É POSSÍVEL REDUZIR O NÚMERO DE QUESTÕES DO ENEM POR MEIO DE UMA TESTAGEM ADAPTATIVA COMPUTADORIZADA?

Alexandre Jaloto[1]

Introdução

O Exame Nacional do Ensino Médio (Enem) surgiu em 1998 como um modelo de avaliação que tinha como referência principal a articulação entre a educação básica e a cidadania (INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, 2009). Desde 2009 a nota do participante é calculada segundo a Teoria de Resposta ao Item (TRI), por meio de quatro instrumentos de 45 itens objetivos — cada um para uma área do conhecimento.

Em testes como o Enem, que são aplicados a uma grande população, é necessário que um participante responda itens que influenciam pouco ou quase nada para a construção de uma medida precisa de seu traço latente, pois existe um grande intervalo de valores de traço latente a ser medido. Assim, para medir de maneira relativamente precisa todas as pessoas, é necessário que o teste contenha muitos itens para contemplar esse espectro — itens fáceis, médios e difíceis.

Uma possibilidade de solucionar essa limitação é o desenvolvimento da Testagem Adaptativa Computadorizada — TAC (WAINER, 2000). Na TAC, o programa busca apresentar ao participante um teste com a menor quantidade de itens possível que produza uma medida com alta precisão. Se o participante acerta o item apresentado, o programa lhe apresenta um mais difícil; se ele erra, lhe é apresentado um mais fácil. A cada item apresentado e respondido, o programa calcula novamente a nota do participante. O processo é concluído quando o critério de parada é atingido. Por exemplo, quando o erro da medida atinge um valor considerado adequado, o teste cessa.

Objetivo

O presente trabalho tem como objetivo verificar a possibilidade de redução do tamanho da prova de Ciências da Natureza e suas Tecnologias (CN) para estimar a proficiência dos participantes da edição de 2015 do ENEM por meio de uma TAC.

Material e Método

As análises foram realizadas a partir dos microdados divulgados pelo INEP referentes à edição de 2015 do ENEM — última edição disponível à época do desenvolvimento da

[1] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP, alexandre.jaloto@inep.gov.br

pesquisa. Foram escolhidos os cadernos de CN aplicados em 24 de outubro que não passaram por adaptações ou traduções — cadernos 235, 236, 237 e 238. Primeiramente, os 45 itens de CN foram calibrados. Para tal, foi sorteada uma amostra de 10.000 participantes estratificada de acordo com o percentil da soma dos acertos — foi utilizado como semente o valor 1000. Os estratos foram definidos pelos pontos dos percentis 20 e 90. Para os estratos inferior e superior foram sorteados 2.500 participantes e para o estrato intermediário, 5.000. A escolha da proporção se fundamenta na necessidade de haver pessoas em todas as faixas da escala. Além disso, é importante que haja muitas respostas de pessoas posicionadas na parte superior da escala para uma boa calibração dos itens de maior dificuldade, pois estes são pouco acertados.

Todo o estudo foi realizado no ambiente R (R CORE TEAM, 2018). A calibração foi realizada com o pacote *mirt* (CHALMERS, 2012), por meio da função *mirt()*.O modelo utilizado foi o logístico de três parâmetros. As distribuições dos parâmetros foram indicadas por meio da função *mirt.model()*: discriminação com distribuição log-normal de média 0 e desvio 0,5; acerto casual com distribuição normal de média -1,386294 e desvio 0,5. Em seguida, as notas dos participantes sorteados foram estimadas — doravante essas medidas serão chamadas de *proficiências originais*.

A simulação da TAC foi construída com o pacote *mirtCAT* (CHALMERS, 2016). O critério de apresentação do item foi o de Máxima Informação, que considera a informação que o item traz. O critério de parada foi o erro da medida — caso o valor fosse menor ou igual a 0,5, a aplicação era encerrada. O vetor de resposta de cada participante foi utilizado para produzir as informações necessárias para a estimação da proficiência em cada rodada de resposta. O trabalho não contempla estudos de verificação de evidência de validade do instrumento.

Resultados e Discussão

A Figura 1 mostra a dispersão das proficiências simuladas em função das proficiências originais e a Figura 2 traz a quantidade de itens apresentados na simulação em função da proficiência estimada na simulação. É possível observar na Figura 1 que as proficiências originais com valores abaixo de -0,6 possuem o mesmo valor estimado na simulação. Isso ocorreu por conta do alto erro da medida nessa região da escala: mesmo apresentando os 45 itens, o erro permanece superior a 0,5. Assim, para os participantes com notas mais baixas, a TAC envolveu a apresentação de 45 itens (Figura 2) — não havendo redução do instrumento. Já para os participantes com notas simuladas entre 0,91 e 2,42, o máximo de itens apresentado foi 26 e a média, 7,2.

Nota-se também a existência de muitas proficiências com valores idênticos nas regiões mais altas da escala (Figura 1). Isso tem relação com a baixa quantidade de itens disponíveis para a aplicação da simulação. Uma vez que o banco da simulação dispõe de 45 itens, em muitos casos foram apresentados os mesmos itens a participantes distintos, que apresentavam o mesmo padrão de resposta para esses itens. Isso ocorreu principalmente quando a simulação cessou com menos de dez itens; por exemplo, todas as proficiências estimadas com seis ou sete itens tiveram frequência maior do que 20.

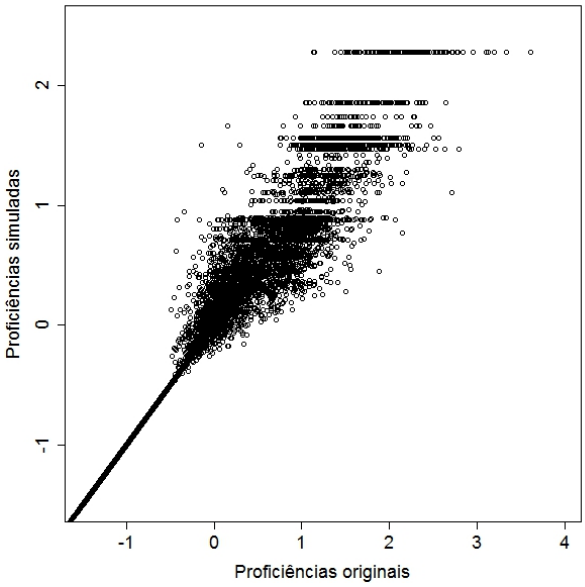

Figura 1 — Dispersão das proficiências simuladas em função das proficiências originais.

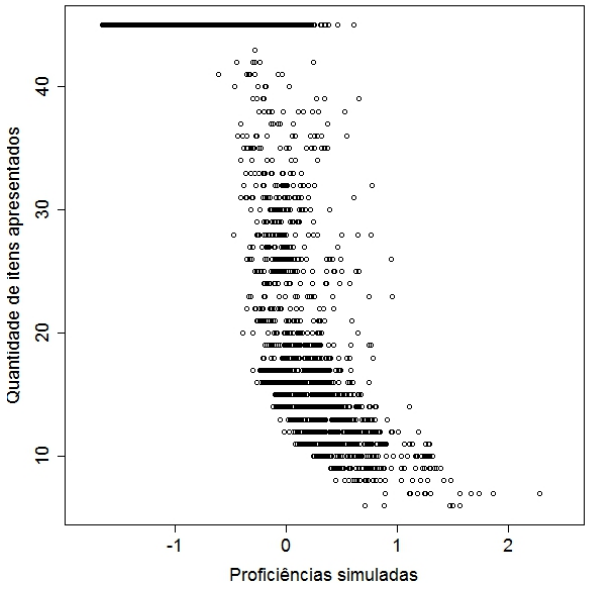

Figura 2 — Quantidade de itens apresentados na TAC em função da proficiência estimada.

Conclusão

O estudo aponta para a potencialidade do uso da TAC para a redução da quantidade de itens do ENEM, sem perder precisão na estimação, para medidas de traço latente de participantes localizados em regiões mais altas da escala, ou seja, no espectro em que se observa uma maior informação — e, consequentemente, um menor erro de medida.

O estudo indica também a necessidade da existência de um banco de itens suficientemente grande a ponto de evitar a exposição dos mesmos itens demasiadamente. Desse modo, estudos futuros devem contar com simulações que possibilitem a apresentação de itens que contemplem os diversos pontos da escala de proficiência. Além disso, novos critérios de parada devem ser testados, como a redução do valor do erro.

Referências

CHALMERS, R. P. mirt: a multidimensional item response theory package for the R environment. **Journal of Statistical Software**, v. 48, n. 6, p. 1–29, 2012. Disponível em:

<https://www.jstatsoft.org/index.php/jss/article/view/v048i06/v48i06.pdf>. Acesso em: 26 set. 2016.

_____. Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. **Journal of Statistical Software**, v. 71, n. 5, p. 1–39, 2016. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v071i05/v71i05.pdf>. Acesso em: 9 maio. 2018.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Exame Nacional do Ensino Médio (ENEM):** textos teóricos e metodológicos. Brasília: MEC/INEP, 2009.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2018. Disponível em: <https://www.R-project.org/>. Acesso em: 9 maio. 2018.

WAINER, H. Introduction and history. In: WAINER, H. (Org.). **Computerized Adaptive Testing:** a primer. 2. ed. Mahwah: Lawrence Erlbaum Associates, 2000. p. 1–22.