294403)
- Revuelta, J., & Ponsoda, V. (1998). A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, 35(4), 311-327.
- Rogers, J., & Swaminathan, H. (1989). A logistic regression procedure for detecting item bias. *Annual meeting of the American Educational Research Association*.
- Sailer, M. O. (2005). crossdes: A package for design and randomization in crossover studies. *Rnews*, 5/2, 24-27. Retirado de <http://cran.r-project.org>

- project.org/doc/Rnews/Rnews_2005-2.pdf
- Salcedo, P., Pinninghoff, M. A., & Contreras, R. (2005). Computerized adaptive tests and item response theory on a distance education platform. Em J. Mira & J. R. Alvarez (Orgs.), *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach, Pt 2, Proceedings* (Vol. 3562, pp. 613-621).
- Samejima, F. (1969a). *Estimation of latent ability using a response pattern of graded scores*. Psychometrika monograph supplement, n. 17, Richmond, VA: Psychometric Society.
- Samejima, F. (1997). Graded response model. Em W. J. van der Linden & R. K. Hambleton (Orgs.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.
- Samejima, F. (1969b). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Richmond, VA: Psychometric Society. Retirado em <http://www.psychometrika.org/journal/online/MN17.pdf>
- Silva, I. B., & Nakano, T. D. C. (2011). Modelo dos Cinco Grandes Fatores da Personalidade: análise de pesquisas. *Avaliação Psicológica*, 10(1), 51-62.
- Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychol Assess*, 17(1), 28-43. doi:10.1037/1040-3590.17.1.28
- Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, 7(2), 201-210. doi:10.1177/014662168300700208
- Stocking, M. L. (1994). Three Practical Issues for Modern Adaptive Testing Item Pools. *Educational Testing Service, Princeton, N.J.REPORT NOETS-RR-94-5*.
- Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing*. Programas e resumos do Annual meeting of the military testing association, Navy personnel research and development center, San Diego, CA.
- Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C. H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, 16, 109-119. doi:10.1007/s11136-007-9169-5
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. Em H. Wainer & H. Braun (Orgs.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Thompson, N. A., & Weiss, D. J. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, 16(1), 1-9.
- Thompson, S., Johnston, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (44). National Center on Educational Outcomes Synthesis Report.
- Urbina, S. (2007). *Fundamentos da Testagem Psicológica*. Porto Alegre: Artes Médicas.
- Valentini, F., & Laros, J. A. (2011). Teoria de Resposta ao Item na Avaliação Psicológica. *Ambiel, RAM, Rabelo, IS, Pacanaro, SV, Alves, G. AS & Leme, IFAS Avaliação Psicológica: guia de consulta para estudantes e profissionais de Psicologia*. São Paulo: Casa do Psicólogo.
- Van Der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, 23(1), 21-29.
doi:10.1177/01466219922031149
- Van Der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22(3), 259-270. Retirado de <http://www.scopus.com/inward/record.url?eid=2-s2.0-0032355085&partnerID=40&md5=9a1f28d6d64031701ea9777eb33d894a>
- Van Der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 195-210.
doi:10.1177/01466219922031329
- van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. Em W. J. van der Linden & C. A. W. Glas (Orgs.), *Computerized adaptive testing. Theory and practice* (pp. 1-25). Boston, MA: Kluwer.
- Van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. Em *Elements of adaptive testing* (pp. 3-30). Springer. Retirado em http://content.schweitzer-online.de/static/content/catalog/newbooks/978/038/785/9780387854595/9780387854595_Excerpt_001.pdf
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, 23(4), 327-345. doi:10.1177/01466219922031446

- Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*(2), 203-226. Retirado de <http://www.jstor.org/stable/1165378>
- Veldkamp, B. P., & Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio, 21*(78), 57-82. doi:10.1590/S0104-40362013005000001
- Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of Answer Feedback and Test Anxiety. *Journal of Educational Measurement, 35*(2), 155. Retirado de <http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=1094021&lang=pt-br&site=ehost-live>
- von, D., A. A., & Wilson, C. (2007). IRT True-Score Test Equating: A Guide Through Assumptions and Applications. *Educational and Psychological Measurement, 67*(6), 940-957. doi:10.1177/0013164407301543
- Wainer, H. (2000). CATs: Whither and whence. *Psicológica, 121*-133.
- Walker, J., Böhnke, J. R., Cerny, T., & Strasser, F. (2010). Development of symptom assessments utilising item response theory and computer-adaptive testing-A practical method based on a systematic review. *Critical Reviews in Oncology/Hematology, 73*(1), 47-67. doi:10.1016/j.critrevonc.2009.03.007
- Wang, T., Hanson, B. A., & Lau, C. M. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23*(3), 263-278. doi:10.1177/01466219922031383
- Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(2), 109-135. doi:10.1111/j.1745-3984.1998.tb00530.x
- Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. doi:10.1007/BF02294627
- Weiss, D. J., & Guyer, R. (2010). Manual for CATSim: Comprehensive simulation of computerized adaptive testing. *St. Paul MN: Assessment Systems Corporation.* Retirado de http://iacat.org/sites/default/files/biblio/CATSIM_Manual.pdf
- Wise, S. L., & Kingsbury, G. G. (2000). Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program. *Psicológica, 21*, 135-155.
- Wouters, H., Zwinderman, A. H., Van Gool, W. A., Schmand, B., & Lindeboom, R. (2009). Adaptive cognitive testing in dementia. *International Journal of Methods in Psychiatric Research, 18*(2), 118-127. doi:10.1002/mpr.283

- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA.
- Yi, Q., Wang, T., & Ban, J.-C. (2001). Effects of Scale Transformation and Test-Termination Rule on the Precision of Ability Estimation in Computerized Adaptive Testing. *Journal of Educational Measurement*, 38(3), 267-292.
doi:10.1111/j.1745-3984.2001.tb01127.x
- Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research*, 22(3), 491-499. doi:10.1007/s11136-012-0179-6
- Zitny, P., Halama, P., Jelinek, M., & Kveton, P. (2012). Validity of cognitive ability tests - comparison of computerized adaptive testing with paper and pencil and computer-based forms of administrations. *Studia Psychologica*, 54(3), 181-194.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2), 223-233. Retirado de http://irt.com.ne.kr/data/Zumbo_LAQ_reprint.pdf

Anexo 1

nome_out	modelo	theta step	items	items	min	max	exp	theta	theta	r_full	full	SEM	cat	SEM	full	SEM	cat	SEM	Alfa	n_items	n_items_sd	n_items_min			
		inic size	sel	first	items	items	ctrl	full	sd full	cat	sd cat	cat	sd cat	ASD	AAD	MSD	RMSD	mean	full_sd	mean	cat_sd				
n_mml_orig	mml	0	1	1	98	8	98	0	0.13	1.29	0.12	1.32	0.99	0.01	0.15	0.04	0.20	0.16	0.07	0.29	0.34	0.95	20.43	6.41	8
n_mml1	mml	-1,1	1	1	98	8	98	0	0.13	1.29	0.13	1.32	0.99	0.00	0.15	0.04	0.20	0.16	0.07	0.28	0.32	0.95	20.53	6.48	8
n_mml2	mml	0	1	5	10	8	98	0	0.13	1.30	0.13	1.32	0.99	0.00	0.15	0.04	0.20	0.16	0.07	0.28	0.32	0.95	20.54	6.39	8
n_mml3	mml	-1,1	1	5	10	8	98	0.45	0.13	1.29	0.12	1.33	0.98	0.01	0.19	0.06	0.25	0.16	0.07	0.33	0.36	0.94	22.01	8.05	8
n_mml4	mml	0	1	5	10	8	98	0.45	0.13	1.30	0.12	1.33	0.98	0.01	0.19	0.06	0.25	0.16	0.08	0.33	0.34	0.94	22.04	7.97	8
n_mml5	mml	0	1	5	10	8	98	0.35	0.13	1.29	0.12	1.33	0.98	0.01	0.20	0.07	0.27	0.16	0.07	0.35	0.34	0.93	21.70	7.31	8
n_mml6	mml	0	1	5	10	8	98	0.3	0.13	1.29	0.12	1.34	0.98	0.01	0.22	0.09	0.30	0.16	0.07	0.36	0.33	0.93	20.70	6.52	8
n_mml7	mml	-1,1	1	5	10	8	98	0.3	0.13	1.29	0.13	1.33	0.98	0.01	0.22	0.09	0.29	0.16	0.08	0.36	0.35	0.93	20.75	6.50	8
n_bey8	beyes_eap	-1,1	1	5	10	8	98	0.3	-0.01	0.78	-0.02	0.77	0.95	0.00	0.19	0.06	0.25	0.15	0.02	0.29	0.02	0.86	21.46	6.24	8
n_bey9	beyes_eap	0	1	5	10	8	98	0.3	-0.01	0.78	-0.02	0.77	0.95	0.00	0.19	0.06	0.25	0.15	0.02	0.29	0.04	0.86	21.35	6.29	8
n_bey10	beyes_eap	0	1	5	10	8	98	0.35	-0.02	0.78	-0.02	0.77	0.96	0.00	0.18	0.05	0.23	0.15	0.02	0.28	0.04	0.87	22.23	7.11	8
n_bey11	beyes_map	0	1	5	10	8	98	0.35	0.19	0.86	0.18	0.84	0.96	0.01	0.18	0.05	0.23	0.15	0.02	0.27	0.04	0.90	21.54	6.97	8
n_bey12	beyes_map	0	1	5	10	8	98	0.3	0.19	0.86	0.18	0.84	0.96	0.01	0.19	0.06	0.25	0.15	0.02	0.28	0.04	0.88	20.79	8	8
n_bey13	beyes_map	-1,1	1	5	10	8	98	0.3	0.19	0.86	0.18	0.83	0.96	0.01	0.20	0.06	0.25	0.15	0.02	0.28	0.04	0.88	20.90	6.27	8
n_wml14	wml	-1,1	1	5	10	8	98	0.3	-0.02	1.13	-0.02	1.16	0.96	0.00	0.00	0.12	0.12	0.16	0.04	0.40	0.19	0.88	21.57	6.78	8
n_wml15	wml	-1,1	1	5	10	8	98	0.35	-0.02	1.12	-0.02	1.16	0.96	0.00	0.24	0.24	0.34	0.16	0.04	0.39	0.19	0.89	22.92	8.12	8
n_wml16	wml	0	1	5	10	8	98	0.35	-0.02	1.13	-0.02	1.15	0.95	0.00	0.25	0.12	0.35	0.16	0.05	0.39	0.19	0.88	22.79	8.08	8
n_mml17	mml	0	1	2	10	8	25	0.35	0.13	1.29	0.13	1.34	0.98	0.00	0.21	0.08	0.28	0.16	0.07	0.35	0.33	0.93	20.09	5.44	8
n_mml18	mml	0	1	2	10	8	25	0.3	0.13	1.29	0.12	1.34	0.98	0.01	0.22	0.09	0.30	0.16	0.07	0.36	0.34	0.93	19.85	5.35	8
n_mml19	mml	-1,1	1	2	10	8	25	0.3	0.13	1.30	0.12	1.35	0.98	0.01	0.22	0.09	0.29	0.16	0.08	0.37	0.36	0.93	19.83	5.38	8
n_mml20	mml	0	1	2	10	8	20	0.35	0.13	1.29	0.13	1.33	0.98	0.00	0.22	0.08	0.29	0.16	0.07	0.35	0.32	0.93	17.70	3.64	8
n_mml21	mml	0	1	2	10	8	20	0.3	0.13	1.29	0.12	1.34	0.98	0.01	0.23	0.09	0.29	0.16	0.07	0.37	0.34	0.92	17.73	3.64	8
n_mml_epm1	mml	0	1	2	10	8	25	0.2	0.13	1.30	0.12	1.36	0.95	0.01	0.32	0.17	0.41	0.16	0.07	0.48	0.37	0.87	8.77	2.23	6
n_mml_epm2	mml	0	1	2	10	8	25	0.15	0.13	1.29	0.12	1.36	0.94	0.01	0.35	0.21	0.46	0.16	0.07	0.52	0.37	0.86	8.82	1.94	4
n_mml_epm3	mml	0	1	2	10	8	25	0.1	0.13	1.30	0.11	1.38	0.92	0.02	0.41	0.31	0.55	0.16	0.08	0.60	0.44	0.81	8.04	1.72	0
n_mml_epm4	mml	0	1	2	10	8	20	0.35	0.13	1.29	0.12	1.35	0.96	0.01	0.28	0.13	0.36	0.16	0.07	0.43	0.35	0.90	8.53	1.98	8

n_itens	p_casos15	p_casos20	p_casos25	p_casos30	n_term1	n_term2	n_term3	n_term4	n_term5	n_term6	n_term7	n_term8	n_term9	p_low_exp	p_high_exp	Precisão fixada		
max																		
44	23.09	48.03	79.80	93.30	2167	5154	1036	0	0	0	0	0	0	0.65	0.18	0.51	0.85	
46	22.74	47.15	79.18	92.86	2146	5093	1118	0	0	0	0	0	0	0.65	0.18	0.51	0.85	
48	22.31	47.33	79.57	92.86	2177	5142	1038	0	0	0	0	0	0	0.65	0.19	0.51	0.85	
47	24.24	44.19	65.68	83.25	74	6985	1278					20	28	0.29	0.00	0.51	0.85	
45	23.54	44.21	65.44	83.58	55	7034	1240											
42	23.63	42.43	65.55	88.14	4	6713	1244					396	396	0.15	0.00	0.52	0.85	
39	23.73	45.65	73.66	95.08	4	6324	1064					967	967	0.09	0.00	0.52	0.85	
38	23.53	45.28	74.17	94.87	3	6265	1130	0	0	0	0	959	959	0.09	0.00	0.52	0.85	
40	19.21	40.89	71.19	94.23	1	6278	1048					1030	1030	0.09	0.00	0.30	0.85	
39	20.47	41.82	71.40	94.11	5	6221	1055					1076	1076	0.09	0.00	0.30	0.85	
43	20.65	39.54	63.68	87.42	3	6353	1212					389	389	0.16	0.00	0.30	0.85	
42	23.14	43.65	67.84	89.84	12	6793	1261					291	291	0.18	0.00	0.33	0.85	
38	22.64	45.21	75.04	95.21	2	6348	1116					891	891	0.10	0.00	0.32	0.85	
39	22.19	44.48	74.15	95.47	3	6388	1096					870	870	0.11	0.00	0.32	0.85	
44	28.53	43.70	72.16	89.05	0	5602	598					2147	2147	0.02	0.00	0.45	0.85	
44	27.69	42.01	65.04	77.70	1	6249	681					1424	1424	0.08	0.00	0.45	0.85	
44	28.43	42.52	65.54	77.70	0	6303	660	6	6			1388	1388	0.08	0.00	0.45	0.85	
25	23.38	43.11	100.00	100.00	0	4156	915	0	0	3332	0	54	54	0.26	0.00	0.52	0.85	
25	23.19	45.00	100.00	100.00	0	4325	899	0	0		2751	0	382	0.15	0.00	0.52	0.85	
25	23.49	44.63	100.00	100.00	0	4308	925	0	0		2709	0	415	0.15	0.00	0.52	0.85	
20	23.30	100.00			0	2478	708					5168	0	3	0.37	0.00	0.52	0.85
20	22.35	100.00			0	2488	640					5157	0	72	0.28	0.00	0.52	0.85
25	96.42	99.51	100.00		7629	479	93	0	7			2	0	147	0.57	0.00		
25	98.13	99.86	100.00		6910	486	78	0	26			1	0	856	0.56	0.00		
19	99.80	100.00			4435	333	56	0	30			3503	3503	0.86	0.00			
20	97.30	100.00			7876	245	111	0	0			125	0	0.73	0.00			