Testes adaptativos para o Enade: uma aplicação metodológica

JEAN PITON-GONÇALVES[I]

http://dx.doi.org/10.22347/2175-2753v12i36.2735

Resumo

O Exame Nacional de Desempenho dos Estudantes avalia o rendimento dos concluintes de cursos da Educação Superior. É realizado por meio de lápis e papel e analisado sob a metodologia da Teoria Clássica de Testes. Buscando minimizar custos de distribuição, aplicação e correção, os Testes Computadorizados são inovadores e, quando utilizado sob a ótica da Teoria de Resposta ao Item, podem ser adaptativos. Um teste adaptativo seleciona os itens dinamicamente conforme o examinado responde ao teste, resultando em um teste individualizado, com a vantagem de ser mais curto que os testes convencionais. Este artigo tem como objetivo avaliar a viabilidade da aplicação metodológica de um teste adaptativo. Partindo de 10.861 respostas de estudantes de Licenciaturas em Matemática e 206.359 testes adaptativos simulados, os resultados mostram que uma seleção por Kullback-Leibler é a que apresenta um teste mais curto e tão preciso quanto os convencionais.

Palavras-chave: Teste Adaptativo. Teoria de Resposta ao Item. Enade.

Submetido em: 21/02/2020
Aprovado em: 27/07/2020

[I] Universidade Federal de São Carlos (UFSCar), São Carlos (SP), Brasil; http://orcid.org/0000-0002-7392-2001; e-mail: jpiton@ufscar.br.

Adaptive tests for Enade: a methodological application

Abstract

National Exam for the Assessment of Student Performance evaluates the performance of graduates of Higher Education courses and is carried out using pencil and paper and analyzed under the methodology of the Classical Test Theory. To minimize costs of distribution, application and correction, Computerized Tests are innovative and, when used under the Item Response Theory, can be adaptive. An adaptive test selects items dynamically as the respondent responds to the test, resulting in an individualized test, with the advantage of being shorter than conventional tests. This article aims to evaluate the viability of the methodological application of an adaptive test. Based on 10,861 responses from undergraduate students in mathematics and 206,359 simulated adaptive tests, the results show that a selection by Kullback-Leibler is the one that presents a shorter and more accurate test than the conventional ones.
Keywords: Adaptive Test. Item Response Theory. National Exam for the Assessment of Student Performance (Enade).

Pruebas adaptativas para el Enade: una aplicación metodológica

Resumen

El Examen Nacional de Desempeño Estudiantil evalúa el rendimiento de los egresados de los cursos de Educación Superior y el examen se realiza con lápiz y papel y se analiza bajo la metodología de la Teoría Clásica de Pruebas. Buscando minimizar los costos de distribución, aplicación y corrección, las Pruebas Computarizadas son innovadoras y, cuando se usan desde la perspectiva de la Teoría de Respuesta al Ítem, pueden ser adaptativas. Una prueba adaptativa selecciona ítems de forma dinámica a medida que el examinado responde a la prueba, lo que da como resultado una prueba individualizada, con la ventaja de ser más corta que las pruebas convencionales. Este artículo tiene como objetivo evaluar la viabilidad de la aplicación metodológica de una prueba adaptativa. Con base en 10.861 respuestas de estudiantes de pregrado en matemáticas y 206.359 pruebas adaptativas simuladas, los resultados muestran que una selección de Kullback-Leibler es la que presenta una prueba más corta y precisa que las convencionales.
Palabras clave: Prueba Adaptativa. Teoría de Respuesta al Ítem. Examen Nacional para la Evaluación del Desempeño del Alumno (Enade).

Introdução

No Brasil, o Exame Nacional de Desempenho de Estudantes (Enade) avalia o rendimento dos concluintes de cursos selecionados da Educação Superior. Atualmente, o exame é realizado em quatro horas por meio de lápis e papel, com correção por ficha ótica para os itens objetivos e correção manual para os dissertativos. Os itens apresentam diferentes níveis de dificuldade e os resultados são analisados com o uso da Teoria Clássica de Testes (TCT) (PASQUALI, 2009). Vista como avanço metodológico em relação à TCT, a Teoria de Resposta ao Item (TRI) (do inglês *Item Response Theory* – IRT) vem se consolidando para apoiar a elaboração de testes mais eficazes e justos com o desempenho do estudante, sendo utilizada há uma década no Exame Nacional do Ensino Médio (Enem), por exemplo. Desde 2005 (VENDRAMINI, 2005), a literatura vem apontando estudos que indicam a possibilidade de adotar-se a TRI enquanto instrumento metodológico para o Enade.

Buscando minimizar custos de distribuição, aplicação e correção de testes, os Testes Computadorizados (do inglês *Computer-Based Testing* - CBT) são inovadores e estão no escopo das Tecnologias Digitais da Informação e da Comunicação. Para aplicação, em larga escala, no cenário brasileiro, pode-se contar com os laboratórios de informática de faculdades e universidades, cujo funcionamento é para o ensino e a pesquisa e atende exigência do Ministério da Educação (MEC). Assim, não haveria necessidade de o exame ser realizado simultaneamente, como ocorre com o *Test of English as a Foreign Language* (TOEFL), que é previamente agendado e pode ser realizado em diferentes datas ao longo do ano.

Apesar desse cenário favorável à elaboração de um "Enade computadorizado", um problema característico dos testes em larga escala é o largo número de itens administrados e o tempo de prova, que provoca a fadiga do estudante e que pode afetar o seu desempenho.

Enquanto solução para testes longos e exaustivos, os Testes Adaptativos selecionam os itens[1] dinamicamente conforme o examinado responde ao teste, mantendo a precisão em relação aos tradicionais e administrando um número de itens menor que 50% em relação a um teste com número fixo de itens (WEISS; KINGSBURY, 1984).

1 De acordo com Osterlind (1998), os itens em um teste não podem ser chamados de questões, pois um item pode assumir outros formatos, que não são necessariamente interrogativos.

Buscando na literatura métodos, critérios e modelos que apoiem uma avaliação adaptativa no Enade, Santana et al. (2017) aplica um teste adaptativo, porém não se apropria dos conceitos de traço latente e da estimação pela TRI. Com isso, a revisão da literatura indica uma lacuna em estudos de testes adaptativos que sigam estritamente a TRI no contexto do Enade. Também não foram encontradas pesquisas que apliquem apenas a TRI em provas de Matemática do Enade.

Com o objetivo de responder a seguinte questão: "é viável a aplicação metodológica de um Teste Adaptativo Computadorizado baseado na TRI munido de itens da prova de Matemática da última edição do Enade?", esta pesquisa propõe e avalia uma aplicação metodológica de um teste adaptativo composto por itens objetivos de Matemática.

Exame Nacional de Desempenho de Estudantes

No Brasil, o Enade avalia o rendimento dos concluintes de cursos selecionados da Educação Superior, em todo o território nacional. Desde sua primeira aplicação, em 2004, integra o Sistema Nacional de Avaliação da Educação Superior (Sinaes) que avalia as instituições, os cursos e o desempenho dos estudantes de graduação.

O exame tem por objetivo geral aferir o estudante quanto (i) ao desempenho previsto na diretriz curricular do seu curso, (ii) às habilidades e às competências necessárias para o desempenho específico de sua profissão ligadas à realidade brasileira e mundial (INEP, 2017). Com isso, obtêm-se o indicador de qualidade denominado Conceito Enade.

O Enade da área de Matemática é realizado trienalmente e avalia o desempenho dos estudantes em cursos de Licenciatura e Bacharelado em Matemática por meio de 40 itens entre discursivos (resposta aberta) ou objetivos (múltipla escolha), sendo: (i) 10 itens de Formação Geral (25% da nota) distribuídos em oito objetivos e dois discursivos e (ii) 30 itens do Componente Específico (75% da nota) distribuídos em 27 objetivos e três discursivos. Ao final do exame, o estudante responde ao questionário de avaliação que versa sobre a dificuldade e a qualidade dos itens. Além disso, o exame é complementado por um questionário preenchido *on-line*, em outro momento, sobre informações socioeconômicas e acadêmicas do estudante. A última edição do Enade (2017) de Licenciatura em Matemática teve 13.340 inscritos em que 10.904 participaram da prova, ou seja, abrangendo 81,7% dos inscritos (INEP, 2017).

Metodologicamente, os itens do Enade apresentam diferentes níveis de dificuldade e toda da análise estatística segue a Teoria Clássica de Testes (TCT), em que o instrumento construído depende diretamente do objeto medido, ou seja, o resultado vai depender fortemente do instrumento utilizado (PASQUALI, 2009). Por exemplo, é como se a medida 1 cm medisse diferentemente em uma regra plástica e em outra de madeira. Dessa forma, o escore[2] depende dos itens que o compõem. Devido ao uso da TCT, o exame apresenta algumas limitações na análise estatística (ANDRADE; TAVARES; VALLE, 2000; SCHER et al., 2014), tais como: (i) os examinados que acertam a mesma quantidade de itens possuem o mesmo escore, independentemente da dificuldade dos itens, (ii) por originarem da proporção de acertos, os índices de dificuldade e de discriminação dependem da amostra e (iii) o erro padrão de medida é o mesmo para todos os escores.

Diferentemente do que ocorre metodologicamente com o Sistema de Avaliação da Educação Básica (Saeb) e o Enem, o Enade não adota a TRI. Com o objetivo de subsidiar o leitor, a próxima seção abordará os elementos essenciais da TRI.

Teoria de Resposta ao Item

A TRI propõe uma modelagem estatístico-matemática para as características latentes do examinado e modela a probabilidade de um indivíduo responder corretamente a um item em função do seu traço latente[3] que é psicometricamente mapeado para um estimador θ. A teoria é regida por dois postulados básicos (PASQUALI; PRIMI, 2003): (i) o desempenho do sujeito em um item pode ser predito a partir de um conjunto de fatores, aptidão ou traço latente θ; o traço latente é a causa e o desempenho o efeito e (ii) a relação entre o desempenho e o traço latente pode ser descrita pela Curva Característica do Item (CCI).

Um requisito da teoria é existência de um Banco de Itens[4] calibrado. Em casos dicotômicos[5], a calibração estima o(s) parâmetro(s) de cada item a partir dos acertos (1) e dos erros (0) do conjunto de examinados. A escolha dos modelos/métodos da TRI depende diretamente do objetivo do teste, do tamanho do

2 Pontuação ou nota do estudante quando interpretada por um sistema de medida.
3 No cenário educacional, o traço latente é concebido com habilidade, competência e/ou proficiência.
4 É um banco de dados que contém os itens e os parâmetros estimados associados a cada um deles.
5 Quando se tem uma única resposta objetiva correta e as demais são distratoras.

banco de itens e do desempenho computacional disponível (PITON-GONÇALVES, 2012).

Essencialmente, para se aplicar um teste baseado na TRI, necessita-se: (i) elaborar itens pautados em habilidades, em competências e no conteúdo específico, (ii) determinar métodos de estimação inferencial para os itens do banco e as habilidades dos examinados e (iii) elaborar uma escala que culminará na "nota" final. Os trabalhos de Lord (1980), Andrade, Tavares e Valle (2000) e Baker (2001) apontam para diversas vantagens da TRI frente a TCT, mas, apesar disso, os testes continuam sendo cansativos para o examinado, uma vez que administram um grande número de itens. No caso do Enem, por exemplo, são 45 itens somente na prova de Matemática e suas Tecnologias, em um total de 180 itens da prova completa.

Testes Adaptativos baseados na TRI

Com o objetivo de diminuir custos de aplicação e melhorar a qualidade e rapidez dos resultados, um Teste Computadorizado avalia eletronicamente os conhecimentos e as habilidades de um examinado, gerenciando todo o processo de avaliação, aplicando e corrigindo os itens, produzindo relatórios de desempenho e/ou resultados, garantindo a segurança e acurácia do teste (PITON-GONÇALVES, 2012).

A constante evolução da psicometria, estatística, educação, computação e da matemática aplicada tem permitido testes que vão além da múltipla escolha simples. Por exemplo, atualmente é possível se ter um teste que capture o tempo de resposta, expressões faciais, gestos e outros comportamentos não verbais do examinado. Além disso, o item pode ser apresentado em outros formatos que não são somente de múltipla escolha, tais como completamento e substituição (PITON-GONÇALVES, 2012).

Os contextos de aplicação de um teste computadorizado são inúmeros: Avaliação Educacional em Larga Escala, recrutamento em recursos humanos, publicidade e marketing, treinamento militar, estudos de depressão e ansiedade, dentre outros. Contudo, responder a longos testes e/ou questionários pode resultar em cansaço e estresse, tanto do ponto de vista físico como psíquico.

Dada a existência de uma metodologia consolidada e efetiva (como a TRI) para a elaboração de provas mais eficazes e justas com o desempenho do estudante, é possível construir um teste computadorizado baseado na TRI, diminuindo a fadiga do examinado em longos e exaustivos testes. Tal possibilidade é viabilizada por um Teste

Adaptativo, compreendido como um Método Alternativo Informatizado (PITON-GONÇALVES, 2004).

O primeiro teste adaptativo foi desenvolvido pelo psicólogo francês Alfred Binet (1857-1911) com o objetivo de diagnosticar o nível de inteligência de uma criança em comparação com sua idade cronológica, analisando a idade mental (WEISS, 1985). O princípio básico foi fornecer um teste composto por itens de diferentes níveis de dificuldade, em que cada item foi selecionado pelo aplicador conforme as respostas (corretas e incorretas). Já na década de 1950, também foram desenvolvidos o Teste Adaptativo de Dois Estágios e o Teste Adaptativo Estratificado (WEISS, 1985), que foram aplicados manualmente, sem o auxílio de computadores.

Os avanços tecnológicos da década de 1960 permitiram propostas de automação de um teste adaptativo, como foi trabalho de Reckase (1974), que resultou na denominação contemporânea de Teste Adaptativo Informatizado ou Computadorizado (do inglês, *Computer Adaptive Test* ou *Computerized Adaptive Testing*), reconhecido na literatura nacional como TAI ou internacionalmente como CAT. É um processo automatizado (ou supervisionado) de gerenciamento, configuração e resultados de um teste, auxiliando nas tarefas de correção, produção automatizada de estatísticas, geração de relatórios em diversas categorias (indivíduo, grupo, validação do instrumento etc.) e controle do tempo de teste.

Um CAT é, essencialmente, um teste computadorizado em que os itens são selecionados dinamicamente conforme o examinado responde ao teste, resultando em um teste personalizado. Ou seja, cada um responde a um conjunto diferente de itens em quantidade e grau de dificuldade, permitindo um teste acurado, válido e fidedigno, com a vantagem de ser mais curto que aqueles que administram os mesmos itens para todos os examinados.

Metodologicamente, o CAT pode seguir as Redes Neurais Artificiais (MASLOVSKYI; SACHENKO, 2015), a Lógica Fuzzy (BADARACCO; MARTÍNEZ, 2013), as Redes Bayesianas (PLAJNER; VOMLEL, 2015) e a TRI. A última é a mais utilizada por pesquisadores e instituições das áreas de psicologia e educação.

O CAT baseado na TRI[6] seleciona o próximo item seguindo um critério de seleção que está ligado aos parâmetros psicométricos dos itens do banco. Quando se utiliza

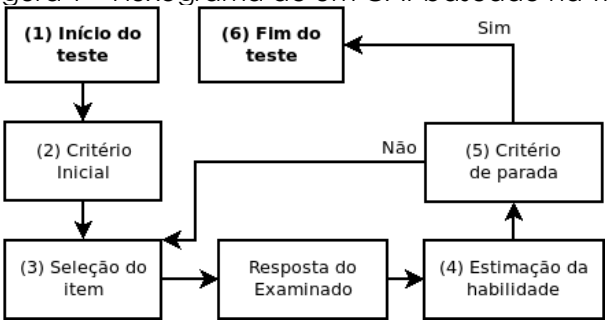
[6] Por questão de convenção textual, adotaremos para as próximas seções que a abreviatura CAT significa o mesmo que CAT baseado na TRI e traço latente como habilidade.

modelos logísticos, por exemplo, a fiabilidade do item se aproxima cada vez mais do nível de habilidade do examinado, uma vez que a dificuldade do item segue a mesma métrica. Esquematicamente, inicia-se o teste respondendo o primeiro item.

Em um exemplo prático ilustrativo, considere um banco com 30 itens distribuídos em 10 fáceis, 10 médios e 10 difíceis. Um examinado A (de baixa habilidade) inicia o teste e responde incorretamente a oito itens fáceis. Neste caso não há sentido continuar o teste, pois a probabilidade de acertar as médias e/ou as difíceis é baixa. Então, o CAT terminaria o teste. Em uma segunda situação, um examinado B (de alta habilidade) inicia o teste e responde corretamente a dois itens fáceis e, em seguida, três de nível médio (já que acertou dois fáceis, pode-se elevar o nível de dificuldade). O CAT selecionará os cinco próximos itens do nível difícil e, com isso, encerra-se o teste, uma vez que os itens que ainda restam no banco são mais fáceis para o nível de habilidade do examinado e não acrescentariam informação ao teste. É importante frisar que a adaptabilidade implica em administrar um teste com menos da metade dos itens em relação a um teste com itens iguais e quantidade fixa (WEISS; KINGSBURY, 1984). Por exemplo, se um teste foi projetado para ter 40 itens, se fosse adaptativo bastariam, no máximo, 20 itens.

O fluxograma da Figura 1 mostra os cinco componentes fundamentais de um CAT. Primeiro, inicia-se o teste (1) e aplica-se (2) um ou mais critérios iniciais; (3) seleciona-se o próximo item e recebe-se a resposta do indivíduo; (4) estima-se a habilidade provisória. (5) decide-se se o teste deve continuar. Se sim, então fim de teste e obtêm-se a habilidade final estimada do teste. Senão, retoma-se o passo (3).

Figura 1 – Fluxograma de um CAT baseado na TRI



Fonte: O autor (2013) adaptado de PITON-GONÇALVES (2012).

Metodologia

O modelo Rasch é de um parâmetro em relação ao item e foi estendido posteriormente para dois (BAKER, 2001). Birnbaum (1968) modificou o modelo logístico

de dois parâmetros e incluiu um terceiro parâmetro assim denominando-o de Modelo Logístico de Três Parâmetros (ML3P). Este modela a probabilidade de acerto mediante a habilidade ϴ do examinado em um dado item de acordo com a seguinte função (BIRNBAUM, 1968):

$$P(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}}$$

Em que *a* é a discriminação que mede a intensidade da variação da região de baixa probabilidade de acerto para a alta probabilidade do modelo de resposta, *b* é a dificuldade e c representa a probabilidade (de acerto casual) de obter uma resposta correta mediante uma baixa habilidade estimada do examinado. No contexto educacional *c* é interpretado como "chute". O ML3P é o mais adequado para as avaliações educacionais, uma vez que considera a probabilidade de acerto casual do examinado, que é uma condição comum nesse cenário de testes objetivos.

A literatura apresenta também outros modelos que são de utilização exclusiva em CAT como, por exemplo, a incorporação do tempo de resposta e o tempo de teste (TIANYOU; HANSON, 2005; LINDEN, 2006).

Quando há mais de uma habilidade ϴ no teste, este passa a ser representado por um vetor habilidade **ϴ** e passa a ser denominado de Teste Adaptativo Computadorizado Multidimensional (MCAT) (PITON-GONÇALVES; ALUISIO, 2015).

Estimação dos parâmetros de item e de habilidade

A calibração dos itens é um procedimento caro e computacionalmente lento (PITON-GONÇALVES; MONZON; ALUÍSIO, 2009) e, quando o modelo adotado é o ML3P, calibrar significa determinar numericamente os parâmetros *a, b* e *c* para cada item, partindo de um banco de respostas de uma amostra de examinados.

Diferentemente dos testes que utilizam somente a TRI, no CAT é necessário (re)estimar a habilidade após cada resposta do examinado, seguindo métodos tradicionais de estimação, tais como (LINDEN; GLASS, 2000): a Máxima Verossimilhança (do inglês, *Maximum Likelihood*), Máximo *a Posteriori* (do inglês *Bayesian Maximum a Posteriori*) e Esperança *a Posteriori* (do inglês *Bayesian Expected a Posteriori* – EAP). O último trata a média da distribuição *a posteriori* e o erro-padrão como o desvio-padrão *a posteriori* e vem sendo utilizado como método para estimar as proficiências dos participantes do Enem (BRASIL, 2012) e do Saeb (KLEIN, 2003).

Seleção de itens

O CAT é caracterizado por selecionar os itens de acordo com a habilidade e resposta do examinado, procedimento inviável em testes via lápis e papel. Os critérios de seleção podem ser divididos em dois grandes paradigmas: o clássico e o moderno. O primeiro contempla a Máxima Informação ou Informação de Fisher (do inglês *Fisher's Information* – MFI) (LORD, 1980), Máxima Informação Ponderada *a Posteriori* (do inglês *Maximum Posterior-Weighted Information*) e Máxima Informação Esperada (do inglês *Maximum Expected Information*) (LINDEN, 1998).

Contemplando os princípios e a axiomática da Divergência de Kullback-Leibler, os critérios modernos buscam a informação baseada em conceitos da Entropia, comparando distribuições de probabilidade. A Informação de Kullback-Leibler (KL) mede a divergência entre duas distribuições de probabilidade e a Informação de Kullback-Leibler com Distribuição *a Posteriori* (KLP) que, essencialmente, considera uma distribuição de densidade a posteriori (HUA-HUA; ZHILIANG, 1996).

Critérios iniciais e de parada

O primeiro item a ser administrado é um ponto fundamental para a capacidade adaptativa do CAT. No caso dicotômico, por exemplo, se todos os examinados iniciarem pelo mesmo item, o teste terá um comportamento seletivo semelhante a uma "árvore binária" de possibilidades de respostas, diminuindo a variabilidade do teste. Outra possibilidade, que apresenta bons resultados, é separar um conjunto de itens que tenham um mesmo nível de dificuldade e escolher o primeiro item aleatoriamente (PARSHALL; SPRAY; KALOHN; DAVEY, 2002). Uma terceira opção é selecionar *n* itens de nível fácil e estimar a habilidade somente após as *n* respostas aos itens, contribuindo para a estimação inicial mais eficiente. Feito isso, parte-se para os itens formais do teste.

Quanto aos critérios de parada, Chun, Weiss e Zhuoran (2018) consideram duas principais abordagens: um número fixo ou variável de itens administrados. Os referidos autores discutem que CAT que administram um número fixo de itens podem apresentar maiores erros na estimação da habilidade, terminando o teste de forma prematura. Por outro lado, testes que possuem um critério de parada variável, podem ser demasiadamente longos, o que contrapõe à motivação da utilização de um CAT.

Para avaliar o desempenho do critério de parada de um CAT de comprimento variável, Babcock e Weiss (2012) sintetizam as principais medidas entre a habilidade

estimada pelo CAT e a verdadeira[7], sendo elas: (i) o coeficiente de correlação linear de Pearson (r), (ii) o coeficiente de correlação por postos de Kendall (τ), que é mais robusto em *outliers*, (iii) a raiz do erro quadrático médio (RMSE) e o (iv) vício ou viés (bias) médio.

Apesar das vantagens do CAT delineadas ao longo desse artigo, Miguel (2017) aponta como desvantagem a capacidade do pesquisador quanto à produção de instrumentos informatizados, pois não é uma habilidade inerente a todo tipo de formação. Em muitos casos, recorre-se aos profissionais da área de informática que detenham sólidos conhecimentos estatístico-matemáticos, uma vez que os algoritmos do CAT devem ser projetados de forma otimizada para que processamento seja correto, seguro (criptografia avançada) e rápido, garantindo que todos os resultados sejam confiáveis e fidedignos, incluindo estudos de convergência numérica (PITON-GONÇALVES; ALUISIO, 2015).

Outra preocupação é quanto ao tempo computacional de processamento dos cálculos, uma vez que os critérios de estimação e seleção são realizados em tempo de execução, fato que não permite falhas ou baixa precisão numérica.

Lacuna e objetivos

Quanto à revisão da literatura no que se refere à aplicação da TRI no Enade, o trabalho de Vendramini (2005) aplica três modelos da TRI para analisar o desempenho de estudantes do exame de 2004 amostralmente. Primi, Hutz e Silva (2011) analisaram a prova de Psicologia do Enade 2006 respondida por 26.613 estudantes, empregando o modelo Rasch na calibração.

Corrêa et al. (2012) apresentam um instrumento de medida que viabiliza o uso da TRI na avaliação do Enade do ponto de vista estratégico. Lopes e Vendramini (2013) aplicaram a TRI na prova de Pedagogia do Enade 2005, partindo de 49.497 participantes, buscando o processo de equalização das provas. Campos (2013) avaliou a gestão do Banco Nacional de Itens para elaboração da prova do Enade, comparando a TCT com a TRI, apresentando vantagens e desvantagens.

Em outro trabalho, Lopes e Vendramini (2015) buscaram as propriedades psicométricas das provas de Pedagogia do Enade 2005 via TRI. Partindo de 230.486

[7] A habilidade estimada pelo CAT após cada item respondido é conhecida como habilidade provisória $\hat{\theta}$. A habilidade verdadeira θ é aquela estimada a partir do teste respondido na sua totalidade. Um CAT ideal seria aquele em que $\theta \approx \hat{\theta}$. A comparação entre θ e $\hat{\theta}$ só faz sentido em estudos de simulação, uma vez que em aplicação com usuários reais θ é desconhecido.

participantes, Scher et al. (2014) analisaram e aplicaram a TRI na prova de Administração e concluíram a viabilidade como instrumento de medida de avaliação. Coelho (2014) e Coelho, Ribeiro Junior e Bonat (2014) aplicaram a TRI com mais de uma habilidade em 94 estudantes de engenharias, 189 de Administração e 436 de Estatística referente ao Enade 2009. O trabalho de Camargo, Camargo, Andrade e Bornia (2016) abordou o desempenho dos estudantes de Ciências Contábeis no Enade edição 2012 por meio da TRI, contemplando 47.098 estudantes.

Com foco no CAT aplicado ao Enade, o trabalho de Santana et al. (2017) foi um estudo com 92 estudantes do curso de Direito a partir do Ambiente Virtual de Aprendizagem MOODLE, utilizando o *plugin Adaptive Quiz*[8], implementado de acordo com o algoritmo de Wright (1988) e o trabalho de Linacre (2000). Resultados mostram que os itens estavam de acordo com os parâmetros de ajuste do modelo Rasch quando os critérios de seleção e parada são realizados pelo algoritmo de Wright, que não se apropria dos conceitos de traço latente e da estimação pela TRI.

A revisão da literatura indicou uma lacuna em pesquisas pautadas nos microdados do Enade, principalmente em edições mais recentes, e que realizem estudos e simulações com CAT totalmente pautados na TRI. Também não foram encontradas pesquisas que apliquem a TRI em provas de Matemática do Enade.

Visando apoiar políticas públicas de avaliação nacional e diante das lacunas supramencionadas, o objetivo desta pesquisa é responder à questão: "é viável a aplicação metodológica de um Teste Adaptativo Computadorizado baseado na TRI munido de itens da prova de Matemática da última edição do Enade?". Este artigo aborda métodos e critérios adequados para a aplicação de um CAT, partindo de itens objetivos da prova de Licenciatura em Matemática de 2017, que é a última edição do exame que apresenta dados disponibilizados para pesquisas científicas.

Um questionamento que pode emergir quanto à viabilidade do CAT é com referência à obtenção de um escore menor para um estudante que respondeu um número maior de itens em comparação com demais estudantes. Veldkamp e Matteucci (2013) confirmam essa possibilidade em CATs e uma solução é fixar o número de itens administrados igualmente para todos.

Para aplicar um CAT no Enade busca-se, em primeiro lugar, estabelecer um critério de parada de forma que se administre um número fixo de itens, sempre menor

8 O *plugin* está descontinuado. Informações em moodle.org/plugins/mod_adaptivequiz.

que a quantidade de itens disponível. Em outras palavras, de modo que o comprimento do teste seja fixo, mas que os examinados respondam a diferentes itens. Em segundo lugar, adotar um critério de seleção que minimize o erro da habilidade (provisória) estimada e que, rapidamente, se aproxime da habilidade verdadeira, implicando em um número menor de itens administrados. Por último, fornecer critérios iniciais que valorizem a adaptabilidade do teste.

Materiais e métodos

Computacionalmente, a Linguagem R foi adotada e implantada no sistema operacional Ubuntu/Linux. Enquanto principais pacotes para as simulações, o *ltm* e *IRToys* apoiaram a estimação dos parâmetros de item e da habilidade; e o *catR* apoiou as simulações de CAT. Além disso, houve o uso de algumas rotinas do RPubs de Castro[9].

A fonte de dados foi coletada dos microdados[10] do Enade edição 2017. Partindo de 146MB (megabytes) de dados distribuídos categoricamente em 3.8102.729 palavras, foram desenvolvidos algoritmos específicos para a extração, filtragem e conversão das respostas de cada participante do curso de Licenciatura em Matemática[11]. Dos 10.904 presentes, foram considerados apenas os estudantes *presentes com resultado válido*[12], resultando em 10.861 examinados que compõem o Banco de Examinados.

Outro material utilizado é a Prova Enade 2017 Matemática Licenciatura[13], em que a parte objetiva é composta por 35 itens, distribuídos em oito itens de Formação Geral e 27 do Componente Específico. Os 35 itens foram analisados e (re)categorizados nessa pesquisa, uma vez que o objetivo foi obter um banco de itens que avaliasse estritamente o conteúdo matemático específico. Dentre os 27 do Componente Específico, 18 itens atendem o critério de conteúdo matemático e os demais são de conhecimentos gerais e pedagógicos. Uma vez que os 18 itens avaliam somente o conteúdo matemático, houve a pressuposição de unidimensionalidade.

9 rpubs.com/castro
10 inep.gov.br/microdados
11 Nos microdados é o código de área 702 na variável CO_GRUPO.
12 Nos microdados é o código 555 na variável TP_PRES. As provas em branco e provas desconsideradas pela Aplicadora e pelo Inep não foram consideradas para a composição do Banco de Examinados.
13 inep.gov.br/web/guest/provas-e-gabaritos3

Para testar a unidimensionalidade dos itens dicotômicos, adotou-se a técnica *Modified Parallel Analysis* (DRASGOW; LISSAK, 1983) que, essencialmente, analisa o segundo autovalor da matriz de correlações tetracóricas dos itens dicotômicos. A referida técnica está implementada no pacote *ltm*, por meio do comando *unidimTest*. Sob o ML3P, resultados mostraram que o segundo autovalor dos dados observados (0.7219) é maior do que o segundo autovalor dos dados sob o modelo assumido (0.4583), confirmando a unidimensionalidade.

Quanto aos métodos implementados, calibrou-se 18 itens (indexados originalmente na prova do 9º ao 26º) no ML3P pelos pacotes *IRToys* e *ltm*, que seguem a metodologia *Latent Variable Modeling* (LTM) (RIZOPOULOS, 2006), adotando o ML3P. Foram estimadas as habilidades verdadeiras de todos os estudantes por EAP. A habilidade inicial do teste foi determinada por proximidade do primeiro quartil dos parâmetros dificuldade do banco, ou seja, adotou-se a habilidade $\theta = 1.0$. O primeiro item foi selecionado a partir de um subconjunto de itens, de forma que o parâmetro dificuldade estivesse no intervalo [1.0, 1.8] e, em seguida, aplicou-se uma seleção aleatória seguindo a metodologia de Parshall, Spray, Kalohn e Davey (2002). A precisão numérica trabalhada foi de 14 dígitos, mas para a apresentação do artigo, truncou-se em cinco.

Simulações computacionais

A Tabela 1 traz os parâmetros estimados, em que a coluna ID Enade é a indexação na prova e ID-BI é no BI do CAT. Os itens Q16 e Q21 foram excluídos por apresentarem, respectivamente, uma discriminação acentuada e negativa[14]. Além disso, Q21 apresentou uma dificuldade muito baixa (-9.75611), uma vez que o adequado é estar entre -3 e 3. O item Q23 foi excluído por apresentar uma dificuldade muito acentuada.

Dessa forma, o BI final contém 15 itens, (re)indexados de um a 15. É importante destacar que os três itens excluídos também foram considerados como "fracos" pela análise clássica de itens do Enade 2017 (INEP, 2017, p. 224).

14 Quando a < 0, significa que a curva logística é "invertida". Ou seja, a probabilidade de acerto de um estudante de alta habilidade é baixa e de baixa habilidade é alta e, por isso, o item deve ser descartado.

Tabela 1 – Parâmetros do Banco de Itens calibrados pela TRI

ID Enade	ID BI	Discriminação	Dificuldade	Chute
Q09	1	3.10960	1.98459	0.21932
Q10	2	1.14930	-0.20234	0.00686
Q11	3	2.28264	1.46511	0.21024
Q12	4	1.81743	1.09326	0.13955
Q13	5	0.80551	0.67563	0.00155
Q14	6	2.64245	2.21457	0.14723
Q15	7	1.54527	1.78883	0.15764
Q16	*	6.76271	2.54769	0.10521
Q17	8	1.03360	2.32181	0.17393
Q18	9	2.40216	2.05515	0.11853
Q19	10	3.31689	2.30568	0.25792
Q20	11	3.04938	1.83017	0.18699
Q21	*	-0.14892	-9.75611	0.12496
Q22	12	2.17605	1.31338	0.17585
Q23	*	0.06117	24.36599	0.04809
Q24	13	3.54401	1.76628	0.18138
Q25	14	0.91780	2.02989	0.18815
Q26	15	0.74573	-0.59699	0.00720

Fonte: O autor (2019).

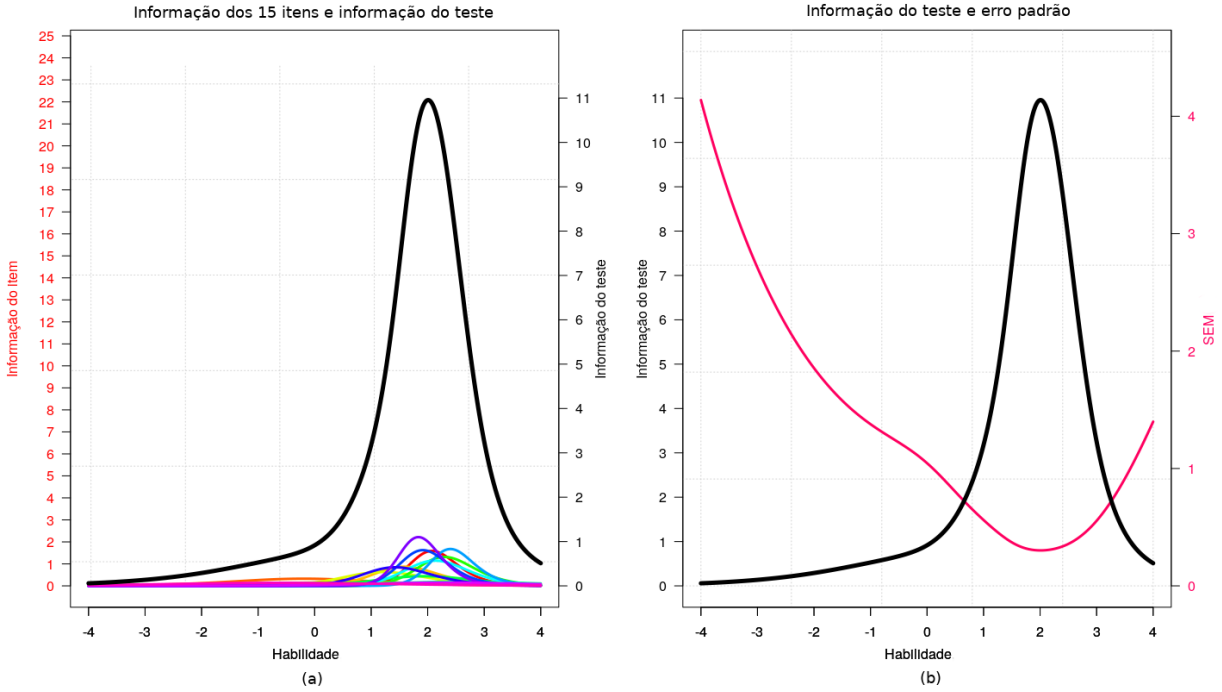
Em uma segunda etapa, quando se aplica todos os itens do teste nos estudantes, o gráfico da Figura 2 traz a Curva de Informação do Teste, obtida por meio da Função de Informação do Teste (BAKER, 2001), que representa todas as informações dos itens somados no *continuum* das habilidades. Seguindo o gráfico (a) da mesma figura, o teste é mais informativo em torno da habilidade igual a 2.0 e os itens, individualmente, em sua maioria, mais informativos entre 1.5 e 2.5. O gráfico (b) comprova que a Função Informação do Teste é inversamente relacionada ao erro-padrão da estimativa das habilidades, que é um resultado esperado.

Com o objetivo de responder à questão de pesquisa, foi estudada a correlação entre as habilidades provisórias e as habilidades verdadeiras, obtidas na administração de todos os itens do BI. Foram simulados 162.915 CAT de maneira que o 1º CAT termina com 1 item administrado, o 2º com 2 itens e assim por diante até o 15º com 15 itens. O critério de seleção escolhido foi o KLP, justificados adiante no experimento final. Os demais critérios são os mesmos do início desta seção.

Quando recorremos à literatura, Weiss e Kingsbury (1984) afirmam que um CAT pode administrar até 50% do comprimento de um teste de comprimento fixo,

mantendo os mesmos níveis de precisão da TRI. As simulações de Babcock e Weiss (2012) e Zhongmin, Chunyan, Yong e Hanwei (2018) consideram que um CAT que estima com precisão a habilidade do examinado, possuem um RMSE menor que 0.22 e um Bias menor que 0.01. No caso do coeficiente de Pearson e Kendall, r > 0.85 significa forte correlação. Partindo destas informações, foram calculadas as quatro medidas dos estimadores ao término de cada CAT, resultando nos dados da Tabela 2.

Figura 2 – Funções de informação dos 15 itens, do teste e o erro-padrão das habilidades



Fonte: O autor (2019).

Ao analisarmos os resultados, é notório que o critério de parada fixo mais adequado é de sete itens, uma vez que as quatro medidas são atendidas em suas precisões mínimas. Enquanto resultado esperado de um CAT que possua um banco de itens corretamente calibrado, sete itens corroboram os 50% apontados pela literatura (WEISS; KINGSBURY, 1984).

Caso o teste termine em seis itens, por exemplo, o RMSE e Kendall não são recomendados. Certamente que o teste poderia terminar em 10 ou 12 itens, porém do ponto de vista de fadiga do examinado é desfavorável. Se o teste for interrompido muito precocemente (quatro itens, por exemplo) as habilidades apresentam um alto RMSE.

Tabela 2 – Medidas entre a habilidade do CAT e a verdadeira em KLP

Itens administrados	Pearson	Kendall	RMSE	Bias médio
1	0.39161	0.22563	0.67060	0.00090
2	0.69604	0.49663	0.52329	0.00612
3	0.80539	0.63015	0.43213	0.00477
4	0.87731	0.71525	0.34982	0.00112
5	0.92076	0.78690	0.28440	0.00340
6	0.94686	0.83108	0.23446	0.00276
7	0.96391	0.86493	0.19416	0.00264
8	0.97588	0.89482	0.15916	0.00246
9	0.98421	0.92428	0.12904	0.00187
10	0.98985	0.94245	0.10359	0.00140
11	0.99350	0.95620	0.08298	0.00136
12	0.99635	0.97077	0.06227	0.00096
13	0.99811	0.98118	0.04480	0.00023
14	0.99952	0.99339	0.02260	0.00014
15	1.00000	1.00000	0.00000	0.00000

Fonte: O autor (2019).

A eficácia na adaptabilidade de um CAT é decorrente de um critério de seleção que escolha um item ótimo mediante a habilidade estimada. Neste caso, por se tratar de um banco de itens considerado "pequeno", selecionar os itens de forma mais eficiente é um dos objetivos fundamentais. Apesar dos bons resultados, sete itens apresentaram como um resultado adequado e indicado para um CAT, mas não se tem (ainda) a comprovação de que KLP é um critério de seleção, de fato, razoável. Nesse contexto, foram comparados quatro métodos de seleção, a saber: MFI, KL, KLP e *Random* (aleatório), em um total de 43.444 testes, de forma que 10.861 participantes realizassem, cada um, um teste que administra sete itens.

Tabela 3 – Medidas entre a habilidade do CAT e a verdadeira nos quatro critérios de seleção

	MFI	KL	KLP	Random
Pearson	0.96179	0.95412	0.96391	0.81664
Kendall	0.86482	0.84706	0.86493	0.58367
RMSE	0.19956	0.21827	0.19416	0.42069
Bias médio	0.00034	0.00082	0.00264	-0.00649

Fonte: O autor (2019).

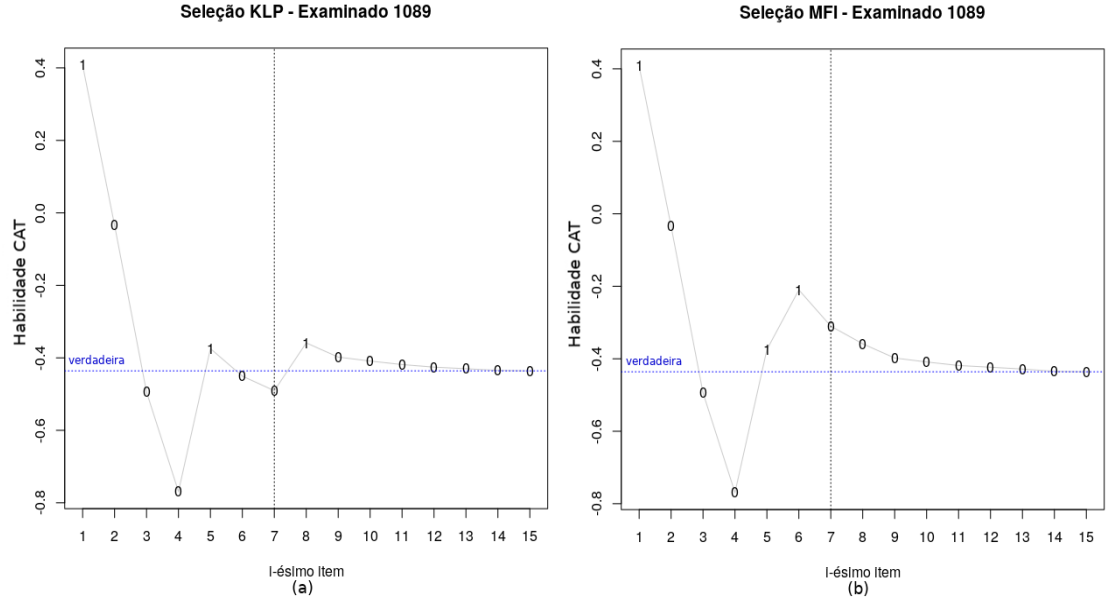
Os resultados da Tabela 3 mostram que KLP apresentou forte correlação (0.96391) em Pearson e Kendall (0.86493). Neste caso, o RMSE de 0.19416 é adequado, assim

como o Bias de 0.00264. Comparando MFI e KLP podemos dizer que são equivalentes (numericamente) em termos de correlação, e o mesmo ocorre com o RMSE. Em contrapartida, o Bias de MFI é menor do que KLP. Portanto, conclui-se que MFI e KLP são aptos a comporem um CAT para o Enade, mas, para esta pesquisa, elegeu-se KLP.

Um método inapropriado é o *Random*, pois não é sensível à habilidade do estudante, já que não considera a habilidade para selecionar o próximo item, aumentando a chance de administrar um item de baixa informação, por exemplo.

Os gráficos da Figura 3 trazem um estudo de caso do estudante rotulado com o número 1089, que exemplifica a capacidade de adaptabilidade do CAT e como se comportam os critérios KLP e MFI a cada escolha. O *setup* do teste foi: iniciar pelo item Q11 do BI (b=1.46511), habilidade inicial θ =1.0 e administrar todos os itens do banco.

Figura 3 – Estudo de caso: estudante 1.089 quando submetido aos critérios KLP e MFI. A sua habilidade verdadeira é -0.43590



Legenda: 1 acertou, 0 errou.
Fonte: O autor (2019).

Conclui-se que KLP selecionou melhor os itens neste caso, pois, se o teste terminasse administrando sete itens, a habilidade final seria -0.49069 e, por outro lado, seria -0.30998 na MFI. É certo que há uma pequena diferença entre a habilidade do CAT e a verdadeira, que é compensada pelo fato de se administrar menos da metade (sete itens) do comprimento total (15 itens).

Considerações finais

A revisão da literatura indica uma lacuna em pesquisas pautadas nos microdados do Enade Licenciatura em Matemática que realizem estudos e simulações com CAT que contemplem a TRI, tanto na estimação dos parâmetros de item, quanto das habilidades do estudante. Para preencher esta lacuna, esta pesquisa partiu: (i) da resposta de 10.861 estudantes de cursos de Licenciatura em Matemática pelo Brasil no que se refere ao Enade 2017, (ii) da análise e calibração, via TRI no ML3P, da prova que estes estudantes responderam, culminando em 15 itens válidos para aplicação em CAT.

Com o objetivo de verificar a viabilidade a metodológica de um CAT baseado na TRI, estudou-se, em primeiro lugar, um critério de parada adequado em situações práticas. Nesse sentido, esta pesquisa buscou um teste que administrasse um número fixo de itens para todos os estudantes. Para isso, foram simulados 206.359 CAT, comparando as habilidades finais no CAT *versus* a habilidade verdadeira e utilizando (i) as medidas Pearson, Kendall, RMSE e Bias médio, (ii) o critério de seleção por KLP e (iii) a estimação da habilidade por EAP. Nessas condições, resultados apontaram que um teste que administra sete itens é adequado, uma vez que as quatro medidas são atendidas em suas precisões mínimas.

Em segundo lugar, houve a necessidade de comprovar a eficiência da escolha do critério de seleção KLP. Para tal, foram comparados os métodos MFI, KL, KLP e *Random* (aleatório), em um total de 43.444 testes, de forma que 10.861 participantes realizassem, cada um, um teste que administra sete itens, corroborando com os apontamentos da literatura (WEISS; KINGSBURY, 1984). Adotando as mesmas medidas da simulação anterior, resultados mostraram que KLP apresentou forte correlação em Pearson e Kendall, e RMSE e Bias são atendidos.

Infere-se que o método aplicado nesta pesquisa pode ser estendido para outros cursos e edições do Enade. Enquanto trabalho futuro, almeja-se ampliar o banco de itens do Enade, contemplando outras edições e, consequentemente, aplicar um CAT *on-line* com estudantes reais.

Referências

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. *Teoria de resposta ao item*: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

BABCOCK, B.; WEISS, D. J. Termination criteria in computerized adaptive tests: variable-length cats are not biased. *In*: CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING, 2009, [Mineapolis]. *Proceedings* […]. [Mineapolis]: [s. n.], 2009. p. 1-21.

BADARACCO, M.; MARTÍNEZ, L. A fuzzy linguistic algorithm for adaptive test in intelligent tutoring system based on competences. *Expert Systems with Applications*, New York, v. 40, n. 8, p. 3073-3086, 2013.

BAKER, F. B. *The basics of item response*. 2. ed. College Park, MD: University of Maryland, 2001.

BAKER, F. B.; KIM, S. *Item response theory*: parameter estimation techniques. 2. ed. New York: CRC Press and Francis Group, 2001.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee´s ability. *In*: LORD, F. M.; NOVICK, M.R. (org.). *Statistical theories of mental test scores*. [Boston]: Addison-Wesley, 1968. p. 397-479.

BRASIL. Ministério da Educação. *Entenda a sua nota no Enem*: guia do participante. Brasília, DF: Inep, 2012. Disponível em: http://www.portal.singular.com.br/arquivos/thumbs/documentos_cursinho/Entenda%20a%20sua%20nota%20do%20ENEM.pdf. Acesso em: 15 ago. 2015.

CAMARGO, R. V. W.; CAMARGO, R. de C. C. P.; ANDRADE, A. F. de; BORNIA, A. C. Desempenho dos alunos de ciências contábeis na prova ENADE/2012: uma aplicação da Teoria da Resposta ao Item. *Revista de Educação e Pesquisa em Contabilidade (REPeC)*, Brasília, DF, v. 10, n. 3, p. 332-355, ago. 2016. DOI: https://doi.org/10.17524/repec.v10i3.1401. Disponível em: http://www.repec.org.br/repec/article/view/1401/1183. Acesso em: 13 jan. 2019.

CAMPOS, F. C. dos S. *Elaboração da prova do ENADE no modelo do banco nacional de itens*. Orientador: Marcel de Toledo Vieira. 2013. 89 f. Dissertação (Mestrado Profissional em Gestão e Avaliação da Educação Pública) - Centro de Políticas Públicas e Avaliação da Educação, Universidade Federal de Juiz de Fora, Juiz de Fora, MG, 2013. Disponível em: http://mestrado.caedufjf.net/wp-content/uploads/2014/02/dissertacao-2011-fernanda-cristina-dos-santos-campos.pdf. Acesso em: 25 ago. 2018.

CHUN, W.; WEISS, D. J.; ZHUORAN, S. Variable-length stopping rules for multidimensional computerized adaptive testing. *Psychometrika*, [New York], v. 84, n. 3, p. 749-771, 2018. Disponível em: https://www.researchgate.net/publication/329380148_Variable-Length_Stopping_Rules_for_Multidimensional_Computerized_Adaptive_Testing. Acesso em: 4 maio 2019.

COELHO, E. C. *Teoria da resposta ao item: desafios e perspectivas em exames multidisciplinares*. Orientador: Paulo Justiniano Ribeiro Júnior. 2014. 189 f. Tese (Doutorado em Ciências) - Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, 2014. Disponível em: https://acervodigital.ufpr.br/bitstream/handle/1884/36872/R%20-%20T%20-%20EDY%20CELIA%20COELHO.pdf?sequence=3&isAllowed=y. Acesso em: 6 jul. 2016.

COELHO, E. C.; RIBEIRO JUNIOR, P. J.; BONAT, W. H. Exame nacional de desenvolvimento de estudantes de estatística - desafios e perspectivas pela tri. *Revista da Estatística UFOP*, Ouro Preto, MG, v. 3, n. 2, p. 323-337, 2014.

CORRÊA, A. C. et al. Modelagem de um instrumento de medida de avaliação do Enade fundamentado na teoria de resposta ao item (tri): desenho para o mees. *In*: COLÓQUIO INTERNACIONAL SOBRE GESTÃO UNIVERSITÁRIA NAS AMÉRICAS, 12., 2012, Veracruz, México. *Anais* [...]. Florianópolis: UFSC, 2012.

DRASGOW, F.; LISSAK, R. I. Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, Washington, DC, v. 68, p. 363-373, 1983.

HUA-HUA, C.; ZHILIANG, Y. A global information approach to computerized adaptive testing. *Applied Psychological Measurement, [S. l.],* v. 20, n. 3, p. 213-229, 1996.

INEP. *Relatório síntese de área*: matemática (bacharelado e licenciatura). Brasília, DF: MEC, 2017. Disponível em: http://download.inep.gov.br/educacao_superior/enade/relatorio_sintese/2017/Matematica.pdf. Acesso em: 22 jun. 2019.

KLEIN, R. Utilização da teoria de resposta ao item no sistema nacional de avaliação da educação básica (Saeb). *Ensaio*: Avaliação e Políticas Públicas em Educação, v. 11, n. 40, p. 283-296, 2003.

LINACRE, J. M. Computer-adaptive testing: a methodology whose time has come. In: CHAE, U. K.; JEON, E.; LINACRE, J. M. (ed.). Development of computerised middle school achievement test. Chicago: Mesa, 2000.

LINDEN, W. J. V. D. Bayesian item selection criteria for adaptive testing. *Psychometrika*, [New York], v. 63, n. 2, p. 201-216, jun. 1998.

LINDEN, W. J. V. D. U*sing response times for item selection in computerized adaptive testing*. Enschede: University of Twente, 2006.

LINDEN, W. J. V. D.; GLAS, C. A. W. (ed.). *Computerized adaptive testing*: theory and practice. [Berlin]: Kluwer Academic Publishers, 2000.

LOPES, F. L.; VENDRAMINI, C. M. M. Equalização de provas acadêmicas via teoria de resposta ao item. *Psico-USF*, Bragança Paulista, SP, v. 18, n. 1, p. 141–150, jan./abr. 2013. Disponível em: https://www.scielo.br/pdf/pusf/v18n1/v18n1a15.pdf. Acesso em: 9 nov. 2017.

LOPES, F. L.; VENDRAMINI, C. M. M. Propriedades psicométricas das provas de pedagogia do enade via tri. *Avaliação*, Campinas, SP; Sorocaba, SP, v. 20, n. 1, p. 27–47, mar. 2015. DOI: https://doi.org/10.590/S1414-40772015000100004. Disponível em: https://www.scielo.br/pdf/aval/v20n1/1414-4077-aval-20-01-00027.pdf. Acesso em: 18 nov. 2017.

LORD, F. M. *Application of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates, 1980.

MASLOVSKYI, S.; SACHENKO, A. Adaptive test system of student knowledge based on neural networks. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT DATA ACQUISITION AND ADVANCED COMPUTING SYSTEMS: TECHNOLOGY AND APPLICATIONS, 8., 2015, Warsaw, Poland. *Annals* […]. Warsaw, Poland: IEEE, 2015. p. 940–944.

MIGUEL, F. K. A utilização da informática nas pesquisas em avaliação psicológica. *Avaliação Psicológica*, São Paulo, v. 16, n. 4, p. 1-3, 2017.

OSTERLIND, S. *Constructiong test items*: multiple-choice, constructed-response, performance, and other formats. [New York]: Kluwer Academic Publishers, 1998.

PARSHALL, C. G.; SPRAY, J. A.; KALOHN, J. C.; DAVEY, T. *Practical considerations in computer-based testing*. New York: Springer-Verlag, 2002.

PASQUALI, L. Psicometria. R*evista da Escola de Enfermagem da USP*, São Paulo, v. 43, n. especial, p. 992-999, dez. 2009. Disponível em: https://www.scielo.br/pdf/reeusp/v43nspe/a02v43ns.pdf. Acesso em: 12 nov. 2011.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria de resposta ao item - tri. *Avaliação Psicológica*, São Paulo, v. 2, n. 2, p. 99–110, 2003. Disponível em: http://pepsic.bvsalud.org/pdf/avp/v2n2/v2n2a02.pdf. Acesso em: 13 dez. 2007.

PITON-GONÇALVES, J. *A integração de testes adaptativos informatizados e ambientes computacionais de tarefas para o aprendizado do inglês instrumental*. Orientadora: Sandra Maria Aluísio. 2004. 141 f. Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) - Programa de Pós-Graduação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, 2004. Disponível em: https://teses.usp.br/teses/disponiveis/55/55134/tde-03052004-160334/publico/dissertacao_Jean_Piton.pdf. Acesso em: 12 nov. 2004.

PITON-GONÇALVES, J. *Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais*. Orientadora: Sandra Maria Aluísio. 2012. 176 f. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional, Universidade de São Paulo, São Carlos, SP, 2012. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-13032013-105955/publico/tese_revisada_final_jean_piton_jan2013.pdf. Acesso em: 2 jun. 2013.

PITON-GONÇALVES, J.; ALUÍSIO, S. M. An architecture for multidimensional computer adaptive test with educational purposes. *In*: SYMPOSIUM ON MULTIMEDIA AND THE WEB, 18., 2012, New York. *Proceedings* […]. New York: ACM, 2012. p. 17-24.

PITON-GONÇALVES, J.; ALUÍSIO, S. M. Teste adaptativo computadorizado multidimensional com propósitos educacionais: princípios e métodos. *Ensaio*: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 23, n. 87, p. 389-414, 2015. DOI: https://doi.org/10.1590/S0104-40362015000100016. Disponível em: https://www.scielo.br/pdf/ensaio/v23n87/0104-4036-ensaio-23-87-389.pdf. Acesso em: 3 jan. 2016.

PITON-GONÇALVES, J.; MONZÓN, A. J.; ALUÍSIO, S. M. Métodos de avaliação informatizada que tratam o conhecimento parcial do aluno e geram provas individualizadas. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 20., 2009, Porto Alegre. *Anais* [...]. Porto Alegre: Sociedade Brasileira de Computação, 2009. p. 1-10.

PLAJNER, M.; VOMLEL, J. Bayesian network models for adaptive testing. *In*: ANNUAL BAYESIAN MODELLING APPLICATIONS WORKSHOP, 12., 2015, Amsterdam. *Annals* […]. Amsterdam: [s. n.], 2015.

PRIMI, R.; HUTZ, C. S.; SILVA, M. C. R. da. A prova do Enade de psicologia 2006: concepção, construção e análise psicométrica da prova. *Avaliação Psicológica*, São Paulo, v. 10, n. 3, p. 271-294, 2011. Disponível em: http://pepsic.bvsalud.org/pdf/avp/v10n3/v10n3a04.pdf. Acesso em: 17 set. 2017.

RECKASE, M. D. An interactive computer program for tailored testing based on the oneparameter logistic model. *Behaviour Research Methods and Instrumentation*, [Berlin], v. 6, n. 2, p. 208-212, 1974.

RIZOPOULOS, D. ltm: an r package for latent variable modelingand item response theory analyses. *Journal of Statistical Software, [S. l.],* v. 17, n. 5, p. 1-25, 2006.

SANTANA, L. F. et al. Avaliação informatiza adaptativa do Enade pelo moodle: evidências de validade. *Informática na educação*: teoria e prática, Porto Alegre, v. 20, n. 2, p. 222–238, maio/ago. 2017. Disponível em: https://seer.ufrgs.br/InfEducTeoriaPratica/article/view/69900/43630. Acesso em: 18 mar. 2019.

SCHER, V. T. *et al.* Uma aplicação da tri na avaliação do Enade do curso de administração. In: COLÓQUIO INTERNACIONAL DE GESTÃO UNIVERSITÁRIA, 14., 2014, Florianópolis. Anais [...]. Florianópolis: [UFSC], 2014. Disponível em: https://core.ac.uk/download/pdf/30408247.pdf. Acesso em: 23 mar. 2019.

VELDKAMP, B. P.; MATTEUCCI, M. Bayesian computerized adaptive testing. *Ensaio*: Avaliação e Políticas Públicas em Educação, Rio de Janeiro. v. 21, n. 78, p. 57-82, jan./mar. 2013. Disponível em: https://www.scielo.br/pdf/ensaio/v21n78/aop_0313.pdf. Acesso em: 12 abr. 2014.

VENDRAMINI, C. M. M. Avaliação multidimensional de desempenho do estudante. *Avaliação*, Campinas, SP; Sorocaba, SP, v. 10, n. 3, p. 27-40, set. 2005. Disponível em: http://periodicos.uniso.br/ojs/index.php/avaliacao/article/view/1314/1304. Acesso em: 10 out. 2013.

TIANYOU, W.; HANSON, B. A. Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, [S. l.],* v. 29, n. 5, p. 323–339, 2005. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.861.8244&rep=rep1&type=pdf. Acesso em: 24 jul. 2007.

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

WEISS, D. J.; KINGSBURY, G. G. Application of computerized adaptive testing to educational problems. *Journal of Education Measurement, [S. l.],* v. 21, p. 361-375, 1984.

WRIGHT, B. *Practical adaptive testing CAT algorithm. Rasch Measurement Transactions, [S. l.],* 1988.

ZHONGMIN, C.; CHUNYAN, L.; YONG, H.; HANWEI, C. *Comparison of algorithms that allow item review in computerized adaptive testing*. [USA]: ACT, 2018.