DOI: 10.46667/renbio.v15i1.633

ANÁLISE E CLASSIFICAÇÃO DAS QUESTÕES DE BIOLOGIA DO ENEM SEGUNDO SUAS CARACTERÍSTICAS PSICOMÉTRICAS

ANALYSIS AND CLASSIFICATION OF BIOLOGY ITEMS IN ENEM ACCORDING TO ITS PSYCHOMETRIC CHARACTERISTICS

ANÁLISIS Y CLASIFICACIÓN DE LOS PROBLEMAS DE BIOLOGÍA EN ENEM SEGÚN SUS CARACTERÍSTICAS PSICOMÉTRICAS

Patrick Vizzotto[1]

Resumo
Usado, entre outras coisas, como seleção para o Ensino Superior, o Exame Nacional do Ensino Médio é uma das maiores provas em larga escala do mundo. Tendo em vista o potencial de influenciar a vida de uma pessoa, aspira-se que o exame tenha condições de realizar uma inferência justa daquilo que se propõe avaliar. O artigo divulga os resultados de uma investigação que estimou a qualidade psicométrica dos itens de Biologia das edições de 2009 a 2019. Foi possível notar que mais da metade dos itens não foram considerados bons para aferir a proficiência que a prova de Ciências da Natureza se propõe a medir. Recomenda-se que estudos subsequentes possam realizar uma análise qualitativa dos aspectos pedagógicos dos itens inadequados.

Palavras-chave: Avaliação; Ensino de Biologia; Psicometria; Microdados; Educação em Ciências.

Abstract
Used, among other things, as a selection for Higher Education, the National High School Exam is one of the largest large-scale exams in the world. In view of the potential to influence a person's life, it is aspired that the exam is able to make a fair inference. The article publishes the results of an investigation that estimated the psychometric quality of biology items from 2009 to 2019. It was possible to notice that more than half of the items were not considered good for measuring the proficiency that the Nature Sciences test proposes to measure. It is recommended that subsequent studies can perform a qualitative analysis of the pedagogical aspects of inadequate items.

Keywords: Evaluation; Biology Teaching; Psychometrics; Microdata; Science Education.

Resumen
Utilizado como una selección para la Educación Superior, el Examen Nacional de Escuela Secundaria es uno de los exámenes a gran escala más grandes del mundo. En vista del potencial de influir en la vida de una persona, se aspira a que el examen sea capaz de hacer una inferencia justa. El artículo publica los resultados de una investigación que estimó la calidad psicométrica de los ítems de biología de 2009 a 2019. Fue posible notar que más de la mitad de los ítems no se consideraron buenos para medir la competencia que la prueba de Ciencias de la Naturaleza propone medir. Se recomienda que estudios posteriores puedan realizar un análisis cualitativo de los aspectos pedagógicos de los ítems inadecuados.

Palabras clave: Evaluación; Enseñanza de la Biología; Psicometría; Microdatos; Educación Científica.

[1] Doutor em Educação em Ciências - Universidade Federal do Rio Grande do Sul (UFRGS). Porto Alegre, RS - Brasil. Professor do Curso de Licenciatura em Ciências Naturais e do Programa de Pós-Graduação em Educação em Ciências e Matemática - Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA). Pará, Brasil. **E-mail:** patrick.fisica@hotmail.com

DOI: 10.46667/renbio.v15i1.633

***

1 Introdução

O Exame Nacional do Ensino Médio (ENEM) é tido como um dos maiores exames do mundo (TRAVITZKI, 2017). Pode ser classificado como um exame de grande escala, mesmo que a participação das pessoas seja voluntária. Concebido, inicialmente, como um instrumento de avaliação do ensino médio, sua finalidade e constituição foram se modificando com o passar das décadas. Hoje em dia, a adesão dos participantes é significativa, tendo em vista que a nota na prova pode garantir o ingresso na maioria das Instituições de Ensino Superior (IES) públicas do país e é condição para pleitear vagas e bolsas de estudo em diversas IES privadas (SOARES; SOARES; SANTOS, 2021).

Um escore satisfatório no exame pode promover diversas vantagens com potencial de impactar o presente e o futuro da pessoa, seja no contexto profissional, seja na vida social. Sendo de grande monta as possibilidades de nortear a vida de um participante, evidencia-se a importância de garantir a sua qualidade no que tange os processos de elaboração e validação.

Validade e fidedignidade são termos empregados em uma área da psicologia chamada "psicometria" que, entre outras coisas, atua nos processos que asseguram a qualidade de um instrumento de avaliação por meio de diretrizes e protocolos de aferição (HUTZ; BANDEIRA; TRENTINI, 2015).

Dentro da área da Educação e Psicologia, uma referência de grande relevância para este fim é o "*Standards for educational and psychological testing*" (AERA; APA; NCME, 2014), que integra os parâmetros psicométricos que os testes da educação e psicologia devem atingir para ser conferida a sua qualidade. Os testes em larga escala consideram uma grande quantidade de indicadores psicométricos recomendados pela literatura para aferir a qualidade dos seus instrumentos, a fim de responder se a avaliação proposta mede aquilo que se propõe.

Assim, levando em conta as três frentes: o impacto significativo que o ENEM pode ter na vida de um participante; a importância de se ter um teste válido e confiável; a existência de parâmetros que auxiliam a atestar essa qualidade, se problematiza: qual é a qualidade das questões de Biologia do ENEM?

De modo a responder tal pergunta de pesquisa, o objetivo desse estudo visou aferir a qualidade psicométrica dos itens de Biologia do ENEM das edições de 2009 a 2019. Os objetivos específicos que guiaram a pesquisa foram: baixar os Microdados do ENEM; filtrar as informações referentes aos itens de Biologia de cada ano; analisar a qualidade psicométrica dos itens e provas; e classificar cada questões de acordo com a sua qualidade.

DOI: 10.46667/renbio.v15i1.633

Um item de Biologia é toda aquela questão da prova de Ciências da Natureza que aborda algum assunto de Biologia, ou ainda, o item que, para escolher a resposta certa, o participante necessite usar saberes da área da Biologia.

Defende-se ser fundamental estudos voltados para esse fim, pois é essencial a busca de meios que assegurem a qualidade de uma prova da importância e do tamanho que é o ENEM. Ademais, usar os Microdados de cada edição possibilita com que se realize inferências a partir de dados reais, ou seja, avalia-se a qualidade do exame a partir de uma análise de comportamento empírico dos itens, respondidos pelos participantes de cada ano. Isso tudo serve a um propósito superior que visa sustentar que o exame contenha de uma qualidade mínima, que consiga promover uma seleção justa, que cumpra o seu objetivo fim de maneira adequada.

Em pesquisas nacionais é possível verificar a existência de diferentes estudos com esse escopo, podendo destacar pesquisas como as de Travitzki (2017) que analisa itens do ENEM do ano de 2009 a 2011 e reflete sobre o conceito de qualidade psicométrica; Pontes Junior et al. (2016) que verificam a qualidade de questões de educação física das edições de 2009 a 2013; Gomes, Golino e Perez (2020), onde estudam a fidedignidade dos escores das edições de 2011; Gonçalves e Almeida (2018) que analisam a dificuldade e discriminação de itens de matemática do ano de 2012; e por fim, destaca-se a pesquisa de Soares, Soares e Santos (2021), onde estudaram diferentes medidas de tendência central dos itens de 2016 a 2018.

Em suma, nota-se, das produções já existentes, o propósito de pesquisadores em apurar particularidades envolvendo a qualidade do ENEM. Assim, acredita-se que uma pesquisa com o foco que está sendo dado aqui, poderá contribuir com a literatura da área ao fornecer análises de itens de Biologia e promover o debate sobre a qualidade do ENEM também para esta área das Ciências da Natureza.

Na sequência, apresenta-se as bases que deram suporte teórico para a pesquisa.

2 Fundamentação teórica

2.1 Análise da qualidade de um teste

Para atestar a qualidade de uma prova como o ENEM a literatura da área pode se basear em critérios estabelecidos por diferentes referenciais teóricos e metodológicos. As recomendações referendadas ao redor do mundo recorrem a uma área da psicologia denominada psicometria (HUTZ; BANDEIRA; TRENTINI, 2015), (PASQUALI, 2017), esfera que estuda aspectos intrínsecos das pessoas. Por isso, o faz através de indicadores que representam tais constructos. De maneira prática, entre outras coisas, a psicometria se ocupa da elaboração e validação de testes para os mais diversos fins. Esses saberes são transpostos para o contexto educacional e, na literatura, é possível encontrar uma série de produções que visam abordar as normatizações com vistas a assegurar a qualidade de um teste nos mais diferentes setores da educação (AERA; APA; NCME, 2014).

DOI: 10.46667/renbio.v15i1.633

Das técnicas mais antigas às mais modernas, um conjunto de análises e procedimentos são realizados para determinar os parâmetros das provas e dos itens. Considera-se que a qualidade dos dados coletados por um questionário, um teste, uma prova etc., dependem, parcialmente, das suas características psicométricas. Normalmente essas características são aferidas por procedimentos estatísticos sofisticados.

2.2 Fidedignidade e Validade

Em suma, a literatura sempre aborda dois conceitos fundamentais no que tange a avaliação da avaliação: a validade e a confiabilidade. Por validade, entende-se a característica que atesta se o instrumento em questão mensura aquilo que se destina a medir. No que lhe concerne, a confiabilidade busca verificar a hipótese de que, se o mesmo instrumento fosse aplicado a uma mesma pessoa em intervalos de tempo diferentes, esta, apresentaria desempenhos estatisticamente semelhantes. Ou seja, validade refere-se mais a atestar a coerência entre a medida e o seu objetivo a medir e a confiabilidade, relaciona-se mais com a reprodutibilidade do instrumento (HUTZ; BANDEIRA; TRENTINI, 2015), (PASQUALI, 2017).

Um dos testes mais utilizados para aferir a confiabilidade é o coeficiente Alfa de Cronbach, pois ele considera as correlações positivas entre os itens de um teste, enquanto considera no seu cálculo o número de itens (HUTZ; BANDEIRA; TRENTINI, 2015). Em geral, a confiabilidade analisa o teste por inteiro, enquanto há outros indicadores psicométricos que analisam item a item.

Já, para aferir a validade de um instrumento ou de um conjunto de itens, busca-se uma avaliação a partir de especialistas da área e do público-alvo para o qual o teste se destina (HUTZ; BANDEIRA; TRENTINI, 2015). Há uma série de procedimentos para a aferição da validade, sendo o índice de validade de conteúdo (VILARINHO, 2015) e o coeficiente de validade de conteúdo (SANTOS, 2018), alguns deles. Basicamente, quando se cria um instrumento, os especialistas e o público-alvo avaliam a qualidade de cada item em aspectos como a adequação semântica, o entendimento, a interpretação, as alternativas de respostas, entre outros. Os melhores itens são classificados e os piores, eliminados. Os que ficam, participam de etapas posteriores, que podem ser testes piloto ou outros procedimentos estatísticos.

2.3 O suporte teórico usado

Diferentes métodos para averiguar o comportamento dos itens podem ser empregados para a avaliação. Existem duas categorias comumente utilizadas para analisar a qualidade de testes e itens: a Teoria Clássica de Testes (TCT) e a Teoria de Resposta ao Item (TRI).

DOI: 10.46667/renbio.v15i1.633

A TCT busca mensurar o desempenho de um participante a partir do total de acertos em um teste, ou seja, considera-se o escore final para determinar as características de qualidade. Para conferir qualidade ao teste é necessário considerar o instrumento na totalidade. Tais procedimentos seguem pressupostos como a imposição de que itens iguais sejam respondidos por todos os participantes, em iguais circunstâncias e com uma quantidade mínima de respondentes (SOARES; SOARES; SANTOS, 2021).

No entanto, apenas com o escore bruto não é possível realizar uma comparação adequada entre indivíduos que obtiveram bons escores em uma prova fácil e pessoas com baixos desempenhos em uma prova difícil, por exemplo. Em função de limitações como essa, mostra-se necessária a existência de indicadores que aferem a dificuldade dos testes e dos itens, bem como, a capacidade de o item discernir aqueles participantes que possuem, daqueles que não possuem o conhecimento averiguado no teste, indicador chamado discriminação.

Já a TRI, no contexto educacional, busca averiguar a proficiência da pessoa para determinado conhecimento e não apenas um escore, como o total de respostas corretas (SOARES; SOARES; SANTOS, 2021). De modo geral, é medido um traço latente (habilidade), aspecto intrínseco que se manifesta por meio da resposta de alguns itens e representam indicadores desse traço latente. Ou seja, a TRI se fundamenta na hipótese de a pessoa acertar ou errar uma questão segundo o seu conhecimento sobre o assunto (TRAVITZKI, 2017).

A TRI não substitui a TCT, elas fazem parte de procedimentos complementares, com seus potenciais e suas limitações. De modo geral, indicadores iguais, aferidos independentemente pelas duas teorias, tendem a apresentar um comportamento semelhante, como os índices de dificuldade e discriminação do item, por exemplo (SOARES; SOARES; SANTOS, 2021). A superioridade da TRI, entre outras coisas, centra-se na possibilidade de comparar proficiências de diferentes pessoas e grupos através de diferentes instrumentos (TRAVITZKI, 2017).

O ENEM, até o ano de 2008, recorria à TCT. A partir de 2009, o exame passou por uma reformulação e a TRI veio em substituição como método de validação dos itens e estimação do desempenho dos participantes. O modelo usado no ENEM é o de 3 parâmetros, em que é descrita a probabilidade de um respondente acertar uma questão dependendo da sua proficiência. Esse modelo, como o nome diz, tem como condição conhecer 3 parâmetros sobre o item: o parâmetro “a” corresponde à discriminação do item; o “b” é a sua dificuldade; e o “c” corresponde a chance de uma pessoa com baixo desempenho responder à questão de maneira correta. Para cada item, espera-se que o parâmetro “a” seja alto. Já para o parâmetro “b” deseja-se que a distribuição seja equilibrada, não tendo itens fáceis nem difíceis como maioria; em simultâneo, deseja-se que o parâmetro “c” seja o menor possível (< 20% para um item com 5 alternativas de resposta) (TRAVITZKI, 2017).

DOI: 10.46667/renbio.v15i1.633

Por fim, considera-se a TCT e a TRI como ferramentas que podem auxiliar a responder à pergunta de pesquisa. Assim, a seção seguinte irá expor quais testes serão empregados para aferir a qualidade dos itens de Biologia, bem como, que critérios devem ser usados para interpretar os resultados.

3 Metodologia

3.1 Caracterização da pesquisa

Esta é uma pesquisa de abordagem quantitativa, de natureza básica, com objetivos de pesquisa descritiva e exploratória, que investiga bancos de dados e os analisa através de técnicas de estatística (ROBAINA et al., 2021; GIL, 2008).

3.2 Os Microdados do ENEM

Os Microdados consistem em arquivos que contém informações sobre as mais diversas avaliações ou pesquisas. Podem conter informações sobre as questões, os desempenhos, informações de caracterização social dos participantes, as alternativas assinaladas etc. Qualquer cidadão pode ter acesso aos Microdados por meio do site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem). Os arquivos são compilados de modo a serem acessados através de determinados softwares estatísticos. No contexto nacional, o INEP é o maior compilador de Microdados voltados para pesquisas na educação, gerando informações sobre o ENEM, a Prova Brasil, o censo da Educação Superior, o Censo Escolar, o PISA (brasileiro), entre outros (TRAVITZKI, 2017).

3.3 Recorte temporal

Os dados analisados foram obtidos a partir dos Microdados do ENEM do ano de 2009 até 2019. O recorte temporal escolhido para a pesquisa se justifica porque, a partir de 2009, o ENEM passou por uma reformulação de grande monta, onde algumas das mudanças foram: a inserção de uma nova matriz de referência; a mudança na estrutura da prova; o acréscimo da TRI como metodologia de validação e geração da nota final dos candidatos; entre outros, como: o aumento do uso do exame para o ingresso no Ensino Superior público, financiamento estudantil e bolsas de estudo para instituições privadas, etc.

3.4 Os indicadores usados

A afim de estabelecer critérios para aferir a qualidade das provas e dos itens, os seguintes indicadores foram utilizados. Essa seção apresentará um breve conceito sobre o teste e a diretriz sobre como ele deve ser interpretado.

DOI: 10.46667/renbio.v15i1.633

Índice de Dificuldade: uma das maneiras de verificar a dificuldade de um item, via TCT, é observar a média de acertos dessa questão. Itens com baixa média de acertos são considerados difíceis, enquanto itens com alta taxa de acertos, são considerados fácies. De acordo com Vilarinho (2015), questões com dificuldade igual ou superior a 0,7 podem ser consideradas fáceis; entre 0,7 e 0,3 são de dificuldade média; e abaixo de 0,3 devem ser consideradas difíceis. Já quando este indicador é observado a partir da TRI, por meio do parâmetro "b", há autores que defendem que os valores adequados para dificuldade do item devem estar entre -3 e +3 (SOARES; SOARES; SANTOS, 2021) ou -4 e +4 (TRAVITZKI, 2017).

Índice de Discriminação: Afere o quanto um item consegue diferenciar pessoas com diferentes níveis de proficiência. Em uma avaliação sobre um determinado constructo, indivíduos com habilidade devem alcançar pontuações diferentes daquelas pessoas sem habilidade. Este índice será analisado nesta pesquisa por meio do parâmetro "a" da TRI. Para interpretá-lo, será usada a seguinte classificação: discriminação boa se o parâmetro "a" for maior ou igual a 0,5; discriminação duvidosa se for entre 0,2 e 0,5; e discriminação ruim se for abaixo de 0,2 (TRAVITZKI, 2017). Índices de discriminação negativos poderiam sinalizar que a probabilidade de responder corretamente uma questão diminui com o aumento da proficiência (PONTES JUNIOR et al., 2016). Baixo valor de discriminação significa que, tanto pessoas com baixa habilidade quanto as com alta habilidade, tem probabilidades semelhantes de acertarem o item. Quanto maior for o valor de discriminação da questão, maior será a contribuição dela para a medida da habilidade (também chamada "proficiência").

Coeficiente Alfa de Cronbach: é uma medida de confiabilidade do teste. Relaciona a correlação entre cada item e a prova toda. Ou seja, o quanto cada questão contribui para a prova, de modo geral. A sua premissa considera que os itens de uma prova são formas paralelas de se mensurar o conhecimento que se deseja medir. Para interpretá-lo, o teste gera um valor entre 0 e 1, sendo os valores mais próximos de 1, indicativos de maior consistência interna. Não há consenso sobre um valor mínimo para atribuir uma consistência como satisfatória. Comumente, usa-se valores a partir de 0,5, 0,6, 0,7. Para esta pesquisa se usará o valor de 0,6, conforme indicado por estudos da área (HUTZ; BANDEIRA; TRENTINI, 2015), (PASQUALI, 2017). Esse coeficiente foi calculado tendo como base as 45 questões de Ciências da Natureza de cada edição.

Correlação item total: é uma medida do item. Indica o quanto ele se correlaciona com o instrumento integral. Itens com alto nível desse indicador sugerem que contribuem fortemente para a consistência interna do instrumento. Por outro lado, itens com baixo valor, provavelmente estão contribuindo para uma menor de consistência interna do instrumento. Em uma validação de questionário, geralmente sugere-se que, itens com baixo valor de correlação item total, sejam eliminados do instrumento, pois, sem eles, a confiabilidade geral tende a melhorar (TRAVITZKI, 2017). A correlação item total foi calculada incluindo todos os 45 itens da prova de Ciências da Natureza.

DOI: 10.46667/renbio.v15i1.633

Coeficiente de correlação bisserial: verifica se aquelas pessoas que obtiveram um desempenho satisfatório no teste, tenderam a assinalar as alternativas corretas dos itens. O cálculo desse indicador acontece através de uma análise de correlação entre o escore das pessoas e a taxa de escolha de cada alternativa. Espera-se que o valor de correlação seja positivo para a alternativa correta e negativo para todas as outras alternativas erradas (também chamadas "distratores"). Esse coeficiente tem potencial de identificar itens com problemas em sua formulação ou com erros no gabarito, pois se a correlação for negativa para a alternativa correta, pode sugerir que o item teve mais acertos vindos de pessoas com baixo desempenho. Ao mesmo tempo, se algum distrator tiver correlação positiva, sinaliza que, por alguma razão, tal alternativa está atraindo mais o grupo de pessoas com bom desempenho. Isso pode configurar um problema de discriminação do item, sendo sugerida a sua remoção ou adequação. Para interpretar esse indicador, nessa pesquisa, será considerado um item bom se a alternativa correta apresentar uma correlação bisserial positiva e igual ou acima de 0,3; duvidoso se o valor for entre 0,15 e 0,30; e ruim se o índice for abaixo de 0,15. Também será classificado como duvidoso o item que apresentar valores positivos para qualquer distrator e/ou se a alternativa correta demonstrar o coeficiente com valor negativo (TRAVITZKI, 2017).

Ajuste do modelo: é um indicador que mostra se o modelo de 3 parâmetros da TRI (empregado no ENEM) se ajusta aos dados analisados, fornecendo as informações de discriminação, dificuldade e potencial de acerto casual de maneira adequada. De modo geral, diz-se que um ajuste não satisfatório não consegue garantir que os parâmetros obtidos sejam invariantes. Ou seja, um modelo com um ajuste inadequado não fornecerá essas informações com precisão, prejudicando a estimativa dos parâmetros, bem como, da proficiência analisada. O teste de ajuste do modelo gerará um p-valor para cada item, que deve ser estatisticamente significativo para sinalizar um ajuste adequado do modelo. Ou seja, o p-valor deve ser menor que 0,05. Itens com p-valor abaixo de 0,05 são considerados bons; já quando o p-valor estiver no intervalo entre 0,05 e 0,10, o item será definido como duvidoso; por fim, um p-valor acima de 0,10 representa um item com ajuste considerado não adequado (TRAVITZKI, 2017).

3.5 Classificação dos itens

De modo a analisar o comportamento empírico das questões de Biologia do ENEM e classificá-las dentro de um padrão de qualidade, usou-se alguns dos indicadores anteriormente mencionados. A classificação aplicada nesta pesquisa seguirá a mesma estrutura proposta por Travitzki (TRAVITZKI, 2017).

DOI: 10.46667/renbio.v15i1.633

Figura 1: Critérios para classificação dos itens.

	TCT		***TRI***	
Classificação do item	*Correlação item total (CORR)*	*Coeficiente bisserial (BISS)*	*Parâmetro de discriminação (A)*	*Ajuste do modelo (FIT)*
Bom	CORR ≥ 0,30	BISS ≥ 0,30	a ≥ 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 ≤ 0,15	BISS ≤ 0,15	a ≤ 0,2	FIT ≥ 0,10

Fonte: Travitzki (2017).

Após a análise de cada questão se reunirá os itens em uma classificação global, baseada na proposta de Travitzki (2017), que os rotulará em: Item duvidoso (indicador global): quando o item for considerado duvidoso em pelo menos 3 dos 4 indicadores, ou ruim em pelo menos 1. Item ruim (indicador global): quando for ruim em pelo menos 2 dos 4 indicadores. Se os itens não se encaixarem nesses parâmetros, então significa que foram considerados bons.

Adicionalmente, para algumas discussões desta pesquisa, se classificará os itens como adequados e não adequados. Os itens adequados são aqueles considerados como bons e os não adequados, por sua vez, são todos aqueles sendo classificados como duvidosos ou como ruins.

3.6 Software usado e processo de amostragem

Para acessar os Microdados do ENEM e calcular os indicadores psicométricos necessários utilizou-se o software R (estatística) (R CORE TEAM, 2018), com os pacotes "*mirt*" (CHALMERS, 2012) para os cálculos da TRI e o "*psych*" (REVELLE, 2014) para as análises via TCT. Tabelas a gráficos foram elaborados através do Microsoft Excel.

Ao somar o número de inscritos no ENEM nas 11 edições analisadas chegou-se ao número de 69.556.698 inscritos. Essa quantidade, demasiadamente grande, mostra a necessidade de se realizar um recorte amostral nos dados. De modo a definir critérios para esse recorte, foram aplicados os seguintes filtros para analisar somente os dados daqueles que: 1) eram participantes da primeira aplicação; 2) responderam a todos os itens; 3) estudaram em escola regular; 4) concluíram ou estavam concluindo o Ensino Médio no ano em que fizeram a prova; 5) estiveram presentes em todas as provas da edição; 6) receberam o caderno azul.

Assim, conforme pode ser analisado na figura abaixo, o recorte acarretou uma redução substancial do número de dados analisados, fixando a amostra a um total de 6.981.815 participantes.

DOI: 10.46667/renbio.v15i1.633

Figura 2: Número de inscritos e amostra selecionada para cada edição.

Ano	*Inscritos*	*Amostra*	*Ano*	*Inscritos*	*Amostra*	*Ano*	*Inscritos*	*Amostra*
2019	5.095.270	512.707	2015	7.746.427	336.720	2011	5.380.856	817.224
2018	5.513.747	543.571	2014	8.722.248	1.030.138	2010	4.626.094	727.174
2017	6.731.341	305.701	2013	7.173.563	1.011.596	2009	4.148.720	571.247
2016	8.627.367	349.597	2012	5.791.065	776.140			

Fonte: Microdados do INEP.

Salienta-se que, no momento de conclusão dessa investigação, os Microdados do ENEM do ano de 2020 infelizmente ainda não haviam sido disponibilizados pelo INEP. Destaca-se também que outros indicadores poderiam ter sido empregados na análise da qualidade dos itens, sendo as opções aqui usadas, fruto de um recorte. Não foram identificados, conforme os gabaritos disponibilizados nos Microdados, questões de Biologia que tenham sido anuladas. Por fim, também se destaca que, na apresentação dos resultados, optou-se por não realizar uma análise pedagógica dos itens, devido ao tamanho do manuscrito, sendo esse passo, uma sugestão fortemente recomendada para pesquisas futuras.

Na sequência, a seção dos resultados apresentará as características psicométricas dos itens de Biologia.

4 Resultados

4.1 Fidedignidade e dificuldade

Abaixo, é possível visualizar, para cada ano, os índices de confiabilidade e dificuldade das provas de Ciências da Natureza.

DOI: 10.46667/renbio.v15i1.633

Figura 3: Confiabilidade e dificuldade das edições analisadas.

Ano	*Alfa de Cronbach*	*Dificuldade da prova de CN*	*Ano*	*Alfa de Cronbach*	*Dificuldade da prova de CN*	*Ano*	*Alfa de Cronbach*	*Dificuldade da prova de CN*
2009	0,720	0,35	*2013*	0,589	0,26	*2017*	0,625	0,27
2010	0,752	0,33	*2014*	0,640	0,28	*2018*	0,580	0,25
2011	0,746	0,33	*2015*	0,658	0,27	*2019*	0,601	0,28
2012	0,764	0,30	*2016*	0,664	0,26			

Fonte: Autor.

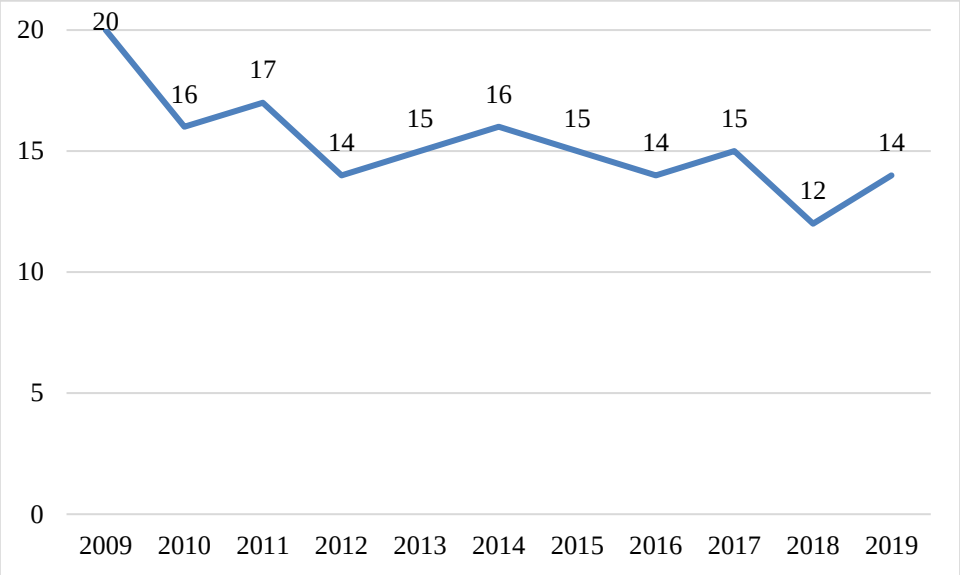
Nas provas de 2013 e 2018, a consistência interna das provas ficou inferior ao recomendado (< 0,60). Com relação a dificuldade das provas, observou-se que todas elas puderam ser classificadas como difíceis, tendo em vista que os índices desse indicador ficaram entre 0,25 e 0,45 (TRAVITZKI, 2017). Além disso, notou-se que a edição de maior dificuldade em Ciências da Natureza foi a de 2018, e, no outro espectro, a de menor dificuldade, foi a de 2009.

4.2 Distribuição das questões de acordo com a qualidade psicométrica

No início, apresenta-se, na figura a seguir, o número de itens de Biologia para as 11 provas do ENEM analisadas.

DOI: 10.46667/renbio.v15i1.633

Figura 4: Itens de Biologia por edição do ENEM.



Fonte: Autor.

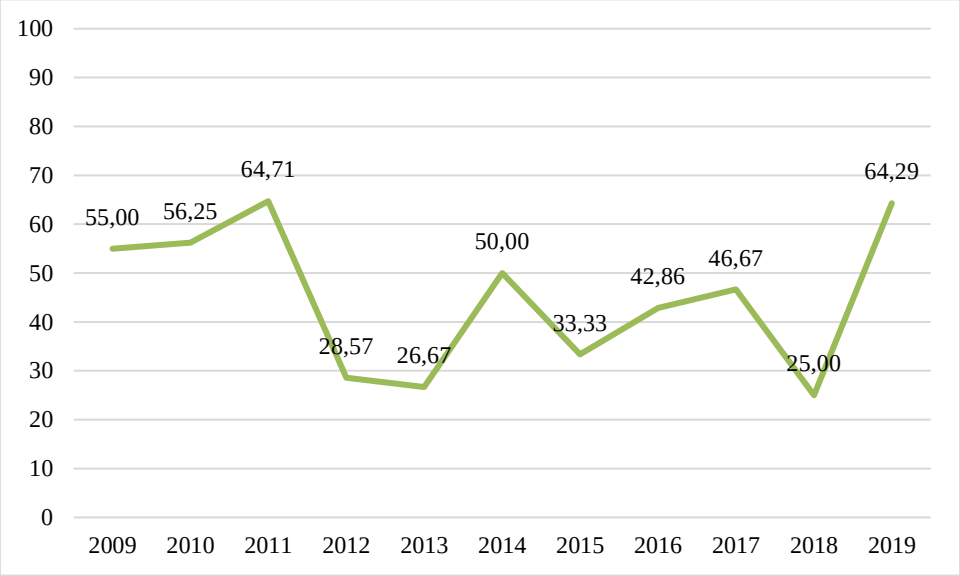
É possível notar que, em sua maioria, as provas de Ciências da Natureza tendem a distribuir as questões de Biologia, Química e Física de maneira proporcional. No entanto, observa-se alguns anos em que esse padrão não é observado, como em 2009 e 2018, onde, respectivamente se teve a maior e a menor quantidade de itens de Biologia.

Os critérios psicométricos aplicados na análise permitiram rotular os itens de Biologia em: itens bons, itens duvidosos e itens ruins. Considerando todas as edições pesquisadas, as provas de Ciências da Natureza tiveram 168 itens de Biologia. De acordo com a classificação empregada 77 (45,83%) foram rotulados como bons; 69 (41,07%) classificados como duvidosos; e 22 (13,10%) considerados ruins. Ou seja, chega-se ao percentual de que 54,17% das questões de Biologia do ENEM das edições de 2009 até 2019 não atendem às especificações de qualidade psicométrica esperadas (AERA; APA; NCME, 2014).

Na figura a seguir o percentual de itens de Biologia considerados bons são apresentados em uma análise longitudinal.

DOI: 10.46667/renbio.v15i1.633

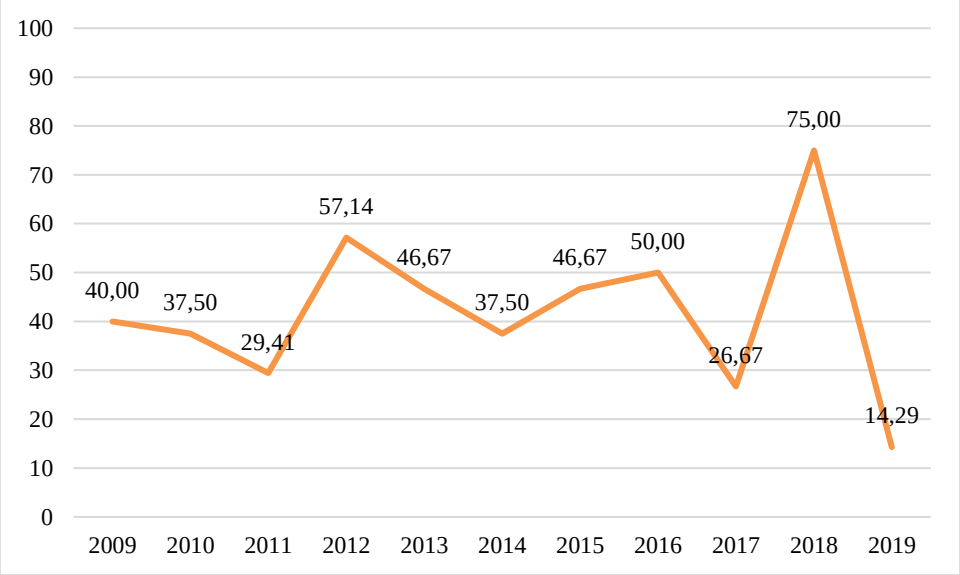
Figura 5: Percentual de itens bons.



Fonte: Autor.

Destaca-se os anos de 2012, 2013 e 2018 pois, o percentual de itens bons não chegou nem a 30% do total de questões de Biologia daquelas edições. Na próxima figura mostra-se a distribuição dos itens duvidosos.

Figura 6: Percentual de itens duvidosos.



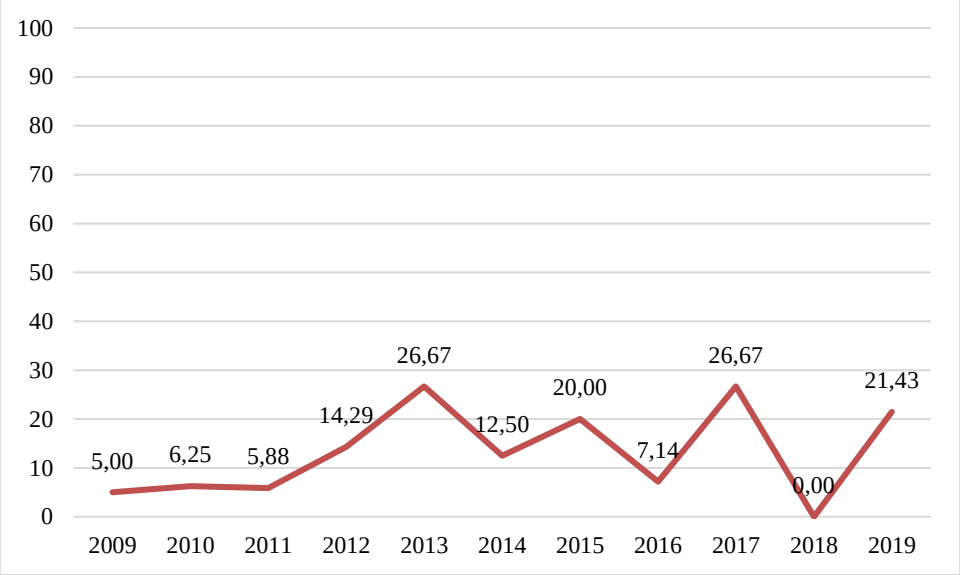
Fonte: Autor.

Já os itens duvidosos também foram numerosos no decorrer das edições. Os destaques dessa categoria vão para os anos de 2012 e 2018, onde mais de 50% das questões de Biologia encaixaram-se nesse grupo.

DOI: 10.46667/renbio.v15i1.633

Por sua vez, as questões ruins podem ser analisadas na figura a seguir.

Figura 7: Percentual de itens ruins.

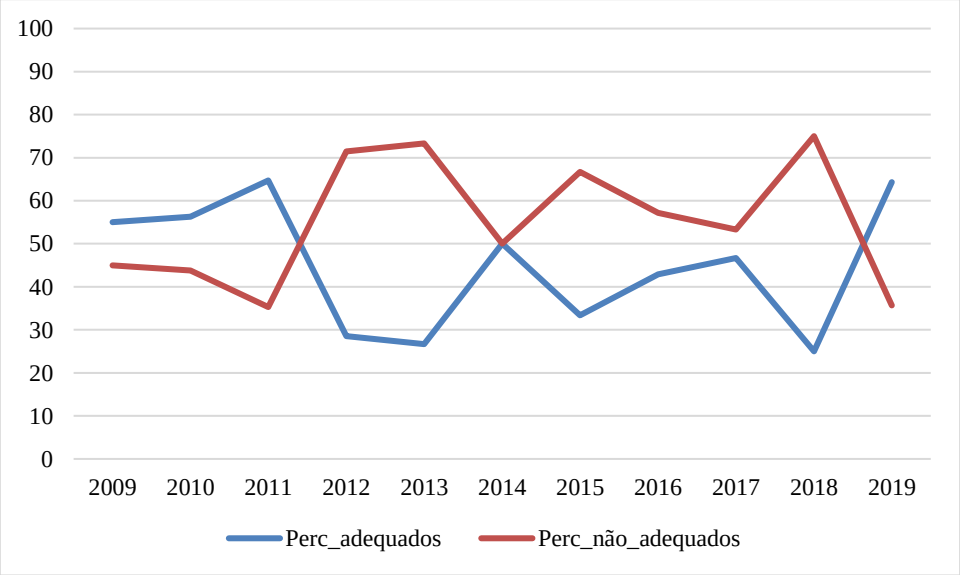


Fonte: Autor.

Finalmente, os itens ruins compuseram mais de 1/4 dos itens de Biologia nas edições de 2013 e 2017. Um destaque positivo foi para o ano de 2018, onde não se identificou itens com essa classificação.

Para resumir a análise, a próxima figura mostra a porcentagem de questões julgadas como adequadas (itens bons) e não adequadas (itens duvidosos e ruins).

Figura 8: Percentual de itens adequados e não adequados.



Fonte: Autor.

DOI: 10.46667/renbio.v15i1.633

As provas de 2012, 2013 e 2018 se destacam pois são edições em que as questões não adequadas tendem a ser maioria, uma vez que, mais de 70% dos itens de Biologia pertencem a esse grupo.

4.3 Disposição das questões conforme sua qualidade

Após uma análise geral, se faz necessário dispor os itens de Biologia de acordo com a classificação que receberam. Assim, a figura abaixo apresenta, por ano do ENEM, quais questões foram rotuladas como boas, duvidosas e ruins.

Figura 9: Questões conforme sua classificação.

Edição	Itens bons	Itens Duvidosos	Itens 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
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

Fonte: Autor.

DOI: 10.46667/renbio.v15i1.633

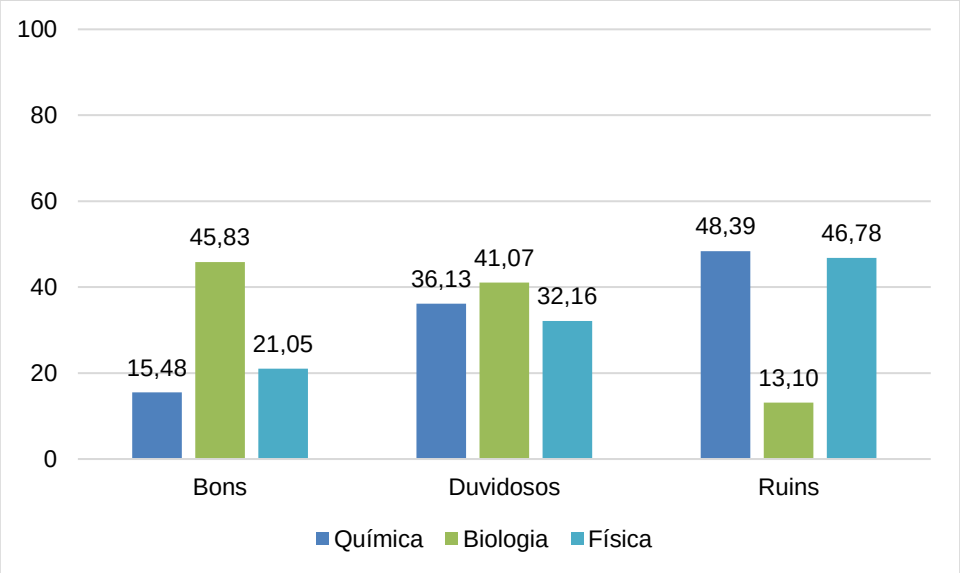
5 As características dos itens de Biologia do ENEM

Deseja-se que a distribuição apresentada na seção anterior possa ser útil para o avanço do estado da arte da área, auxiliando professores que utilizem itens do ENEM em suas práticas, bem como, para pesquisadores que investigam avaliações em larga escala. Um item, em uma avaliação em larga escala, tem potencial de ser útil à avaliação se cumprir com o seu objetivo de aferição. Conforme debatido anteriormente, uma das maneiras de conferir qualidade a uma questão é analisar as suas características de validade e confiabilidade.

De modo geral, é possível perceber que são fundamentais pesquisas posteriores que busquem aprofundar a avaliação dos itens, principalmente, aqueles classificados como não adequados, uma vez que, nesse manuscrito, devido à quantidade de questões, uma análise pedagógica dos mesmos tornou-se inviável. Não obstante, alguns pontos importantes para discussão desses resultados podem ser ressaltados. Em um primeiro momento, acredita-se ser importante contrapor a categorização das questões de Biologia com a mesma classificação feita para itens da Física e Química.

Uma vez que a análise foi realizada para todos os 45 itens de cada edição, sendo o recorte aqui apresentado, voltado apenas para as questões de Biologia, dispõem-se dos dados de categorização para os demais itens também. Para fundamentar a reflexão dessa seção, essa comparação geral foi realizada e pode ser observada na figura a seguir.

Figura 10: Classificação de qualidade dos itens das Ciências da Natureza para as edições de 2009 até 2019.



Fonte: Autor.

DOI: 10.46667/renbio.v15i1.633

Os itens de Biologia foram os que mais obtiveram classificações boas dentre as três disciplinas. Por outro lado, quando comparados o percentual de itens duvidosos, as questões de Biologia ficam ligeiramente na frente, sendo seguida pela Química e pela Física. Finalmente, quando comparada a categoria de itens rotulados como ruins, a Biologia mostrou-se com a menor quantidade de itens nessa classificação. Quase metade dos itens da Química e da Física encaixam-se nessa classificação, enquanto que, para a Biologia, essa percentagem fica em 13%.

Os desfechos observados nessa pesquisa mostram que é imperativo investigações adicionais que tenham como foco examinar os itens ruins e duvidosos, de modo a proceder com análises pedagógicas dessas questões, dentro de uma perspectiva qualitativa.

É importante destacar que o INEP não disponibiliza, nos Microdados, informações sobre os três parâmetros de cada item. Ou seja, os próprios pesquisadores precisam estimar essas informações, via TRI, a partir do índice de acertos de cada participante. Isso significa dizer que pode haver uma ligeira diferença entre os valores aqui observados e aqueles empregados pelo INEP para calcular a nota de cada pessoa.

Levando em conta os resultados obtidos nessa investigação e o rigor metodológico da TRI, questiona-se: de que maneira a qualidade dos itens de Biologia do ENEM podem ser melhorados? Em que esfera as melhorias devem ocorrer? Na elaboração dos itens? Na validação? Deve haver um maior pré-teste? Deve-se investir na formação técnica dos elaboradores das questões? De modo geral, os horizontes para dar continuidade a investigações com esse escopo são diversos, sendo fortemente recomendada a ampliação de investigações voltadas para a qualidade do ENEM.

É relevante destacar o comprometimento de autarquias como o INEP, pois protagonizam uma função substancial ao proporcionar acesso aos Microdados das avaliações em larga escala do Brasil. Tal ação auxilia pesquisadores a refletirem e colocarem em prática pesquisas com problemáticas voltadas para pensar a relevância, função, dificuldades e desafios de exames da magnitude e importância do ENEM em nosso país.

Por fim, salienta-se a relevância que o ENEM tem para cada cidadão que deseja ingressar no Ensino Superior. Com isso, destaca-se que investigações voltadas para avaliar a qualidade desse exame não tem como intenção desqualificar a prova e os seus princípios metodológicos. Ao invés disso, busca-se, ao reconhecer o valor do exame para a sociedade, promover reflexões sobre o seu funcionamento e maneiras de superar os obstáculos de uma implementação com qualidade.

DOI: 10.46667/renbio.v15i1.633

6 Considerações finais

Analisou-se os resultados de uma investigação que aferiu aspectos da qualidade psicométrica do ENEM. Constatou-se que, 54,17% dos itens de Biologia das edições de 2009 a 2019, não puderam ser classificados como itens adequados à prova de Ciências da Natureza.

Salienta-se que uma análise psicométrica mensura unicamente um recorte dos Microdados disponibilizados pelo INEP, enquanto estima, de forma autônoma, os parâmetros da TRI empregados na aferição.

Isto posto, destaca-se que a presente pesquisa não tem como meta concluir o debate. Pelo contrário, visa incentivar a importância de estudos subsequentes que deem continuidade a investigações como esta, de modo a replicar as observações aqui exibidas, ao mesmo tempo em que, busquem analisar os itens a partir de diferentes perspectivas metodológicas, como em uma análise qualitativa, por exemplo.

Referências

American Educational Research Association (AERA); American Psychological Association (APA); National Council on Measurement in Education (NCME). **Standards for educational and psychological testing.** Washington, DC: APA, 2014.

CHALMERS, R. Philip. Mirt: A multidimensional item response theory package for the R environment. **Journal of statistical Software**, v. 48, n. 1, p. 1-29, 2012.

Gil, Antônio Carlos. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2008.

GOMES, Cristiano Mauro Assis; GOLINO, Hudson Fernandes; DE SOUZA PERES, Alexandre José. Fidedignidade dos escores do Exame Nacional do Ensino Médio (Enem). **Psico**, v. 51, n. 2, p. e31145, 2020.

HUTZ, Claudio Simon; BANDEIRA, Denise Ruschel; TRENTINI, Clarissa Marceli. **Psicometria.** Porto Alegre: Artmed, 2015.

PONTES JUNIOR, José Airton; SILVA, Ana Géssica; TAVARE, Erisvan; ARAUJO, Leandro Sousa; BASTOS, Fernando Cunha.; CRUZ, Francisca Nimara Inácio; ALMEIDA, Leandro Silva. Aspectos psicométricos dos itens de Educação Física relacionados aos conhecimentos de Esporte e Saúde no Exame Nacional do Ensino Médio (Enem). **Motricidade,** v. 12, n. 1, p. 12-21, 2016.

PASQUALI, Luiz. **Psicometria:** teoria dos testes na psicologia e na educação. Petrópolis: Vozes, 2017.

DOI: 10.46667/renbio.v15i1.633

PITON-GONÇALVES, Jean; ALMEIDA, André Marcos. Análise da dificuldade e da discriminação de itens de Matemática do ENEM. **REMAT: Revista Eletrônica da Matemática,** v. 4, n. 2, p. 38-53, 2018.

R Core Team. **R: a language and environment for statistical computing.** Vienna: R Foundation for Statistical Computing, 2018.

REVELLE, William. **Psych: procedures for personality and psychological research**. R package version 1.4.3. 2014. Disponível em: http://cran.r-project.org/web/packages/psych/psych.pdf. Acesso em: set. 2021.

ROBAINA, José Vicente Lima; FENNER, Roniere dos Santos; MARTINS, Léo Anderson Meira; BARBOSA, Renan de Almeida; SOARES, Jeferson Rosa (Org.). **Fundamentos teóricos e metodológicos da pesquisa em educação em ciências.** Curitiba: Bagai, 2021.

SANTOS, Fernando de Almeida. Redução da escala tendência empreendedora geral (TEG-FIT) a partir do coeficiente de validade de conteúdo (CVC) e teoria da resposta ao item (TRI). **Revista Eletrônica de Ciência Administrativa**, v. 17, n. 2, p. 192-207, 2018.

SOARES, Talita Emidio Andrade; SOARES, Denilson Junio Marques; DOS SANTOS, Wagner. Medidas de Tendência Central: Análise da Qualidade das Questões do ENEM de 2016 a 2018. **Jornal Internacional de Estudos em Educação Matemática**, v. 14, n. 1, p. 119-128, 2021.

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

VILARINHO, Ana Paula Lima. **Uma proposta de análise de desempenho dos estudantes e de valorização da primeira fase da OBMEP.** 2015. 98 f. Dissertação (Mestrado Profissional em Matemática). Programa de Mestrado Profissional em Matemática em Rede Nacional, Universidade de Brasília, Brasília, 2015.

Recebido em setembro de 2021.
Aprovado em março 2022.

Revisão gramatical realizada por: Joana Furini.
E-mail: joanatextos@gmail.com