



AVALIAÇÃO EM LARGA ESCALA NA AMÉRICA LATINA
EVALUACIÓN EN LARGA ESCALA EN LATINOAMÉRICA
LARGE-SCALE EVALUATION IN LATIN AMERICA

<https://doi.org/10.18222/eae.v35.10972>

QUALIDADE DOS ITENS DE CIÊNCIAS DA NATUREZA DO ENEM: PANORAMA DE 2009 A 2022

 PATRICK ALVES VIZZOTTO¹

¹ Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Diamantina-MG, Brasil;
patrickvizzotto@gmail.com

RESUMO

O uso do Exame Nacional do Ensino Médio para o ingresso ao ensino superior conferiu-lhe um papel significativo na vida dos indivíduos. Diante disso, é fundamental que a prova apresente características mínimas de qualidade, contribuindo para uma seleção equitativa dos candidatos. O objetivo deste artigo foi analisar determinados indicadores psicométricos de qualidade, referentes às edições de 2009 a 2022. O estudo teve como foco a classificação dos itens de Química, Biologia e Física como bons, duvidosos e ruins. Para isso, estimaram-se os parâmetros dos componentes por meio da Teoria da Resposta ao Item. Os resultados revelaram que muitos itens não apresentaram adequação para aferir a proficiência pretendida no exame. Recomenda-se, para estudos futuros, uma análise pedagógica das questões.

PALAVRAS-CHAVE PSICOMETRIA • ENSINO DE CIÊNCIAS • AVALIAÇÃO EDUCACIONAL • EXAME NACIONAL DO ENSINO MÉDIO.

COMO CITAR:

Vizzotto, P. A. (2024). Qualidade dos itens de Ciências da Natureza do Enem: Panorama de 2009 a 2022. *Estudos em Avaliação Educacional*, 35, Artigo e10972.
<https://doi.org/10.18222/eae.v35.10972>

CALIDAD DE LOS ÍTEMS DE CIENCIAS NATURALES DEL ENEM: PANORAMA DE 2009 A 2022

RESUMEN

El uso del Exame Nacional do Ensino Médio [Examen Nacional de la Escuela Secundaria] para la admisión a la educación superior le otorga un papel importante en la vida de las personas. Por lo tanto, es fundamental que la prueba presente características mínimas de calidad, contribuyendo para una selección equitativa de los candidatos. El objetivo de este artículo fue analizar determinados indicadores de calidad psicométrica, referidos a las ediciones de 2009 a 2022. El estudio se centró en clasificar los ítems de Química, Biología y Física como buenos, dudosos y malos. Para ello se estimaron los parámetros de los componentes por medio de la Teoría de Respuesta al Ítem. Los resultados revelaron que muchos ítems no eran adecuados para medir la competencia prevista en el examen. Se recomienda para futuros estudios, un análisis pedagógico de las preguntas.

PALABRAS CLAVE PSICOMETRÍA • ENSEÑANZA DE CIENCIAS • EVALUACIÓN DE LA EDUCACIÓN • EXAME NACIONAL DO ENSINO MÉDIO.

QUALITY OF THE NATURAL SCIENCES ITEMS IN THE ENEM: OVERVIEW FROM 2009 TO 2022

ABSTRACT

The use of the Exame Nacional do Ensino Médio [National High School Exam] as a gateway to higher education grants it a significant role in individuals' lives. In light of this, it is essential that the exam demonstrates minimum quality standards, contributing to an equitable selection of candidates. The aim of this article was to analyze specific psychometric quality indicators concerning the editions from 2009 to 2022. The study focused on classifying the Chemistry, Biology, and Physics items as good, questionable, or poor. For this purpose, the parameters of the components were estimated using Item Response Theory. The results showed that many items did not adequately assess the proficiency intended in the exam. Future studies are encouraged to include a pedagogical analysis of the questions.

KEYWORDS PSYCHOMETRICS • SCIENCE TEACHING • EDUCATIONAL EVALUATION • EXAME NACIONAL DO ENSINO MÉDIO.

Recebido em: 27 FEVEREIRO 2024

Aprovado para publicação em: 11 DEZEMBRO 2024



Este é um artigo de acesso aberto distribuído nos termos da licença Creative Commons do tipo BY-NC.

INTRODUÇÃO

A avaliação educacional, especialmente em larga escala, desempenha um papel crucial no cenário acadêmico contemporâneo, fornecendo informações importantes sobre o desempenho dos alunos, a relevância das políticas educacionais, entre outros. Na América Latina, essa prática tem sido objeto de intensos debates e reflexões, dada a sua relevância para o desenvolvimento e aprimoramento dos sistemas educacionais na região (Bauer et al., 2015). Nesse contexto, o presente estudo explora a temática da avaliação em larga escala na América Latina, com foco nos desafios e implicações encontradas nas questões de Ciências da Natureza do Exame Nacional do Ensino Médio (Enem).

O Enem já foi reconhecido como um dos maiores exames do planeta em termos de número de inscritos (Travitzki, 2017). Apesar de sua participação ser voluntária, o exame pode ser considerado de larga escala devido à sua ampla participação. Criado como uma maneira de avaliar o ensino médio, ao longo dos anos, sua estrutura e propósito foram modificados. Atualmente, sua alta adesão acontece porque a nota obtida na avaliação pode garantir acesso ao ensino superior público em todo o país, além de servir como requisito para concorrer a vagas e financiamentos em instituições privadas (Soares et al., 2021).

O bom desempenho no Enem pode proporcionar diferentes benefícios aos participantes, moldando suas perspectivas presentes e futuras em relação à educação e à ascensão social. Dada a influência significativa que a prova pode ter na trajetória dos participantes, é crucial assegurar sua qualidade, tanto na distribuição segura quanto na elaboração dos itens que fazem parte do teste. Sobre este último, é essencial que a prova possua certo nível de qualidade psicométrica (Hutz et al., 2015).

Uma das frentes de trabalho da psicometria é o desenvolvimento e validação de testes, a fim de garantir os parâmetros de qualidade dos instrumentos utilizados (Hutz et al., 2015). No âmbito internacional, uma referência relevante é o *Standards for educational and psychological testing* (American Educational Research Association [Aera] et al., 2014), que estabelece padrões para os parâmetros de qualidade que os testes da educação e da psicologia devem seguir.

Considerando o impacto do Enem na vida das pessoas e a relevância de uma prova possuir características de validade e confiabilidade, surge a questão: qual a qualidade das questões de Ciências da Natureza do Enem?

Este artigo pretende aferir indicadores psicométricos de questões de Ciências da Natureza do Enem entre 2009 e 2022. Especificamente, pretende-se: 1) obter acesso aos microdados do Enem de 2009 a 2022; 2) filtrar informações alusivas às questões da prova de Ciências da Natureza; 3) calcular indicadores psicométricos das questões; e 4) classificar as questões conforme suas características de qualidade.

Um estudo como este pode ser importante para a área, uma vez que permite refletir sobre a qualidade do exame com base em dados reais e verificar se alguns dos critérios de qualidade da prova são atendidos. A importância dessa temática reside também na necessidade de compreender e analisar criticamente os processos e resultados das avaliações educacionais, especialmente em um país marcado por desigualdades socioeconômicas e educacionais (Faillace et al., 2022). A avaliação em larga escala desempenha um papel fundamental na formulação de políticas públicas, na tomada de decisões pedagógicas e na promoção da equidade e qualidade da educação. Portanto, examinar de forma sistemática e rigorosa a qualidade dos itens dessa avaliação mostra-se essencial.

No âmbito geral, merecem destaque os estudos conduzidos por Travitzki (2017), que discute aspectos da qualidade da prova e examina as edições entre 2009 e 2011; Pontes et al. (2016), que analisam a qualidade de questões de Educação Física nos anos de 2009 a 2013; Gomes et al. (2020), que mensuram a fidedignidade dos escores de participantes na edição de 2011; Piton-Gonçalves e Almeida (2018), que examinam os índices de dificuldade e discriminação dos exames de Matemática de 2012; e também o estudo realizado por Soares et al. (2021), que exploram as medidas de tendência central das questões de 2016 a 2018.

Ao refletir sobre as pesquisas já realizadas, torna-se evidente o interesse de pesquisadores por investigar aspectos relacionados à qualidade da prova. Portanto, este artigo pode contribuir para enriquecer esse debate, ao analisar fatores que influenciam a garantia da qualidade tanto das provas quanto dos itens.

Nesse contexto, este estudo destaca-se ao oferecer uma análise detalhada de alguns aspectos associados à qualidade das questões de Biologia, Física e Química do Enem para um intervalo temporal de 14 anos de prova. Ao investigar o tema, busca-se identificar padrões, tendências e desafios específicos enfrentados nessas disciplinas. Dessa forma, contribui não apenas para a compreensão do panorama educacional na região, mas também para o desenvolvimento de práticas avaliativas mais justas e de maior fidedignidade.

Do ponto de vista social, o manuscrito se destaca por apresentar à literatura uma análise crítica e detalhada dos itens de Ciências da Natureza do Enem, especialmente em um contexto de profundas desigualdades educacionais no Brasil. Ao avaliar rigorosamente a qualidade psicométrica das questões, este estudo proporciona subsídios essenciais para o aprimoramento do exame, que desempenha um papel decisivo na democratização do acesso ao ensino superior. Uma prova que falha em medir com precisão as habilidades e conhecimentos dos candidatos pode perpetuar ou até ampliar as desigualdades existentes. Assim, ao identificar inadequações e sugerir melhorias, o artigo não apenas enriquece o debate acadêmico, mas também oferece uma base sólida para políticas educacionais que visem à equidade e à justi-

ça social. Em última análise, contribui para assegurar que o Enem cumpra seu papel social de maneira efetiva, promovendo uma seleção mais justa e representativa.

Portanto, a seguir, discutem-se as bases deste estudo.

FUNDAMENTAÇÃO TEÓRICA

A noção de qualidade pode variar dependendo do referencial adotado. Por exemplo, uma avaliação pode se constituir como uma das maneiras de refletir sobre a educação. Foi para esse propósito que o Enem foi concebido há mais de vinte anos.

Na área de Educação em Ciências, alguns autores destacam a importância de exames desse tipo, enquanto outros apontam críticas, expondo aspectos negativos desse tipo de prova (Silveira et al., 2015; Nascimento et al., 2018). Dessa perspectiva, é justo o interesse em garantir que provas como o Enem atendam a um certo grau de exigência para cumprir sua meta, que, hoje em dia, além de refletir sobre o sistema educacional, também serve como um caminho para o ensino superior.

Para aferir a qualidade de uma prova como essa, as recomendações internacionais recorrem à psicometria, uma área que estuda fatores intrínsecos dos indivíduos e as diretrizes para elaboração e validação de exames. Nessa área abordam-se os conceitos de validade e fidedignidade. A validade verifica se o instrumento mede aquilo que se pretende medir, enquanto a fidedignidade analisa se o teste produz resultados consistentes com o passar do tempo (Hutz et al., 2015).

Diferentes métodos são usados, a fim de aferir indicadores de qualidade de questões nesse tipo de exame. Duas categorias amplamente utilizadas para a análise são a Teoria Clássica de Testes (TCT) e a Teoria da Resposta ao Item (TRI).

A TCT visa a medir o desempenho de um indivíduo com base em seu índice de acertos. Isso significa que a nota final dos participantes determina os indicadores de qualidade do teste. De modo a garantir a qualificação do exame, é preciso analisar o teste na íntegra. Os procedimentos seguem certos pressupostos. Um deles é a exigência de que os indivíduos respondam aos mesmos itens sob as mesmas condições, com um número mínimo de respondentes. Porém, apenas o escore bruto não permite um paralelo apropriado entre pessoas que alcançaram bons resultados em uma prova fácil e aquelas que apresentaram menor nota em um exame difícil. Por isso, são necessários indicadores que avaliem a dificuldade das provas e questões, assim como a aptidão do item em distinguir participantes com e sem conhecimento, indicador chamado de discriminação.

Em contrapartida, a TRI busca avaliar a proficiência de um indivíduo em determinado conhecimento, não se limitando a um escore final de acertos. A TRI afere um traço latente (habilidade) que é observado por meio das respostas aos itens, representando parâmetros de tal traço latente. Assim, a TRI baseia-se no

pressuposto de que a pessoa acerta ou erra um item conforme seus saberes sobre a temática abordada.

Até o ano de 2008, o Enem utilizava a TCT. A partir de 2009, a prova foi reestruturada e adotou a TRI para avaliação das questões e estimação de desempenho. O modelo utilizado em cada questão da prova é o de três parâmetros em função da proficiência do indivíduo. Os três parâmetros do modelo são: parâmetro de discriminação (“a”), parâmetro de dificuldade (“b”) e probabilidade de acerto ao acaso (“c”). Deseja-se que o parâmetro “a” seja alto, o “b” tenha uma distribuição equilibrada e um parâmetro “c” mínimo (por exemplo, menor que 20% (ou 1/5) no caso de uma questão com cinco alternativas de resposta) (Travitzki, 2017).

Por fim, tanto a TCT quanto a TRI são referenciais capazes de auxiliar a responder às perguntas desta pesquisa. A seção seguinte detalha quais indicadores foram empregados para aferir a qualidade das questões de Ciências da Natureza, assim como os critérios para compreendê-los.

INDICADORES DE QUALIDADE

O indicador “dificuldade” permite avaliar o quão difícil um item é para os participantes. Na TCT, a dificuldade de uma questão é determinada observando-se a média de acertos. Questões as quais poucas pessoas acertaram são classificadas como difíceis. Por outro lado, itens em que muitas pessoas obtiveram acerto são fáceis. Segundo Vilarinho (2015), itens com dificuldade igual ou superior a 0,7 são considerados fáceis; entre 0,7 e 0,3 são de dificuldade média; e abaixo de 0,3 são considerados difíceis. Na TRI, o parâmetro “b” é usado para avaliar este indicador, com valores considerados adequados pela literatura quando variam entre -3 e +3 (Soares et al., 2021) ou -4 e +4 (Travitzki, 2017).

A “discriminação” avalia a capacidade de uma questão diferenciar participantes com níveis de habilidade diferentes. Denominada de parâmetro “a” na TRI, uma discriminação é considerada adequada quando o seu valor é no mínimo 0,5, duvidosa se estiver entre 0,2 e 0,5, e ruim se for menor que 0,2 (Travitzki, 2017). Uma discriminação negativa pode sugerir que a chance de resposta correta diminui com o aumento da proficiência (Pontes et al., 2016). Quanto maior a discriminação de uma questão, maior poderá ser a contribuição dela na medição da competência avaliada.

A “correlação item-total” mede a associação de uma questão com o escore total do exame. Questões com alta correlação contribuem significativamente para a confiabilidade do teste, ao mesmo tempo que questões com baixa correlação podem prejudicar essa consistência (Travitzki, 2017). Para o presente estudo, a correlação item-total foi aferida considerando todos os 45 itens de cada prova.

O “coeficiente de correlação bisserial” é um parâmetro que examina se os participantes que tiveram uma boa performance na prova tendem a escolher as respostas corretas das questões. Esse cálculo é realizado através da análise da correlação entre o escore dos indivíduos e a frequência de escolha de cada resposta do item. Idealmente, deseja-se que a correlação seja positiva para a resposta correta e negativa para as incorretas. Esse indicador tem a capacidade de localizar questões problemáticas, com aspectos a melhorar na formulação ou no gabarito, uma vez que uma correlação negativa para uma resposta certa pode significar que a questão foi mais acertada por pessoas com menor escore (Travitzki, 2017). No entanto, se uma resposta incorreta possuir uma correlação positiva, isso indica que essa alternativa está sendo mais escolhida pelo grupo de indivíduos com escores satisfatórios. Essa disparidade pode indicar um problema de discriminação da questão, sugerindo sua revisão ou remoção. Na interpretação desse indicador, um item é bom se o bisserial da resposta certa for no mínimo 0,3; duvidoso se estiver entre 0,15 e 0,30; e ruim se for inferior a 0,15 (Travitzki, 2017).

Por fim, o “ajuste do modelo” é outro indicador importante que avalia se o modelo de três parâmetros da TRI se ajusta adequadamente aos dados. Um ajuste insatisfatório pode comprometer a precisão dos parâmetros e das estimativas de proficiência. Nesse indicador é gerado um valor de *p* para cada questão. Esse necessita ser menor que 0,05 (estatisticamente significante). Questões com *p*-valor abaixo desse limite são consideradas adequadas; entre 0,05 e 0,10 são duvidosas; e acima de 0,10 sugerem uma inadequação no seu ajuste (Travitzki, 2017).

METODOLOGIA

Este é um estudo quantitativo, exploratório quanto aos seus objetivos, básico quanto à sua natureza, e usa estatística descritiva e inferencial para investigar dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) (Robaina et al., 2021; Gil, 2008). Os microdados são arquivos com os mais completos dados sobre pesquisas e/ou avaliações. Eles contêm informações sobre os itens das provas, os acertos e erros dos participantes, suas informações socioeconômicas, etc. Os microdados do Inep¹ são de acesso público e qualquer indivíduo pode baixá-los para realizar seus estudos.

As informações analisadas nesta pesquisa referem-se ao Enem de 2009 a 2022. Esse recorte justifica-se, pois, conforme referido, o Enem sofreu uma grande reestruturação em 2009, que alterou, entre outras coisas: a matriz de referência; a estrutura do teste; a inclusão da TRI. É relevante destacar que, até o momento de

1 <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados>

finalização deste artigo, os microdados de 2023 ainda não haviam sido disponibilizados pelo Inep.

Classificação dos itens

Os itens agrupados como pertencentes à área de Física são os que abordaram exclusivamente temas relacionados à Física ou que exigiram o emprego de saberes de Física para analisar as alternativas do item. O mesmo critério foi adotado para as questões categorizadas na Biologia e na Química.

Para avaliar o desempenho de cada questão analisada e categorizá-las com base em padrões de qualidade, foram utilizados os indicadores mencionados anteriormente. A classificação adotada neste estudo baseou-se na abordagem feita por Travitzki (2017).

A Tabela 1 apresenta os elementos empregados para categorização das questões segundo a sua qualidade.

TABELA 1
Critérios de classificação das questões

Classificação do item	TEORIA CLÁSSICA DOS TESTES (TCT)		TEORIA DA RESPOSTA AO ITEM (TRI)	
	Correlação item-total (CORR)	Coefficiente bisserial (BISS)	Parâmetro de discriminação (a)	Ajuste do modelo (p-valor)
Bom	$CORR \geq 0,30$	$BISS \geq 0,30$	$a \geq 0,5$	$FIT < 0,05$
Duvidoso	$0,15 < CORR < 0,30$	$0,15 < BISS < 0,30$	$0,2 < a < 0,5$	$0,05 < FIT < 0,10$
Ruim	$CORR \leq 0,15$	$BISS \leq 0,15$	$a \leq 0,2$	$FIT \geq 0,10$

Fonte: Travitzki (2017).

Após a verificação das questões, elas foram agrupadas em uma classificação global, conforme a sugestão de Travitzki (2017), que as organiza como:

- *item duvidoso* (indicador global): quando a questão for considerada duvidosa em pelo menos três dos quatro indicadores, ou ruim em pelo menos um.
- *item ruim* (indicador global): quando o item for classificado como ruim em pelo menos três dos quatro indicadores, ou anulado do gabarito.
- *item bom* (indicador global): se as questões não se enquadrarem nas duas classificações.

Programa utilizado e recorte amostral

A fim de obter acesso aos microdados e estimar os parâmetros necessários, usou-se o *software* R (R Core Team, 2018), com os pacotes “mirt” (Chalmers, 2012) para a TRI e o “psych” (Revelle, 2024) para a TCT.

Ao longo das 14 edições examinadas, encontrou-se um total de 82.190.898 inscritos. Dada essa magnitude, tornou-se necessário realizar um recorte nos dados. Como critérios para essa delimitação foram usados os dados daqueles que: 1) participaram da primeira aplicação; 2) assinalaram todas as questões; 3) frequentaram escolas regulares; 4) haviam terminado o ensino médio no ano do exame; 5) não faltaram a nenhuma prova da edição; 6) responderam ao caderno azul.

TABELA 2
Número de inscritos e amostra filtrada por edição

ANO	INSCRITOS	AMOSTRA	ANO	INSCRITOS	AMOSTRA
2009	4.148.720	571.247	2016	8.627.367	349.597
2010	4.626.094	727.174	2017	6.731.341	305.701
2011	5.380.856	817.224	2018	5.513.747	543.571
2012	5.791.065	776.140	2019	5.095.270	512.707
2013	7.173.563	1.011.596	2020	5.776.800	538.867
2014	8.722.248	1.030.138	2021	3.389.100	442.037
2015	7.746.427	336.720	2022	3.468.300	474.056

Fonte: Inep (2024).

Assim, como ilustrado na Tabela 2, o recorte resultou em uma redução significativa no número de informações analisadas, usando uma amostra de 10,3% (8.436.775) dos indivíduos.

Validação dos dados gerados

Os parâmetros da TRI foram calculados para cada ano analisado de forma independente. Devido à falta de disponibilidade, por parte do Inep, dos parâmetros das questões para a maioria das edições, foi realizada uma análise paralela entre os dados gerados nesta pesquisa e os parâmetros estimados por outro estudo da literatura, visando a verificar a consistência dos dados obtidos neste trabalho.

Para esse propósito, foi destacado o estudo de Rubini (2019), que calculou os parâmetros do Enem para as edições de 2009 a 2017, com base nos dados da população de participantes de cada exame. Tendo em vista que neste trabalho foi selecionada uma amostra, a comparação dos dados foi considerada apropriada para avaliar a fidedignidade dos dados gerados por este estudo, pois esse recorte faz parte do todo, considerando o mesmo período de tempo do referido estudo.

Com isso, notou-se que os parâmetros da TRI (de 2009 a 2017) gerados nesta pesquisa e no estudo de Rubini (2019) não são exatamente idênticos, mas são, na maioria, equivalentes (correlação de Pearson entre os parâmetros dos dois estudos:

parâmetro a: 0,994, $p < 0,000$; parâmetro b: 0,998, $p < 0,000$; parâmetro c: 0,990, $p < 0,000$). Portanto, esse resultado contribuiu para validar a consistência dos dados estimados neste estudo.

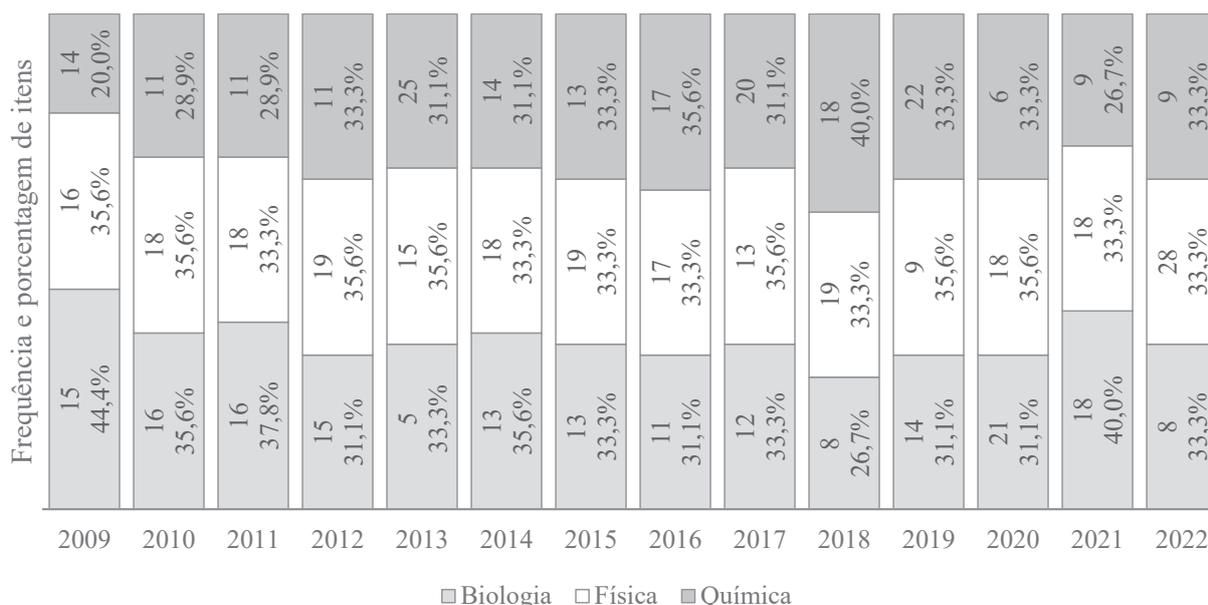
A seção seguinte apresenta o resultado dos indicadores psicométricos das questões.

RESULTADOS

Visão geral dos dados

Em um primeiro momento, apresenta-se, na Figura 1, a proporção de itens de Biologia, Física e Química da prova de Ciências da Natureza para cada edição do Enem analisada.

FIGURA 1
Distribuição dos itens de Biologia, Física e Química para cada edição do Enem

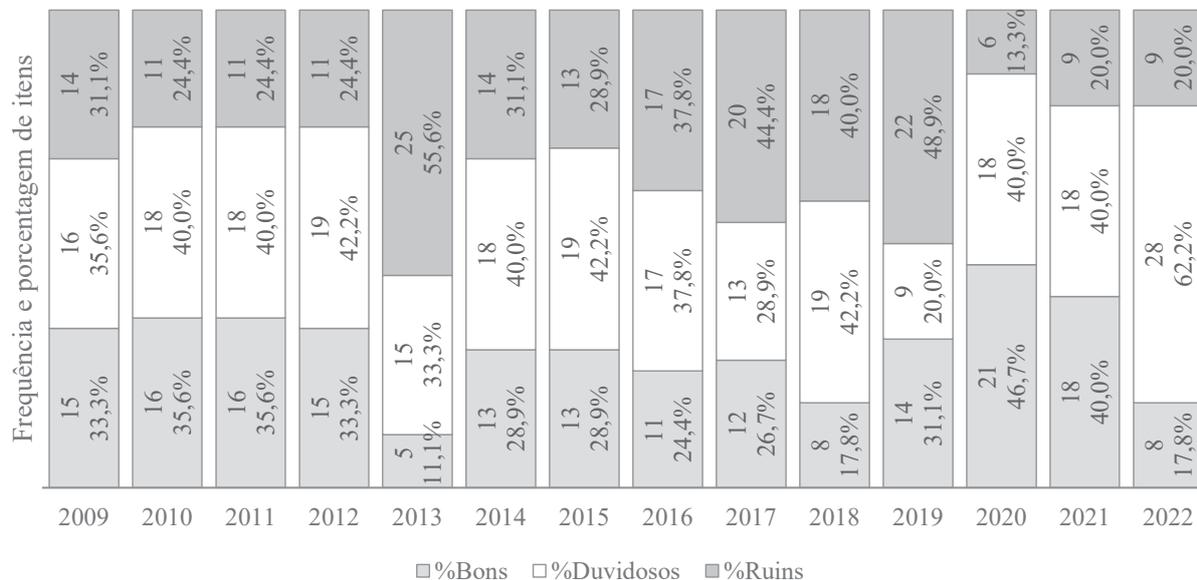


Fonte: Elaboração do autor com dados da pesquisa.

Na Figura 1, pode-se notar que apenas nas edições de 2015 e 2022 a distribuição foi de 15 itens por disciplina. Nos demais anos, a quantidade de itens de cada tema teve ligeira variação, acentuando-se nas edições de 2009 e 2021 (contendo mais itens de Biologia) e em 2018 (com mais itens de Química). Isso significa que, na maioria das edições, a distribuição dos itens pode ser considerada equilibrada, o que ajuda a mostrar que, nas análises seguintes, se uma das disciplinas apresentar maior percentual de itens ruins ou duvidosos, essa divergência não pode ser explicada tão somente pela diferença de quantidade de itens para cada tema.

Ainda realizando uma análise panorâmica dos dados, a Figura 2 mostra, para cada edição, a quantidade de itens classificados em cada uma das categorias de qualidade.

FIGURA 2
Distribuição da qualidade dos itens por edição analisada



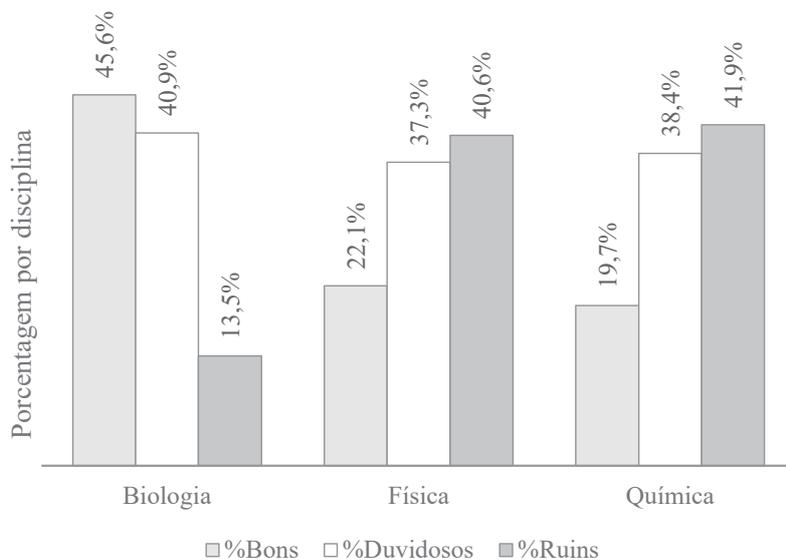
Fonte: Elaboração do autor com dados da pesquisa.

Nota-se que a edição com maior porcentagem de itens considerados ruins foi a de 2013, na frente de 2019 e 2017. Já o ano de 2022 deteve a maior porcentagem de itens duvidosos. Por sua vez, observa-se que em 2013 foi o ano com menor porcentagem de itens bons. Apenas em 2020 a porcentagem de itens bons foi superior tanto à porcentagem de itens duvidosos quanto à de ruins. Na maior parte das demais edições, a porcentagem de itens duvidosos foi superior à de itens bons.

Na Figura 2 já é possível notar com maior detalhamento o comportamento empírico dos itens a partir da classificação realizada. Esta análise sinaliza que a área de Ciências da Natureza pode possuir grande quantidade de itens “problemáticos”, que demandam uma análise qualitativa de maior monta em suas estruturas pedagógicas.

Em um terceiro momento, pode-se observar, na Figura 3, um aprofundamento das análises a partir de cada disciplina da área de Ciências da Natureza. Nela, apresenta-se a porcentagem de itens bons, duvidosos e ruins para Biologia, Física e Química, somadas todas as edições analisadas.

FIGURA 3
Distribuição da qualidade dos itens para Biologia, Física e Química



Fonte: Elaboração do autor com dados da pesquisa.

Na Figura 3, é possível observar que Biologia detém a maior porcentagem de itens bons (23,5 p.p. acima da segunda maior) e a menor de itens ruins (27,1 p.p. abaixo da segunda menor), comparada à Física e Química. No entanto a porcentagem de itens duvidosos em relação às duas outras disciplinas, além de ser maior, tem magnitude expressiva (40,9%). Por outro lado, Química apresentou o pior conjunto de itens em termos de qualidade, com a menor porcentagem de itens bons e a maior de itens ruins, em relação às outras duas disciplinas. Física variou ligeiramente quando comparada à Química, mostrando comportamento similar.

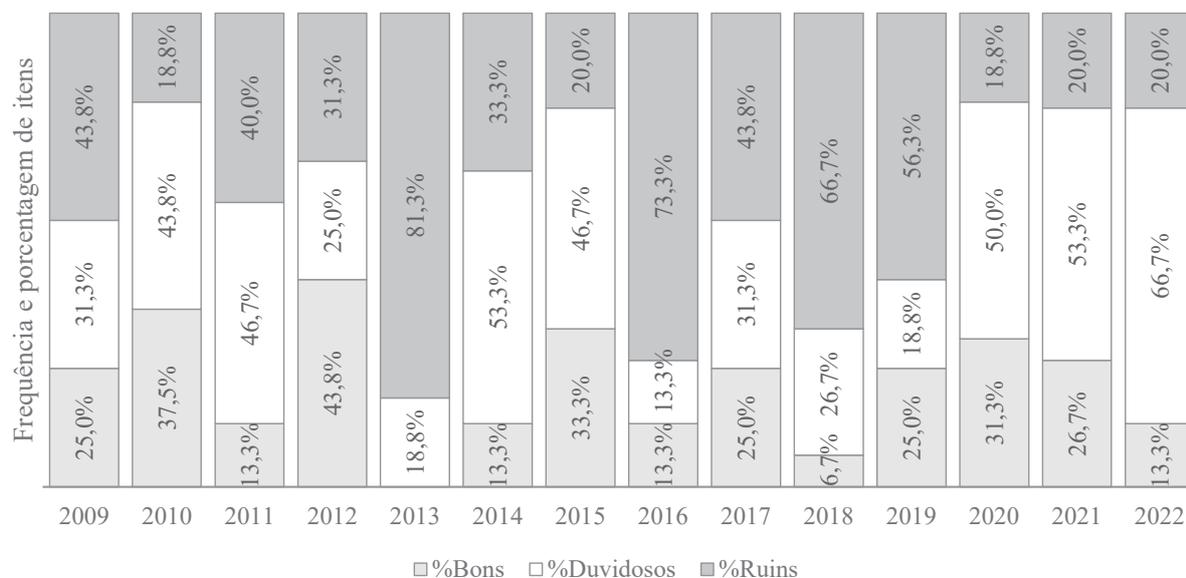
Essa análise aponta que os itens de Física e Química são aqueles que necessitam receber maior atenção no que tange a uma análise pedagógica de maior detalhamento. Isso porque é importante compreender quais aspectos desses itens mais contribuem para a classificação observada.

Não obstante, é fundamental reconhecer que ambas as disciplinas são consideradas desafiadoras por grande parte dos alunos (Quadros et al., 2011). Com isso, destaca-se que, na presente classificação, esse fato pode não ter influenciado no resultado obtido, uma vez que apenas o parâmetro de discriminação da TRI foi levado em conta. Ou seja, a metodologia que classificou os itens para este artigo não levou em consideração o parâmetro de dificuldade de cada um deles. Esse indicador, em particular, será objeto de estudo individual nas páginas a seguir.

Análise da qualidade das questões por disciplina

Após observar a qualidade das questões de cada disciplina a partir de uma classificação geral, considerando todas as edições juntas, a Figura 4 apresenta uma análise individual para a disciplina de Física.

FIGURA 4
Distribuição da qualidade dos itens de Física de 2009 a 2022



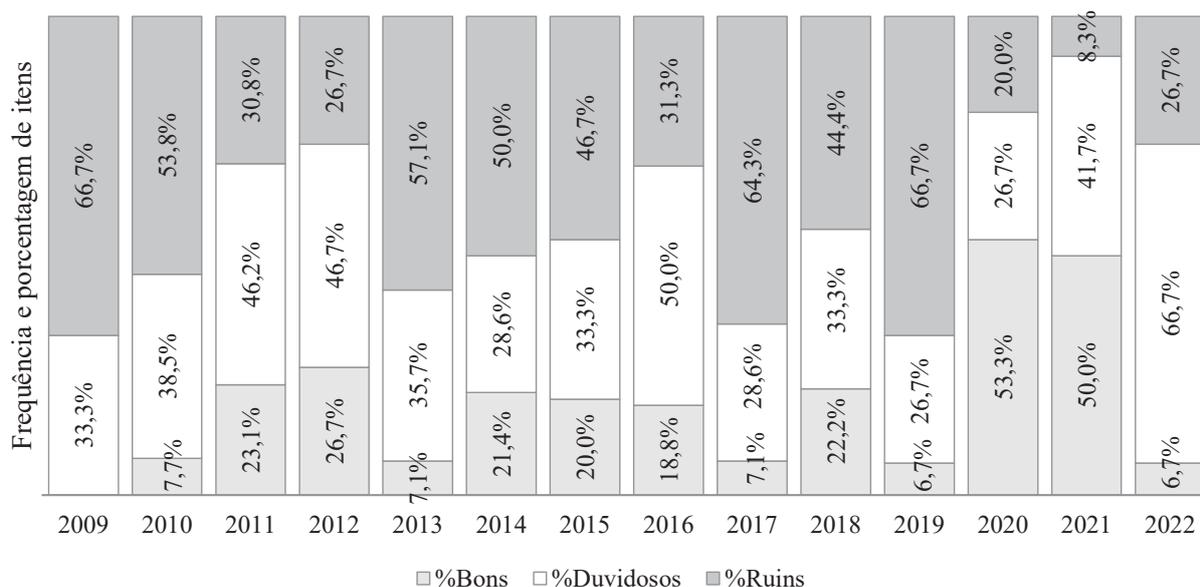
Fonte: Elaboração do autor com dados da pesquisa.

A partir da Figura 4 é possível perceber que apenas no ano de 2012 o percentual de itens bons foi superior tanto aos itens duvidosos quanto aos ruins. Destaca-se o ano de 2013, em que nenhuma questão foi considerada boa e mais de 80% dos seus itens foram classificados como ruins. As edições de 2014, 2020, 2021 e 2022 tiveram pelo menos 50% dos seus itens considerados duvidosos. Mais da metade dos itens foram considerados ruins nas edições de 2016, 2018 e 2019 (73,3%, 66,7% e 56,3%, respectivamente). A fim de complementar a análise e fornecer subsídios para pesquisas futuras, o Apêndice A apresenta, para cada edição, a lista de questões de Física (Tabela A1) classificadas segundo sua qualidade.

Por sua vez, uma análise detalhada também foi feita para as questões de Química. A Figura 5 apresenta seus resultados.

FIGURA 5

Distribuição da qualidade dos itens de Química de 2009 a 2022

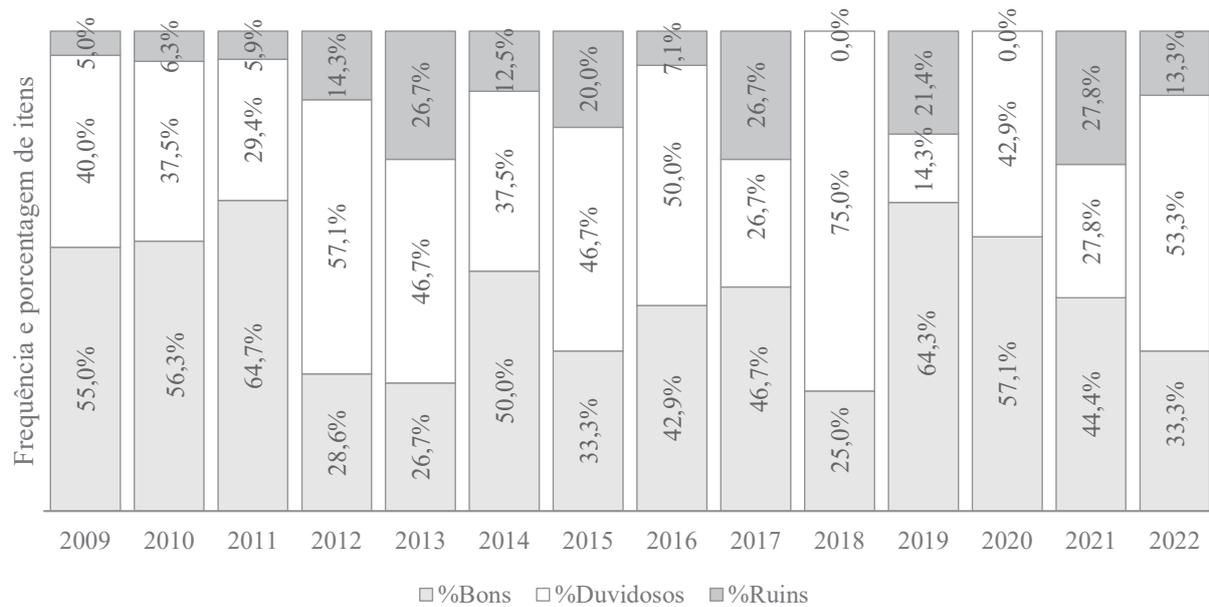


Fonte: Elaboração do autor com dados da pesquisa.

Em Química também houve uma edição do Enem em que nenhum item foi considerado bom (2009). Nas edições de 2020 e 2021, o percentual de questões boas superou tanto o de duvidosas quanto o de ruins. No entanto, esse padrão não se repetiu em nenhuma outra edição analisada. Por outro lado, em 2009, 2010, 2013, 2014, 2017 e 2019, os itens ruins formaram 50% ou mais do total de questões de cada prova. O Apêndice A expõe também uma lista de questões de Química (Tabela A2) classificadas segundo sua qualidade.

Por fim, realizou-se igual análise para os itens de Biologia, que pode ser conferida na Figura 6.

FIGURA 6
Distribuição da qualidade dos itens de Biologia de 2009 a 2022



Fonte: Elaboração do autor com dados da pesquisa.

Biologia apresentou um maior percentual de itens bons em relação às outras duas disciplinas em todas as edições do Enem, com exceção de 2012. As edições de 2009, 2010, 2011, 2014, 2017, 2019, 2020 e 2021 tiveram maior percentual de questões boas em relação ao percentual de duvidosas e de ruins.

Destacam-se os anos de 2018 e 2020, nos quais não houve itens ruins. Não obstante, nota-se que em 2012, 2013, 2015, 2016, 2018 e 2022 o percentual de itens duvidosos foi superior tanto ao de itens bons quanto ao de ruins. Mesmo assim, na visão geral, os itens de Biologia superaram os de Física e Química em termos de qualidade a partir dos parâmetros empregados na análise. O Apêndice A exibe, por sua vez, a lista de questões de Biologia (Tabela A3) classificadas segundo sua qualidade.

Dificuldade dos itens

Outra informação obtida a partir da análise empregada é referente ao nível de dificuldade dos itens. Com base no parâmetro “b” da TRI, classificaram-se as questões em ordem de dificuldade.

A fim de apresentar uma síntese, a Tabela A4 lista os dez itens mais fáceis e os dez itens mais difíceis da prova de Ciências da Natureza de cada edição.

Deseja-se que este levantamento possa servir de subsídio para pesquisas futuras que tenham como foco aprofundar investigações acerca das características de itens considerados fáceis e/ou difíceis nas provas de Ciências da Natureza do Enem. Uma análise pedagógica pode contribuir para verificar os assuntos, formatos, abordagens, contextualizações, entre outros aspectos que possam ter em comum questões fáceis ou difíceis.

Acredita-se que esse diagnóstico contribua significativamente para aprimorar o processo de ensino e aprendizagem desses conteúdos, permitindo a elaboração de estratégias didáticas com maior direcionamento às necessidades dos estudantes.

Na sequência, os resultados apresentados são discutidos à luz do referencial teórico da área.

DISCUSSÕES

Ao observar a Figura 2, que apresenta a classificação dos itens por qualidade em cada edição, torna-se evidente que há uma proporção significativa de itens considerados ruins ou duvidosos em várias edições do Enem. Essa constatação estimula reflexões importantes sobre a validade e confiabilidade dos resultados obtidos por meio de avaliações em larga escala. Afinal, a qualidade dos itens tem um impacto direto na capacidade de mensurar de forma precisa o conhecimento e as habilidades dos estudantes.

A análise por disciplina, como evidenciada nas Figuras 4, 5 e 6, permitiu uma compreensão mais aprofundada dos desafios específicos enfrentados por cada disciplina da área. É evidente que tanto Física quanto Química apresentam itens com padrões consistentes de baixa qualidade em diversas edições do Enem, com percentuais significativos de questões classificadas como ruins ou duvidosas. Essa tendência fornece subsídios para discussões acerca dos processos de elaboração e revisão dos itens nessas disciplinas. Por outro lado, os itens de Biologia mostraram um panorama de maior otimismo em termos de qualidade, com uma proporção maior de questões classificadas como boas. Isso pode indicar, entre outras coisas, uma maior adequação dos conteúdos e abordagens pedagógicas utilizadas nessa disciplina, bem como maior sucesso no processo de elaboração e revisão dos itens.

Na literatura da área de ensino de ciências, há uma série de estudos que investigaram os fatores que afetam a qualidade das questões utilizadas e que podem contribuir para explicar os resultados encontrados nesta pesquisa. Para ilustrar, pode-se destacar o artigo de Bassalo (2011), o qual identificou que a maioria das questões de Física analisadas continha “pegadinhas”, informações irrelevantes e dados essenciais ausentes para sua resolução. Adicionalmente, o autor relata que algumas questões apresentavam mais de uma resposta correta.

Lopes (2015) observou a falta de congruência entre os temas tradicionalmente ensinados no ensino médio e os abordados nas questões das avaliações. Enquanto isso, Duarte et al. (2013) realizaram uma análise pedagógica de uma questão de alto grau de dificuldade do Enem de 2009. Eles notaram que essa questão demandava um conhecimento altamente específico de física e exigia uma memorização particular de uma relação matemática para ser resolvida com êxito.

O estudo conduzido por Silveira et al. (2015) identificou deficiências na elaboração das questões de Física e ressaltou a importância da qualidade dos itens para garantir a validade dos exames. Eles destacaram a presença de falhas nos enunciados das questões, nas opções de resposta e na ênfase excessiva na necessidade de uma interdisciplinaridade não crítica.

Na área da Química, em especial, destaca-se o estudo conduzido por Núñez e Ramalho (2017), que investigaram os equívocos e desafios enfrentados por um grupo de alunos ao realizar o Enem. Os autores destacaram um desempenho insatisfatório dos estudantes em relação aos erros e dificuldades de aprendizagem, especialmente relacionados à compreensão da estrutura, propriedades e aplicações de substâncias e materiais, interpretação de reações e cálculos estequiométricos, entre outros aspectos.

No âmbito da Biologia, o estudo realizado por Mancini et al. (2017) analisou os itens de Biologia presentes no Enem de 2009 a 2014, utilizando a Taxonomia de Bloom Revisada. Os autores buscaram classificar as exigências cognitivas e o nível de conhecimento demandados por cada item. Os resultados revelaram que 84% das questões se enquadram em domínios de baixa ordem cognitiva, enquanto apenas 16% abordavam domínios de alta ordem cognitiva.

No entanto, é importante ressaltar que a qualidade dos itens não deve ser avaliada apenas com base em critérios de dificuldade e discriminação, como delimitado nesta pesquisa. Outros aspectos, como relevância, clareza, alinhamento com os objetivos educacionais, entre outros, também desempenham um papel fundamental na construção de avaliações confiáveis e justas. Portanto, uma análise de maior abrangência, com vistas a incluir também tais aspectos, mostra-se necessária para pesquisas futuras, a fim de garantir a qualidade e equidade dos processos avaliativos.

Também é importante ressaltar que esta análise se baseou apenas em alguns indicadores psicométricos. Contudo, a validade dos itens, por exemplo, não foi objeto de investigação, o que sugere outra lacuna para estudos futuros. Assim, é possível que itens considerados duvidosos possam ter certa relevância para a prova, uma vez que a distribuição dos tópicos abordados deve ser considerada de maneira abrangente.

Ao mesmo tempo, destaca-se novamente que, para a maioria das edições, os parâmetros de cada item não são disponibilizados nos microdados do Enem. Isso significa que esta pesquisa teve de calcular esses parâmetros a partir do índice de acertos dos participantes, o que pode resultar em alguma discrepância entre os resultados obtidos e os parâmetros utilizados pelo Inep para calcular a nota de cada participante. Para mitigar esse potencial viés, conforme apresentado, foi realizada uma comparação entre os parâmetros gerados por esta pesquisa e os parâmetros

da pesquisa de Rubini (2019), observando-se uma correlação de Pearson considerada muito forte. Isso indica uma confiabilidade satisfatória para as conclusões obtidas a partir das análises realizadas neste manuscrito.

Ao analisar os resultados desta pesquisa, e considerando a rigidez que a TRI e outros procedimentos psicométricos utilizados pelo Inep podem oferecer para a qualidade de uma prova, surgem reflexões como: quais aspectos do Enem requerem maior atenção para melhorar a qualidade das questões de Ciências da Natureza? Elaboração dos itens? Revisão pedagógica? Processo de calibração? As possibilidades de investigações são diversas, representando sugestões de estudo potencialmente valiosas para pesquisadores da área da educação e ensino de ciências.

De modo geral, os resultados desta pesquisa fornecem dados relevantes para a compreensão dos desafios enfrentados na elaboração e aplicação de avaliações em larga escala no Brasil e na América Latina, especialmente no que diz respeito às disciplinas de Ciências da Natureza. Essas reflexões são fundamentais para embasar políticas e práticas educacionais que visem não apenas a mensuração do conhecimento dos estudantes, mas também o aprimoramento da qualidade do ensino e da aprendizagem nas escolas do país.

Ao examinar mais profundamente os resultados da pesquisa, é importante considerar o contexto mais amplo da educação na América Latina. A região enfrenta uma série de desafios estruturais e socioeconômicos que impactam diretamente a qualidade e equidade do sistema educacional como um todo (Rosa & Silva, 2017). Dessa forma, acredita-se que os resultados dessas avaliações em larga escala refletem não apenas as características intrínsecas das questões de Biologia, Física e Química, mas também as dinâmicas mais amplas que moldam o cenário educacional latino-americano.

Além disso, é importante destacar a necessidade de uma abordagem ampla e inclusiva na elaboração e aplicação das avaliações em larga escala. Muitas vezes, as avaliações padronizadas tendem a privilegiar determinados tipos de conhecimento e habilidades, deixando de fora aspectos importantes do currículo e da experiência educacional dos estudantes (Souza, 2019). Isso pode levar a uma visão limitada e distorcida do desempenho dos alunos, bem como a uma marginalização de certos grupos e formas de conhecimento, como aqueles relacionados à cultura local e à diversidade étnica.

Não obstante, é fundamental reconhecer que as avaliações em larga escala têm sido frequentemente utilizadas como ferramentas de prestação de contas e tomada de decisões políticas na América Latina. No entanto, essa ênfase excessiva na avaliação quantitativa pode levar a uma cultura de ensino centrada na preparação para os testes, em detrimento de uma abordagem mais ampla e significativa da educação (Moreira, 2018). Portanto, é essencial que as avaliações em larga escala

sejam complementadas por outros instrumentos de avaliação, como avaliações formativas e qualitativas, que forneçam uma visão mais abrangente e contextualizada do progresso dos alunos e do desempenho das escolas. Ao mesmo tempo, é importante destacar que o Enem pode ser também um importante instrumento de avaliação de políticas públicas, e não ser instrumentalizado como uma ferramenta de culpabilização da escola e/ou do docente pelo fracasso escolar dos alunos (Cavalcanti et al., 2018).

Por fim, é essencial que as avaliações em larga escala sejam vistas como parte de um processo mais amplo de melhoria da qualidade da educação. Isso requer um compromisso contínuo com a equidade, a inclusão e a justiça social, bem como uma abordagem colaborativa e participativa que envolva todos os atores da educação.

CONSIDERAÇÕES FINAIS

O presente estudo teve como objetivo analisar a qualidade dos itens de Biologia, Física e Química do Enem. Nesse sentido, a pesquisa procurou contribuir para uma compreensão mais aprofundada dos desafios e implicações encontrados nas disciplinas de Ciências da Natureza, fornecendo reflexões para a literatura da área e para o debate acadêmico sobre aspectos de qualidade da avaliação educacional no Brasil.

Os principais resultados da pesquisa revelaram padrões consistentes de baixa qualidade dos itens de Física e Química em diferentes edições do Enem, contrastando com o desempenho de grande parte das questões de Biologia. Esses achados ressaltam a necessidade de estudos futuros com enfoque em uma análise qualitativa dos aspectos pedagógicos de cada item e dos processos de elaboração e revisão das questões dessas disciplinas.

No entanto, é importante reconhecer as lacunas do estudo, incluindo a falta de consideração de outros aspectos importantes da avaliação educacional, como a relevância dos itens, a clareza das questões e o alinhamento deles com os objetivos educacionais, representados pela matriz de referência do Enem. Além disso, a análise se limitou às disciplinas de Biologia, Física e Química, deixando de fora as demais áreas do conhecimento que também são avaliadas no Enem.

Destaca-se o importante papel desempenhado por instituições como o Inep, que desenvolvem um trabalho crucial ao disponibilizar os microdados e uma variedade de outros documentos para acesso público. Essa postura transparente possibilita a realização de numerosas pesquisas na área da educação, permitindo análises, questionamentos e reflexões sobre o papel, a importância, a intencionalidade e os desafios das avaliações em larga escala existentes no Brasil. Os resultados provenientes dessas investigações podem servir como embasamento teórico para fundamentar políticas públicas que ajudem a assegurar que exames de larga escala, como o Enem, possam cumprir seu objetivo.

É crucial ressaltar a importância que o Enem possui para cada cidadão que almeja ingressar no ensino superior. Investigações voltadas para avaliar a qualidade do exame não têm a intenção de desacreditá-lo, mas, sim, de reconhecer sua relevância e magnitude, promover reflexões que visem a aprimorá-lo e estimular discussões sobre como enfrentar os desafios associados à elaboração de uma avaliação em larga escala com validade e confiabilidade.

Por fim, é importante ressaltar que uma análise psicométrica, conforme conduzida neste artigo, apesar da solidez do método empregado, examina apenas uma parte dos dados com base nas informações oficiais disponíveis. Isso significa que, devido às limitações metodológicas, epistemológicas e práticas, uma avaliação mais abrangente deve contemplar também outros aspectos, muitos dos quais qualitativos, como uma análise pedagógica das questões, por exemplo.

Assim, este estudo não tem a intenção de encerrar a discussão sobre a qualidade dos itens do Enem, mas, sim, incentivar a realização de mais pesquisas com esse propósito em diversas áreas, utilizando uma variedade de instrumentos, métodos e referências teóricas.

REFERÊNCIAS

- American Educational Research Association (Aera), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Aera.
- Bassalo, J. M. F. (2011). Questões de física do Enem/2009. *Caderno Brasileiro de Ensino de Física*, 28(2), 325-355. <http://dx.doi.org/10.5007/2175-7941.2011v28n2p325>
- Bauer, A., Alavarse, O. M., & Oliveira, R. P. de (2015). Avaliações em larga escala: Uma sistematização do debate. *Educação e Pesquisa*, 41(especial), 1367-1382. <http://dx.doi.org/10.1590/S1517-9702201508144607>
- Cavalcanti, C. J. de H., Nascimento, M. M., & Ostermann, F. (2018). A falácia da culpabilização do professor pelo fracasso escolar. *Revista Thema*, 15(3), 1064-1088. <https://periodicos.ifsul.edu.br/index.php/thema/article/view/1059>
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29. <https://doi.org/10.18637/jss.v048.i06>
- Duarte, L. P. A. de A., Gonçalves, W. P., Jr., & Barroso, M. F. (2013). Estudo de um item quantitativo na prova de Física do Enem 2009. In *Anais do 20. Simpósio Nacional de Ensino de Física – SNEF* (pp. 1-9). Universidade de São Paulo. https://www.if.ufrj.br/~marta/enem/producao_enem/Snef2013_T0455-1.pdf
- Faillace, H. F. P., Britto, I. L. M., & Costa, F. da S. (2022). O desempenho dos alunos do ensino médio no Enem 2019 e a desigualdade social: Um estudo utilizando análise exploratória e técnicas de agrupamento. *Cadernos do IME – Série Estatística*, 53, 33-64. <https://doi.org/10.12957/cadest.2022.72568>
- Gil, A. C. (2008). *Métodos e técnicas de pesquisa social* (3ª ed.). Atlas.
- Gomes, C. M. A., Golino, H. F., & Peres, A. J. de S. (2020). Fidedignidade dos escores do Exame Nacional do Ensino Médio (Enem). *Psico*, 51(2), 1-10. <https://doi.org/10.15448/1980-8623.2020.2.31145>

- Hutz, C. S., Bandeira, D. R., & Trentini, C. M. (Orgs.). (2015). *Psicometria*. Artmed.
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2024). *Microdados do Enem*. <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem>
- Lopes, J. C. (2015). *As questões de física do Enem 2011* [Dissertação de mestrado]. Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação em Ensino de Física.
- Mancini, G. V., Marques, A. C., Jr., & Cintra, E. P. (2017). Análise dos itens de biologia presentes no Enem. In *Anais do 10. Congreso Internacional sobre investigación en Didáctica de las Ciencias* (pp. 1479-1484). Universitat Autònoma de Barcelona. <https://raco.cat/index.php/Ensenanza/article/view/337020/427890>
- Moreira, M. A. (2018). Uma análise crítica do ensino de Física. *Estudos Avançados*, 32(94), 73-80. <https://doi.org/10.1590/s0103-40142018.3294.0006>
- Nascimento, M. M., Cavalcanti, C., & Ostermann, F. (2018). Uma busca por questões de Física do Enem potencialmente não reprodutoras das desigualdades socioeconômicas. *Revista Brasileira de Ensino de Física*, 40(3), 1-18. <http://dx.doi.org/10.1590/1806-9126-RBEF-2017-0237>
- Núñez, I. B., Ramalho, B. L. (2017). Os itens de Química do Enem 2014: Erros e dificuldades de aprendizagem. *Acta Scientiae*, 19(5), 799-816.
- Piton-Gonçalves, J., & Almeida, A. M. (2018). Análise da dificuldade e da discriminação de itens de Matemática do Enem. *REMAT: Revista Eletrônica da Matemática*, 4(2), 38-53. <https://doi.org/10.35819/remat2018v4i2id3060>
- Pontes, J. A. F., Jr., Silva, A. G. da, Tavare, E. D., Sousa, L. A., Bastos, F. A. C., Cruz, F. N. I. da, & Almeida, L. S. (2016). Aspectos psicométricos dos itens de Educação Física relacionados aos conhecimentos de Esporte e Saúde no Exame Nacional do Ensino Médio (Enem). *Motricidade*, 12(1), 12-21. <https://www.redalyc.org/pdf/2730/273050666003.pdf>
- Quadros, A. L. de, Silva, D. C. da, Andrade, F. P. de, Aleme, H. G., Oliveira, S. R., & Silva, G. de F. (2011). Ensinar e aprender Química: A percepção dos professores do Ensino Médio. *Educar em Revista*, (40), 159-176. <https://doi.org/10.1590/S0104-40602011000200011>
- R Core Team. (2018). *R: A language and environment for statistical computing* (p. 10). R Foundation for Statistical Computing. <https://www.r-project.org/>
- Revelle, W. (2024). *Psych: Procedures for Personality and Psychological Research*. R package version 1.4.3. Northwestern University. <https://cran.r-project.org/web/packages/psych/psych.pdf>
- Robaina, J. V. L., Fenner, R. dos S., Martins, L. A. M., Barbosa, R. de A., & Soares, J. R. (Orgs.). (2021). *Fundamentos teóricos e metodológicos da pesquisa em educação em ciências*. Bagai. <http://educapes.capes.gov.br/handle/capes/585938>
- Rosa, G. A. da, & Silva, D. Q. da (2017). Educação popular na América Latina e questão social: Da desigualdade à resistência. *Educação*, 42(2), 319-332. <https://doi.org/10.5902/1984644424126>
- Rubini, G. M. (2019). *O que o Enem revela sobre a aprendizagem em Física na Educação Básica* [Tese de doutorado, Universidade Federal do Rio de Janeiro]. Pemat. <https://pemat.im.ufrj.br/index.php/pt/producao-cientifica/teses/2019/228-o-que-o-enem-revela-sobre-a-aprendizagem-em-fisica-na-educacao-basica>
- Silveira, F. L. da, Barbosa, M. C. B., & Silva, R. da (2015). Exame Nacional do Ensino Médio (Enem): Uma análise crítica. *Revista Brasileira de Ensino de Física*, 37(1), 1-5. <http://dx.doi.org/10.1590/S1806-11173710001>
- Soares, T. E. A., Soares, D. J. M., & Santos, W. (2021). Medidas de Tendência Central: Análise da qualidade das questões do Enem de 2016 a 2018. *Jornal Internacional de Estudos em Educação Matemática*, 14(1), 119-128. <https://jjeem.pgsscogna.com.br/jjeem/article/view/7979>

Souza, J. E. P. de (2019). Avaliação em Larga Escala, o senso comum, críticas e ponderações. *Revista do Instituto de Políticas Públicas de Marília*, 5(2), 139-156. <https://doi.org/10.36311/2447-780X.2019.v5n2.11.p139>

Travitzki, R. (2017). Avaliação da qualidade do Enem 2009 e 2011 com técnicas psicométricas. *Estudos em Avaliação Educacional*, 28(67), 256-288. <https://doi.org/10.18222/eae.v28i67.3910>

Vilarinho, A. P. L. (2015). *Uma proposta de análise de desempenho dos estudantes e de valorização da primeira fase da OBMEP* [Dissertação de mestrado, Universidade de Brasília]. Repositório Institucional da UnB. <http://icts.unb.br/jspui/handle/10482/19335?mode=full>

APÊNDICE A

TABELA A1

Itens de Física listados segundo classificação e edição do Enem

EDIÇÃO DO ENEM	CLASSIFICAÇÃO		
	Bons	Duvidosos	Ruins
2009	5, 14, 19, 20	24, 30, 31, 32, 37	17, 18, 27, 35, 38, 39, 45
2010	52, 56, 58, 59, 63, 89	47, 48, 54, 68, 78, 81, 84	50, 67, 70
2011	66, 86	54, 56, 60, 63, 70, 77, 84	46, 67, 73, 74, 78, 80
2012	47, 50, 54, 61, 64, 71, 72	77, 78, 83, 88	55, 60, 67, 73, 74
2013	-	48, 52, 57	46, 61, 65, 66, 72, 75, 76, 79, 82, 83, 85, 87, 89
2014	55, 76	50, 57, 66, 68, 72, 84, 87, 90	46, 62, 64, 67, 82
2015	63, 68, 82, 85, 86	49, 50, 57, 65, 75, 79, 88	53, 64, 70,
2016	57, 74	55, 88	47, 49, 54, 59, 63, 64, 66, 77, 82, 84, 86
2017	101, 105, 107, 115	99, 120, 128, 129, 133	93, 103, 108, 110, 112, 127, 131
2018	97	92, 108, 112, 115	103, 104, 120, 122, 123, 125, 128, 129, 131, 134
2019	94, 98, 100	92, 109, 119	102, 106, 111, 117, 121, 126, 131, 132, 134, 135
2020	91, 113, 126, 133, 134	93, 98, 99, 106, 107, 109, 116, 130	105, 111, 114
2021	99, 100, 102, 108	92, 105, 107, 109, 120, 125, 128, 131	94, 126, 133
2022	96, 128	93, 95, 100, 103, 111, 115, 116, 123, 135	112, 119, 134

Fonte: Elaboração do autor com dados da pesquisa.

TABELA A2**Itens de Química listados segundo classificação e edição do Enem**

EDIÇÃO DO ENEM	CLASSIFICAÇÃO		
	Bons	Duvidosos	Ruins
2009	-	12, 26, 29	15, 23, 34, 36, 43, 44
2010	53	65, 79, 80, 83, 85	55, 69, 72, 73, 74, 77, 82
2011	79, 85, 90	52, 58, 59, 72, 75, 81	50, 55, 62, 83
2012	46, 53, 69, 82	49, 58, 59, 63, 76, 86, 90	66, 70, 79, 84
2013	69	47, 54, 58, 81, 90	49, 51, 64, 68, 71, 74, 77, 86
2014	56, 80, 86	48, 54, 78, 83	51, 58, 59, 65, 70, 77, 88
2015	51, 58, 59	52, 62, 73, 81, 84	55, 60, 71, 76, 77, 80, 90
2016	46, 67, 89	50, 51, 52, 53, 60, 68, 72, 85	58, 70, 76, 78, 81
2017	126	97, 114, 130, 134	95, 102, 104, 106, 113, 119, 121, 122, 124
2018	113, 121, 126, 132	93, 95, 96, 99, 135	102, 105, 109, 114, 116, 118, 124, 130
2019	104	103, 105, 113, 129	91, 95, 108, 112, 118, 120, 122, 124, 128, 130
2020	92, 94, 97, 112, 124, 129, 131, 132	96, 100, 102, 110	120, 123, 135
2021	95, 97, 113, 122, 124, 135	101, 103, 117, 130, 134	115
2022	131	91, 97, 101, 108, 109, 117, 120, 124, 127, 129	94, 106, 107, 113

Fonte: Elaboração do autor com dados da pesquisa.

TABELA A3**Itens de Biologia listados segundo classificação e edição do Enem**

EDIÇÃO DO ENEM	CLASSIFICAÇÃO		
	Bons	Duvidosos	Ruins
2009	1, 2, 4, 6, 8, 10, 11, 13, 16, 21, 22	3, 7, 9, 25, 28, 33, 40, 42	41
2010	46, 49, 60, 61, 62, 75, 86, 88, 90	51, 57, 64, 71, 76, 87	66
2011	47, 48, 49, 57, 64, 68, 69, 71, 76, 82, 89	51, 53, 61, 87, 88	65
2012	51, 52, 80, 81	48, 56, 65, 68, 75, 85, 87, 89	57, 62
2013	50, 70, 78, 80	53, 59, 60, 63, 67, 84, 88	55, 56, 62, 73
2014	49, 52, 60, 69, 71, 75, 81, 85	47, 63, 73, 74, 79, 89	53, 61
2015	47, 48, 67, 72, 89	46, 54, 56, 61, 78, 83, 87	66, 69, 74,
2016	48, 65, 71, 79, 87, 90	56, 61, 62, 73, 75, 80, 83	69
2017	92, 94, 100, 109, 111, 118, 132	96, 98, 125, 135	91, 116, 117, 123

(continua)

(continuação)

EDIÇÃO DO ENEM	CLASSIFICAÇÃO		
	Bons	Duvidosos	Ruins
2018	98, 100, 127	94, 101, 106, 107, 110, 111, 117, 119, 133	-
2019	93, 96, 97, 99, 101, 107, 114, 115, 123	116, 133	110, 125, 127
2020	95, 101, 104, 115, 117, 118, 119, 122	103, 108, 121, 125, 127, 128	-
2021	91, 93, 96, 104, 110, 111, 127, 132	106, 112, 114, 121, 123	98, 116, 118, 119, 129
2022	92, 104, 110, 130, 132	98, 99, 114, 118, 122, 125, 126, 133	102, 121

Fonte: Elaboração do autor com dados da pesquisa.

TABELA A4**Top 10 itens mais fáceis e mais difíceis de cada edição do Enem²**

ANO	ITENS MAIS FÁCEIS	ITENS MAIS DIFÍCEIS	ANO	ITENS MAIS FÁCEIS	ITENS MAIS DIFÍCEIS
2009	31, 27, 1, 13, 14, 8, 20, 22, 5, 4	43, 30, 39, 36, 35, 18, 41, 34, 38, 23	2016	49, 79, 89, 46, 87, 90, 71, 65, 67, 48	72, 54, 80, 59, 84, 47, 70, 78, 86, 63
2010	55, 70, 72, 49, 52, 46, 56, 88, 75, 59	69, 82, 48, 66, 76, 50, 67, 71, 73, 74	2017	131, 119, 110, 101, 115, 100, 94, 132, 118, 109	121, 102, 108, 127, 113, 103, 112, 123, 124, 106
2011	74, 83, 89, 82, 85, 71, 57, 69, 51, 47	59, 55, 50, 62, 77, 80, 46, 67, 65, 78	2018	105, 128, 123, 113, 100, 97, 115, 98, 127, 117	124, 91, 122, 125, 129, 107, 118, 104, 109, 131
2012	52, 71, 50, 80, 72, 51, 69, 47, 53, 63	66, 87, 67, 70, 62, 73, 74, 55, 84, 57	2019	126, 111, 120, 106, 110, 104, 123, 96, 107, 115	131, 95, 112, 129, 109, 102, 121, 117, 122, 127
2013	83, 64, 87, 89, 52, 70, 84, 51, 53, 50	74, 82, 62, 48, 58, 86, 85, 55, 77, 79	2020	132, 92, 122, 124, 113, 118, 104, 127, 112, 97	125, 106, 116, 98, 130, 105, 123, 111, 114, 135
2014	60, 66, 70, 86, 78, 67, 63, 68, 47, 46	84, 88, 64, 77, 66, 82, 70, 59, 61, 62	2021	118, 112, 127, 111, 99, 132, 100, 93, 122, 124	133, 105, 106, 115, 123, 119, 94, 98, 126, 116
2015	60, 66, 70, 86, 78, 67, 63, 68, 47, 46	52, 80, 88, 76, 71, 77, 64, 90, 53, 55	2022	112, 113, 134, 119, 94, 99, 131, 104, 132, 130	129, 93, 107, 123, 121, 126, 106, 102, 116, 100

Fonte: Elaboração do autor com dados da pesquisa.

- 2 O critério utilizado para essa classificação foi o parâmetro de dificuldade obtido pela TRI. Os itens, tanto os "mais fáceis" quanto os "mais difíceis", estão dispostos em ordem crescente de dificuldade dentro de cada edição do Enem.