A formatação da prova afeta o desempenho dos estudantes? Evidências do Enem (2016)[1]

Leonardo Barichello[2]
ORCID: 0000-0001-9372-454X
Rita Santos Guimarães[3]
ORCID: 0000-0002-6324-7436
Dalson Britto Figueiredo Filho[4]
ORCID: 0000-0001-6982-2262

Resumo

Este artigo analisa o impacto da posição em que as questões são apresentadas sobre o desempenho dos estudantes no Exame Nacional do Ensino Médio (Enem) no Brasil em 2016. A partir de uma amostra de 4.427.790 casos, calculamos o índice de acerto por questão para os diferentes cadernos de prova da área de Matemática e suas Tecnologias. Os resultados indicam a presença do efeito fadiga na prova do Enem 2016, ou seja, a ordem de apresentação das questões afeta a proporção de respostas corretas, que diminui à medida que o item é apresentado mais próximo do final da prova. As evidências exploratórias também sugerem que o efeito fadiga se manifesta tanto em estudantes de baixo quanto de alto desempenho. Por exemplo, a posição do item reduziu o índice de acerto em até 18 pontos percentuais, controlando pelo nível de desempenho. Este artigo faz a primeira avaliação empírica do efeito fadiga no Enem e os resultados representam uma contribuição para a literatura sobre influências não cognitivas em avaliação e são úteis para fundamentar estudos mais sistemáticos sobre o impacto do efeito fadiga em testes padronizados de larga escala, inclusive para além do caso específico analisado. Ao final, sugerimos medidas que podem mitigar esse efeito no Enem.

Palavras-chave

Efeito fadiga – Testes padronizados – Enem – Microdados – Desempenho educacional.

1- O presente trabalho foi realizado com apoio parcial da Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp) – processo número 2019/17135-2. Agradecemos também ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) a partir do Programa de Excelência Acadêmica (Proex).

2- Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Jundiaí, SP, Brasil. Contato: leonardo.barichello@ifsp.edu.br
3- Universidade Estadual de Campinas, Campinas, SP, Brasil. Contato: guimaraes.rita@gmail.com.
4- Universidade Federal de Pernambuco, Recife, PE, Brasil. Contato: dalson.figueiredofo@ufpe.br.

https://doi.org/10.1590/S1678-4634202248241713por

*Does the structure of the test affect the performance of students? Evidences from the Enem (2016)**

Abstract

This paper analyzes the impact of the position of questions on students' performance on the National Secondary Education Examination (Enem) in 2016. From a sample of 4,427,790 cases, we calculated the hit rate per question for the different workbooks in the Mathematics and its Technologies test. The results indicate presence of the fatigue effect on the 2016 Enem, that is, the order in which the questions are presented affects the proportion of correct answers, which is diminished as an item is presented closer to the end of the test. The exploratory evidence also suggests that the fatigue effect is manifested in students of both low and high performance. For example, the position of an item reduced the hit rate up to 18%, controlling for performance level. This paper conducts the first empirical evaluation of the fatigue effect during the Enem. The results contribute to the literature on the non-cognitive influences in evaluation, being useful to substantiate more systematic studies on the fatigue effect's impact on large-scale standardized tests, beyond the case analyzed. At the end, we suggest measures that can mitigate this effect during the Enem.

Keywords

Fatigue effect – Standardized tests – Enem – Microdata – Educational performance.

Introdução

O gerenciamento adequado do tempo é frequentemente apontado como um fator importante para explicar o desempenho em testes padronizados de larga escala (RODRIGUES, 2007; WOYCIEKOSKI; HUTZ, 2009). Em particular, tanto a ansiedade quanto a ausência de estratégias específicas de resolução de questões são elementos que podem afetar negativamente a performance geral do candidato (GONZAGA; ENUMO, 2018). No Brasil, por exemplo, durante os dias que antecedem o Exame Nacional do Ensino Médio (Enem), a efetiva administração do tempo é tema recorrente em sítios de notícias, blogs e até mesmo na seção de perguntas e respostas do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)[5].

No primeiro dia do Enem, os candidatos devem responder, em quatro horas e trinta minutos, noventa questões de múltipla escolha, sendo metade sobre Ciências Humanas e suas Tecnologias e metade sobre Ciências da Natureza e suas Tecnologias. Isso significa que, em média, cada questão deve ser resolvida em três minutos. No segundo dia, além das

5 - Ver: https://agenciabrasil.ebc.com.br/educacao/noticia/2016-11/enem-administrar-bem-o-tempo-e-fundamental-na-hora-da-prova e http://portal.inep.gov.br/enem/perguntas-frequentes.

noventa questões, desta vez sobre Linguagens, Códigos e suas Tecnologias e Matemática e suas Tecnologias, há uma redação e um acréscimo de meia hora no tempo máximo de duração da prova. As estratégias de não investir tanto tempo em questões que talvez estejam além do alcance do candidato e de não gastar tempo demais no começo da prova, perdendo a chance de acertar alguma questão fácil eventualmente colocada no final, são importantes e podem ser melhoradas com orientação e prática.

Sassaki *et al*. (2018) analisaram a influência de fatores não cognitivos (como cansaço e posição das questões) no desempenho dos estudantes no Programa Internacional de Avaliação de Estudantes, o Pisa[6] (sigla em inglês). Os autores examinaram a diferença de desempenho em uma mesma questão que, por motivo de aleatorização do teste, foi apresentada aos estudantes em momentos diferentes da prova. Especificamente, os alunos resolveram quatro blocos de questões (cada um com duração estimada de 30 minutos), e compostos pelo mesmo conjunto de questões, porém com ordem definida ao acaso. Sendo assim, um estudante pode iniciar o teste com uma questão, mas outro pode receber essa mesma questão ao final do bloco. Os resultados indicam que o índice de acerto em uma questão cai à medida que o item é apresentado mais tardiamente. E, em termos comparativos, os estudantes brasileiros são mais suscetíveis a esse efeito.

O primeiro resultado obtido por Sassaki *et al*. (2018) é reportado na literatura psicométrica internacional como efeito fadiga (*fatigue effect* ou *test fatigue*, em inglês) e já foi identificado em diferentes avaliações que aleatorizam (pelo menos parcialmente) a ordem de apresentação das questões para os candidatos (ALBANO, 2013; BORGHANS; SCHILS, 2012; DAVIS; FERDOUS, 2005; MEYERS; MILLER; WAY, 2008)[7]. Este artigo procura contribuir com esse debate a partir de uma análise exploratória do efeito fadiga no maior teste padronizado de larga escala no Brasil. Com uma amostra de 4.427.790 casos, examinamos a variação do percentual de acertos por questão e por tipo de prova com foco na área de Matemática e suas Tecnologias do Enem de 2016. A escolha desse exame se mostra especialmente relevante, já que os estudantes brasileiros foram os mais suscetíveis ao efeito fadiga dentre todos os participantes na edição 2015 do Pisa, conforme reportado por Sassaki *et al*. (2018).

Esquematicamente, o restante do artigo está organizado em quatro partes. Na próxima seção, discutimos estudos que investigaram se e como a fadiga pode afetar o desempenho de estudantes em testes padronizados. Depois disso, apresentamos as principais características do desenho desta pesquisa com o objetivo de aumentar a transparência e garantir a replicabilidade dos resultados. Em particular, explicamos o processo de coleta e tratamento dos dados e descrevemos como se dá o ordenamento das questões na prova do Enem. Em seguida, apresentamos a análise exploratória e, na última parte, as conclusões.

6- Programa Internacional de Avaliação do Estudante, desenvolvido pela Organização para a Cooperação e Desenvolvimento Econômico (OCED) desde 1999 e aplicado a estudantes de cerca de 15 anos para fins de comparação entre sistemas do sistema educacional de mais de setenta países. Mais informações em: http://www.oecd.org/pisa/.

7- Efeito análogo pode ser observado na pesquisa de *survey*, em que a fadiga de painel representa uma das principais ameaças à qualidade das informações em levantamentos longitudinais. Em casos extremos, como bem aponta Lavrakas (2008), a fadiga de painel pode produzir atrito na amostra e elevar a proporção de não resposta. Por sua vez, a fadiga do respondente é um fenômeno bem documentado na literatura e ocorre quando o entrevistado cansa da entrevista. Um dos efeitos do cansaço é a emissão de respostas inconsistentes e até mesmo aleatórias (HERZOG; BACHMAN, 1981).

Efeito fadiga e desempenho em testes padronizados

Desde a década de 1980, houve um crescimento na utilização de avaliações de larga escala no Brasil (ALAVARSE, 2013). De maneira geral, essas avaliações têm caráter exclusivamente somativo, ou seja, são utilizadas para fins de seleção e certificação de estudantes ou *accountability* de sistemas educacionais (NEVO, 2011). São exemplos dessas avaliações no contexto brasileiro o Enem, a Prova Brasil, o Sistema Nacional de Avaliação da Educação Básica (Saeb) e o Exame Nacional de Desempenho dos Estudantes (Enade – avalia o rendimento dos concluintes dos cursos de graduação).

O número de participantes nessas avaliações exclui diversas metodologias que poderiam ser utilizadas nesse processo e força outras características pela sua dificuldade logística de implementação, incluindo elaboração, aplicação e correção (BAUER; ALAVARSE; OLIVEIRA, 2015). Por conta disso, é comum que avaliações de larga escala sejam compostas por questões objetivas, pautadas por uma lista de habilidades ou conteúdos bem delimitada e aplicadas em contextos muito controlados.

Mesmo com tais restrições, essas avaliações têm como objetivo aferir o conhecimento dos estudantes, ou seja, espera-se que o resultado de cada estudante seja um reflexo de suas capacidades cognitivas. Porém, mesmo os defensores dos métodos empregados em avaliações de larga escala reconhecem que fatores não cognitivos interferem na medição obtida por esse tipo de instrumento. Por exemplo, Borghans, Meijers e Ter Weel (2008) mostram que características de personalidade, como motivação intrínseca e baixa aversão a risco, influenciam o desempenho quando variáveis como restrição de tempo e recompensa são modificadas.

Davis e Ferdous (2005) testaram se havia diferença na performance de estudantes americanos de 10 e 12 anos em questões conforme a posição em que elas apareciam em uma avaliação. Segundo os autores, a hipótese implicitamente adotada por formuladores de avaliação desse tipo é de que a posição da questão não influencia a dificuldade geral do teste. Porém, ao determinarem a dificuldade das questões usando a Teoria da Resposta ao Item (TRI) em diferentes cadernos de uma avaliação (compostos pelas mesmas questões, mas em ordens diferentes), os autores encontraram diferenças significativas, de modo que quanto mais próxima do início da avaliação uma questão é apresentada, menor é a sua dificuldade.

Meyers, Miller e Way (2008) e Albano (2013) identificaram o mesmo efeito em outros contextos. Entretanto, os autores foram além da mensuração do impacto e propuseram modelos que incorporam esse efeito ao cálculo do desempenho seguindo a abordagem da TRI, ou seja, propuseram modelos que consideram a posição em que uma questão foi apresentada aos participantes como um de seus parâmetros.

No relatório técnico do Pisa realizado em 2000, Adams e Wu (2002) identificaram uma diferença entre o nível de dificuldade registrado para uma mesma questão quando esta é apresentada em momentos diferentes da avaliação. A diferença é significativa ao ponto de um fator de correção ter sido aplicado para obtenção da nota final de cada país. Embora os autores não tenham explicitamente relacionado essa diferença ao efeito fadiga, o fenômeno parece estar relacionado a ele. Mais recentemente, Borghans e Schils (2012),

ao analisarem dados de edições seguintes do Pisa, reportaram o mesmo efeito apontado por Davis e Ferdous (2005): o desempenho de estudantes cai à medida que uma mesma questão é apresentada mais tardiamente em uma avaliação, independentemente do nível de dificuldade do item.

Em trabalho semelhante, Marchioni (2017) investigou o efeito fadiga no Pisa e os resultados indicam que a fadiga afeta mais estudantes sul-americanos do que discentes de outros países. É interessante complementar essa observação com um dado já presente no relatório técnico do Pisa (2000). Comparativamente, os autores concluem que a variação da dificuldade de uma questão é "bastante estável" (ADAMS; WU, 2002, p. 157) quando se considera os vários países participantes. Entretanto, o Brasil aparece em primeiro lugar em termos do número médio de questões que os estudantes sequer chegaram a tentar responder.

Sassaki *et al*. (2018) examinaram dados da edição 2015 do Pisa com foco no efeito fadiga. O desenho de pesquisa se beneficiou de informações detalhadas graças à aplicação digital da avaliação para um subconjunto dos estudantes. A conclusão foi de que os brasileiros de fato são mais afetados pelo efeito fadiga do que estudantes de outros países, tanto de alto desempenho quanto de países com desempenho comparável ao nosso.

A interferência desse tipo de variável levanta questões importantes, especialmente em um momento em que avaliações de larga escala têm sido usadas em diversos níveis e com propósitos de grande relevância social, como a seleção de estudantes para o acesso ao ensino superior no caso do Enem, foco deste trabalho.

Estudos sobre o Enem

O Enem foi criado em 1998, com o objetivo de avaliar a qualidade da aprendizagem no Ensino Médio. Em 1999, passou a ser usado como critério para ingresso no Ensino Superior e, em 2008, foi substancialmente reformulado, passando a ser fundamentado por preceitos da Teoria da Resposta ao Item (TRI), para servir como principal mecanismo de acesso a diversas universidades, incluindo a maior parte das instituições federais do país (TRAVITZKI, 2017). O Enem é realizado anualmente, e o conteúdo da prova é dividido em dois dias diferentes. Atualmente, o resultado pode ser utilizado também para ingresso em algumas universidades internacionais, para obtenção de bolsas de estudo em universidades particulares e como critério para certificação de conclusão do Ensino Médio[8].

Desde 1998, os dados gerados a partir da aplicação da prova, contendo não apenas as respostas, mas também informações detalhadas dos participantes como indicadores socioeconômicos, geográficos, trajetória escolar, entre outros, são disponibilizados publicamente no formato de microdados no sítio eletrônico do Inep.

Analiticamente, a natureza desagregada das informações oferece uma oportunidade singular para investigar fenômenos de larga escala relacionados ao final da Educação Básica no Brasil. Travitzki (2017), por exemplo, usou os microdados das edições de 2009 e 2011 para fazer uma meta-avaliação da confiabilidade das provas, empregando as mesmas técnicas que fundamentam a elaboração do Enem. O autor concluiu que a

8- Informação disponível em: https://enem.inep.gov.br.

prova de matemática de 2009 apresentou confiabilidade insuficiente e diversas questões demonstraram comportamento empírico fora do esperado nas duas edições.

Todavia, o efetivo uso dos microdados na pesquisa aplicada, seja para meta-avaliação do exame, seja para a investigação de outros fenômenos, ainda é reduzido. Uma busca realizada em janeiro de 2020, no portal SciELO, a partir dos termos "enem" e "microdados", retorna apenas quatro artigos e nenhum deles aborda temática similar à que propusemos explorar neste texto.

Lima *et al.* (2019) apresentam uma revisão sistemática da literatura sobre artigos que utilizam os dados fornecidos pelo Inep acerca do Enem e Enade. Com auxílio do Google Scholar (https://scholar.google.com), de abrangência mais ampla que o Scielo, Lima *et al.* (2019) chegaram a um conjunto de 54 trabalhos, publicados entre 2005 e 2016, sendo que 17 destes se referiam ao Enem. Essas produções foram agrupadas em quatro categorias de acordo com a natureza dos seus objetivos: conteúdo/conhecimento, administrativo, desempenho/rendimento e teste/desenvolvimento de ferramentas.

Os artigos do primeiro grupo investigaram aspectos relacionados ao conteúdo das questões do Enem. Os do segundo focaram em aspectos ligados a gestão e acesso. Já os trabalhos do terceiro investigaram o desempenho de grupos específicos ou a relação do desempenho com outras variáveis. Por fim, os artigos do quarto enfatizaram o desenvolvimento de ferramentas ou metodologias que facilitem a interpretação e uso das informações disponíveis nos microdados.

Apesar de os artigos do terceiro grupo terem examinado aspectos relacionados ao desempenho dos estudantes, o foco era analisar e comparar o desempenho de estudantes de acordo com alguma variável geográfica ou socioeconômica. Portanto, nenhum dos artigos analisados por Lima *et al.* (2019) aborda tema similar ao deste texto.

O trabalho apresentado por Toffoli (2019), publicado posteriormente à revisão de Lima *et al.* (2019), se aproxima do nosso em termos do potencial de contribuir para a melhoria do Enem como um exame de seleção. Segundo a autora,

> Estudos sobre as avaliações em larga escala são importantes para identificar etapas do processo que não estão funcionando como esperado e também para validar as etapas cujos resultados são adequados. Em ambos os casos, o objetivo dos estudos deve ser a melhoria dos processos a cada edição do exame. (TOFFOLI, 2019, p. 4).

Toffoli (2019) analisa as qualidades psicométricas de cada questão e da prova como um todo, e suas conclusões são bastante preocupantes, dado que os resultados do Enem são mais do que simples medidas e possuem consequências sociais muito relevantes (TOFFOLI *et al.*, 2016). Seus resultados podem ser sintetizados na seguinte constatação:

> Na prova de matemática do Enem 2015, o item mais fácil obteve o parâmetro da dificuldade em b = –1,98, e aproximadamente 23,4% dos participantes possuem habilidades menores do que esse valor, indicando que esses indivíduos não souberam responder nenhum entre os 45 itens, só acertando itens ao acaso, ou popularmente falando, no chute. (TOFFOLI, 2019, p. 21).

No seu estudo, Toffoli (2019) considerou apenas as respostas em um dos cadernos de prova do Enem, ou seja, seria impossível notar o efeito fadiga na análise de seus dados.

Travitzki (2017) chega um pouco mais perto do nosso objeto de estudo ao discutir seus resultados. O autor aponta a presença de diversas questões nas edições de 2009 e 2011 do Enem que seriam consideradas inadequadas de acordo com critérios da Teoria Clássica dos Testes e observa que boa parte delas se concentrou na parte final da prova. O autor oferece duas explicações para esse fenômeno:

> Uma possível explicação é que os itens anteriores demandassem muito trabalho, levando os candidatos a uma maior exaustão mental (ou menor tempo disponível) ao final da prova. Outra possível explicação, não excludente, é que boa parte dos candidatos tenha deixado de fazer as questões por considerá-las excessivamente difíceis, sendo assim mais produtivo investir o tempo de prova nos outros itens. (TRAVITZKI, 2017, p. 281).

Note que o autor não trata do efeito fadiga propriamente dito, mas sugere que o cansaço pode afetar o desempenho dos participantes nas questões que são apresentadas no final da prova, a ponto de serem consideradas de qualidade duvidosa em termos psicométricos.

As ressalvas em relação ao Enem colocadas por Toffoli (2019), aliada à grande relevância social da avaliação (TOFFOLI *et al*., 2016), indicam a importância de que o exame seja colocado sob escrutínio. Embora não tenha sido o foco principal de sua análise, as explicações sugeridas por Travitzki (2017) já destacam a relevância de considerar a posição das questões na prova do Enem ao analisar a qualidade psicométrica da prova. Além disso, a conclusão de Sassaki *et al*. (2018), de que os estudantes brasileiros são mais suscetíveis à variação da posição das questões do que estudantes de outros países, embora tenha sido observada na prova do Pisa, reforçam ainda mais a relevância de analisar o efeito fadiga em um exame como o Enem. Por fim, os escassos estudos identificados nesta revisão salientam a originalidade do foco deste trabalho.

Metodologia

A nossa fonte exclusiva de informação foram os microdados do Enem 2016, mais especificamente os dados referentes à prova de Matemática e suas Tecnologias[9]. A escolha de limitar o ano de aplicação e o conteúdo se justifica por dois motivos: as dificuldades computacionais presentes na manipulação de enormes bases de dados e a natureza exploratória do trabalho.

No que se refere à dificuldade computacional, a extração inicial retornou uma planilha com 8.627.368 linhas, o que inviabiliza não apenas o uso de softwares populares

9- O conjunto de dados utilizado para realização da análise apresentada neste texto não está mais disponível da forma como utilizamos devido à mudança de política de compartilhamento de dados do Inep. Entretanto, os arquivos que geramos para realização da análise dos dados estão disponíveis integralmente no portal Open Science Foundation e podem ser acessados em: https://osf.io/ev39z/?view_only=27d3b73665a04f079e98da507d0ac67b.

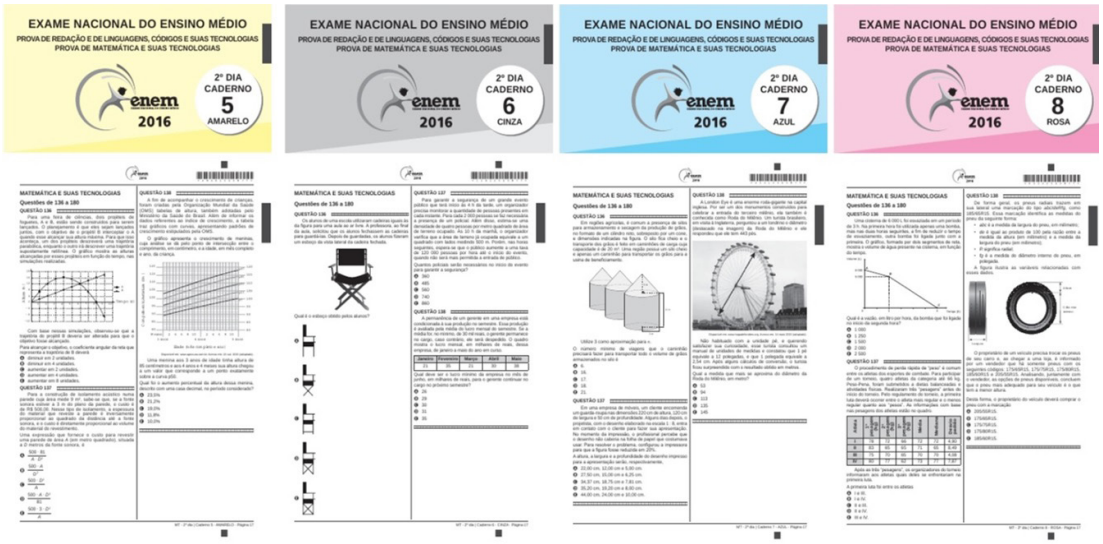
como o Microsoft Excel[10], mas também apresenta desafios em softwares de análise estatística para fins de pesquisa em computadores com configurações de mercado (8 GB de memória RAM).

No que se refere ao aspecto exploratório do estudo, a opção por Matemática e suas Tecnologias se justifica pela afinidade dos dois primeiros autores com a área, tanto em relação à prática pedagógica quanto no que diz respeito à agenda de pesquisa específica no tema de educação matemática. Por fim, o ano de 2016 foi escolhido por ser o conjunto de dados mais atual disponível quando as primeiras ideias deste artigo foram discutidas pelos autores.

A ordem das questões nos cadernos do Enem

Toda aplicação do Enem é composta por quatro cadernos de prova (visualmente diferenciados pela cor) com as mesmas questões, mas apresentadas em ordem diferente. A Figura 1 mostra a primeira página da prova de Matemática e suas Tecnologias para cada um dos quatro cadernos da aplicação principal de 2016.

Figura 1 – Cadernos de prova por cor



Fonte: Inep, 2016. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/provas-e-gabaritos. Acesso em: 14 abr. 2022.

Operacionalmente, o principal objetivo da diversificação de cores é dificultar cópia ou troca de cartões de respostas, uma vez que estudantes sentados próximos devem receber cadernos de cores diferentes. Todavia, não encontramos justificativas nos documentos

10 - De acordo com a Microsoft, a capacidade total do Excel é de 1.048.576 linhas e 16.384 colunas, ver: https://support.microsoft.com/en-us/office/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3.

oficiais sobre como os cadernos são montados. Detectamos ainda que o conteúdo de cada página é fixo, mas a ordem das páginas dentro de uma determinada área do conhecimento é alterada de um caderno para o outro. Isso faz com que certas questões estejam sempre próximas, independentemente da cor do caderno, mas uma dada página pode figurar em diferentes posições, configuração similar ao que ocorre no Pisa, que foi utilizado por Sassaki *et al*. (2018) para identificar o efeito fadiga.

Conforme discutido anteriormente, essa variação pode afetar a chance de responder corretamente à questão, o que indicaria a interferência de um fator não cognitivo no desempenho dos estudantes. Esse é o fenômeno que vamos investigar, em caráter exploratório, neste estudo.

Os microdados do Enem 2016

O formato de microdados adotado pelo Inep é analiticamente versátil, mas apresenta duas limitações relevantes diante do que nos propusemos a investigar. A primeira delas, já mencionada, é o tamanho dos arquivos. A segunda se refere ao formato das respostas de cada participante, uma vez que elas não são dadas como campos independentes, mas como uma sequência de caracteres indicando qual foi a alternativa assinalada pelo estudante seguida por uma cadeia de caracteres com o gabarito das respectivas questões. Tecnicamente, esse formato exige um pré-processamento dos dados antes que seja possível de fato aferir informações acerca das respostas dadas pelos participantes às questões que compuseram as provas de cada área de conhecimento contemplada no Enem.

Por conta disso, antes de importarmos o conjunto de dados em um software de análise estatística, fizemos um pré-processamento por meio de um script desenvolvido em linguagem C. O primeiro passo foi selecionar apenas estudantes que fossem de interesse para a pesquisa. Nossa escolha inclui apenas participantes com presença regular nos dois dias da prova, que não estavam prestando como treineiros, com situação regular no Ensino Médio e que fizeram um dos quatro cadernos de prova usados na primeira aplicação[11]. Com esses critérios, ficamos com uma amostra de 4.427.790 participantes.

Depois disso, alteramos o formato em que as respostas às questões são armazenadas nos microdados. As cadeias de caracteres com as respostas a cada questão da área de Matemática e suas Tecnologias foram convertidas em 45 variáveis binárias, uma para cada questão, indicando se o estudante a acertou (1) ou errou (0). Além disso, separamos os participantes por cadernos de prova (cores) incluindo também informações socioeconômicas e demográficas. A decisão de incluir esses campos veio da possibilidade de considerá-los em análises subsequentes e, como nossa base de dados será disponibilizada publicamente, outros pesquisadores poderão se beneficiar de um conjunto mais amplo de informações.

Depois de devidamente tratados, os dados foram analisados com auxílio do software R, versão 3.4.4[12].

11- É possível que haja mais de uma aplicação do Enem, seja por motivos de segurança ou para contemplar grupos específicos. Entretanto, a primeira aplicação costuma abranger a imensa maioria dos participantes.

12- R é um software livre para tratamento estatístico de dados. Disponível em: https://cran.r-project.org/.

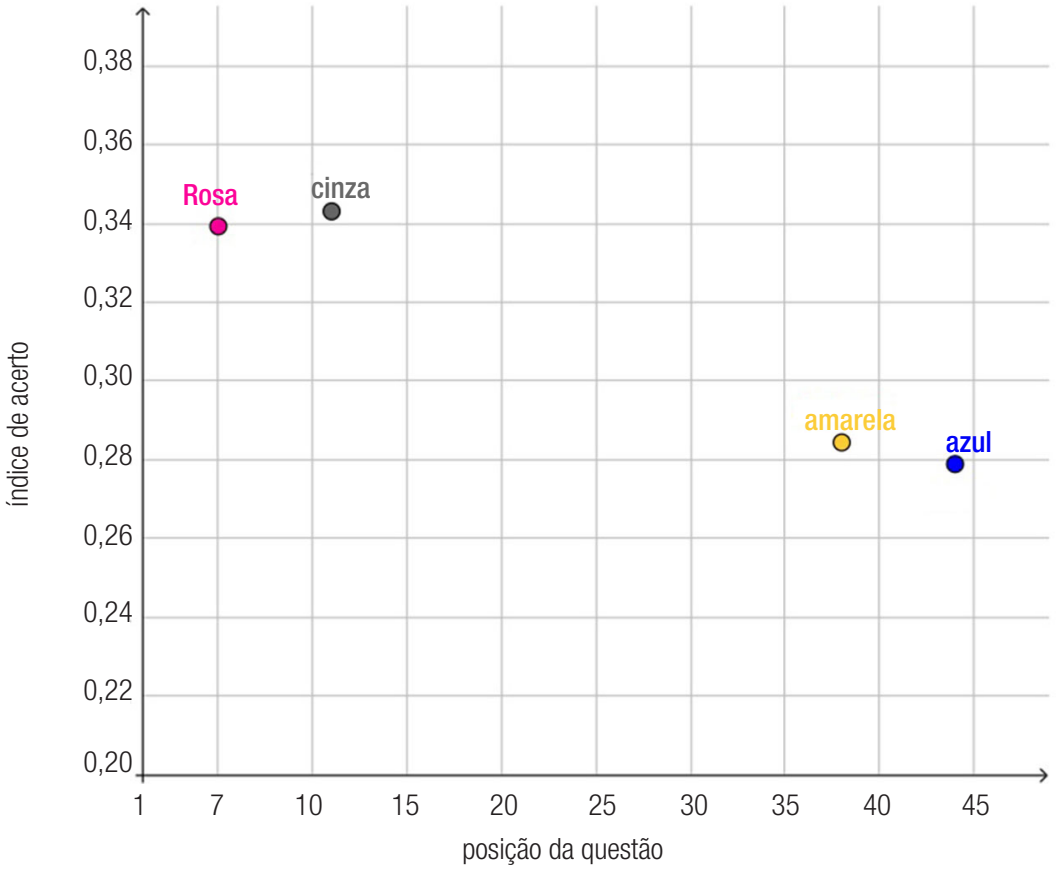
Análise

Efeito fadiga

A exploração dos dados teve início com a identificação da posição de cada uma das questões (discriminadas pelo código identificador numérico adotado nos microdados do Inep) nos quatro cadernos da aplicação principal do Enem 2016. Em seguida, foi calculada a taxa de acerto de cada questão em cada um dos cadernos separadamente.

O Gráfico 1 mostra a variação da taxa de acerto para a questão com código 88786 em cada um dos cadernos, identificados pelas suas cores.

Gráfico 1 – Porcentagem de acertos da questão 88786 em função da posição em que foi apresentada



Fonte: Elaboração própria, 2020.

O valor em que estamos interessados é o percentual de acerto por questão. Teoricamente, se não existe efeito fadiga, a frequência de respostas corretas deve ser igual, independentemente da posição da questão. Todavia, como pode ser observado, a questão 88786 apareceu no final da prova nos cadernos Azul e Amarelo com taxa média

de acerto de aproximadamente 28%. Quando a mesma questão é apresentada no início da prova (cadernos Rosa e Cinza), o percentual médio de acertos sobe para cerca de 34%. A diferença de seis pontos percentuais equivale a uma variação de 21,4% e é suficientemente alta para levantar dúvidas sobre a validade e a confiabilidade do teste como instrumento de avaliação de conhecimentos e habilidades de natureza cognitiva.

Para estimar em que medida esse fenômeno afeta outras questões, criamos um indicador especialmente desenhado para capturar essa tendência, que chamaremos de EF. O EF de uma questão é calculado a partir da diferença entre os índices de acerto quando maximizamos a distância das posições dessa questão. No exemplo apresentado no Gráfico 1, a maior distância é 39, que é a diferença absoluta entre 44 (caderno Azul) e 5 (caderno Rosa). Assim, basta calcular a diferença absoluta entre 33,9% (índice de acerto no caderno Rosa) e 27,9% (índice de acerto no caderno Azul) para encontrar um EF de 6,0%. A Tabela 1 apresenta a frequência de acertos geral, a maior distância e o EF de todas as questões.

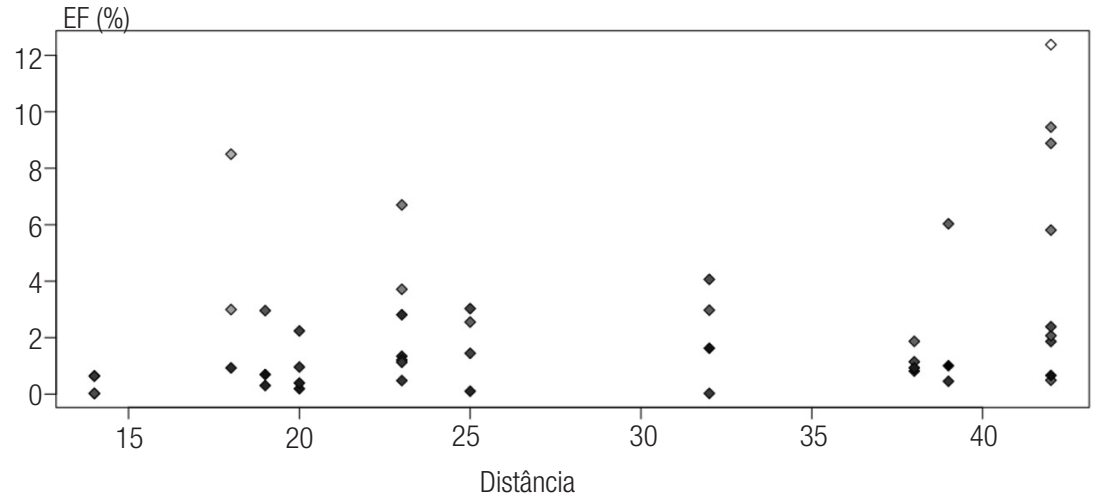
Tabela 1 – Frequência de acertos geral, maior distância e EF das questões de Matemática e suas Tecnologias no Enem 2016

ID questão	Acertos geral (%)	Maior distância	EF (%)	ID questão	Acertos geral (%)	Maior distância	EF (%)
38786	16,99	14	0,02	25285	31,04	25	2,55
87262	26,86	14	0,03	16644	25,15	25	3,03
42692	23,11	14	0,64	48223	23,47	32	0,02
8476	16,26	14	0,64	44243	15,51	32	1,62
60291	20,51	18	0,93	17264	30,35	32	2,97
42955	42,93	18	3,00	96833	27,32	32	4,06
45081	46,36	18	8,50	53278	16,05	38	0,82
95265	19,14	19	0,30	85018	17,34	38	0,93
39762	15,11	19	0,69	32686	25,73	38	1,14
25723	28,93	19	2,96	27005	31,81	38	1,87
30029	16,68	20	0,19	60315	21,36	39	0,45
53721	17,95	20	0,39	39198	11,57	39	1,00
83906	22,42	20	0,96	88786	31,05	39	6,03
85050	24,17	20	2,24	11472	30,03	42	0,49
10052	22,33	23	0,48	59795	14,71	42	0,66
83234	22,35	23	1,13	24747	25,17	42	1,86
32969	25,64	23	1,21	42706	29,03	42	2,07
29844	19,30	23	1,34	96774	25,75	42	2,39
83152	20,61	23	2,81	37515	35,16	42	5,81
53219	37,12	23	3,71	18364	36,01	42	8,88
32808	37,42	23	6,70	30865	37,87	42	9,46
40660	19,31	25	0,10	32221	62,53	42	12,38
83608	24,87	25	1,45				

Fonte: Elaboração própria, 2020.

As informações presentes na Tabela 1 também podem ser visualizadas no Gráfico 2.

Gráfico 2 – EF e distância para cada questão (n=45) da prova de Matemática e suas Tecnologias do Enem 2016



Fonte: Elaboração própria, 2020.

Cada ponto representa uma das 45 questões, a coordenada X representa a maior distância, a coordenada Y representa o EF e a cor representa o índice de acerto daquela questão (quanto mais claro, mais alto o índice de acerto). Primeiramente, vale a pena notar que o EF chega a 12% e esse caso ocorre em uma questão cuja distância chegou a 42 (ponto no canto superior direito do Gráfico 2). Além disso, temos sete questões com EF maior do que 5%.

A fim de explorar a forma como as duas variáveis, maior distância e EF, se relacionam, aplicamos uma regressão linear a esse conjunto de dados. O coeficiente padronizado da reta obtida é 0,326 (p-valor=0,029; n=45), ou seja, se a distância entre a mesma questão aumentar em 10 posições, devemos observar uma variação média de acerto de 3,26%[13]. O coeficiente de determinação (r^2), que é comumente interpretado como o total da variância da variável dependente explicada pelo conjunto das variáveis explicativas, foi de 0,106, o que significa dizer que a distância entre os itens do caderno de prova é responsável por cerca de 10% da variação do EF.

Para garantir resultados mais robustos, invertemos o raciocínio de modo a identificar o EF a partir da menor distância em que uma questão ocorreu entre os

13- Utilizamos um modelo linear de mínimos quadrados ordinários (MQO), tendo como variável dependente o EF e como variável independente a distância máxima das posições das questões nos respectivos cadernos de prova. Para garantir resultados mais robustos, estimamos um novo modelo controlando a relação entre EF e distância pela taxa geral de acerto das questões. O impacto da distância entre as questões continua positivo ($\beta_1 = 0,167$) com valor p de 0,051.

diferentes cadernos[14]. A expectativa é de que o EF calculado dessa maneira seja bastante baixo, já que estamos comparando questões praticamente na mesma posição nos diferentes cadernos de prova. De fato, as distâncias mínimas variaram entre 1 e 13, não obtivemos nenhum EF maior que 3%, e a grande maioria das questões (34) apresentaram EF abaixo de 1%.

Apesar de não ser um efeito homogêneo, existem questões que têm índice de acerto consideravelmente menor quando apresentadas no final do caderno de prova do que quando aparecem no início. Uma possível explicação para esse fenômeno, alinhada ao que Davis e Ferdous (2005) sugerem, é que o participante cansa ao longo do exame e, ao se deparar com uma questão que exigiria mais esforço cognitivo, acaba tendo um desempenho pior do que se a mesma questão estivesse no início da prova.

O efeito fadiga que encontramos com o auxílio do indicador EF na edição de 2016 do Enem nos parece equivalente ao fenômeno identificado por Sassaki *et al*. (2018) na edição de 2015 do Pisa. A abordagem adotada neste texto, que inovou ao incorporar a posição das questões nos diferentes cadernos de prova, complementa os resultados encontrados por Travitzki (2017) no que se refere à adequação das questões das edições de 2009 e 2011 do Enem. O autor identificou, após analisar apenas um dos cadernos de prova, que havia uma concentração maior de questões com comportamento anômalo no final da prova e sugeriu como explicação para o fenômeno o cansaço dos participantes e a extensão da prova. O comportamento do EF que observamos aqui oferece novos elementos que podem, inclusive, ajudar a compreender as qualidades psicométricas de questões do Enem dentro da abordagem da TRI.

Grupos diferentes sofrem fadiga em questões diferentes

No Gráfico 2, nota-se ainda que questões com baixa frequência de acerto geral têm, em sua esmagadora maioria, baixo EF, o que sugere alguma interação entre essas variáveis.

Das 34 questões com até 31% de acerto, 26 têm EF menor do que 2% e as 8 restantes têm EF de, no máximo, 4%. Ou seja, questões mais difíceis (que tiveram baixo índice de acerto) parecem causar EF mais baixo.

Essas observações, acompanhadas do fato de o Enem ter um público participante bastante heterogêneo, nos fizeram considerar como seria o EF para participantes com desempenhos diferentes.

Para isso, dividimos os participantes em grupos baseados no número de acertos na prova de Matemática e suas Tecnologias, não na sua nota final após o uso dos parâmetros da TRI empregada no Enem. Como as questões do Enem são de múltipla escolha com cinco alternativas, espera-se que um participante acerte, em média, nove questões, mesmo respondendo aleatoriamente. Assim, dividimos nossa amostra em cinco grupos a partir do número total de acertos, como apresentado na Tabela 2.

14- Por motivo de espaço, esses não foram inseridos no texto, mas podem ser facilmente obtidos a partir da tabela disponível em: https://osf.io/ev39z/?view_only=27d3b73665a04f079e98da507d0ac67b.

Tabela 2 – Distribuição dos participantes por grupo

Grupo	Número total de acertos	Participantes, n (%)
1	0 – 9	1.612.119 (36,41)
2	10 – 18	2.426.686 (54,80)
3	19 – 27	305.620 (6,90)
4	28 – 36	75.075 (1,70)
5	37 – 45	8.290 (0,19)

Fonte: Elaboração própria, 2020.

Investigamos como o EF afeta os participantes dos diferentes grupos. Por exemplo, será que um participante com muitos acertos (pertencente ao grupo 5), que deve ter um alto nível de conhecimento e uma boa preparação para fazer a prova, é pouco afetado pelo efeito fadiga? Analogamente, será que um participante que escolhe muitas das respostas aleatoriamente (pertencente ao grupo 1), que deve ter um baixo nível de conhecimento e pouca preparação para fazer a prova, é mais afetado pelo efeito fadiga?

Depois de separados os grupos, selecionamos as cinco questões com maior EF para cada um deles, como visto na Tabela 3. Esta tabela não contém 25 questões porque algumas delas ocorreram para mais de um grupo, como a 30865, que está entre as cinco questões com maior EF para os grupos 1, 2 e 3. Em destaque cinza estão as cinco questões que causaram maior efeito fadiga naquele grupo.

Esta tabela deve ser lida por linha. Por exemplo, a questão 96833 foi incluída por ter sido a questão com maior EF para o grupo 3 (17,94%). Entretanto, ela causou EF quase nulo para os grupos 1 e 5, mesmo tendo aparecido com distância 32.

Tabela 3 – As cinco questões com maior EF por grupo de desempenho

ID questão	Maior distância	EF Grupo 1 (%)	EF Grupo 2 (%)	EF Grupo 3 (%)	EF Grupo 4 (%)	EF Grupo 5 (%)
32808	23	3,57	7,40	9,60	4,88	2,49
30865	42	4,63	11,12	16,05	6,03	1,51
45081	18	6,00	8,43	4,67	1,87	0,72
18364	42	6,68	8,43	20,85	10,22	1,80
32221	42	11,98	11,36	5,18	1,57	0,01
88786	39	1,12	8,62	18,39	8,17	2,44
59795	42	1,95	1,05	17,02	14,84	3,50
96833	32	0,00	5,74	17,94	8,03	1,07
17264	32	1,38	2,79	15,18	10,33	2,68
96774	42	0,37	3,16	15,66	11,87	2,81
53278	38	0,20	0,63	12,39	12,36	4,38
85018	38	0,30	0,01	11,28	17,02	8,27
10052	23	0,34	0,08	01,28	1,69	5,10
95265	19	0,60	0,36	01,44	2,40	6,18
53721	20	0,76	0,57	3,70	8,04	6,87
60315	39	0,59	0,63	3,58	4,17	9,08

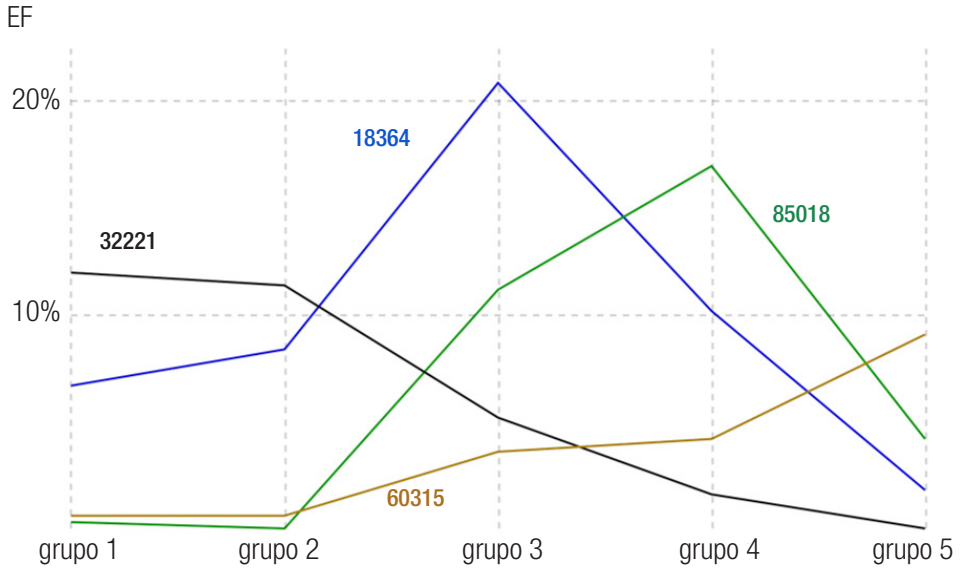
Fonte: Elaboração própria, 2020.

As questões que atingiram maior EF nos grupos 2 e 4 têm valores próximos, mas não são as mesmas. A questão de maior EF no grupo 2 (32221) tem índice de 11,36% para este grupo e apenas 1,57% para os participantes do grupo 4. Já a questão 85018 tem o maior EF para o grupo 4 (17,02%) e 0,01% para o grupo 2. Comparativamente, o grupo 3 é o mais afetado, chegando ao pico de 20,85% para a questão 18364. Essa questão foi a segunda no caderno Azul e a penúltima no caderno Rosa, e os índices de acerto foram 75,5% e 54,7%, respectivamente, entre os participantes do grupo 3. Uma diferença de acerto bastante significativa, considerando que se trata da mesma questão, apenas apresentada em ordem diferente.

Considerando os três grupos centrais, em que se concentram 63,4% dos participantes, com exceção de uma questão (45081), todas apresentam distância 32 ou mais e o EF é alto na maior parte dos casos. Complementarmente, os grupos 1 e 5 demonstram EF geral menor.

Todos os grupos são afetados, mas grupos diferentes são afetados em questões diferentes. Apesar de haver alguma intersecção entre as questões de grupos vizinhos, seu valor varia consideravelmente à medida que os grupos se afastam. O Gráfico 3 ilustra a variação do EF para quatro questões que foram escolhidas por terem apresentado o maior EF em cada grupo.

Gráfico 3 – Variação do EF de quatro questões para os diferentes grupos de desempenho



Fonte: Elaboração própria, 2020.

Os grupos 1 e 2 têm EF máximo na mesma questão (32221), e pode ser observado que esse valor diminui à medida que o desempenho dos participantes aumenta, sendo uma questão que praticamente não afeta os participantes do grupo 5. Já a questão 18364 figura entre os maiores EFs dos grupos 1 e 2 e atinge o ápice no grupo 3. Porém, ela já não está entre as cinco maiores para o grupo 4 e tem EF bastante baixo, no grupo 5. O pico do EF do grupo 4 foi na questão 85018, que não atrapalhou muito os respondentes dos grupos 1,

2 e 5. Por fim, a questão que causou maior EF no grupo 5, 60315, exibiu impacto residual em todos os outros grupos.

Retomando as perguntas colocadas no início desta seção, observamos que os estudantes dos grupos 1 e 5 sofrem menos influência do efeito fadiga. Comparativamente, o grupo 3 é o mais sensível a esse efeito. Foge ao nosso conhecimento artigos que tenham analisado dados do Enem considerando grupos de diferentes desempenhos. As diferenças que identificamos entre os grupos sugerem que essa abordagem é relevante para estudos que busquem entender o desempenho dos participantes e/ou avaliar a qualidade do exame, como os conduzidos por Toffoli (2019) e Travitzki (2017).

EF e a dificuldade das questões

Uma possível explicação para a diferença entre os EFs de diferentes grupos em uma mesma questão, observada na seção anterior, é que as questões que tiveram EF muito baixo em um determinado grupo têm frequência de acerto muito alta ou muito baixa naquele grupo. Isto é, por um lado, os participantes acertam uma questão, independentemente de sua posição no caderno de prova, se ela for fácil para aquele grupo (alta frequência de acerto). Por outro lado, se a questão for difícil para aquele grupo, o fato de ela aparecer no início ou final do caderno de provas também não altera a frequência de acertos, pois, muito provavelmente, os participantes daquele grupo não conseguiram se engajar cognitivamente na resolução do item. Consequentemente, esperamos que questões com EF alto tenham frequência de acerto em uma faixa intermediária dentro daquele grupo.

Para explorar essa relação, criamos a Tabela 4, que apresenta a frequência de acertos para as mesmas questões da Tabela 3 (os destaques em cinza ainda indicam as questões que tiveram maior EF para o grupo).

Tabela 4 – Frequência de acerto das questões dentro de cada grupo

	Frequência de acertos entre estudantes do grupo (%)				
ID questão	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
32808	21,52	41,86	73,27	83,34	92,63
30865	20,81	41,99	79,93	93,23	98,13
45081	28,33	51,69	86,51	92,61	96,07
18364	26,82	36,39	66,37	90,69	98,12
32221	44,50	69,59	92,31	96,55	98,91
88786	14,91	34,15	75,41	89,56	96,00
59795	9,30	13,09	39,47	73,76	90,83
96833	12,33	29,60	71,09	89,84	97,52
17264	20,44	31,87	55,83	82,98	95,17
96774	16,37	26,63	52,90	80,80	95,33
53278	9,97	15,90	34,80	66,81	88,97
85018	12,87	17,90	26,30	52,04	76,91
10052	15,65	25,46	28,18	36,55	61,33
95265	14,50	21,70	22,09	22,60	30,31
53721	12,77	19,37	26,38	42,77	71,47
60315	15,17	24,12	26,38	38,71	74,14

Fonte: Elaboração própria, 2020.

Começando pelo grupo 5, a questão 95265 se destaca como a questão mais difícil em toda a prova, considerando inclusive as questões que não estão apresentadas na tabela e, como constatado por estar em cinza, faz parte do conjunto de questões com maior EF. Além dela, as outras questões em cinza para o grupo 5 apresentam frequência de acertos intermediária. Note que as primeiras onze questões da tabela apresentaram frequência de acerto acima de 88% para esse grupo e EF baixo.

Três faixas de dificuldade podem ser vistas de forma mais acentuada nas questões cinzas para os grupos 3 e 4: quando a frequência de acerto fica muito alta, a questão deixa de apresentar alto valor de EF para aquele grupo e o mesmo pode ser dito quando a frequência de acertos é muito baixa.

Note que, para os grupos 1 e 2, as frequências de acerto mais altas coincidem (em grande parte) com questões com maior EF, já que os índices de acerto nesses grupos nunca atingem valores realmente altos (máximo de 45% no grupo 1 e 70% no grupo 2). Por outro lado, o grupo de questões difíceis (baixa frequência de acertos) é bastante grande e apresenta EF baixíssimo.

Uma possível interpretação para esse fenômeno é de que as questões muito difíceis (baixa frequência de acerto) causam pouca fadiga porque os participantes não investem muito tempo tentando resolvê-las, ou não conseguem se engajar cognitivamente com a sua resolução. Por outro lado, as questões com frequência de acerto alta (muito fáceis) também não causam fadiga porque os participantes conseguem resolvê-las mesmo localizadas no final da prova, quando eles estão cansados. Já as questões na faixa intermediária de frequência de acertos seriam aquelas nem fáceis demais, que podem ser resolvidas rapidamente, nem difíceis demais, ficando além do alcance do candidato, e que por isso salientam o efeito da fadiga causada por uma avaliação com a extensão do Enem.

Essa interpretação se alinha ao que foi sugerido por Travitzki (2017) para explicar o comportamento anômalo para algumas questões do Enem. Nossa análise fundamenta essa sugestão ao demonstrar que o efeito fadiga afeta de forma diferente grupos com desempenhos distintos de acordo com a dificuldade de cada questão.

Conclusões

Este artigo apresenta a primeira evidência empírica de que o efeito fadiga se manifesta no Enem. A partir de uma amostra de mais de 4 milhões de respostas presentes nos microdados da edição 2016 da prova de Matemática e suas Tecnologias, encontramos uma queda substancial no desempenho de estudantes em uma dada questão quando ela figurou mais próxima do fim da prova. Essa evidência é compatível tanto com resultados obtidos em outros países (BORGHANS; SCHILS, 2012), quanto com a experiência de estudantes brasileiros em outros exames (SASSAKI *et al*., 2018).

Em seguida, aprofundamos o entendimento sobre o efeito fadiga considerando como ele se manifesta para diferentes participantes, agrupados de acordo com o número total de acertos obtidos durante a prova. Nossos resultados sugerem que todos os grupos são afetados pelo efeito fadiga, mas grupos com desempenhos diferentes apresentam o efeito em questões distintas.

Por fim, ao considerarmos a dificuldade das questões para cada grupo, identificamos a tendência de que as questões com maior EF são aquelas que estão em uma faixa intermediária de dificuldade. Nossa tentativa de explicação para este fenômeno é que as questões mais fáceis não chegam a ser desafiadoras para esses participantes e, portanto, eles conseguem resolvê-las com a mesma efetividade quando aparecem no começo ou no fim da prova. No outro extremo, questões mais difíceis são percebidas como além do seu alcance pelo estudante, o que resulta em pouco envolvimento cognitivo com a questão, reduzindo o impacto da fadiga. Até o limite de nosso conhecimento, esse resultado é inédito na literatura acadêmica. Essa explicação também está alinhada com o comportamento mais errático na resposta de estudantes para questões presentes no final da prova das edições de 2009 e 2011 do Enem identificado por Travitzki (2017).

Do ponto de vista metodológico, a variação no EF que identificamos na seção "Grupos diferentes sofrem fadiga em questões diferentes" sugere que a abordagem que adotamos, de agrupar estudantes por desempenho, deve ser considerada em estudos nessa área, pois permite um olhar mais refinado para fenômenos como os identificados por Sassaki *et al*. (2018), Toffoli (2019) e Travitzki (2017).

Vale salientar que este estudo tem duas limitações, advindas da seleção intencional de um ano específico (2016) e de apenas uma área do conhecimento (matemática). Reconhecemos essa restrição, mas salientamos que nada impede que os achados deste estudo sejam replicados em outros contextos. Por exemplo, a série temporal pode ser ampliada, outras áreas do conhecimento podem ser incluídas e outras interações podem ser exploradas, como a diferença no EF por sexo ou condição socioeconômica do estudante.

Do ponto de vista psicométrico, o efeito que identificamos pode comprometer as notas atribuídas pela Teoria da Resposta ao Item. Embora o impacto acumulado do efeito fadiga na nota final dos estudantes ainda demande investigação sistemática em estudos futuros, consideramos prudente, dada a relevância social que esses resultados podem ter (TOFFOLI *et al*., 2016), que medidas que visem mitigar esse efeito sejam levadas em consideração pelo Inep.

A literatura sobre a Teoria da Resposta ao Item em avaliações padronizadas de larga escala é bastante rica internacionalmente (TOFFOLI, 2019), podendo fornecer caminhos e opções para a mitigação desse efeito. Entretanto, destacamos três opções que surgiram ao longo da elaboração deste trabalho: adaptação da solução adotada na edição de 2000 do Pisa (ADAMS; WU, 2002) de recálculo dos parâmetros de dificuldade de cada questão para cada um dos cadernos separadamente; utilização de modelos que considerem a posição da questão como um dos parâmetros no processamento dos resultados via Teoria da Resposta ao Item (ALBANO, 2013); e uso de outras formas de criação de cadernos de prova diferentes entre si, como a aleatorização das alternativas de cada questão em vez das questões.

Referências

ADAMS, Ray; WU, Margaret (ed.). **PISA 2000 technical report.** Paris: Organisation for Economic Co-operation and Development, 2002. Disponível em: https://www.oecd.org/pisa/data/33688233.pdf. Acesso em: 23 abr. 2022.

ALAVARSE, Ocimar Munhoz. Desafios da avaliação educacional: ensino e aprendizagem como objetos de avaliação para a igualdade de resultados. **Cadernos Cenpec**, São Paulo, v. 3, n. 1, p. 135-153, 2013.

ALBANO, Anthony D. Multilevel modeling of item position effects: modeling item position effects. **Journal of Educational Measurement**, Washington, v. 50, n. 4, p. 408-426, 2013.

BAUER, Adriana; ALAVARSE, Ocimar Munhoz; OLIVEIRA, Romualdo Portela. Avaliações em larga escala: uma sistematização do debate. **Educação e Pesquisa**, São Paulo, v. 41, esp., p. 1367-1382, 2015.

BORGHANS, Lex; MEIJERS, Huub; TER WEEL, Bas. The role of noncognitive skills in explaining cognitive test scores. **Economic Inquiry**, Hoboken, v. 46, n. 1, p. 2-12, 2008.

BORGHANS, Lex; SCHILS, Trudie. The leaning tower of Pisa: decomposing achievement test scores into cognitive and noncognitive components. *In*: SOCIETY OF LABOR ECONOMISTS CONFERENCE, 17., 2012, Chicago. **Anais…** Chicago: SOLE, 2012. Disponível em: http://conference.nber.org/confer/2012/SI2012/ED/Borghans_Schils.pdf. Acesso em: 23 abr. 2022.

DAVIS, Jeff; FERDOUS, Abdullah. **Using item difficulty and item position to measure test fatigue**. Washington: American Institutes for Research, 2005.

GONZAGA, Luiz Ricardo Vieira; ENUMO, Sônia Regina Fiorim. Lidando com a ansiedade de provas: avaliação e relações com o desempenho acadêmico. **Boletim [da] Academia Paulista de Psicologia**, São Paulo, v. 38, n. 95, p. 266-277, 2018.

HERZOG, Regula; BACHMAN, Jerald. Effects of questionnaire length on response quality. **The Public Opinion Quarterly**, Oxford, v. 45, n. 4, p. 549-559, 1981.

LAVRAKAS, Paul. **Encyclopedia of survey research methods**. Thousand Oaks: Sage, 2008.

LIMA, Priscila da Silva Neves *et al.* Análise de dados do Enade e Enem: uma revisão sistemática da literatura. **Avaliação**, Campinas, v. 24, n. 1, p. 89-107, 2019.

MARCHIONI, Cynthia G. **Habilidades no cognitivas en América Latina**: una medición desde pruebas estandarizadas. 2017. Tese (Mestrado em Economia) – Universidad Nacional de La Plata, La Plata, 2017.

MEYERS, Jason; MILLER, Edward; WAY, Walter. Item position and item difficulty change in an IRT-based common item equating design. **Applied Measurement in Education**, Abingdon, v. 22, n. 1, p. 38-60, 2008.

NEVO, David. Evaluation in education. *In*: SHAW, Ian; GREENE, Jennifer; MARK, Melvin (ed.). **The SAGE handbook of evaluation**. Thousand Oaks: Sage, 2011. p. 442-460.

RODRIGUES, Margarida Maria Mariano. **Avaliação educacional sistêmica na perspectiva dos testes de desempenho e de seus resultados**: estudo do SAEB. 2007. Tese (Doutorado em Psicologia) – Universidade de Brasília, Brasília, DF, 2007.

SASSAKI, Alex Hayato *et al*. **Por que o Brasil vai mal no PISA?** Uma análise dos determinantes do desempenho no exame. São Paulo: Insper, 2018. (Policy paper; n. 31).

TOFFOLI, Sônia Ferreira Lopes. Análise da qualidade de uma prova de matemática do Exame Nacional do Ensino Médio. **Educação e Pesquisa**, São Paulo, v. 45, e187128, 2019.

TOFFOLI, Sônia Ferreira Lopes *et al*. Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. **Educação e Pesquisa**, São Paulo, v. 42, n. 2, p. 343-358, 2016.

TRAVITZKI, Rodrigo. Avaliação da qualidade do Enem 2009 e 2011 com técnicas psicométricas. **Estudos em Avaliação Educacional**, São Paulo, v. 28, n. 67, p. 256-288, 2017.

WOYCIEKOSKI, Carla; HUTZ, Claudio Simon. Inteligência emocional: teoria, pesquisa, medida, aplicações e controvérsias. **Psicologia:** reflexão e crítica, Porto Alegre, v. 22, n. 1, p. 1-11, 2009.

Recebido em: 31.07.2020
Revisado em: 24.11.2020
Aprovado em: 10.02.2021

Leonardo Barichello é licenciado em matemática pela Universidade Estadual de Campinas (Unicamp), mestre em educação matemática pela Universidade Estadual de São Paulo Júlio de Mesquita Filho (Unesp), *campus* Rio Claro, e doutor em educação pela Universidade de Nottingham (Inglaterra). É professor do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), *campus* Jundiaí.

Rita Santos Guimarães é bacharel e licenciada em matemática pela Unicamp, mestre em ensino de ciências exatas pela Universidade Federal de São Carlos (UFSCar) e doutora em educação pela Universidade de Nottingham (Inglaterra). É bolsista de pós-doutorado da Fapesp no Instituto de Matemática, Estatística e Computação Científica da Unicamp.

Dalson Britto Figueiredo Filho é bacharel em ciências sociais, mestre e doutor em ciência política pela Universidade Federal de Pernambuco (UFPE). É professor adjunto do Departamento de Ciências Políticas da UFPE.