https://doi.org/10.18222/eae.v0ix.7216

TESTE ADAPTATIVO INFORMATIZADO DA PROVINHA BRASIL – LEITURA: RESULTADOS E PERSPECTIVAS

RODRIGO TRAVITZKI[I]
OCIMAR MUNHOZ ALAVARSE [II]
DOUGLAS DE RIZZO MENEGHETTI[III]
ÉRICA MARIA DE TOLEDO CATALANI[IV]

[I] Universidade São Francisco (USF), Campinas-SP, Brasil; *r.travitzki@gmail.com*
[II] Universidade de São Paulo (USP), São Paulo-SP, Brasil; *ocimar@usp.br*
[III] Centro Universitário da Fundação Educacional Inaciana "Padre Sabóia de Medeiros" (Centro Universitário FEI), São Bernardo do Campo-SP; *douglasrizzo@fei.edu.br*
[IV] Universidade de São Paulo (USP), São Paulo-SP, Brasil; *ericamtc@usp.br*

RESUMO

Este artigo descreve um Teste Adaptativo Informatizado (TAI) da Provinha Brasil – Leitura, com base na Teoria da Resposta ao Item. Detalham-se o funcionamento e o desenvolvimento do algoritmo. O TAI foi aplicado com o uso de tablets a 1.983 alunos dos 1o e 2o anos do ensino fundamental, em 15 escolas da Rede Municipal de Ensino de São Paulo. Os resultados confirmam a qualidade dos itens da Provinha Brasil, do trabalho realizado nas escolas e, sobretudo, do TAI. Em relação à gestão do tempo de prova, conclui-se que há uma associação positiva entre proficiência e tempo, mas só até certo ponto; os alunos tendem a demorar mais nos itens mais difíceis; essa tendência é mais intensa nos alunos mais proficientes, confirmando a hipótese de que eles tendem a gerir melhor o tempo de prova.

PALAVRAS-CHAVE TESTE ADAPTATIVO INFORMATIZADO • AVALIAÇÃO DE COMPETÊNCIA • PROVINHA BRASIL • TEORIA DA RESPOSTA AO ITEM.

TEST ADAPTATIVO INFORMATIZADO DEL PROVINHA BRASIL – LEITURA: RESULTADOS Y PERSPECTIVAS

RESUMEN

Este artículo describe un Test Adaptativo Informatizado (TAI) del la Provinha Brasil– Leitura, con base en la Teoría de la Respuesta al Ítem. Se detalla el funcionamiento y el desarrollo del algoritmo. El TAI fue aplicado con el uso de tablets a 1.983 alumnos del primero y segundo año de la enseñanza primaria, en 15 Escuelas de la Red Municipal de Enseñanza en San Pablo, Brasil. Los resultados confirman la calidad de los ítems del la Provinha Brasil, del trabajo realizado en las escuelas y, especialmente, del TAI. En relación con la administración del tiempo del examen, se concluye que hay una asociación positiva entre proficiencia y tiempo, pero solo hasta cierto punto; los alumnos tienden a demorarse más en los ítems más difíciles; esta tendencia es más intensa en los alumnos más proficientes, confirmando la hipótesis de que ellos tienden a administrar mejor el tiempo del examen.

PALAVRAS-CHAVE TEST ADAPTATIVO INFORMATIZADO • EVALUACIÓN DE COMPETENCIA • PROVINHA BRASIL • TEORÍA DE LA RESPUESTA AL ÍTEM.

A COMPUTERIZED ADAPTIVE TEST OF PROVINHA BRASIL - LEITURA: RESULTS AND PERSPECTIVES

ABSTRACT

This article describes a Computerized Adaptive Test (CAT) of Provinha Brasil – Leitura,[1] based on Item Response Theory. We detail the operation and development of the algorithm. The CAT was administered by means of tablet computers to 1,983 students in the 1st and 2nd grades of primary education, in 15 schools of the Municipal Education System of São Paulo. Results confirm the quality of Provinha Brasil's items, of the work done in schools and, mainly, of the CAT. As to the management of test time, we found a positive association between proficiency and time, but only to a certain extent; students tend to take longer on the more difficult items; this tendency is stronger in more proficient students, thus confirming the hypothesis that they tend to manage test time better.

KEYWORDS COMPUTERIZED ADAPTIVE TESTING • COMPETENCY ASSESSMENT • PROVINHA BRASIL • ITEM RESPONSE THEORY.

1 Provinha Brasil – Leitura is a Brazilian standardized reading assessment test for primary education students.

INTRODUÇÃO[2]

Os testes padronizados, sobretudo por seu emprego nas avaliações externas, têm se tornado cada vez mais presentes em escolas públicas e, progressivamente, pode se aventar, inserem-se na cultura escolar brasileira como expressão de políticas educacionais que, mediante usos variados dos resultados, utilizam-se desses testes. Contudo, para que esse tipo de política possa contribuir para a melhoria da qualidade da educação, há diversos desafios a serem enfrentados. Alguns dizem respeito à interpretação dos resultados e sua utilização em estratégias de gestão de redes de ensino e de escolas e mais ainda no cotidiano da sala de aula. Outros se referem a dificuldades logísticas relacionadas à segurança e ao gerenciamento de grandes quantidades de papel. Um terceiro tipo de desafio diz respeito à qualidade técnica dos testes enquanto instrumentos de medida. O Teste Adaptativo Informatizado (TAI) é uma tecnologia de avaliação da aprendizagem capaz de contribuir para a superação desses desafios, principalmente os dos dois últimos tipos.

A ideia de um teste que se adapte à proficiência de cada indivíduo remonta à década de 1970, embora ainda não tenha se popularizado plenamente, sendo muito incipiente no Brasil. Nesse tipo de teste, cada indivíduo responde a um conjunto de itens, selecionado em função de seu domínio – sua proficiência – daquilo que está sendo avaliado no decorrer da aplicação do teste de tal modo que o teste só se define completamente ao final de sua realização, podendo-se, portanto, encontrar testes que diferem de respondente para respondente em decorrência de suas diferenças de proficiência. No início do teste, a cada respondente é apresentado um item, o qual pode, por exemplo, ser selecionado aleatoriamente ou escolhido de acordo com uma condição pré-determinada. Posteriormente, durante a realização do teste, ele se adapta a cada indivíduo, dependendo de seu desempenho (representado por seus acertos e erros), principalmente através da seleção de itens mais difíceis ou mais fáceis, e tem o objetivo tanto de otimizar a precisão da estimação de proficiência de cada indivíduo quanto de reduzir o tempo de aplicação, em contraste com os testes em papel ou apresentados em computador, considerados lineares, que desde o início estão prontos com uma determinada quantidade de itens, mesmo que apresentados sequencialmente. Essa é a essência de um teste adaptativo, que pode apresentar diversas variações e níveis de complexidade. Quando esse tipo de teste é aplicado em formato digital, mediante processos informatizados, denomina-se Teste Adaptativo Informatizado.

2 Parte deste trabalho foi apresentada no congresso "Teaching, Learning and E-learning (IAC-TLEl)" de 2018, em Viena. Agradecemos aos diversos profissionais da educação que colaboraram para o projeto piloto, ao Núcleo Técnico de Avaliação da Secretaria Municipal de Educação de São Paulo, à Faculdade de Educação da Universidade de São Paulo (USP), ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) e à Organização das Nações Unidas para a Educação, a Ciência e a Cultura (Unesco).

Comparado a um teste linear convencional, o TAI apresenta algumas vantagens, como permitir: 1) igualar a precisão de medida para diferentes níveis de proficiência;[3] 2) manter a precisão do teste reduzindo significativamente o número de itens; e 3) aumentar a precisão do teste caso se mantenha o número de itens (BARRADA, 2012). Evidentemente, a viabilidade de um TAI depende, como qualquer metodologia de avaliação, de um banco de itens com grande variabilidade e densidade de dificuldade em relação à escala de proficiência.

Em um teste adaptativo, após a resposta dada ao primeiro item, são selecionados itens de acordo com as respostas anteriores. Assim, a cada resposta, adiciona-se alguma informação sobre a medida de proficiência do respondente, processo pelo qual se diminui a incerteza (erro) da estimativa de proficiência. Em outras palavras, o instrumento de medida – o teste – torna-se mais preciso a cada resposta. Adicionalmente, a escolha de itens em tempo real permite a coleta de informações mais qualificadas sobre a proficiência dos alunos em pontos menos explorados da escala, como os extremos superior e inferior.

Para se compreender melhor essa ideia, basta imaginar um teste linear apresentado a um aluno que tenha baixa proficiência e tenha acertado seis itens fáceis que lhe foram apresentados no conjunto de itens do teste. Nessa situação, o acerto de mais um item fácil adicionaria pouca informação ao conjunto, uma vez que já se tenha considerável certeza de que o aluno tem um domínio menor dentre as habilidades do construto em questão – o objeto do qual se busca estimar o domínio. Esse novo item não deverá alterar a estimativa da proficiência, tampouco reduzirá a incerteza da estimativa. Esta é uma das grandes vantagens dos testes adaptativos: otimizar a coleta de informação sobre a proficiência do aluno, evitando realizar medidas, mediante a apresentação de mais itens, desnecessárias, pois

> Em um formato TAI, a seleção de itens e a estimativa de proficiência seguem de mãos dadas. A eficiência na estimativa de proficiência está fortemente relacionada à seleção de itens apropriados para um indivíduo. De maneira circular, a adequação dos itens para um indivíduo depende em grande parte da qualidade das estimativas de proficiência intermediárias.[4] (LINDEN; GLAS, 2010, p. 4, tradução nossa)

3 Tendo em vista que, em testes convencionais, há uma imprecisão maior nas estimativas de proficiência nos extremos inferior e superior da escala.

4 Do original: *"Within a CAT format, item selection and ability estimation proceed hand in hand. Efficiencies in ability estimation are heavily related to the selection of appropriate items for an individual. In a circular fashion, the appropriateness of items for an individual depends in large part on the quality of interim ability estimates".*

No TAI da Provinha Brasil – Leitura, objeto deste artigo, utilizou-se um banco de 39 itens fornecido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) e buscou-se construir um teste com menos itens e maior precisão para cada aluno, em média, do que o teste linear original aplicado em papel composto de 20 itens de múltipla escolha. Ademais, além de diminuir a quantidade de itens e aumentar a precisão, procurou-se resolver um problema que não se encontrava abordado na literatura, ou seja, que, em uma avaliação da aprendizagem com função formativa, o critério de parada do teste levasse em conta não apenas a incerteza relativa à proficiência estimada, mas também a incerteza relativa ao nível de proficiência atribuído ao examinando, ou seja, sua classificação em um dos segmentos da escala de proficiência. Sobretudo porque a cada nível de proficiência corresponde uma interpretação pedagógica, e a perspectiva do teste é que seu resultado seja utilizado por professores em suas atividades cotidianas, quando é mais importante saber a descrição do que o aluno é capaz de realizar naquele nível do que receber um número – a proficiência estimada –, que apenas permitiria comparar o desempenho dos alunos entre si. A partir dessa informação mais qualitativa, os professores podem planejar atividades didáticas para atingir patamares superiores. Adicionalmente, a avaliação – julgamento sobre o desempenho do aluno – é feita segundo critérios que relacionam desempenho e ano de escolarização com esses níveis de interpretação pedagógica.

Se a ciência dos testes adaptativos ainda está evoluindo, revelando seu grande potencial, há também consideráveis desafios quanto à complexidade dos procedimentos estatísticos, em comparação a dos testes lineares (LINDEN; GLAS, 2010). É necessário, por exemplo, que haja um ou mais critérios para a seleção de itens, e a escolha de um critério pode depender de fatores como os objetivos do teste, a extensão do banco de itens e a distribuição das dificuldades de seus itens. É importante considerar, ainda, questões relativas à segurança e sustentabilidade do banco de itens, que podem ser otimizadas com diferentes técnicas de controle da taxa de exposição do item (BARRADA, 2010). Além disso, o TAI precisa de um ou mais critérios de parada. Os critérios normalmente utilizados são: 1) alcançar um número predeterminado de itens; 2) alcançar um mínimo de incerteza na estimativa da proficiência; e 3) alcançar limite mínimo de informação que um novo item adicionaria à estimativa da proficiência (BARRADA, 2012).

Como ponderado anteriormente, a qualidade e o tamanho do banco de itens são um aspecto importante e desafiador, cuja solução depende dos objetivos do teste e de iniciativas específicas de elaboração de itens. Para Barrada (2012), há quatro objetivos gerais para o TAI, aos quais é possível dar maior ou menor importância: 1) confiabilidade da estimativa da proficiência; 2) segurança do banco de itens; 3) restrições de conteúdo; e 4) manutenção do banco de itens. Alguns desses objetivos

se contrapõem, como, por exemplo, (1) em relação a (2) e (4). Com efeito, há um *trade-off* entre precisão da medida e segurança do banco de itens,[5] que pode ser minimizado com outros métodos de seleção de itens, em detrimento da seleção do item mais informativo a cada momento do teste (GEORGIADOU; TRIANTAFILLOU; ECONOMIDES, 2007).

O TAI descrito neste artigo é uma versão adaptativa informatizada da Provinha Brasil, um instrumento padronizado disponibilizado pelo Ministério da Educação (MEC), criado em 2007 (BRASIL, 2007a), para avaliação da proficiência em leitura, cujos resultados são expressos numa escala de cinco níveis, numerados de 1 a 5. O TAI aqui descrito foi aplicado em 15 escolas de ensino fundamental da Rede Municipal de Ensino de São Paulo. O artigo tem uma primeira seção na qual há uma breve descrição da Provinha Brasil – Leitura, depois é focalizado o funcionamento do algoritmo do TAI, que é o componente responsável pela parte "adaptativa" do teste. Na terceira seção, são tratadas as simulações computacionais utilizadas na construção do algoritmo e no ajuste inicial de alguns parâmetros. Na seção seguinte, são apresentados os resultados da aplicação do TAI, que confirmaram a validade do instrumento. Na quinta seção, é tratada a aplicação experimental do TAI, na qual é analisada, entre outros resultados, uma questão mais geral a partir dos dados coletados, relativa à forma de os alunos gerirem o tempo de prova. Na conclusão, estão condensados os resultados do TAI da Provinha Brasil – Leitura e algumas indicações para a continuidade de pesquisas assemelhadas.

A PROVINHA BRASIL – LEITURA

O processo de alfabetização e letramento é especialmente importante nos anos iniciais do ensino fundamental, momento em que professores destinam grande parte do ensino para o desenvolvimento da competência leitora e escritora das crianças (SOARES, 2016). Os indicadores nacionais têm mostrado resultados insatisfatórios no desenvolvimento dessas competências para a totalidade da população brasileira em idade escolar. Nesse contexto, a construção do TAI voltou-se para a Provinha Brasil – Leitura,[6] desenvolvida pelo Inep para os alunos do 2º ano do ensino fundamental a partir do Plano de Desenvolvimento da Educação (PDE) (BRASIL, 2007b), em consonância com as recomendações de organismos internacionais para a "década da alfabetização" (2003 a 2012) (GONTIJO, 2012).

5 A segurança do banco de itens se refere, por exemplo, à taxa de exposição dos itens. Quanto mais um item é apresentado a uma população, menos informativo ele tende a ser. Especialmente se o teste proporciona recompensas (*high stakes*).

6 A Provinha Brasil incorporou a competência em Matemática a partir de 2011, mas o projeto do TAI concentrou-se na leitura, considerando sua importância no processo de escolarização.

A construção de uma versão adaptativa e informatizada da Provinha Brasil – Leitura decorreu de projeto desenvolvido pelos pesquisadores do Grupo de Estudos e Pesquisas em Avaliação Educacional (Gepave), vinculado à Faculdade de Educação da Universidade de São Paulo (Feusp), em parceria, no momento inicial do projeto, com o Inep e, durante todo seu desenvolvimento, com a Secretaria Municipal de Educação de São Paulo (SME-SP), mediante seu Núcleo Técnico de Avaliação (NTA), envolvendo gestores centrais, regionais (supervisores escolares), diretores e coordenadores pedagógicos de escola, professores e alunos.[7] Dar a devida importância à questão da leitura no desenvolvimento de um projeto pedagógico consequente requer, entre outros elementos, cuidar dos procedimentos avaliativos (SOARES, 2016), considerando a perspectiva formativa do uso de seus resultados, sendo a Provinha Brasil – Leitura um instrumento construído nessa perspectiva, no qual, com base numa Matriz de Referência, são consideradas as habilidades concernentes à alfabetização, compreendida como desenvolvimento da compreensão das regras de funcionamento do sistema de escrita alfabética, e ao letramento inicial, entendido como apreensão de possibilidades de usos e funções sociais da linguagem escrita. Alfabetização e letramento, no escopo da Provinha Brasil, são abordados como processos complementares e paralelos, e os itens que compõem os testes buscam abarcá-los no espectro de dificuldade do teste. Importante salientar que os itens são pré-testados nacionalmente, além de a matriz estar referenciada em vários documentos do MEC voltados para a formação de docentes dos anos iniciais do ensino fundamental. Com isso, buscou-se garantir tanto a fidedignidade dos resultados quanto a validade para fins de utilização de seus resultados, condição para que possam ser integrados plenamente no processo de ensino nos primeiros anos da escolarização.

Aplicada desde 2008 e descontinuada em 2016, a Provinha Brasil compõe-se de dois testes, sendo o primeiro para aplicação em março (início do ano letivo) e o segundo, em outubro (final do ano letivo), ambos contendo 20 itens de múltipla escolha elaborados segundo uma matriz de especificações para leitura, pré-testados, conforme normas estatísticas, calibrados por especialistas do Inep e disponibilizados aos professores do país, conjuntamente com orientação para aplicação, "correção" – na verdade, processamento das respostas – e interpretação dos resultados.

Esses resultados iniciais, constituídos pela quantidade de acertos dos alunos nos testes, são, posteriormente, expressos numa escala de proficiência em leitura com cinco níveis de proficiência, com uma interpretação pedagógica para cada nível e com sugestões de trabalho docente para avanços dos alunos. A qualidade desse processo está apoiada no pré-teste dos itens para que se possa efetuar essa

7 Mais detalhes do projeto podem ser encontrados em Alavarse *et al.* (2018) e Catalani (2019).

associação entre quantidade de acertos e proficiência, tendo sido confirmada, no caso do TAI em epígrafe, em discussões com mais de 100 professores das escolas participantes do projeto.

Na versão em papel da Provinha Brasil, os próprios professores aplicam o teste, tabulam as respostas e interpretam os resultados, sendo todos os itens do teste disponibilizados para os professores e gestores a cada aplicação. Essa disponibilização, segundo opiniões de professores, permite maior compreensão dos resultados dos alunos, como se apreende num estudo de caso no município de Camaragibe, que descreveu uma interessante apropriação do instrumento, destacando ainda a importância do trabalho constante dos professores e gestores para os bons resultados relatados (MORAIS; LEAL; ALBUQUERQUE, 2009). Ainda que não seja o foco deste trabalho, é importante registrar dissensos em relação à concepção de leitura implícita na matriz de referência e na própria estrutura do teste da Provinha Brasil, como encontrado em Gontijo (2012). Entretanto, como apontado, são constatadas potencialidades da Provinha Brasil no apoio ao trabalho docente, pois o resultado da aplicação não se resume a um número, como seria a contagem de acertos ou mesmo o nível de proficiência do aluno, pois, no material de aplicação de cada teste, cada nível é acompanhado de uma descrição do que o aluno é capaz de fazer em termos de alfabetização e letramento inicial, e, ainda, são feitas algumas sugestões de intervenção docente para que o aluno possa avançar para o nível superior.

Importante salientar que, no banco de itens utilizado, cada item estava associado a um descritor da Matriz de Leitura, isto é, a cada elemento dessa matriz que, em conjunto, descreve o que seria o construto "leitura", cuja proficiência se pretende avaliar e que, nos documentos da Provinha Brasil, é considerada uma

> [...] atividade que depende de processamento individual, mas se insere num contexto social e envolve [...] capacidades relativas à decifração, à compreensão e à produção de sentido. A abordagem dada à leitura abrange, portanto, desde capacidades necessárias ao processo de alfabetização até aquelas que habilitam o(a) estudante à participação ativa nas práticas sociais letradas, ou seja, aquelas que contribuem para o seu letramento. Isso implica que o(a) estudante desenvolva, entre outras habilidades, as de ler palavras e frases, localizar informações explícitas em frases ou textos, reconhecer o assunto de um texto, reconhecer finalidades dos textos, realizar inferências e estabelecer relações entre partes do texto. (BRASIL, 2016, p. 9)

Nesses termos, a Provinha Brasil, além de se constituir em material sob maior controle dos professores para fins de sua aplicação e tratamento das respostas,

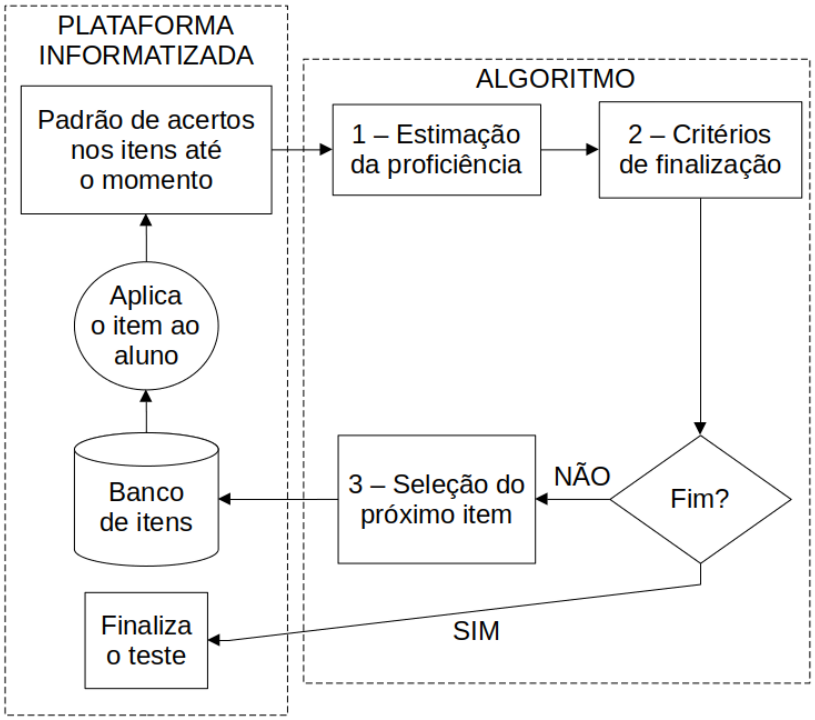
explicita seus fundamentos, o escopo de seus resultados interpretados e, ainda, fornece alternativas de desdobramentos didáticos, o que favorece o debate nas escolas, inclusive por ser mais transparente do que outras avaliações da aprendizagem externas.

O FUNCIONAMENTO DO ALGORITMO DO TAI

São diversos os desafios para viabilizar um teste de avaliação da aprendizagem. No caso de um TAI, além dos tradicionais, relacionados à qualidade dos itens e validade do teste como um todo, dois outros são adicionados: ser informatizado e adaptativo, visando a promover benefícios de ordem logística e de precisão de medida de proficiência. Para enfrentar esses dois desafios, optamos por construir dois módulos relativamente independentes: uma interface amigável para o usuário (criada com tecnologia *web* utilizando Java como principal linguagem de programação) e um sistema de processamento estatístico e psicométrico em tempo real (desenvolvido na linguagem R). O primeiro módulo permite ao teste ser informatizado, enquanto o segundo o torna adaptativo. A interface será aqui referida como "plataforma informatizada", enquanto o processamento estatístico será denominado de "algoritmo".

A plataforma informatizada é responsável pela exibição de itens aos respondentes e pela captura automática das respostas. Funciona *online* e foi acessada através de uma *intranet* para todas as escolas participantes. A apresentação dos itens foi realizada através do uso de *tablets* conectados à rede *Wi-Fi* das escolas. A leitura das comandas dos itens aos alunos, que na versão original da Provinha deve ser feita pelo professor aplicador, foi feita mediante essa plataforma e disponibilizada individualmente aos alunos, ao usarem os *tablets*, equipados com fones de ouvido e API de Fala do Google. A Figura 1 representa, de forma simplificada, os componentes do TAI da Provinha Brasil – Leitura e suas inter-relações.

FIGURA 1 – Representação simplificada dos componentes do TAI da Provinha Brasil – Leitura e de suas inter-relações



Fonte: Elaboração dos autores.

O algoritmo tem como fundamento teórico a Teoria da Resposta ao Item (TRI), descrita, entre outros autores, por Baker (2001), e seu objetivo é proporcionar uma dinâmica adaptativa à plataforma informatizada de testes. Mais especificamente, buscou-se otimizar a precisão da medida e minimizar o número de itens do teste, evitando perdas na validade do instrumento e no tempo de processamento computacional.

O algoritmo engloba três componentes, cada qual caracterizado por uma sequência de ações, descritas a seguir.

1) Estimação de proficiência:
   a) recebe como *input* da plataforma as respostas de cada aluno – configurando um padrão de acertos – e a distribuição *a priori* da proficiência, além dos parâmetros do banco de itens;
   b) estima a proficiência e o erro padrão, utilizando um método bayesiano que incorpora a distribuição *a priori*.
2) Critério de finalização do teste:
   a) verifica se o teste alcançou o limite máximo de itens;
   b) verifica se o erro padrão é menor do que o limite máximo definido;
   c) verifica se o nível de proficiência foi identificado de forma confiável;
   d) se ao menos uma das três condições anteriores for verdadeira, e se o teste já ultrapassou o limite mínimo de itens, envia um *output* para a plataforma terminar o teste; caso contrário, segue com o terceiro passo.

3) Seleção do próximo item do teste:
    a) identifica o descritor da matriz menos representado entre os itens aplicados ao aluno até o momento;
    b) busca os itens desse descritor no banco de itens;
    c) busca o item mais informativo desse subconjunto, levando em conta a proficiência estimada na etapa 1;
    d) retorna o item selecionado como *output* para a plataforma (juntamente com a proficiência e erro padrão estimados, para que a plataforma possa incluí-los como informação na próxima vez que acionar o algoritmo, depois que o item selecionado for respondido).

A estimação da proficiência é realizada com base na distribuição esperada *a posteriori* – *Expected a Posteriori* (EAP) (BOCK; MISLEVY, 1982) – com 21 pontos de quadratura. Os critérios para seleção de itens incluem: 1) a Máxima Informação de Fisher (MFI) (BARRADA, 2010); e 2) a seleção equilibrada de itens de cada eixo de conteúdo da matriz.

A seleção equilibrada de itens entre os descritores da matriz de referência tem o objetivo de evitar a perda de validade da prova devido à seleção dos itens promovida pelo TAI. De fato, trata-se de um cuidado importante para evitar um efeito colateral do critério de seleção de itens por máxima informação de Fisher. O algoritmo mantém sempre uma proporção semelhante de itens de cada eixo de conteúdo para garantir que o teste adaptativo represente a matriz desejada. Contudo, embora tenha sido implementada no algoritmo, a seleção equilibrada de itens não foi aplicada nessa fase piloto, em virtude do número reduzido de itens no banco.

Para determinação do fim de teste, o algoritmo utiliza um critério misto, levando em conta três regras:

a) número de itens do teste (mínimo de oito e máximo de 20 itens);
b) limite permitido de incerteza (erro padrão menor que 35 pontos);
c) grau de confiança na determinação do nível de proficiência da escala da Provinha Brasil – Leitura (85% de confiança, segundo nossas simulações).

Os dois primeiros critérios são amplamente utilizados em testes adaptativos (BARRADA, 2012). O terceiro critério (c) desenvolvido para o TAI da Provinha Brasil – Leitura constitui uma modificação do critério de parada utilizado nas avaliações com finalidade de classificar os sujeitos. A classificação é usada em situações de avaliação somativas, nas quais uma decisão do tipo sim ou não deve ser tomada, mais comum em exames e certificações (BABCOCK; WEISS, 2012; WEISS; KINGSBURY, 1984). Trata-se, aparentemente, de uma contribuição significativa desse projeto para o estado da arte do TAI.

SIMULAÇÕES DO TAI DA PROVINHA BRASIL – LEITURA

Esta seção apresenta as tecnologias utilizadas e a metodologia de simulação empregada no desenvolvimento do TAI da Provinha Brasil – Leitura.

Software e *hardware*

O algoritmo foi escrito na linguagem de programação R, especializada em estatística, de código aberto e livre. As simulações também foram realizadas nessa linguagem. Para ambas as finalidades, foram testados os pacotes: *catR* (MAGIS; RAÎCHE, 2012), *PP* (REIF, 2019) e *irtoys* (PARTCHEV, 2016). As simulações foram inicialmente realizadas em 2016 e refeitas em fevereiro de 2018, com novas versões dos pacotes, a fim de preservar a qualidade dos resultados apresentados. As versões dos pacotes utilizadas nas simulações de 2018, aqui apresentadas, foram: *irtoys* 0.2.0; *PP* 0.6.1; *catR* 3.13.

O computador utilizado para a realização das simulações foi um *notebook* com processador i7 de 4 núcleos de 2.50 GHz e 8 Gb de memória, com o Sistema Operacional Linux Mint. Não foi utilizado processamento paralelo ou aceleração via GPU.

Simulações

Para desenvolvimento do algoritmo, foram testados, via simulação, cinco métodos de seleção de itens (além da seleção aleatória) e sete métodos para estimação da proficiência. Os métodos foram testados quanto à precisão e velocidade. Além disso, as simulações permitiram o ajuste de dois parâmetros no algoritmo: o erro padrão máximo e o valor crítico do intervalo de confiança.

As simulações se fundamentam na função logística de dois parâmetros da TRI (ANDRADE; TAVARES; VALLE, 2000), que descreve a probabilidade de um indivíduo com proficiência conhecida acertar um item com parâmetros conhecidos. Embora tenham sido testados diferentes bancos, os resultados aqui descritos se referem ao banco de itens fornecido pelo Inep, composto de 39 itens com dois parâmetros definidos (dificuldade e discriminação).

Para cada situação, foram realizadas 1.000 simulações, cada uma com 1.000 participantes (com distribuição normal de proficiência, de média 500 e desvio padrão 100) respondendo a 20 dos 39 itens do banco original do Inep.

Para estimação da proficiência foram comparados quatro métodos, sendo um deles testado em vários pacotes, totalizando sete métodos na prática. Dois métodos se fundamentam no princípio da verossimilhança: a busca da máxima verossimilhança (LORD, 1980) e da verossimilhança ponderada (WARM, 1989). Os outros dois métodos utilizam estatística bayesiana: a distribuição esperada *a posteriori* (EAP, do inglês *Expected a Posteriori)* (BOCK; MISLEVY, 1982) e o estimador modal (BIRNBAUM, 1969).

A seguir, listamos os sete métodos comparados, provenientes dos três pacotes:

1) ML: máxima verossimilhança (pacote *catR*);
2) WL: verossimilhança ponderada (do pacote *catR*);
3) BM: estimador bayesiano modal (do pacote *catR*);
4) EAP: método EAP (função thetaEst do pacote *catR*);
5) eapC: método EAP (função eapEst do pacote *catR*);
6) eapI: método EAP (do pacote *irtoys*);
7) eapP: método EAP (do pacote *PP*).

Os métodos para seleção de itens testados, provenientes do pacote *catR* (MAGIS; RAÎCHE, 2012), foram os seguintes:

a) *random*: seleção aleatória de itens do banco;
b) MFI: seleciona o item com maior informação para a proficiência estimada até então, a partir da função de informação do item (ANDRADE; TAVARES; VALLE, 2000);
c) bOpt (A regra de Urry): seleciona o item com nível de dificuldade mais próximo da proficiência estimada até então;
d) thOpt (estratificação por Máxima Informação): adaptação do método MFI com objetivo de aumentar a segurança do banco de itens;
e) O método progressivo (REVUELTA J.; PONSODA, 1998): o item é selecionado segundo dois elementos, um relativo à Máxima Informação e outro aleatório. Ao longo do teste, o elemento aleatório vai perdendo a importância. Isso promove maior segurança do banco de itens; e
f) O método proporcional (BARRADA, 2010): o item é selecionado segundo probabilidades relacionadas à Informação de Fisher, também com objetivo de promover maior segurança do banco de itens.

Critérios para finalização de teste

O principal objetivo na elaboração do critério para finalização foi proporcionar um teste que tenha menos itens e maior precisão do que um teste similar, porém não adaptativo. Para tanto, foram considerados três critérios simultaneamente.

Em primeiro lugar, foi definido de antemão um limite máximo de 20 itens e um mínimo de oito. O limite mínimo garante a aplicação de ao menos quatro itens de cada um dos dois eixos da Provinha.[8] Já o limite máximo garante a finalização do teste nos patamares do teste em papel da Provinha Brasil, mesmo que os demais critérios não sejam alcançados.

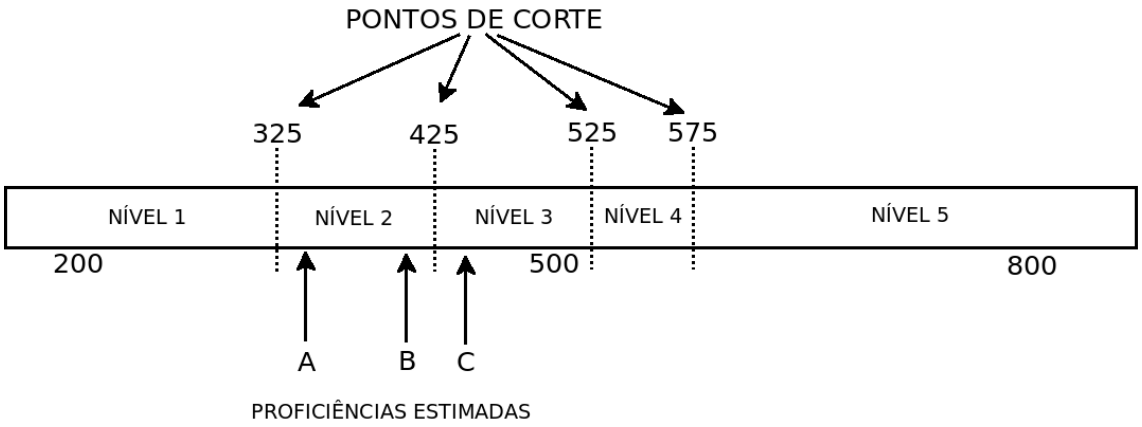
8 Os dois eixos utilizados na Provinha Brasil são: 1) apropriação do sistema de escrita; e 2) leitura. Contudo, em virtude do tamanho do banco disponível, optamos por não utilizar esse critério no piloto.

Em segundo lugar, foi definido um erro máximo na estimativa da proficiência, mensurado pelo erro padrão. Esse erro começa alto e vai diminuindo ao longo do teste, conforme mais itens vão sendo respondidos. O critério de finalização consiste em determinar um limite máximo permitido para o erro padrão, de acordo com os itens disponíveis, a população-alvo e os objetivos. E, finalmente, o critério "confiabilidade do nível de proficiência", descrito a seguir.

Confiabilidade do nível de proficiência

Este critério busca otimizar o tamanho da prova, garantindo a alocação correta do aluno no nível de proficiência. Ele determina o encerramento do teste a partir do momento em que a proficiência e seu intervalo de confiança estão inteiramente contidos em um único dos cinco níveis de proficiência definidos para a Provinha Brasil (Figura 2). Vale ressaltar que os pontos que dividem a escala da Provinha Brasil em cinco níveis, também denominados pontos de corte, resultaram de um processo psicométrico (ancoragem) associado à análise pedagógica dos itens realizada por especialistas e educadores. É a descrição dos níveis – resultante desse processo – que permite a interpretação da nota do aluno (proficiência) com base na Teoria da Resposta ao Item. Embora esse método tenha sido identificado posteriormente na literatura como método de parada com objetivo de classificação, não foi encontrado o uso para fins de avaliação formativa na literatura sobre testes adaptativos.

FIGURA 2 – Pontos de corte e níveis de proficiência na escala da Provinha Brasil – Leitura



Fonte: Elaboração dos autores.

Com a adição desse critério de fim de teste, o TAI da Provinha Brasil pode ser encerrado mesmo que o erro na estimativa da proficiência de um aluno seja relativamente alto, desde que o intervalo que contenha a proficiência esteja completamente contido em um dos cinco níveis da Provinha Brasil – Leitura. Afinal, a Provinha não é um teste de seleção ou certificação, em que a precisão da

proficiência é mais relevante. O mais importante é saber se o nível de proficiência do aluno foi adequadamente identificado, priorizando assim o diagnóstico pedagógico e, consequentemente, a intervenção para melhoria da alfabetização. Em outras palavras, esse critério contribui para a interpretação e avaliação, que vão além da simples mensuração.

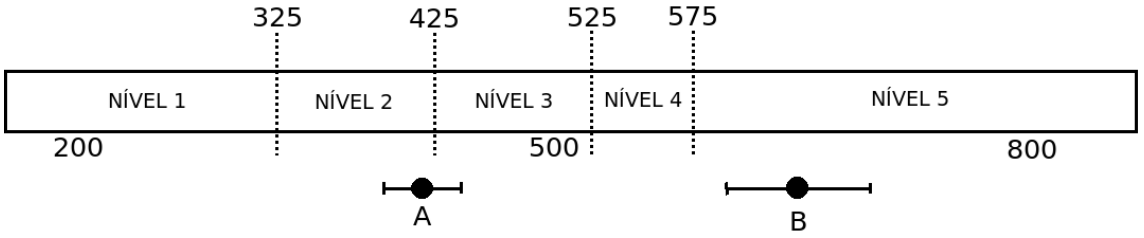
A título de exemplo, na Figura 2 é possível observar que a proficiência aferida para o aluno B está mais próxima da proficiência aferida para o aluno C do que a aferida para o aluno A. Contudo, levando-se em conta os pontos de corte definidos pela interpretação pedagógica da escala, os alunos A e B estão no mesmo nível da escala, enquanto o aluno C pertence ao nível seguinte. Em termos pedagógicos, isso significa que, supostamente, os alunos A e B demonstraram domínio de habilidades que exigem intervenções similares, enquanto o aluno C demonstrou domínio mais amplo que exige outras intervenções.

Cabe ressaltar que há sempre alguma incerteza na estimação da proficiência de um aluno, qualquer que seja o método utilizado. Tal incerteza depende principalmente da quantidade de itens respondidos e também da proximidade entre a dificuldade do item e a proficiência do respondente. Supondo a normalidade da distribuição de proficiência, a confiabilidade da estimativa (seu intervalo de confiança) pode ser determinada a partir do erro padrão. O intervalo de confiança é definido por um limite mínimo e um máximo e pode ser obtido com diferentes graus de confiança, multiplicando-se o erro padrão pelo valor crítico, que, por sua vez, depende do grau de confiança desejado. Para o algoritmo foi definido um grau de confiança de 85%, que corresponde a um valor crítico de 1,44 (FERREIRA, 2005). Isso significa que, quando o teste termina por esse critério, há 85% de chance de a proficiência verdadeira do aluno estar dentro do nível identificado pelo teste, segundo a Teoria da Resposta ao Item.

A Figura 3 ilustra a utilidade do critério de finalização do teste por confiabilidade do nível de proficiência como complemento ao critério de erro padrão máximo. O aluno B tem um erro (intervalo de confiança) maior do que o aluno A, porém seu nível de proficiência já foi estimado com segurança (nível 5) depois que ele acertou cinco itens em um teste adaptativo. O aluno A, por sua vez, respondeu a 11 itens nesse teste adaptativo, mas ainda não foi classificado de forma confiável, podendo pertencer aos níveis 2 ou 3. Afinal, o erro da estimativa não depende apenas do número de itens apresentados, mas também do padrão de acertos de cada aluno e dos parâmetros dos itens respondidos. Mais ainda, a quantidade e posicionamento dos pontos de corte interferem fortemente nesse critério. Com efeito, o aluno B foi beneficiado pela adição desse terceiro critério de finalização de teste: terminando mais rapidamente sem perda na precisão do teste, dado que a finalidade prática da Provinha Brasil – Leitura é fornecer uma medição confiável da proficiência

do aluno para que os professores, considerando os cinco níveis de proficiência, possam tomar decisões pedagógicas a partir de uma informação mais fidedigna. Destaca-se, sem entrar no mérito dessa decisão, que o Inep definiu como desejável que cada aluno esteja pelo menos no nível 4 ao final do 2º ano.

FIGURA 3 – Representação dos intervalos de confiança de duas proficiências estimadas



Fonte: Elaboração dos autores.

RESULTADOS DAS SIMULAÇÕES

Esta seção apresenta os resultados obtidos via simulação na etapa de construção do algoritmo e ajuste dos parâmetros básicos.

Métodos de seleção de itens e estimação de proficiência

Os métodos de seleção de itens e estimação de proficiência foram testados quanto à sua precisão e velocidade de processamento. Em relação à seleção de itens, todos os métodos se mostraram suficientemente rápidos. O Teste t revelou que todos os critérios apresentam menor erro do que a ausência (*random*, seleção aleatória de itens), mas que as diferenças entre os métodos não são significativas ($p < 0,05$) na maioria das simulações. Assim sendo, levando em conta a literatura especializada, optamos por incluir no algoritmo a seleção pelo método da Máxima Informação de Fisher. Contudo, se o TAI da Provinha Brasil – ou outro TAI – vier de fato a se consolidar como política pública, será necessário rever essa escolha técnica, pois ela não leva em conta a segurança e sustentabilidade do banco de itens. Os métodos conhecidos como progressivo ou proporcional poderiam ser mais adequados nesse caso.

Em relação aos métodos de estimação da proficiência, não foi observada diferença significativa ($p < 0,05$) na precisão dos métodos com menor erro (BM, EAP, eapC, eapI, WL) para um teste de 20 itens, com população de proficiência média igual a 500. Contudo, vale notar que, em populações com média diferente da esperada (600 ou 400, por exemplo), o método WL se mostrou mais preciso do que os outros, o que se deve à sua menor dependência de uma estimativa *a priori* da população.

Observou-se ainda que o método EAP do pacote *PP* (eapP) apresentou erro muito maior do que os outros. Isso ilustra outra função importante das simulações

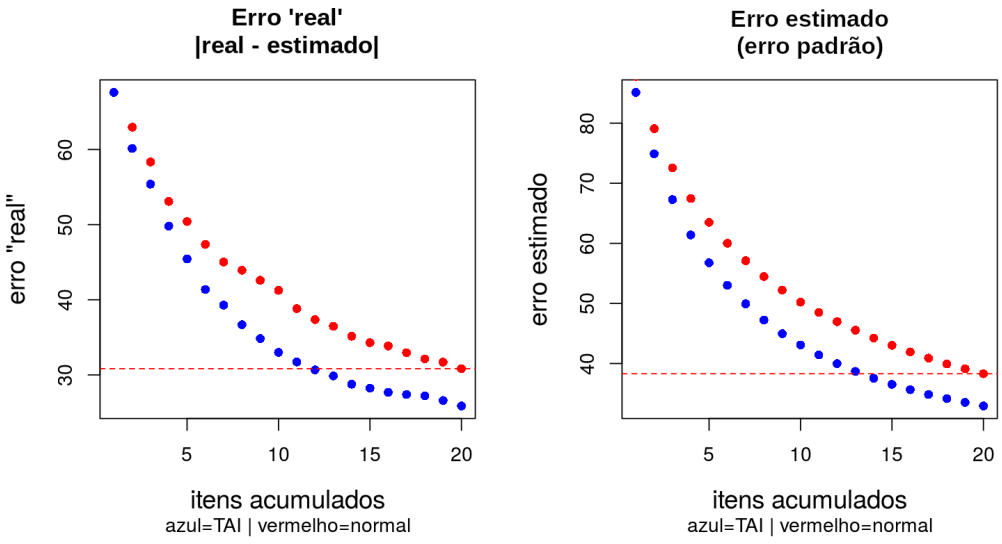
para o desenvolvimento de algoritmos, que é prevenir o uso de pacotes de qualidade duvidosa, com resultados inconsistentes. Tal cuidado é especialmente importante quando se trabalha com *software* livre, mas não deixa de ser necessário com *softwares* proprietários.

Em suma, levando em conta a precisão e a velocidade simultaneamente, o método escolhido para a estimação da proficiência foi o EAP do pacote *irtoys*. Para a seleção do próximo item, foi escolhido o método MFI, embora seja necessária uma revisão dessa escolha, caso haja necessidade de se preservar a segurança do banco de itens.

Critérios para fim de teste: erro padrão máximo

Para determinar o erro padrão máximo a ser aceito pelo TAI, consideramos o objetivo de produzir um teste que tenha, em média, maior precisão e menor tamanho do que um teste convencional, não adaptativo. A Figura 4 mostra os erros de estimativa em testes convencionais e adaptativos, segundo as simulações realizadas. O erro estimado é aquele que pode ser obtido pelo algoritmo do TAI a cada novo padrão de respostas do examinando. O erro "real", por sua vez, não pode ser obtido pelo algoritmo. Só o conhecemos no contexto da simulação.

FIGURA 4 – Erros de estimativa em testes convencionais (normal) e adaptativos (TAI) de diferentes tamanhos de até 20 itens, segundo as simulações realizadas



Fonte: Elaboração dos autores.
Nota: A linha vermelha pontilhada indica o erro do teste normal com 20 itens.

Nos dois gráficos da Figura 4, a linha pontilhada de cor vermelha marca o erro alcançado pelo teste convencional depois de 20 itens. Isso corresponde a um erro padrão de 38 e uma diferença de proficiências de 31. Com efeito, os gráficos mostram que um teste adaptativo entre 12 e 13 itens tende a produzir estimativas

tão precisas quanto um teste convencional de 20 itens. Com base nessa observação, é possível delimitar um ponto médio; ou seja, um erro padrão correspondente a um teste convencional de 15 a 19 itens. Tal ajuste depende do equilíbrio desejado entre precisão e tamanho do teste em cada situação de avaliação.

É importante destacar que nos dois métodos de detecção do erro houve essa semelhança, entre 12 e 14 itens, confirmando a qualidade da estimativa obtida pelo *software* utilizado. Caso não houvesse tal semelhança, o algoritmo utilizado não teria qualquer garantia de corresponder à realidade.

Enfim, com o objetivo de proporcionar um teste menor e mais preciso, foi escolhido o limite máximo para o erro padrão: 35 pontos na escala da Provinha Brasil, que corresponderia a um teste adaptativo com 16 itens. Posteriormente, os resultados da aplicação do teste confirmaram essa previsão.

Confiabilidade do nível de proficiência

Para delimitação dos níveis de proficiência, foram utilizados os pontos de corte da Provinha Brasil – Leitura (Figura 2). Para o mesmo conjunto de dados pode-se obter intervalos de confiança mais ou menos largos, dependendo do nível de confiança que se deseja. A cada nível de confiança corresponde um valor crítico. Para ajuste do melhor valor crítico para o algoritmo da Provinha Brasil, foram testados quatro níveis de confiança.

Nota-se na Tabela 1 que, como seria esperado, quanto maior o nível de confiança desejado, menos testes serão finalizados segundo o critério da confiabilidade do nível de proficiência. É importante saber, contudo, quantos testes foram corretamente finalizados (comparando o nível de proficiência "real" e o estimado) em cada caso. Verificamos quantos dos testes finalizados por esse critério teriam sido bem-sucedidos na determinação do nível de proficiência do aluno.

TABELA 1 – Testes finalizados pelo critério de confiabilidade do nível de proficiência segundo o nível de confiança, em 1.000 simulações

NÍVEL DE CONFIANÇA	VALOR CRÍTICO	TESTES FINALIZADOS	FINALIZAÇÕES CORRETAS	TAXA DE ACERTO (%)
80%	1,28	354	297	83,9
85%	1,44	242	207	85,5
90%	1,645	114	109	95,6
95%	1,96	62	62	100,0

Fonte: Elaboração dos autores.

Esses resultados mostram que, na escala da Provinha Brasil – Leitura, há razoável correspondência entre o nível de confiança definido pelo valor crítico e o nível de confiança do critério confiabilidade do nível de proficiência. No algoritmo, definimos o valor crítico de 1,44.

RESULTADOS DA APLICAÇÃO EXPERIMENTAL DO TAI

Foram realizadas 1.983 aplicações do TAI da Provinha Brasil – Leitura, abarcando 823 alunos do 1º ano e 1.160 alunos do 2º ano do ensino fundamental, distribuídos em 80 turmas de 15 escolas de ensino fundamental da Rede Municipal de Ensino de São Paulo.

TABELA 2 – Estatísticas descritivas da aplicação experimental do TAI da Provinha Brasil – Leitura a alunos do 1º e 2º ano do ensino fundamental da Rede Municipal de Ensino de São Paulo

ESTATÍSTICAS		PROFICIÊNCIA	ACERTOS	ITENS	DURAÇÃO	APLICAÇÕES
Total	Média	462,72	0,55	17,33	11,52	1983
	Erro padrão	1,84	0,00	0,05	0,12	
	Desvio padrão	81,93	0,20	2,18	5,53	
1º Ano	Média	416,83	0,44	16,74	10,88	823
	Erro padrão	2,14	0,01	0,06	0,20	
	Desvio padrão	61,41	0,15	1,60	5,75	
2º Ano	Média	495,28	0,63	17,75	11,98	1160
	Erro padrão	2,32	0,01	0,07	0,16	
	Desvio padrão	79,04	0,19	2,43	5,32	

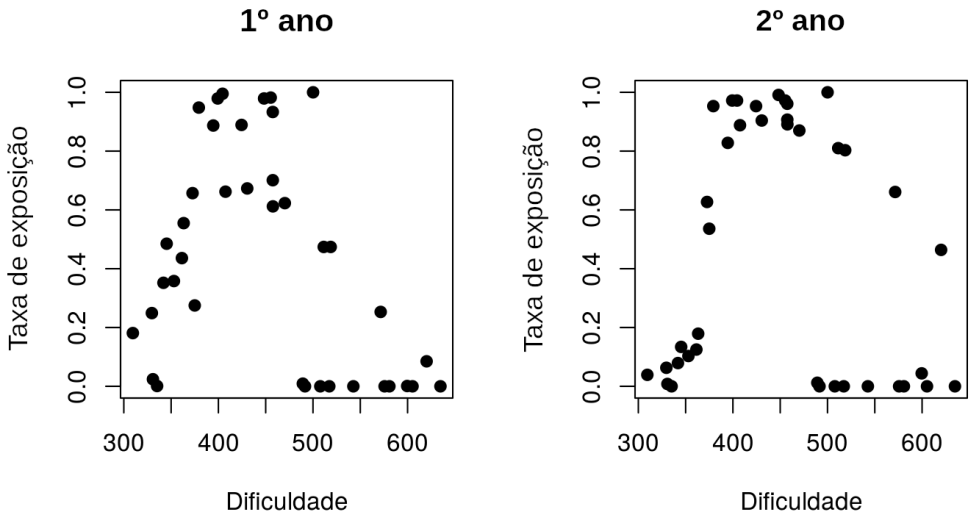
Fonte: Elaboração dos autores.

A Tabela 2 descreve quatro aspectos importantes da aplicação do TAI. Nota-se que a proficiência média estimada para o 2º ano (495,28 pontos) apresenta magnitude adequada, considerando-se que a escala da Provinha Brasil (projetada para o 2º ano) tem média 500 e desvio padrão 100. O desvio padrão (79,04), por sua vez, foi menor do que 100, o que reflete uma variância menor da população analisada em relação à população para qual a Provinha Brasil foi delineada, ou seja, alunos de todo Brasil. Nota-se, ainda, que a média do 1º ano foi menor do que a média da escala, o que também seria esperado, dado que os itens foram produzidos e calibrados para o 2º ano. Esses resultados confirmam, portanto, não apenas a qualidade do algoritmo utilizado no TAI, mas também a qualidade dos itens produzidos pelo Inep.

Em relação à taxa de exposição dos itens, o TAI de fato utilizou mais alguns itens do que outros. No total, foram aplicados apenas 31 itens para os alunos. Levando em conta que o banco possui 39 itens, conclui-se que houve aproveitamento de 79,5% do banco de itens. A Figura 5 busca relacionar a taxa de exposição e a dificuldade do item. Mais uma vez, os resultados empíricos corroboram os resultados esperados via simulação: além de uma taxa maior nos itens de dificuldade

média, observa-se também que, no 2º ano, houve um uso maior dos itens mais difíceis, e um uso menor dos itens mais fáceis. Nota-se, ainda, que a grande maioria dos itens não utilizados se encontra na porção superior da escala.

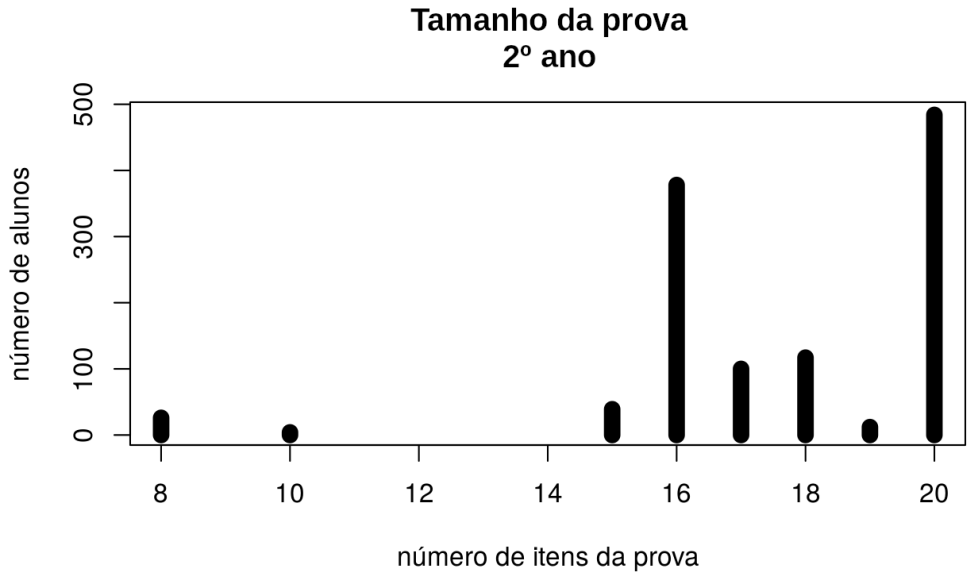
FIGURA 5 – Relação entre taxa de exposição e dificuldade dos itens para alunos do 1º e 2º anos do ensino fundamental



Fonte: Elaboração dos autores.

Em relação ao número de itens apresentados aos alunos, nota-se que boa parte dos testes terminou com 16 itens (Figura 6), como esperado pelas simulações e pela escolha do limite máximo de 35 pontos para o erro padrão da estimativa da proficiência. Além disso, alguns testes terminaram com dez ou menos itens, o que ocorreu em virtude do outro critério para finalização de teste, a confiabilidade do nível de proficiência. Nesse sentido, os resultados confirmam a importância desse critério como complemento ao critério do erro padrão máximo. Por outro lado, o elevado número de testes com 20 itens não foi previsto pelas simulações, revelando limitações no modelo utilizado. Levando em conta que o algoritmo visa a manter certo grau mínimo de incerteza, é possível que haja efeitos reais não previstos pelo modelo unidimensional de dois parâmetros da TRI, aqui utilizado como base das simulações.

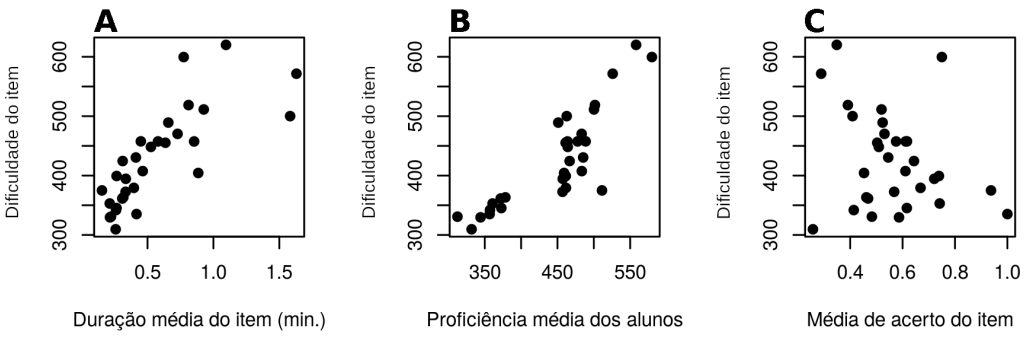
FIGURA 6 – Número de aplicações do TAI da Provinha Brasil – Leitura a alunos de 2º ano do ensino fundamental, segundo o tamanho da prova



Fonte: Elaboração dos autores.

Como esperado, os itens mais difíceis foram apresentados a alunos com maior proficiência, como mostra a tendência da Figura 7B. Da mesma forma, os itens mais difíceis demoraram mais para ser respondidos pelos alunos (Figura 7A). Nota-se ainda que não houve relação linear entre dificuldade e média de acerto nos itens (Figura 7C), uma propriedade dos testes adaptativos marcadamente diferente dos testes convencionais, nos quais se observam alta correlação entre a média de acerto no item e sua dificuldade estimada via TRI. De modo geral, os três resultados confirmam o que seria esperado em um TAI.

FIGURA 7 – Relação entre a dificuldade (parâmetro b) dos itens e três aspectos de sua aplicação: A) o tempo médio para resolução do item; B) a proficiência média dos alunos que receberam o item; C) a média de acerto no item. Alunos de 1º e 2º ano do ensino fundamental



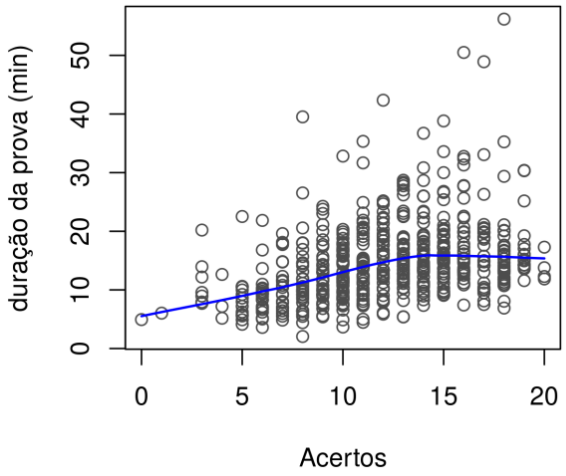
Fonte: Elaboração dos autores.

Gestão do tempo e proficiência

O aspecto temporal da execução do teste também foi investigado, comparando alunos de diferentes proficiências. A hipótese inicial é que os alunos mais proficientes tendem a gerir melhor o tempo de prova.

Em primeiro lugar, os resultados dos testes não adaptativos[9] mostram uma associação positiva entre a proficiência e a duração da prova (Figura 8). Ou seja, há certa tendência de que alunos mais proficientes fiquem mais tempo fazendo a prova. A regressão linear confirma que essa relação é significativa ($p<0,001$). Por outro lado, parece haver um limite para essa tendência, pois, a partir de 14 acertos, os alunos parecem não precisar de mais tempo para obter bons resultados. Observa-se inclusive uma fraca tendência de queda a partir desse ponto, mas a regressão linear indica que ela não é significativa ($p = 0,42$).

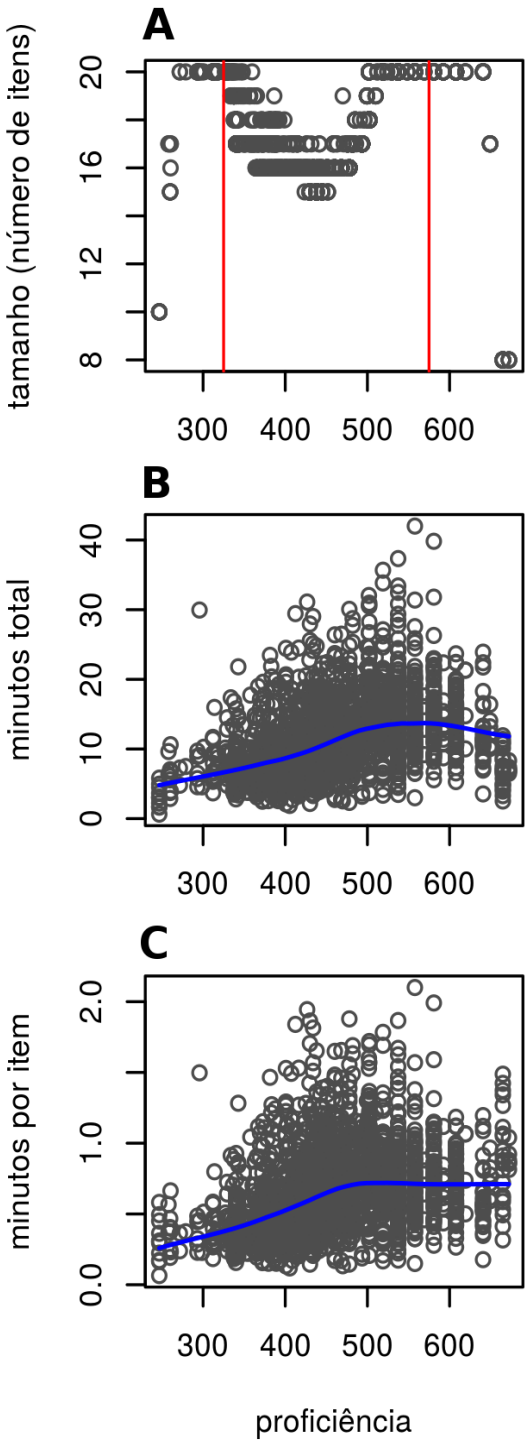
FIGURA 8 – Relação entre o número de acertos e a duração da Provinha Brasil – Leitura (versão eletrônica não adaptativa). Alunos de 2º ano do ensino fundamental



Fonte: Elaboração dos autores.

9 Os resultados e amostra dos testes eletrônicos não adaptativos – realizados nesse projeto como uma espécie de controle – estão detalhados em Alavarse *et al.* (2018).

FIGURA 9 – Relação entre proficiência e (A) tamanho, (B) duração total e (C) duração ponderada. Alunos de 1º e 2º ano do ensino fundamental



Fonte: Elaboração dos autores.
Nota: Linhas vermelhas: pontos de corte mínimo e máximo. Linhas azuis: tendência não linear calculada com estimador M.

Os resultados do TAI são semelhantes. É importante levar em conta que o TAI tem tamanhos variados (em número de itens), diferentemente do teste não adaptativo, sempre com 20 itens. Assim sendo, a análise da relação entre proficiência

e duração da prova é um pouco mais complexa no caso do TAI. A Figura 9 resume as principais informações necessárias nesse sentido.

Nota-se que os menores testes (com oito ou dez itens) correspondem a alunos nos extremos da escala apresentada na Figura 9A. Nesses casos, a prova finalizou pelo critério de *confiabilidade do nível de proficiência*. Mas a maior parte das finalizações geradas pelo algoritmo adaptativo se baseou no critério de *confiabilidade da proficiência*, que acabou surtindo efeito na faixa mais mediana da escala de proficiência – em que o erro de estimativa costuma ser menor. Essa distribuição na escala de proficiência confirma o caráter complementar dos dois critérios de finalização de teste em um TAI.

Na Figura 9B, nota-se uma tendência de crescimento até certo ponto. A partir de 600 pontos, há um declínio suave. Contudo, essa figura representa a duração total do teste, mas o TAI tem diferentes tamanhos. É possível que os alunos com mais de 600 pontos tenham ficado pouco tempo na prova, pois o algoritmo finalizou o teste rapidamente. De fato, quando observamos a Figura 9C, essa hipótese se confirma, já que o declínio desaparece quando analisamos o tempo por item, em vez do tempo total de prova.

Nota-se ainda que a tendência observada na Figura 9C é semelhante à observada na Figura 8. Com efeito, essa relação entre proficiência e tempo investido na prova (uma associação positiva até certo ponto, a partir do qual se estabiliza) se mostra consistente nos dois tipos de teste, adaptativo e não adaptativo. Um detalhe importante é que o tempo de prova deixa de aumentar quando a proficiência é um pouco maior do que a média observada: 463 pontos na Figura 9C e 13 acertos na Figura 8. A interpretação desses fatos não é clara, mas sua repetição nos dois casos indica um terreno fértil para investigação.

Foi analisado também um segundo aspecto da relação entre gestão do tempo e proficiência. Para tanto, selecionaram-se dois estratos da amostra: o terço inferior e o terço superior dos alunos em termos de proficiência. A Tabela 3 compara os dois estratos em algumas características dos itens aplicados.

TABELA 3 – Média e coeficiente de variação da duração de aplicação dos itens, segundo os estratos definidos pelo terço inferior e superior dos alunos de 1º e 2º ano do ensino fundamental em termos da proficiência

ESTRATOS	DURAÇÃO MÉDIA (MINUTOS)	COEFICIENTE DE VARIAÇÃO DA DURAÇÃO	NÚMERO TOTAL DE ITENS RESPONDIDOS
Terço inferior	0,45	0,91	29
Terço superior	0,68	0,74	22
P valor (Teste t pareado)	0,01	0,01	---

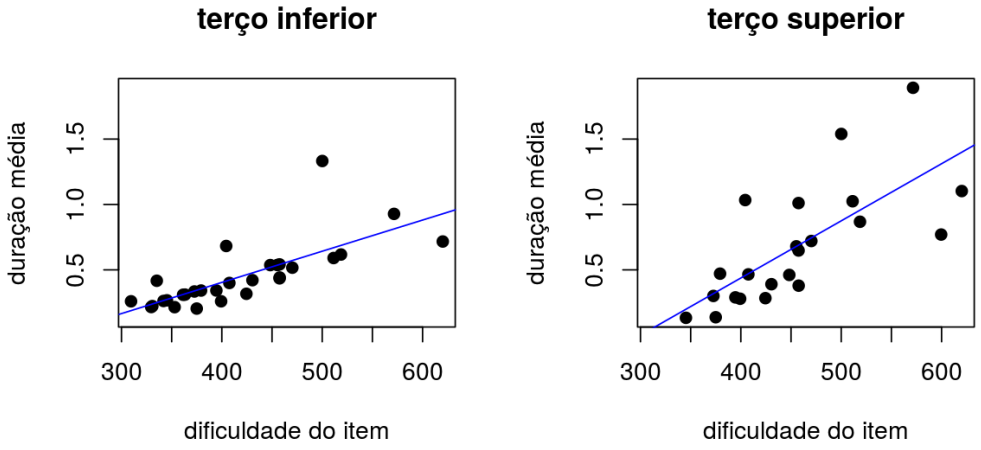
Fonte: Elaboração dos autores.

Nota: A significância da diferença entre os estratos é apontada pelo P valor do Teste t pareado.

Para verificar se as diferenças nas médias dos dois estratos descritas na Tabela 3 são significativas, foi realizado o Teste t pareado com 95% de confiança. Nota-se que os alunos do estrato superior tendem a ficar mais tempo resolvendo os itens, o que confirma os resultados anteriores. Além disso, o coeficiente de variação indica que há menor variabilidade no estrato superior; ou seja, esses alunos tendem a gerir o tempo de forma mais homogênea em cada item, sugerindo certa racionalidade quanto ao uso do tempo, e talvez também uma capacidade de concentração mais constante.

Por fim, há um ponto a ser verificado, partindo da relação observada entre a duração média e a dificuldade dos itens (Figura 7A): os alunos tendem a demorar mais nos itens mais difíceis, o que faz sentido e pode ser considerado parte da capacidade de gerir o tempo de prova. Mas em que medida isso se diferencia nos dois estratos? A Figura 10 mostra que, no estrato superior, a relação entre duração e dificuldade do item é mais intensa do que no estrato inferior. A regressão linear confirma essa observação.

FIGURA 10 – Relação entre duração (em minutos) e dificuldade (na escala Provinha Brasil – Leitura) dos itens em cada estrato. Alunos de 1º e 2º ano do ensino fundamental



Fonte: Elaboração dos autores.
Nota: Linha azul estimada por regressão linear (p<0,001).

Em suma, nossos resultados sugerem que os alunos mais proficientes tendem a gerir melhor o tempo nos seguintes aspectos: a) investem mais tempo na prova, de um modo geral; b) investem mais tempo nos itens mais difíceis, e menos tempo nos mais fáceis; c) investem o tempo de forma mais homogênea em cada item – o que poderia estar relacionado, já no campo da especulação, a uma capacidade de concentração mais constante.

CONCLUSÃO

Os resultados obtidos confirmam a qualidade do TAI da Provinha Brasil – Leitura e dos métodos utilizados para desenvolver o algoritmo, partindo do arcabouço conceitual da TRI, com base em simulações e *software* livre. Com efeito, este texto pode servir como referência para a construção de outros testes adaptativos em moldes semelhantes. Foram observadas diversas características esperadas em um TAI, como a redução no tamanho médio das provas e no tempo médio de resolução, o uso diferencial dos itens e a ausência de correlação entre dificuldade e média de acerto dos itens. O TAI da Provinha Brasil – Leitura se mostrou também um instrumento útil para ser aplicado a diferentes populações de alunos, podendo provavelmente incluir também o 3º ano, especialmente com a ampliação do banco com itens que pudessem atender a alunos com maior proficiência, como é de se esperar nesse ano; embora possa haver aqueles que tenham desempenho parecido com seus colegas de anos anteriores. Para os alunos do 1º ano, pelo fato de a aplicação ter ocorrido no final do ano letivo, e como realizariam o Teste 1 no início do próximo ano letivo, o banco de itens foi adequado, além do objeto de avaliação do TAI – alfabetização e letramento inicial – ser relativo a uma competência que é trabalhada desde o 1º ano e que é cumulativa. Esse caráter mais cumulativo e menos restrito a cada série ficou evidenciado no espectro das proficiências médias observadas no 1º e 2º anos, confirmando, em termos psicométricos, a qualidade da própria Provinha Brasil – Leitura, ainda que desenvolvida pelo Inep para uso prioritário dos professores do 2º ano do ensino fundamental. Vale, também, destacar que nenhum aluno do 1º ano manifestou "surpresa" com o conteúdo dos itens respondidos, considerando que todas as aplicações foram acompanhadas e registradas em relatórios.

O critério de finalização de teste proposto neste trabalho (confiabilidade do nível de proficiência) também se mostrou eficaz e potencialmente útil para testes com escala discretizada (em níveis), que priorizem a avaliação (não apenas a mensuração) e a interpretação pedagógica, seja com finalidade formativa ou somativa. Cabe notar que o grau de influência desse critério depende diretamente do número de pontos de corte e sua distribuição na escala de proficiência.

Além disso, a análise da duração do teste confirma a hipótese de que os alunos mais proficientes tendem a gerir melhor o tempo de prova. Em primeiro lugar, há uma associação positiva entre proficiência e tempo de prova, que se estabiliza na porção superior da escala de proficiência. Tal associação foi observada tanto nos testes adaptativos quanto não adaptativos. Outra tendência geral observada é de os itens mais difíceis demorarem mais para serem respondidos. Comparando os terços superior e inferior dos alunos na escala de proficiência, observou-se que: a) o terço superior demora mais tempo respondendo aos itens do teste; b) o terço superior apresenta menor variação no tempo dedicado a cada item, o que poderia estar

relacionado – especulamos – a certa constância na capacidade de concentração; c) o terço superior apresenta uma inclinação maior na relação entre dificuldade do item e duração de aplicação, ou seja, esses alunos tendem a dedicar mais tempo aos itens mais difíceis e menos tempo aos itens mais fáceis. Esse aspecto da gestão do tempo de prova foi encontrado em toda a população, mas está especialmente presente nos alunos mais proficientes.

É importante explicitar algumas limitações deste trabalho. Em primeiro lugar, não foi incluído o aspecto do engajamento pessoal do examinando nas simulações, apenas sua proficiência e os parâmetros dos itens do banco. No entanto, é esperado que haja um engajamento maior dos alunos em um teste adaptativo, especialmente dos alunos nos extremos superior e inferior da escala de proficiência. Em segundo lugar, o método de estimação utilizado (EAP) é bayesiano, produzindo resultados menos precisos quando a população não apresenta média próxima da esperada *a priori*. Nesse sentido, recomenda-se um método não bayesiano, como a verossimilhança ponderada (WL), caso não haja informação confiável sobre a população. Ou, ainda, pode-se empregar um método misto, iniciando o teste com WL e terminando com EAP. Em terceiro lugar, o banco de itens utilizado foi bastante reduzido, e, como vimos, a taxa de exposição de alguns itens se mostrou bastante alta, enquanto outros sequer foram aplicados. Para superar tal limitação, recomenda-se outro método de seleção de itens, pois a Máxima Informação de Fisher tende a gerar esse tipo de resultado, especialmente se associada a um modelo logístico com parâmetro a (inclinação) constante, como foi o caso. Outra limitação deste estudo, particularmente importante, é o tamanho reduzido do banco de itens, sendo aconselhável um banco com pelo menos 100 itens para um desempenho eficaz dos testes adaptativos.

Por fim, cabe mencionar possíveis aprimoramentos futuros, como a inclusão de métodos para otimizar a segurança do banco de itens (controlando a taxa de exposição, por exemplo), a transformação do TAI em um Teste de Múltiplos Estágios (TME), em que os itens são selecionados em blocos, ou mesmo a possibilidade de se pré-testar novos itens durante a aplicação do teste.

REFERÊNCIAS

ALAVARSE, O.; CATALANI, E.; MENEGHETTI, D.; TRAVITZKI, R. Teste Adaptativo Informatizado como recurso tecnológico para alfabetização inicial. *Revista Iberoamericana de Sistemas, Cibernética e Informática*, v. 15, n. 3, p. 68-78, 2018.

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

BABCOCK, B.; WEISS, D. J. Termination criteria in Computerized Adaptive Tests: do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, v. 1, n. 1, p. 1-18, Dec. 2012.

BAKER, F. B. *The basics of Item Response Theory*. 2nd ed. Washington: ERIC Clearinghouse on Assessment and Evaluation, 2001.

BARRADA, J. R. A method for the comparison of Item Selection Rules in Computerized Adaptive Testing. *Applied Psychological Measurement*, v. 34, n. 6, p. 438-452, 2010.

BARRADA, J. R. Tests adaptativos informatizados: una perspectiva general. *Anales de Psicología*, v. 28, n. 1, p. 289-302, 2012.

BIRNBAUM, A. Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, v. 6, p. 258-276, 1969.

BOCK, R. D.; MISLEVY, R. J. Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, v. 6, n. 4, p. 431-444, 1982.

BRASIL. Ministério da Educação. Portaria Normativa n. 10, de 24 de abril de 2007. *Diário Oficial da União*, Brasília, 26 abr. 2007a. Disponível em: http://portal.mec.gov.br/arquivos/pdf/provinha.pdf. Acesso em: 20 ago. 2020.

BRASIL. Presidência da República. Casa Civil. Subchefia para Assuntos Jurídicos. Decreto n. 6.094, de 24 de abril de 2007. Dispõe sobre a implementação do Plano de Metas Compromisso Todos pela Educação, pela União Federal, em regime de colaboração com Municípios, Distrito Federal e Estados, e a participação das famílias e da comunidade, mediante programas e ações de assistência técnica e financeira, visando a mobilização social pela melhoria da qualidade da educação básica. *Diário Oficial da União*, Brasília, p. 5, 25 abr. 2007b. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2007/Decreto/D6094.htm. Acesso em: 20 ago. 2020.

BRASIL. Ministério da Educação. *Provinha Brasil*: avaliando a alfabetização: guia de apresentação, correção e interpretação dos resultados: Teste 2. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2016.

CATALANI, Érica Maria Toledo. *Teste Adaptativo Informatizado da Provinha Brasil*: a construção de um instrumento de apoio para professores(as) e gestores(as) de escolas. 201. 282 f. Tese (Doutorado em Educação) – Faculdade de Educação, Universidade de São Paulo, São Paulo, 2019.

FERREIRA, D. F. *Estatística básica*. Lavras, MG: Ed. UFLA, 2005.

GEORGIADOU, E. G.; TRIANTAFILLOU, E.; ECONOMIDES, A. A. A Review of Item Exposure Control Strategies for Computerized Adaptive Testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment*, v. 5, n. 8, p. 1-38, 2007.

GONTIJO, C. M. M. Avaliação da alfabetização: Provinha Brasil. *Educação e Pesquisa*, v. 38, n. 3, p. 603-622, 2012.

LINDEN, W. J.; GLAS, C. A. W. *Elements of Adaptive Testing*. New York: Springer, 2010.

LORD, F. *Applications of Item Response Theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates, 1980.

MAGIS, D.; RAÎCHE, G. Random generation of response patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, v. 48, n. 8, p. 1-31, 2012.

MORAIS, A. G.; LEAL, T. F.; ALBUQUERQUE, E. B. C. “Provinha Brasil”: monitoramento da aprendizagem e formulação de políticas educacionais. *Revista Brasileira de Política e Administração da Educação*, v. 25, n. 3, p. 301-320, maio/ago. 2009.

PARTCHEV, I. *irtoys*: A Collection of Functions Related to Item Response Theory (IRT). S.l.: The Comprehensive R Archive Network, 2016. Disponível em: https://cran.r-project.org/package=irtoys. Acesso em: 19 mar. 2020.

REIF, M. *PP*: Estimation of person parameters for the 1,2,3,4-PL model and the GPCM. S.l: GitHub, 6 abr. 2019. Disponível em: https://github.com/manuelreif/PP. Acesso em: 19 mar. 2020.

REVUELTA J.; PONSODA, V. A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, v. 35, n. 4, p. 311-327, 1998.

SOARES, M. *Alfabetização*: a questão dos métodos. São Paulo: Contexto, 2016.

WARM, T. A. Weighted likelihood estimation of ability in item response theory. *Psychometrika*, v. 54, n. 3, p. 427-450, Sept. 1989.

WEISS, D. J.; KINGSBURY, G. G. Application of Computerized Adaptive Testing to educational problems. *Journal of Educational Measurement*, v. 21, n. 4, p. 361-375, Winter 1984.

NOTA: Este artigo foi produzido colaborativamente pelos autores. Rodrigo Travitzki foi responsável pela criação do algoritmo, Ocimar Munhoz Alavarse executou a coordenação geral do projeto, Douglas De Rizzo Meneghetti criou o *software* de aplicação da prova eletrônica e Érica Maria de Toledo Catalani realizou a análise psicométrica dos resultados.

Recebido em: 19 MARÇO 2020
Aprovado para publicação em: 16 SETEMBRO 2020