A Ordem das Questões Afeta o Desempenho dos Estudantes? Evidências do ENEM 2018 e 2019

Does the Formatting of the Test Affect Student Performance? Evidence from ENEM 2018 and 2019

Gabriel Henrique Fernandes da Silva (*in memoriam*)
IMECC/Unicamp
g234847@dac.unicamp.br

Mayra Picoloto Papani
IMECC/Unicamp
m184537@dac.unicamp.br

Rita Santos Guimarães
IMECC/Unicamp
rita@ime.unicamp.br

Resumo
Nosso objetivo neste estudo é verificar se a ordem das questões nos diferentes cadernos de prova do ENEM pode afetar o número de acertos em cada questão da prova de Matemática e suas Tecnologias, fenômeno chamado de efeito fadiga. Trata-se de uma replicação e extensão da investigação realizada por Barichello, Guimarães e Figueiredo Filho (2021) com os microdados da edição 2016 do ENEM, incluindo agora os dados referentes à prova de Matemática e suas Tecnologias das edições de 2018 e 2019. A nossa análise de dados levou em conta uma amostra de 3.366.761 participantes no ENEM para a edição de 2018 e 3.161.373 para a edição de 2019. A partir desses dados, calculamos a diferença entre o índice de acerto de uma mesma questão nos dois cadernos de prova com a maior diferença relativa entre as posições em que a questão foi apresentada e analisamos essa informação considerando o desempenho dos estudantes e a dificuldade da questão. Primeiramente, concluímos que nas edições 2018 e 2019 também é possível identificar o efeito fadiga em diversas questões. Além disso, mostramos que as considerações feitas pelos autores do artigo original acerca da maneira como o efeito fadiga age em grupos de estudantes com diferentes desempenhos se manteve nessas duas edições da prova, corroborando seus resultados. Consideramos as conclusões deste estudo de grande relevância dada a importância social do ENEM no Brasil e a escassez de estudos sobre as suas qualidades como avaliação de larga escala.

Palavras-chave: efeito fadiga; avaliação de larga escala; ENEM; matemática.

Abstract
Our aim in this study is to verify if the order in which a question is presented in the different configurations of ENEM in mathematics is affecting the proportion of students' score in each question; such a phenomenon is called fatigue effect. This study is a replication and an extension of a study by Barichello, Guimarães e Figueiredo Filho (2021) in which they analysed data from 2016, now we looked at the data from ENEM examation in 2018 and 2019, still only considering the area of Mathematics and its Technologies. We analysed a sample of 3.366.761 participants of ENEM 2018 and 3.161.373 of ENEM 2019. Initially, we calculated the maximum distance in which the same question was presented in each configuration of the exame. We then calculated the difference in score for this maximum distance. We also took into account the students' scores and how difficult each question was. Firstly, we were able to identify the fatigue effect for many questions in 2018 and 2019 editions of ENEM in the mathematics exam. Secondly, similar observations, as the ones

suggested by the authors of the initial study, when participants were grouped according to performance were possible for the two new editions. This extension of the results from 2016 to data from ENEM 2018 and 2019 are very important due to the social relevance of this exam for Brazilian students and also due to the lack of other studies that considered the effectiveness of ENEM as a large scale evaluation and selection exam.

Keywords: fatigue effect; large scale exam; ENEM; mathematics.

Introdução

Avaliações de larga escala como o ENEM[1] e o PISA[2] tem como objetivo aferir as habilidades cognitivas dos seus participantes. Neste sentido, espera-se que a influência de fatores não cognitivos seja nula ou muito pouco relevante. Porém, ao investigar os microdados da edição 2015 do PISA, com o objetivo de compreender as razões que poderiam explicar o baixo desempenho dos estudantes brasileiros no exame, Sassaki, Pietra, Filho e Komatsu (2018) se depararam com um fenômeno que, segundo os dados, parecia afetar com mais intensidade os estudantes brasileiros do que os de outros países: o índice de acerto de uma questão caía à medida que uma mesma questão era apresentada mais tardiamente na prova.

Esse fenômeno, embora nunca explorado em exames brasileiros, é documentado na literatura internacional. Trata-se do efeito fadiga. Em linhas gerais, o efeito fadiga faz com que o índice de acerto de uma questão, em exames que fazem algum tipo de aleatorização na ordem em que as questões são apresentadas aos estudantes, caia à medida que esta é apresentada mais próxima do final do exame, tendo sido identificado em diferentes exames e países (ADAMS e WU, 2002; ALBANO, 2013; BORGHANS e SCHILS, 2012; DAVIS e FERDOUS, 2005; MEYERS, MILLER e WAY, 2008).

A identificação desse tipo de fenômeno em avaliações levanta preocupações sobre a validade destas como instrumentos que buscam aferir características cognitivas dos participantes. Essas preocupações se tornam ainda mais pertinentes quando tratamos de provas com grandes implicações sociais, como o ENEM, uma vez que seus resultados não são utilizados simplesmente para relatar como um indivíduo se saiu ao responder alguns itens em determinado momento, mas para fundamentar decisões como o acesso às escassas vagas e aos auxílios financeiros para ingresso e permanência no ensino superior público e privado no Brasil (TOFFOLI, ANDRADE, BORNIA e QUEVEDO-CAMARGO, 2016).

[1] ENEM: Exame Nacional do Ensino Médio, para mais informações acessar: Enem — Inep (www.gov.br)
[2] PISA: Programa Internacional de Avaliação de Estudantes, para mais informações acessar: Pisa — Inep (www.gov.br)

A partir disso, Barichello, Guimarães e Figueiredo Filho (2021) investigaram a ocorrência do efeito fadiga no ENEM. Graças à disponibilização dos microdados pelo INEP[3], os autores puderam analisar o índice de acerto das questões de Matemática e suas Tecnologias nos quatro cadernos (compostos pelas mesmas 45 questões, mas cujas páginas são ordenadas de forma diferente entre cada caderno) utilizados na aplicação principal da edição 2016 do exame.

Primeiramente, Barichello, Guimarães e Figueiredo Filho (2021) identificaram a ocorrência do efeito fadiga em diversas questões do exame. Em seguida, os autores estenderam a análise agrupando os estudantes por desempenho e considerando o efeito fadiga em cada grupo. Essa análise permitiu aos autores concluir que todos os grupos de estudantes parecem ter sofrido o efeito fadiga, embora em diferentes questões. A partir dessa observação, Barichello, Guimarães e Figueiredo Filho (2021) consideraram a dificuldade das questões (medida a partir do índice de acertos da referida questão entre os estudantes de cada grupo) e concluíram que o efeito fadiga parece atuar mais intensamente em questões que se localizam em uma zona de dificuldade intermediária: questões muito difíceis ou muito fáceis para um grupo de estudantes com certo desempenho geral em matemática apresentam efeito fadiga muito baixo, mas questões intermediárias apresentam efeito fadiga que pode checar a 15 pontos percentuais.

Essa conclusão é alarmante, pois aponta para a possibilidade de que um componente totalmente externo ao estudante (o caderno que lhe foi dado) afete o seu desempenho em um exame que pode decidir seu futuro acadêmico e profissional.

Entretanto, como os próprios autores admitem, o estudo baseou-se apenas nos índices de acertos, nas questões da área de Matemática e suas Tecnologias da edição 2016 do ENEM. O que relatamos neste artigo são os resultados obtidos na replicação desse estudo com as questões da área de Matemática e suas Tecnologias das edições de 2018 e 2019.

Objetivo

O objetivo mais amplo dos nossos estudos é replicar a análise realizada por Barichello, Guimarães e Figueiredo Filho (2021) para todas as áreas do conhecimento nas

[3] INEP: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, para mais informações acessar: Inep (www.gov.br)

edições de 2018 e 2019 do ENEM. Entretanto, neste texto, reportaremos os resultados iniciais da nossa análise focada nas questões da área de Matemática e suas Tecnologias.

A escolha dos anos foi feita por se tratarem das edições mais recentes do ENEM com microdados disponíveis no momento em que nossos estudos foram iniciados. Já a escolha da área foi feita por conta da familiaridade dos autores com ela, facilitando as etapas iniciais da análise, mais exploratórias e qualitativas.

Métodos

Por se tratar de uma replicação, os métodos adotados neste estudo seguiram os que foram empregados no estudo inicial de Barichello, Guimarães e Figueiredo Filho (2021). Isso foi possível tanto por conta da colaboração direta com os autores, quanto pela disponibilização de todas as ferramentas necessárias em repositório público (omitido para fins de anonimização). Entretanto, consideramos adequado descrever, ao menos em linhas gerais, as principais etapas e conceitos utilizados ao longo de nossa análise.

A primeira etapa consiste no pré-processamento dos microdados, quando excluímos os estudantes com presença não regular em algum dos dias do exame e os que não resolveram um dos quatro cadernos da aplicação principal, resultando em uma população de 6.528.134 estudantes. Em seguida, separamos os dados em quatro arquivos de acordo com os cadernos e incluímos apenas as variáveis de interesse para a análise.

Então, esses quatro arquivos são processados no software R[4] de forma a extrairmos a posição e o índice de acerto de cada questão em cada um dos cadernos. A partir dessas duas informações, definimos as seguintes variáveis para cada questão:

- Distância: a maior distância entre duas posições de uma mesma questão em cadernos diferentes;
- EF: a diferença absoluta entre o índice de acerto nos dois cadernos que resultaram na maior distância.

Em seguida, os estudantes são divididos em 5 grupos de acordo com o número de questões que acertaram na prova: grupo 1, de 0 a 9 questões; grupo 2, de 10 a 18 questões e assim sucessivamente. Por fim, o cálculo do EF de cada questão é feito também dentro de cada grupo.

[4] Software de análise estatística livre e gratuito, disponível em: r-project.org/

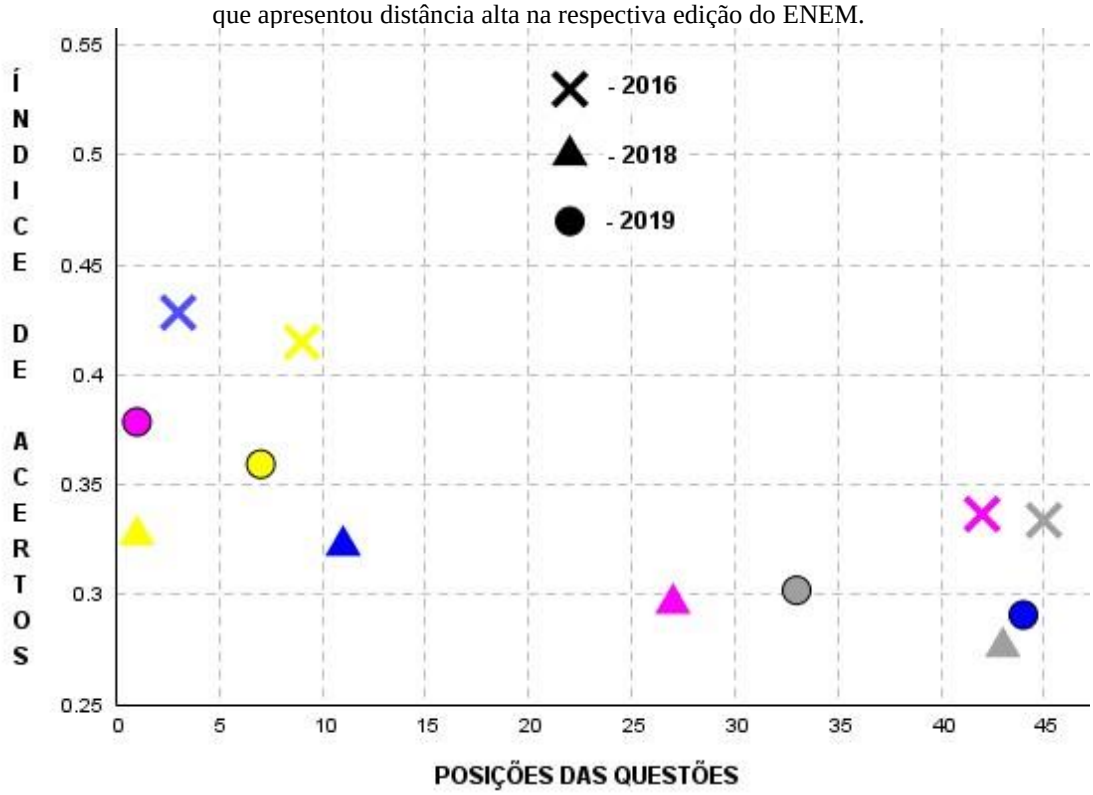
Os dados brutos, os scripts utilizados e as tabelas com as variáveis descritas acima para cada questão também podem ser acessados integralmente no endereço eletrônico anterior (vide pastas 2018 e 2019 para os dados analisados neste texto e 2016 para os dados do artigo inicial).

Análise de dados

Nossos dados englobam para o ano de 2018 um total de 3.366.761 candidatos e para 2019 um total de 3.161.373 candidatos. O motivo desses valores serem inferiores ao total de inscritos no ENEM é que decidimos descartar casos como os de estudantes que não compareceram em todos os dias da prova, estudantes que deixaram a prova em branco ou aqueles que prestaram o exame em data alternativa em relação à principal divulgada pelo Ministério da Educação.

A primeira análise que fizemos foi verificar se havia diferença no índice de acerto de uma questão quando ela apresentava uma grande diferença na posição entre cada caderno. Para fins de exemplificação, apresentamos a seguir as questões de código: 30865 (ano de 2016), 111524 (2018) e 111518 (2019). O critério para escolha destas questões em específico foi por estarem entre as questões que apresentam uma grande distância em diferentes cadernos. A cor de cada ponto representa a cor do caderno onde a questão aparece e o eixo informa sobre a “posição da questão” e o “índice de acertos” para cada uma delas.

Gráfico 1: Índice de Acertos versus Posições das Questões de uma questão de cada ano (2016, 2018 e 2019) que apresentou distância alta na respectiva edição do ENEM.



Fonte: elaboração própria usando microdados do ENEM/INEP (2021).

Como pode ser observado ao analisar as questões indicadas pela mesma forma no gráfico, assim como apontado em Barichello, Guimarães e Figueiredo Filho (2021), conforme uma questão está mais distante do início da prova, o seu índice de acerto tende a cair, chegando a variar de cerca de 38% (questão de 2019 no caderno rosa) de acerto até cerca de 29% (mesma questão no caderno azul). Neste caso, o EF dessa questão é 9% (38 – 29).

Os dados de 2018 mostram que na aplicação do exame daquele ano, 12 das 45 questões apresentaram EF acima de 3%. O mesmo ocorre para os anos de 2019 e 2016. Desta maneira, podemos corroborar as conclusões de Barichello, Guimarães e Figueiredo Filho (2021) de que a ordem das questões afeta o índice de acerto de boa parte dessas questões pelos estudantes, ou seja, o efeito fadiga afetou o desempenho dos estudantes no exame nas três edições – 2016, 2018 e 2019 – na área de Matemática e suas Tecnologias.

Vejamos agora se o comportamento observado pelos autores do artigo inicial acerca da ocorrência de efeito fadiga em questões diferentes quando agrupamos estudantes de acordo com o número de acertos na prova também pode ser identificado nos dados de 2018 e 2019.

Tabela 1: Divisão dos participantes por número de acertos na prova de matemática do ENEM de 2018 e 2019.

		2018		2019	
Grupos	**N° de Acertos**	**N° de Participantes**	**%**	**N° de Participantes**	**%**
1	0 a 9	1.252.522	37,20	1.297.549	41,04
2	10 a 18	1.806.531	53,66	1.549.321	49,01
3	19 a 27	262.012	7,78	241.175	7,63
4	28 a 36	42.771	1,27	66.234	2,10
5	37 a 45	2.925	0,09	7.094	0,22

Fonte: elaboração própria (2021).

Após a separação desses grupos, analisamos o efeito fadiga em cada um deles. Começamos criando uma planilha com o ID de todas as questões da prova de matemática do ENEM da base de microdados do INEP, suas respectivas posições em cada caderno na aplicação da prova e calculamos a máxima distância que se obteve, as frequências de acertos das questões para cada grupo e o EF em cada uma das questões nos cinco grupos[5]. Com as tabelas de todos os grupos em mãos, selecionamos as cinco questões com maior EF em cada grupo. Consideramos que isto seria interessante para considerarmos o que ocorre com o comportamento de uma mesma questão em mais de um grupo.

A seguir, temos estas tabelas para os anos de 2018 e 2019. As questões marcadas em cinza são as cinco com maior efeito fadiga em cada grupo. Podemos notar é que nenhuma destas duas tabelas possuem 25 questões, uma vez que o conjunto com as questões que apresentaram maiores efeitos fadigas para um grupo tem interseção com o conjunto definido de igual forma para os demais grupos.

Tabela 2: As 5 questões com maior EF por grupo de desempenho – 2018.

ID questão	Maior distância	EF G1 (%)	EF G2 (%)	EF G3 (%)	EF G4 (%)	EF G5 (%)
111609	23	7,88	5,97	1,59	0,42	0,79
32953	35	2,84	8,78	12,39	4,38	0,99
111718	30	5,28	7,44	4,37	3,12	1,99
98294	31	6,31	9,39	9,31	4,64	1,8
111512	28	2,7	7,07	5,61	1,19	0,33

[5] Por limite de espaço no artigo, não seria viável trazermos esta tabela na íntegra, ao leitor que desejar ver a tabela de dados na íntegra, acessar: https://osf.io/ev39z/.

111692	31	1,84	7,32	10,39	3,19	1
111487	30	0,68	4,1	12,46	4,09	0,35
62736	30	1,63	6,48	12,49	6,83	1,32
98070	35	1,23	4,44	13,48	9,65	0,95
61172	25	1,75	0,15	11,58	9,17	0,84
96201	30	1,62	2,66	0,11	7,37	1,09
111477	28	0,06	1,31	5,34	9,97	6,6
5990	30	0,67	0,7	0,58	8,91	7,08
85892	35	1,15	1,72	0,2	2,92	6,23
32189	28	0,76	0,09	3,23	3,64	7,23
111628	19	0,07	0,85	4,38	6,49	7,28

Fonte: elaboração própria (2021).

Esses dados devem ser lidos na horizontal, por exemplo, em 2018 a questão de ID 32953 apresentou a maior distância (35). Para os grupos G1, G2 e G3, esta questão esteve entre as 5 questões de maior efeito fadiga e por isso a célula está destacada em cinza nas três colunas. Os valores do EF são: 2,84%, 8,78% e 12,39%, respectivamente. Para os grupos G4 e G5, esta questão não esteve entre as suas cinco questões de maior EF. Além disso, notamos que nos primeiros três grupos o EF dessa questão foi gradativamente aumentando, enquanto que após isto (para os outros dois grupos), o EF diminuiu. Barichello, Guimarães e Figueiredo Filho (2021) também detectaram essa variação no EF entre diferentes grupos.

Tabela 3: As 5 questões com maior EF por grupo de desempenho – 2019.

ID questão	Maior distância	EF G1 (%)	EF G2 (%)	EF G3 (%)	EF G4 (%)	EF G5 (%)
86432	12	5,49	6,04	2,99	1,25	1,42
117725	12	4,69	4,93	0,59	1,21	0,13
11151	43	4,12	11,88	17,93	7,90	1,47
83754	18	2,54	3,29	6,91	5,53	0,23
11763	15	2,16	2,66	0,81	0,08	0,42
11169	38	0,03	7,45	13,71	3,04	0,60

83994	43	0,71	5,87	11,17	5,48	2,24
117767	40	0,66	0,29	13,10	11,01	3,96
111593	39	0,18	2,26	12,73	10,35	2,46
117950	18	0,47	0,45	3,21	5,90	1,90
17642	40	1,73	0,62	1,82	2,90	9,11
89634	10	0,18	0,14	0,69	4,85	4,25
8401	16	0,92	1,48	1,01	4,91	3,56
9779	15	0,44	0,60	0,81	0,79	2,68

Fonte: elaboração própria (2021).

Vale salientar que os grupos intermediários (G2, G3 e G4) parecem ser mais afetados pelo EF, como é possível ser visto em nossos dados observando que os valores nessas colunas são os mais altos nas duas tabelas anteriores. Além disso, embora o efeito fadiga em uma questão seja diferente nos diferentes grupos, todos os grupos demonstram algum grau deste efeito em algumas questões.

Buscando caracterizar as questões que causam mais EF em cada grupo, voltamos nossa análise para os índices de acerto de cada uma delas. Para gerar os dados que apresentaremos a seguir, separamos os alunos de acordo com os critérios tomados para os cinco grupos e calculamos seus percentuais de acerto por questão. Exibimos esses dados para as mesmas questões das tabelas 2 e 3 para podermos apresentar a seguir as análises e conclusões que esses dados possibilitam.

Tabela 4: Índice de acerto das 5 questões com maior EF por grupo – 2018.

ID	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
111609	56,58	81,5	97,37	99,02	99,45
32953	18,95	36,39	80,58	95,43	98,8
111718	23,8	40,61	42,63	44,75	69,88
98294	30,98	53,96	78,48	90,23	97,26
111512	20,48	43,11	84,81	97,16	99,49
111692	11,53	27,77	72,91	91,16	96,51

111487	17,98	31,86	75,42	94,46	97,71
62736	13	28,18	64,84	89,19	97,57
98070	11,61	22,03	50,22	80,69	96,14
61172	11,69	21,03	49,95	73,88	92,24
96201	16,66	21,34	18,87	44,85	86,87
111477	15,55	25,37	31,26	41,18	75,32
5990	12,47	20,23	23,27	37,04	81,03
85892	14,16	25,15	31,72	31,95	50,43
32189	12,75	20	30,38	50,19	78,87
111628	16,39	22,8	25,55	35,18	64,27

Fonte: elaboração própria (2021).

As tabelas 4 e 5 devem ser lidas como as anteriores. Por exemplo, na tabela 5, a questão de ID 111518 está entre as cinco questões de maior EF para os grupos G1, G2, G3 e G4 mas não para o grupo G5. Porém, agora trazemos o índice de acerto desta questão em cada um dos grupos. Neste caso, o índice de acerto, como esperado por conta da forma como definimos os cinco grupos, aumenta a partir de 17,17% no grupo G1 até 98,49% no grupo G5.

Tabela 5: Índice de acerto das 5 questões com maior EF por grupo – 2019.

ID	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
86432	46,12	68,54	87,74	93,03	96,60
117725	35,36	64,13	91,06	96,42	99,38
111518	17,17	38,07	71,73	89,3	98,49
83754	17,05	28,61	48,05	75,13	93,42
117635	36,88	61,45	88,56	95,92	99,08
111696	12,48	29,12	79,32	96,48	99,42
83994	11,7	33,75	75,99	88,78	95,74
117767	17,38	23,94	40,85	69,08	91,56
111593	17,23	30,05	51,72	72,79	92,99

117950	13,72	18,38	23,73	52,59	84,51
17642	14,32	19,07	13,89	18,09	55,05
89634	10,13	14,54	13,25	33,29	81,73
8401	9,95	12,36	12,79	36,01	71,26
9779	10,83	15,13	25,26	58,07	86,86

Fonte: elaboração própria (2021).

Podemos perceber que a conclusão de Barichello, Guimarães e Figueiredo Filho (2021) acerca da dificuldade das questões se relacionarem com o EF se confirma nos dados de 2018 e 2019. O grupo G5, por exemplo, é um grupo que acerta pelo menos 37 das 45 questões, o que pode ser interpretado como um grande domínio dos conteúdos matemáticos exigidos na prova. Desta forma, questões muito fáceis (como a questão 111609 de 2018), independente de suas posições na prova são pouco afetados pelo EF, afinal, mesmo se estiverem cansados, estes alunos conseguem resolver o item e assinalar a alternativa correta. Já para o grupo G1, que é composto por candidatos que acertaram entre 0 e 9 questões, as questões muito difíceis (como a questão 9779 de 2019) também não demonstram efeito fadiga significativo já que devem estar fora de alcance do conhecimento desses estudantes não importando a posição em que são apresentadas na prova.

Portanto, questões muito difíceis apresentam um EF menor para o grupo G1, enquanto que questões muito fáceis apresentam EF menor para o grupo G5. Ainda é possível estender esse raciocínio para justificar porque os mais atingidos pelo EF são os estudantes dos grupos intermediários (G2, G3 e G4), como já observado por Barichello, Guimarães e Figueiredo Filho (2021).

Conclusão

Este estudo, ao replicar as análises dos dados da prova de Matemática e suas Tecnologias de Barichello, Guimarães e Figueiredo Filho (2021) para as provas de 2018 e 2019, concluiu que o efeito fadiga se manifesta no ENEM também nessas edições, ou seja, o desempenho dos estudantes em uma determinada questão é afetado significativamente pela posição em que a questão é apresentada em cada caderno de prova.

Além disso, ao separarmos os candidatos por grupo de desempenho, concluímos também que todos os grupos são afetados pelo efeito fadiga, mas o efeito ocorre em questões distintas para grupos distintos.

Por fim, ao analisar a dificuldade das questões para cada grupo, concluímos que as questões que apresentam maior EF são aquelas que estão em uma faixa intermediária de dificuldade para cada grupo. Concordamos com a explicação sugerida em Barichello, Guimarães e Figueiredo Filho (2021) de que as questões fáceis para um estudante serão resolvidas (corretamente) independentemente de onde aparecem na prova, seja no começo ou no final, não causando efeito fadiga; por outro lado as questões difíceis para este estudante também não apresentam efeito fadiga por estarem além do seu domínio inviabilizando um envolvimento cognitivo na resolução; por fim, nas questões intermediárias o estudante sofrerá influência do efeito fadiga, uma vez que se tratava de uma questão não fácil o suficiente para ele responder sem muitos esforços, mas também não difícil a ponto do estudante não se sentir capaz nem de tentar resolvê-la.

Desta forma, estendemos os achados de Barichello, Guimarães e Figueiredo Filho (2021) para as edições do ENEM 2018 e 2019, mostrando que não se trata de uma excepcionalidade da edição 2016 desse exame.

Nossas conclusões se somam aos trabalhos de Toffoli (2019) e Travitzki (2017) ao se debruçar sobre as qualidades do ENEM como um exame de larga escala e de grande impacto com o intuito de compreender e sugerir aprimoramentos para tal. A influência do efeito fadiga reforça a necessidade de aprofundamento de estudos dessa natureza, uma vez que se trata de algo com impacto potencialmente significativo para os resultados do ENEM.

Ademais, nossos próximos passos pretendidos são: analisar as questões da disciplina de Matemática e suas Tecnologias através de um olhar mais qualitativo das mesmas, para podermos entender melhor o perfil das questões do exame e como estes diferentes perfis influenciam o efeito fadiga. Além disso, já estamos trabalhando na expansão de nossa análise acerca do EF para outras áreas do conhecimento do ENEM.

Agradecimentos

Gabriel Henrique Fernandes da Silva agradece o apoio financeiro da CAPES através da bolsa PICME sem o qual este projeto não seria viável. Rita Santos Guimarães agradece o financiamento da FAPESP, processo número 2019/17135-2.

Referências

ADAMS, R.; WU, M. **PISA 2000 Technical Report**. Paris: Organisation for Economic Co-operation and Development (OECD), 2002. Disponível em: <https://www.oecd.org/pisa/data/33688233.pdf>.

ALBANO, A. D. Multilevel Modeling of Item Position Effects: Modeling Item Position Effects. **Journal of Educational Measurement**, v. 50, n. 4, p. 408–426, dez. 2013.

BARICHELLO, L; GUIMARÃES, R. S.; FIGUEIREDO FILHO, D. B. A formatação da prova afeta o desempenho dos estudantes? Evidências do ENEM (2016). **Educação e Pesquisa**, no prelo, 2021.

BORGHANS, L.; SCHILS, T. Decomposing Achievement Test Scores Into Measures of Cognitive and Noncognitive Skills. **SSRN Electronic Journal**, p. 1 – 36, dez. 2018.

DAVIS, J.; FERDOUS, A. **Using Item Difficulty and Item Position to Measure Test Fatigue**American Institues for Research, , 2005. Disponível em: <10.1037/e539872012-001>. Acesso em: 16 abr. 2019

MEYERS, J. L.; MILLER, G. E.; WAY, W. D. Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. **Applied Measurement in Education**, v. 22, n. 1, p. 38–60, 31 dez. 2008.

SASSAKI, A. H. et al. Por que o Brasil vai Mal no PISA? Uma Análise dos Determinantes do Desempenho no Exame. **Policy Paper**, v. 31, p. 27, 2018.

TOFFOLI, S. F. L. et al. Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. **Educação e Pesquisa**, v. 42, n. 2, p. 343–358, jun. 2016.

TOFFOLI, S. F. L. Análise da qualidade de uma prova de matemática do Exame Nacional do Ensino Médio. **Educação e Pesquisa**, v. 45, p. 1–19, 2019.

TRAVITZKI, R. Avaliação da qualidade do Enem 2009 e 2011 com técnicas psicométricas. **Estudos em Avaliação Educacional**, v. 28, n. 67, p. 256–288, 28 abr. 2017.