# Análise da qualidade de uma prova de matemática do Exame Nacional do Ensino Médio

Sônia Ferreira Lopes Toffoli[1]
http://orcid.org/0000-0002-4841-4670

## Resumo

O Exame Nacional do Ensino Médio (Enem) é uma avaliação em larga escala que exerce forte influência nas políticas educacionais, nos currículos nos diversos níveis de ensino e também no futuro dos indivíduos que estão sendo avaliados. Esses fatos evidenciam a importância de examinar as diversas variáveis envolvidas na construção, aplicação e pontuação da avaliação. Pode-se, por meio de estudos, identificar etapas do processo que não estão funcionando como esperado e também validar as etapas cujos resultados são adequados. Em ambos os casos, busca-se a melhoria dos processos a cada edição do exame. Neste estudo, foram realizadas análises dos itens da prova de matemática do Enem 2015, por meio de dois modelos da Teoria de Resposta ao Item (TRI), o logístico de três parâmetros com estudos sobre a discriminação, a dificuldade e a probabilidade de acerto casual e o modelo de resposta nominal, com estudos da dificuldade e do comportamento de cada uma das opções de resposta dos itens. Aproximadamente um terço dos itens obteve parâmetros dentro dos padrões de qualidade. Quanto à dificuldade, o item considerado o mais fácil obteve apenas 64,1% de acertos. Além disso, apenas dois itens obtiveram o parâmetro de dificuldade situado abaixo do ponto médio da escala e aproximadamente 23,4% dos participantes tiveram suas habilidades estimadas abaixo do valor correspondente ao item mais fácil na escala, indicando que essas pessoas não souberam responder a nenhum dos 45 itens. Por meio dos diversos métodos utilizados, constatou-se que a prova de matemática do Enem 2015 possui qualidade muito aquém do desejável.

## Palavras-chave

Avaliação em larga escala – Enem – Teoria de Resposta ao Item (TRI) – Validade.

**1-** Universidade Estadual de Londrina (UEL), Londrina, PR, Brasil. Contato: sonialopes@uel.br.

DOI: http://dx.doi.org/10.1590/S1678-4634201945187128

# *Quality analysis of a Mathematics test in National Secondary Education Exam*


## *Abstract*

*The National Secondary Education exam (*Enem*) is large-scale evaluation with strong influence on educational policies, on the curricula of several levels of schooling as well as an impact on the future of the individuals taking the test. These factors show the importance of looking into a number of variants involved in the preparation, the application and the score of such evaluation. By means of studies, it is possible to identify steps in the process that are not working as expected and also to validate the steps whose results are adequate. In both cases, the intent is to improve the process in every edition of the exam. In this study, analyses were conducted for mathematics test in* Enem *2015, by means of two model from the Item Response Theory (IRT), the logistics of three parameters with studies on discrimination, difficulty and probability of casual hits and the model of nominal response, with studies on the difficulty and behavior of each response option in every item. Approximately one third of the items showed parameters within the quality standards. Regarding difficulty, the easiest item had only 64.1 percent of hits. In addition, only two items had the difficulty parameters lower than the mean point of the scale and approximately 23.4 percent of participants had their skills estimated below the value corresponding to the easiest item in the scale. This indicates that these individuals could not answer any of the 45 items. By utilizing several methods, it was found that the mathematics test in* Enem *2015 shows a quality level much lesser than desirable.*



## Keywords

*Large-scale evaluation – National Secondary Education Exam (*Enem*) – Item Response Theory (IRT) – Validity.*


## Introdução

O Exame Nacional do Ensino Médio (Enem) é de responsabilidade do Ministério da Educação, com execução pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Foi criado em 1998 com o objetivo principal de avaliar o desempenho dos alunos egressos do ensino médio e proporcionar uma avaliação nacional da educação. Em 2009, o Enem passou por uma reformulação, possibilitando também a utilização dos resultados individuais como mecanismo de seleção para o acesso à educação superior e programas de concessão de bolsas de estudos e financiamento estudantil do Governo Federal, como o Programa Universidade para Todos (ProUni) e o Programa de Financiamento Estudantil (FIES). O número de participantes do Enem tem

aumentado a cada ano. Os candidatos são motivados pela possibilidade de ingressar em uma das universidades públicas de prestígio do país ou pela oportunidade de custeio dos estudos. A edição de 2015 contou com mais de oito milhões de estudantes inscritos.

As avaliações em larga escala exercem uma forte influência sobre as políticas educacionais, sobre os currículos e sobre o futuro dos indivíduos que estão sendo avaliados. Esses fatos evidenciam a importância de examinar as diversas variáveis envolvidas nessas avaliações, aumentando, assim, a exigência de instrumentos de medidas apropriados. O desenvolvimento de instrumentos confiáveis e padronizados é um campo vasto que ainda requer muitas pesquisas e empenho dos profissionais.

É comum as avaliações em larga escala utilizarem um banco de itens para a elaboração de seus testes, que consiste em uma base de itens que relaciona os elementos de cada item como: enunciado, conteúdo medido, grau de dificuldade, entre outros (MOREIRA JUNIOR, 2011). Essas estatísticas podem ser fornecidas por métodos da Teoria Clássica de Testes (TCT), mas uma tendência mundial, que ganhou grande repercussão a partir de meados do século XX, é a Teoria de Resposta ao Item (TRI), que consiste em um conjunto de modelos matemáticos desenvolvidos para representar as relações entre as características do respondente (habilidades) e as características do item (dificuldade, discriminação e acerto ao acaso). Os modelos da TRI estabelecem que quanto maior a habilidade do indivíduo, maior será a probabilidade de acerto no item. Além disso, a TRI permite que tanto os itens como a habilidade das pessoas sejam dispostos em uma mesma escala, facilitando as análises e possibilitando estudos variados, além de acarretar uma classificação mais justa e coerente das pessoas (ANDRADE; TAVARES; VALLE, 2000).

Atualmente, os modelos da Teoria de Resposta ao Item (TRI) são utilizados em um grande número de testes internacionais, tais como o *Graduate Record Examination* (GRE), o *Test of English as a Foreign Language* (TOEFL), o *Programme for International Student Assessment* (PISA), entre outros (MOREIRA JUNIOR, 2011; VIEIRA, 2017).

Também no Brasil, a TRI tem oferecido suporte a programas de avaliação em larga escala, como o Sistema de Avaliação da Educação Básica (Saeb) (BRASIL, 2017b), o Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (Saresp) e o Exame Nacional do Ensino Médio (BRASIL, 2017a).

Os modelos da TRI que normalmente são utilizados nas avaliações escolares baseiam-se em dados dicotômicos: modelos logísticos de um, dois e três parâmetros. Nesses modelos, as respostas são categorizadas como corretas ou incorretas. No entanto, nem todas as possibilidades de respostas podem ser modeladas por um modelo dicotômico. Por exemplo, para capturar as informações de respostas na escala Likert ou atribuir notas em respostas parcialmente corretas, são necessários modelos de respostas politômicas.

Quando os itens são corrigidos de maneira dicotômica, a pontuação atribuída é 1 para a alternativa correta e 0 para todas as respostas incorretas ou distratores. Nesse contexto, um item com respostas politômicas considera a opção correta e as incorretas, podendo melhorar a estimativa da habilidade avaliada, fornecendo mais informações sobre o nível de compreensão do examinando. Os modelos politômicos permitem que cada opção de resposta ocasione em ganho de informações sobre o traço latente avaliado (BOCK, 1972; DE AYALA; SAVA-BOLESTA, 1999).

Os equívocos dos estudantes na resolução de problemas provocam variações nas respostas, e, provavelmente, as alternativas incorretas podem aumentar as oportunidades para a estimativa do traço latente que está sendo examinado, fornecendo mais informações sobre o nível de compreensão do examinando (DE AYALA; SAVA-BOLESTA, 1999). Além disso, é vantajoso poder diferenciar entre as respostas de um indivíduo que optou por uma resposta incorreta daquele que escolheu aleatoriamente qualquer uma delas. Uma visão geral dos modelos politômicos pode ser encontrada em De Ayala (2009).

O Modelo de Resposta Nominal (MRN) de Bock (1972) é indicado quando as respostas dos itens não possuem ordenação, como em testes de múltipla escolha, e também quando o teste não possui uma resposta correta, como em testes opinião ou de satisfação.

O Enem faz uso de um banco de itens pré-testados para compor as suas provas e, para isso, seleciona novos itens aplicando pré-testes de modo sistemático a alunos do ensino médio em todo o país (BRASIL, 2011). Mesmo com essa metodologia, os itens utilizados nas provas do Enem frequentemente apresentam alguns problemas, talvez ocasionados pela extensão do Brasil e pela diversidade de suas regiões e de seu povo. Em um exame com a dimensão do Enem existem variáveis que devem ser consideradas e que muitas vezes não estão diretamente relacionadas com o instrumento de medidas. Tais variáveis, implícitas ou não ao teste, podem prejudicar a validade do exame. Daí a necessidade de constantes avaliações da qualidade de seus instrumentos, principalmente pelas consequências sociais e pessoais que podem decorrer de um exame dessa natureza.

Estudos sobre as avaliações em larga escala são importantes para identificar etapas do processo que não estão funcionando como esperado e também para validar as etapas cujos resultados são adequados. Em ambos os casos, o objetivo dos estudos deve ser a melhoria dos processos a cada edição do exame (KANE, 2013; SCARAMUCCI, 2011). Esses estudos devem buscar a validade do teste como um todo.

O conceito de validade foi proposto inicialmente na década de 1920 por Kelley (1927): "um teste é válido se mede o que pretende medir"[2]. Esse conceito, assim definido, é centro de muitas críticas, principalmente por não considerar o significado ou mesmo as consequências sociais e políticas do uso desses resultados (SCARAMUCCI, 2011).

Na década de 1980, Messick (1989, p. 13) propôs uma definição de validade que atualmente é considerada o modo moderno de entender esse conceito (SCARAMUCCI, 2011). Consiste em saber se as interpretações e ações sobre os resultados dos testes são justificadas, tanto com base nas evidências científicas como nas consequências sociais e éticas da utilização de teste.

> A validade é um julgamento avaliativo integrado do grau em que as evidências empíricas e teóricas apoiam a adequação e a qualidade das inferências e ações com base nos resultados de testes ou outros meios de avaliação[3].

**2-** *A test is valid if it measures what it purports to measure* (KELLEY, 1927, p. 14, tradução nossa).

**3-** *Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment* (MESSICK, 1989, p. 13, tradução nossa).

Borsboom e colaboradores (2004, p. 1061) propuseram uma concepção de validade mais simples, pelo menos no sentido de sua verificação, que defendem ser teoricamente superior às posições existentes na literatura. Estabeleceram que um teste é válido para a medição de um atributo se, e somente se, satisfaz as seguintes questões: (a) o atributo que se deseja medir existe? (b) as variações no atributo causalmente produzem variações nos resultados da medida? Justificam essa definição do seguinte modo: "Se alguma coisa não existe, então não se pode medir. Se ela existe, mas causalmente não produz variações nos resultados do procedimento de medição, a medida não é eficiente ou está medindo algo completamente diferente".

McNamara (2000, p. 48) caracteriza a validade como uma avaliação do próprio teste e a define como o processo para investigar os procedimentos pelos quais decisões são tomadas a partir das inferências feitas sobre os resultados do teste. Segundo o autor,

> A validação de um teste envolve o pensar na lógica do teste, especialmente em seu *design* e em suas intenções, e também envolve olhar para as evidências empíricas – os fatos – que emergem dos dados advindos de um julgamento do teste ou de administrações operacionais. Se não houver procedimentos de validação disponíveis, há potencial para parcialidades e injustiças. Esse potencial é significativo em proporção ao que está em jogo[4].

As inferências sobre os resultados do teste frequentemente vão muito além dos desempenhos observados. Os resultados dos testes não são utilizados simplesmente para relatar como um indivíduo se saiu ao responder alguns itens sob certas condições. Ao contrário, as pontuações do teste são usadas para apoiar afirmações diversas, que muitas vezes não são evidentes nas avaliações. É necessário avaliar a plausibilidade das afirmações com base nos resultados dos testes para validar as interpretações e utilizações desses resultados (KANE, 2013).

O objetivo deste estudo é realizar uma análise dos itens da prova de matemática do Enem 2015 por meio de dois modelos da TRI, o logístico de três parâmetros (ML3), com estudos sobre discriminação, dificuldade e probabilidade de acerto casual, e o modelo de resposta nominal (MRN), com estudos sobre a dificuldade e o comportamento de cada uma das opções de resposta dos itens. Além da obtenção de informações sobre os itens da prova de matemática do Enem, essas análises são úteis como meio de promover entendimento desses modelos. Por meio dessas análises, é possível obter informações sobre o instrumento de medida e sua validade.

## Modelos

### O modelo logístico de três parâmetros

O modelo logístico de três parâmetros (ML3) considera a dificuldade do item, a discriminação e a probabilidade de acerto ao acaso (LORD, 1980).

**4-** *Test validation similarly involves thinking about the logic of the test, particularly its design and its intentions, and also involves looking at empirical evidence -- the hard facts -- emerging from data from test trials or operational administrations. If no validation procedures are available there is potential for unfairness and injustice. This potential is significant in proportion to what is at stake* (McNAMARA, 2000, p. 48, tradução nossa).

A probabilidade de um indivíduo optar pela alternativa correta no item é dada por

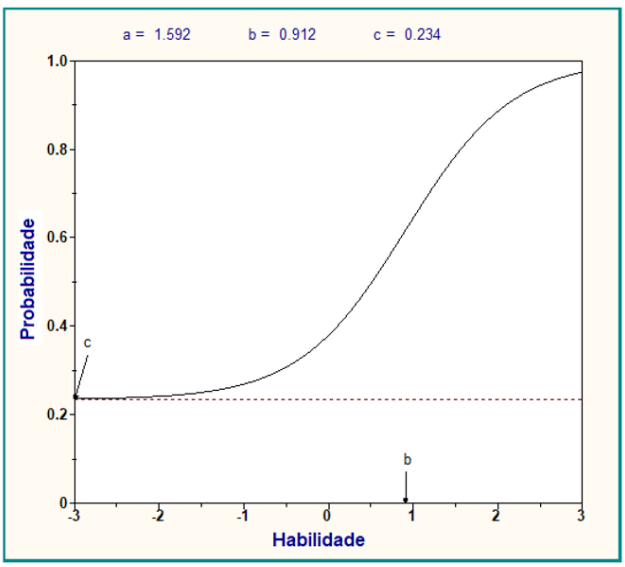
$$P(U_{ij} = 1|\theta_j) = C_i + (1 - C_i)\frac{1}{1+e^{-a_i(\theta_j - b_i)}} \qquad (1)$$

com $i = 1,2... I$ e $j = 1,2 ... n.I$ é o número de itens e $n$ o número de indivíduos da população.

- $U_{ij}$ é uma variável dicotômica que assume os valores 1, quando o indivíduo $j$ responde corretamente o item $i$, ou 0 quando o indivíduo $j$ não responde corretamente o item $i$;
- $\theta_j$ representa a habilidade (traço latente) do j-ésimo indivíduo;
- $b_i$ é o parâmetro de dificuldade do item $i$;
- $a_i$ é o parâmetro de discriminação do item $i$; e
- $c_i$ é o parâmetro da resposta ao acaso (chute).

Nesse modelo, a relação existente entre a probabilidade de um indivíduo responder corretamente a um item e os parâmetros desse item é uma função crescente denominada Curva Característica do Item (CCI). Para exemplificar, a Figura 1 exibe o gráfico da curva característica de um item considerado eficiente na prova de matemática do Enem 2015.

**Figura 1 –** Curva característica do item 4 –ML3




Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

O parâmetro $b$ é uma medida da dificuldade do item, e é dado na mesma unidade da habilidade (eixo horizontal). Observa-se na Figura 1 que quanto maior a habilidade do candidato, maior é a probabilidade (eixo vertical) de este responder corretamente o item. O valor do parâmetro $b$ corresponde à habilidade necessária para uma probabilidade de

acerto igual a (1 + *c*) / 2. A escala utilizada é a (0,1), ou seja, a média é igual a zero e o desvio padrão igual a 1.

O parâmetro *a* é proporcional à inclinação da reta tangente ao gráfico no ponto de inflexão (ponto em que a concavidade do gráfico muda de sinal). Valores pequenos de *a* indicam que o item tem pouco poder de discriminação, isto é, candidatos com habilidades diferentes têm probabilidades parecidas de responder corretamente o item. O parâmetro *c* representa a probabilidade de uma pessoa com habilidade baixa responder corretamente o item, é a probabilidade de acerto ao acaso (chute).

Traçando-se uma linha vertical em uma habilidade, na intersecção desta linha com a CCI, obtém-se a probabilidade de uma pessoa com aquela habilidade responder corretamente o item. Quanto mais para a direita está a CCI, mais difícil é o item.

## O modelo de resposta nominal

O modelo de resposta nominal (MRN) foi introduzido por Bock (1972) e é considerado o modelo da TRI mais geral, pois não exige ordenação das categorias para medir diferentes graus da característica que está sendo avaliada. Esse modelo utiliza toda a informação contida nas respostas dos indivíduos, e não apenas se o item foi respondido corretamente ou não.

O modelo de resposta nominal estabelece que a probabilidade de um indivíduo *j* optar, no item *i*, pela alternativa *k*, é dada por

$$P(U_{ijk} = 1|\theta_j) = e^{a_{ik}(\theta_j - b_{ik})} / \sum_{h=1}^{mi} e^{a_{ih}(\theta_j - b_{in})} \qquad (2)$$

com $i = 1,2, \ldots, I$, $j = 1,2, \ldots, n$, e $k = 1,2, \ldots m_i$.

- $a_{ik}$ é o parâmetro de discriminação e
- $b_{ik}$ é o parâmetro da dificuldade, ambos relacionados com a alternativa *k* do item .

Para cada $\theta_j$, a soma das probabilidades sobre as $m_i$ opções de resposta é 1, isto é, $\sum_{K=1}^{mi} P(U_{ijk}|\theta_j) = 1$. Note que o símbolo *k* fixa a categoria de resposta, não implicando que essas são ordenadas. São impostas duas restrições para estimar os parâmetros: A soma dos $a_{ik}$ e dos $b_{ik}$ são zero, isto é, $\sum_{K=1}^{mi} a_{ik} = \sum_{K=1}^{mi} b_{ik} = 0$.

Nesse modelo, o parâmetro da discriminação *a* pode ser negativo, embora, para a alternativa supostamente correta, sejam esperados valores positivos de *a*, assim como o parâmetro *b* deve ser maior para essa alternativa, pois indivíduos com maior habilidade devem ter maior probabilidade de escolhê-la. O MRN fornece informações sobre a relação entre a probabilidade de escolha de cada alternativa e da habilidade do indivíduo em relação ao traço latente avaliado.

No MRN, os valores dos parâmetros *a* e *b*, de cada alternativa devem ser analisados juntos para uma correta interpretação das respostas dos participantes do teste e, consequentemente, do comportamento dos itens. A curva característica do item (CCI) é uma ferramenta importante para auxiliar na interpretação desses parâmetros e consiste

na representação gráfica de um conjunto de curvas, cada uma delas correspondendo à probabilidade de escolha de uma das alternativas do item, como uma função da habilidade do examinando. Essas curvas são comumente denominadas em inglês por *option response function* (ORF), *category probability curves*, *category response function* ou *option caracteristc curves* (DE AYALA, 2009). Neste trabalho, a opção é denominá-las por curva característica da alternativa (CCA).

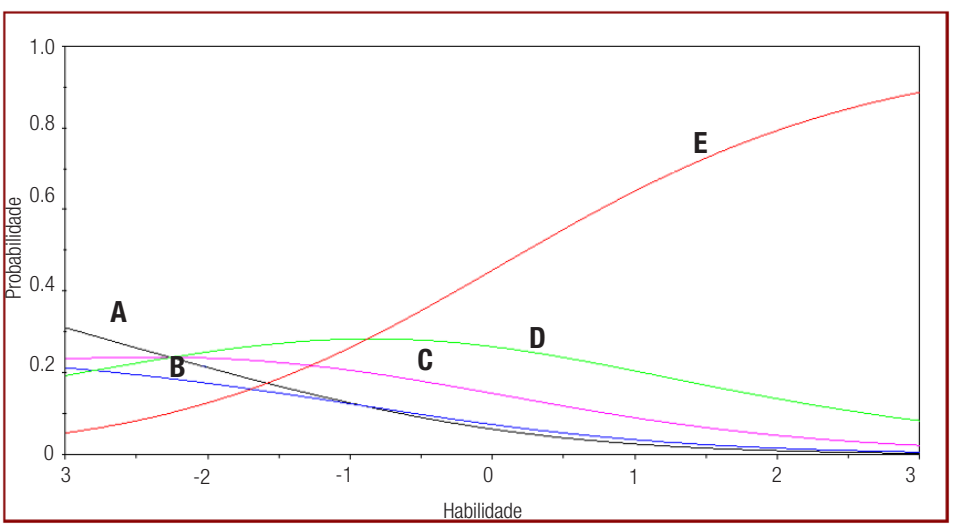
No MRN, não existe alternativa correta, as respostas são modeladas conforme a habilidade dos participantes e a probabilidade de escolherem cada uma das alternativas. Desse modo, é considerada como correta a alternativa que possui maior probabilidade de ser escolhida pelos participantes de habilidade mais alta do teste.

Para a alternativa supostamente correta, a CCA correspondente deverá ser uma função crescente, pois, quanto maior a habilidade dos participantes do teste, maior deverá ser a probabilidade de optarem por essa alternativa. Essa opção de resposta também deverá ter os maiores valores dos parâmetros *a* e *b*.

Para a alternativa supostamente incorreta (a alternativa mais distante da correta), a CCA deverá ser decrescente, $a < 0$, e o valor do parâmetro *a* deverá ser o menor entre todas as alternativas. As outras alternativas são consideradas parcialmente corretas, as CCAs são funções que crescem até um ponto e depois decrescem, tendendo a zero nos extremos (DE AYALA, 2009).

A Figura 2 traz um esboço da curva característica do item 4 da prova de matemática do Enem 2015, caderno azul, o mesmo item que anteriormente serviu de ilustração para o ML3.

**Figura 2 –** Curva característica do item 4 – MRN




Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

Os parâmetros desse item para as cinco alternativas de respostas são expostos na Tabela 1.

**Tabela 1 –** Parâmetros do item 4 – MRN

| | A | B | C | D | E |
|---|---|---|---|---|---|
| *a* | -0,49 | -0,31 | -0,11 | 0,15 | 0,76 |
| b | -0,90 | -0,73 | -0,02 | 0,55 | 1,09 |
| | | | | | Gabarito da prova: E |

Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

A alternativa correta do item é a de letra E, que também é a alternativa considerada como correta pelo MRN, pois, quanto maior a habilidade do respondente, maior é a probabilidade de ele escolher essa alternativa. Os valores dos parâmetros *a* e *b* dessa alternativa também são os maiores entre todas as outras. A alternativa considerada a mais incorreta é a alternativa A, pois, ao contrário da correta, a probabilidade de um indivíduo escolher essa alternativa é menor quanto maior for a sua habilidade. A alternativa D pode ser considerada a segunda melhor opção. Observa-se no gráfico (Figura 2) que indivíduos com habilidade entre aproximadamente -2,3 e -0,8 têm probabilidade maior de escolher essa alternativa.

Os pontos de interseção entre as curvas merecem análises de especialistas da área do traço latente avaliado, uma vez que escondem informações sobre o nível de habilidade dos indivíduos. Algumas questões sobre as relações entre as características implícitas ou não em cada uma das alternativas e a capacidade do respondente podem ser formuladas, ou então, questões sobre qual o conhecimento que um indivíduo possui a mais do que o outro para que este opte por uma alternativa mais correta.

Essas e outras informações que podem ser obtidas sobre os participantes da avaliação, e também sobre os itens, tornam o MRN uma ferramenta importante em diversas áreas, mas principalmente na área educacional.

## Métodos

### Estimação dos parâmetros

Os parâmetros dos itens, segundo o ML3, foram estimados com a utilização do programa BILOG-MG 3.0 (ZIMOWSKI; MURAKI; MISLEVY; BOCK, 2002) e as suas opções *default*, como o método da máxima verossimilhança marginal nas estimativas. O número de ciclos para a estimativa foi fixado em quinhentos, valor exagerado, pois a convergência costuma ocorrer com menos de 25 ciclos e o número de Newton ciclos foi fixado em vinte. Esse valor também é maior do que o necessário normalmente.

A estimação dos parâmetros, segundo o MRN, foi feita utilizando o programa MULTILOG 7.0 (THISSEN, 1991), também pelo método da máxima verossimilhança marginal e as opções *default* do programa. O número de ciclos para a estimativa de máxima verossimilhança marginal foi fixado em 150. Esse valor é também exagerado, pois a convergência normalmente ocorre em menos de 25 ciclos.

### Dados

O Enem 2015 contou com 8.478.096 candidatos inscritos, dos quais, cerca de 1,9 milhão deixaram de comparecer às provas. A prova de matemática do Enem é composta por 45 itens de múltipla escolha com cinco alternativas cada um e são elaboradas quatro montagens diferentes, utilizando-se os mesmos itens: caderno amarelo, caderno azul, caderno cinza e caderno rosa.

Neste estudo, para a calibração dos parâmetros do ML3, foram utilizadas as respostas de 999.999 participantes da prova que responderam à prova de matemática do caderno azul. Esse número é o máximo aceito pelo programa utilizado. Para a calibração dos parâmetros do MRN, foram utilizadas cinquenta mil respostas, das quais foram excluídas aquelas com todos os itens em branco e as com todas as respostas assinaladas com a mesma alternativa. Após essas exclusões, restaram um total de 49.922 respostas válidas para o ensaio. A prova de matemática do caderno azul foi escolhida de modo aleatório.

## Resultados

### Análise da prova pela TCT

Alguns índices fornecidos pela TCT, como a porcentagem de acerto no item e a correlação bisserial (CB), são calculados preliminarmente para serem utilizados na calibração dos itens pela TRI, mas também podem auxiliar na análise da qualidade dos itens. A CB é calculada para cada alternativa do item e estabelece a correlação entre a resposta correta e a habilidade do examinando. Isso indica que, para um item com bom poder de discriminação, a alternativa correta deve ser escolhida pelos indivíduos de maior proficiência, resultando em um valor positivo e alto para a alternativa correta e valores negativos para as alternativas incorretas (ANDRADE; KLEIN, 2005).

Análises baseadas nesse índice são constantemente associadas a inclusão ou não do item em bancos de itens, indicando inicialmente se um item pode ser considerado eficiente ou se tem condições de ser melhorado (MOREIRA JUNIOR, 2011; SOARES, 2005; VEY, 2011). Não existe uma regra única relacionando os valores desse índice e a eficiência do item para discriminar os indivíduos quanto às habilidades as quais o item foi desenvolvido para medir, mas alguns intervalos de valores para a excelência da qualidade dos itens são constantemente sugeridos nas pesquisas. O item é considerado pobre em relação a discriminação se a CB for menor do que 0,15, se a CB estiver entre 0,15 e 0,3, o item é considerado razoável, e se a CB for maior do que 0,3, ele possui um bom poder de discriminação (SOARES, 2005; TRAVITZKI, 2017).

A análise dos valores desse índice não deve ser o único critério para a inclusão ou não do item em um banco, mas é constantemente citado como uma primeira condição após o pré-teste (MOREIRA JUNIOR, 2011; VEY, 2011; SOARES, 2005). Também nessa fase, deve-se analisar esse índice para as alternativas incorretas que devem ser negativos e razoavelmente distantes de zero.

Para os itens com a CB da alternativa correta no intervalo 0,15 < CB <0,3, muitas vezes são sugeridas revisões pedagógicas na tentativa de melhoria do item, essas revisões

devem também ser baseadas nesses índices dos distratores, mas utilizar os itens reescritos sem um novo pré-teste é perigoso, uma vez que não existem garantias de melhorias do item, além de novos problemas que podem ter sido inseridos. Entretanto, os pesquisadores concordam que os itens com a correlação bisserial menor do que 0,15 não devem ser incluídos em bancos de itens.

Na Tabela 2, são expostas a correlação bisserial para a alternativa correta (coluna 3) e a porcentagem de respostas corretas para cada item (coluna 2) da prova de matemática do Enem 2015, caderno azul.

**Tabela 2 –** Estatísticas dos itens pela TCT

| Item | %corretas | Bisserial | Item | %corretas | bisserial | Item | %corretas | Bisserial |
|---|---|---|---|---|---|---|---|---|
| 01 | 29,0 | 0,169 | 16 | 10,5 | -0,032 | 31 | 39,5 | 0,298 |
| 02 | 25,4 | 0,221 | 17 | 15,3 | 0,443 | 32 | 25,9 | 0,344 |
| 03 | 35,6 | 0,462 | 18 | 23,8 | 0,429 | 33 | 13,0 | -0,077 |
| 04 | 44,0 | 0,354 | 19 | 11,2 | 0,016 | 34 | 33,3 | 0,165 |
| 05 | 64,1 | 0,315 | 20 | 33,3 | 0,260 | 35 | 11,7 | 0,173 |
| 06 | 19,6 | 0,298 | 21 | 28,3 | 0,166 | 36 | 21,4 | 0,147 |
| 07 | 60,0 | 0,342 | 22 | 20,4 | 0,245 | 37 | 20,3 | 0,261 |
| 08 | 18,0 | 0,382 | 23 | 38,3 | 0,378 | 38 | 26,8 | 0,410 |
| 09 | 34,0 | 0,286 | 24 | 14,1 | 0,220 | 39 | 23,9 | 0,087 |
| 10 | 28,0 | 0,284 | 25 | 29,3 | 0,479 | 40 | 20,5 | 0,103 |
| 11 | 28,0 | 0,258 | 26 | 12,7 | 0,011 | 41 | 46,6 | 0,390 |
| 12 | 18,8 | 0,134 | 27 | 17,5 | 0,119 | 42 | 32,9 | 0,373 |
| 13 | 35,1 | 0,426 | 28 | 22,5 | 0,136 | 43 | 19,5 | 0,345 |
| 14 | 24,0 | 0,241 | 29 | 20,9 | 0,037 | 44 | 20,6 | 0,087 |
| 15 | 24,1 | 0,034 | 30 | 23,9 | 0,149 | 45 | 17,7 | 0,036 |

Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

Quanto à CB, essa prova possui um terço dos itens em cada um dos três intervalos estabelecidos para análises da discriminação do item. Isso significa que apenas quinze itens (33,3%) podem ser considerados bons com os coeficientes da CB maiores do que 0,3, outros quinze itens (33,3%) com esse índice entre 0,15 e 0,3 e quinze itens (33,3%) com a CB menor do que 0,15. Entre os últimos, a prova apresenta dois itens com desempenho muito ruins (itens 16 e 33) por terem a correlação bisserial negativa para a alternativa correta. Além disso, esses itens foram considerados muito difíceis pelos participantes, e obtiveram apenas 10,5% e 13,0% de acertos, respectivamente.

Um item pode ser considerado difícil pelos participantes da avaliação e ter um bom poder de discriminação, como é o caso do item número 17. Obteve apenas 15,3% de acertos, mas a CB para a alternativa correta ficou em torno de 0,44, indicando a sua eficiência.

Observa-se, também, que a maior porcentagem de acertos foi para o item 5, com 64,1%. Um item com a porcentagem de acertos entre 60% e 70% é considerado de nível médio em relação à dificuldade, portanto, essa prova não possui itens fáceis.

## Análise da prova pelo ML3

Para análises dos itens foram realizadas as estatísticas obtidas pelo ML3 que considera a dificuldade (*b*), a discriminação (*a*) e a probabilidade de acerto casual (*c*). Na TRI, todos os parâmetros dos itens são calculados em uma mesma métrica e, a partir desses parâmetros, pode ser elaborada uma escala de habilidade para ordenar os desempenhos das pessoas, tornando possível a interpretação pedagógica dos valores das habilidades (Andrade; Tavares; Valle, 2000). Interpretar a escala significa escolher alguns pontos ou níveis da escala e descrever os conhecimentos e habilidades que as pessoas demonstraram possuir quando situados na vizinhança desses pontos.

Os parâmetros do ML3 são apresentados na Tabela 3.

**Tabela 3 –** Parâmetros dos itens pelo ML3

| Item | a | b | c | Item | a | b | C | Item | a | b | c |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 01 | 0,838 | 2,695 | 0,194 | 16 | 0,085 | 31,859 | 0,200 | 31 | 1,587 | 1,370 | 0,264 |
| 02 | 2,154 | 2,116 | 0,212 | 17 | 4,177 | 1,817 | 0,109 | 32 | 1,217 | 1,688 | 0,113 |
| 03 | 2,244 | 0,968 | 0,170 | 18 | 2,060 | 1,553 | 0,133 | 33 | 0,085 | 31,859 | 0,200 |
| 04 | 1,592 | 0,912 | 0,234 | 19 | 0,043 | 48,433 | 0,002 | 34 | 3,803 | 2,033 | 0,309 |
| 05 | 1,072 | -0,659 | 0,005 | 20 | 1,340 | 1,759 | 0,224 | 35 | 3,035 | 2,541 | 0,103 |
| 06 | 3,035 | 1,972 | 0,157 | 21 | 1,114 | 2,624 | 0,221 | 36 | 1,979 | 2,607 | 0,192 |
| 07 | 1,260 | -0,198 | 0,116 | 22 | 2,942 | 2,094 | 0,173 | 37 | 1,174 | 2,390 | 0,123 |
| 08 | 2,433 | 1,891 | 0,123 | 23 | 1,993 | 1,148 | 0,233 | 38 | 2,267 | 1,478 | 0,164 |
| 09 | 2,615 | 1,644 | 0,276 | 24 | 1,603 | 2,710 | 0,108 | 39 | 3,562 | 2,585 | 0,230 |
| 10 | 1,778 | 1,827 | 0,202 | 25 | 2,764 | 1,230 | 0,165 | 40 | 3,934 | 2,409 | 0,194 |
| 11 | 3,332 | 1,843 | 0,238 | 26 | 0,028 | 69,452 | 0,001 | 41 | 1,689 | 0,579 | 0,188 |
| 12 | 2,613 | 2,523 | 0,172 | 27 | 1,937 | 2,878 | 0,159 | 42 | 3,376 | 1,381 | 0,243 |
| 13 | 2,687 | 1,167 | 0,221 | 28 | 2,698 | 2,424 | 0,207 | 43 | 2,424 | 1,912 | 0,141 |
| 14 | 2,184 | 2,084 | 0,196 | 29 | 3,692 | 2,766 | 0,204 | 44 | 0,236 | 6,002 | 0,011 |
| 15 | 0,787 | 5,122 | 0,223 | 30 | 3,604 | 2,296 | 0,223 | 45 | 0,249 | 11,461 | 0,128 |

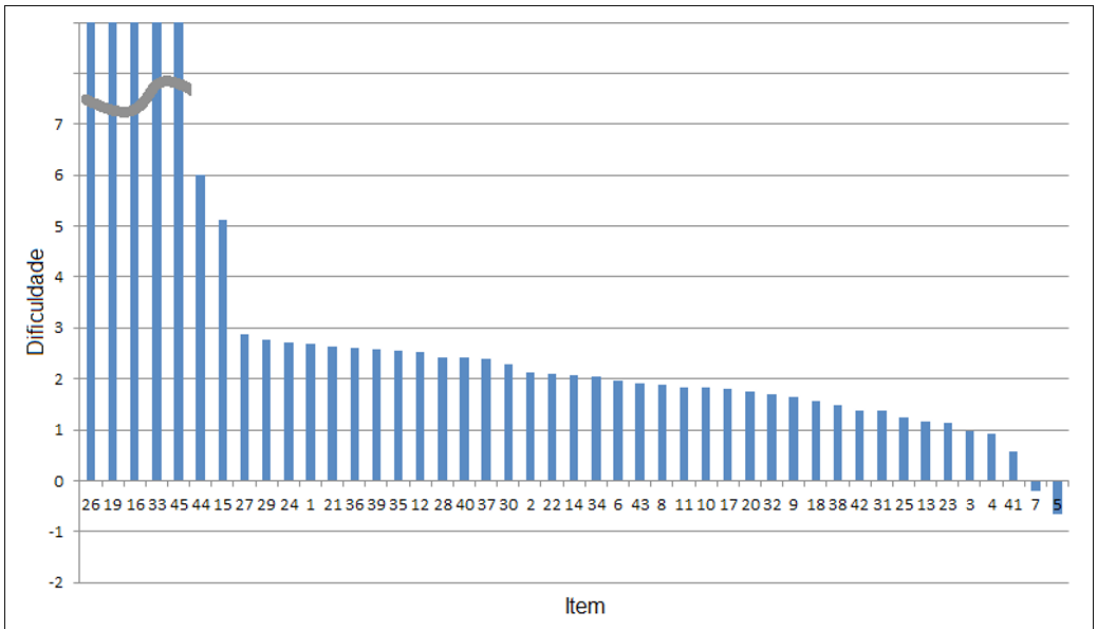
Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

Os parâmetros da dificuldade dos itens são exibidos na terceira coluna da Tabela 3. A escala de habilidades neste ensaio varia entre -3 e 3. Nesse intervalo devem ser alocadas a dificuldade dos itens e também a habilidade das pessoas. Para valores fora desse intervalo, não são possíveis análises sobre o comportamento dos itens ou da habilidade das pessoas que os responderam, mas, os cinco itens considerados mais difíceis, números 16, 19, 26, 33 e 45, possuem o parâmetro da dificuldade (b) entre 11,46 e 69,45, muito afastados desse intervalo.

De modo geral, a prova foi considerada difícil pelos participantes. Apenas os itens 5 e 7 tiveram o parâmetro da dificuldade avaliado abaixo do ponto médio da escala. Para esses itens, a dificuldade é de aproximadamente -0,7 e -0,2, respectivamente.

O gráfico exposto na Figura 3 ilustra o comportamento dos itens da prova quanto à dificuldade. As barras representando as dificuldades dos itens 16, 19, 26, 33 e 45 foram truncadas para uma melhor visualização do gráfico em relação aos outros itens.

**Figura 3 –** Dificuldade dos itens – Caderno azul




Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

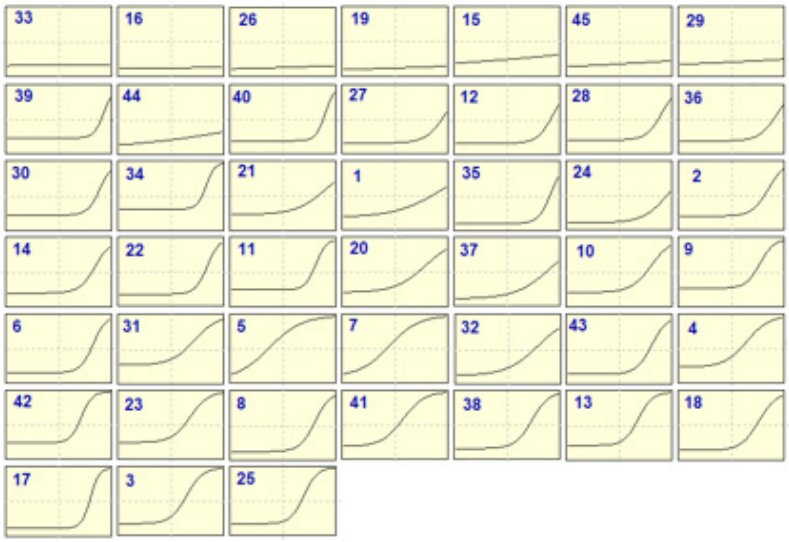
Observa-se que a maioria dos itens são para pessoas de habilidades entre 1 e 3, restando poucos itens para quem possui nível de habilidade entre -1 e 1, considerada mediana, e nenhum para quem possui pouca habilidade (abaixo de -1). É importante que o instrumento discrimine entre indivíduos de diferentes níveis de habilidades, inclusive diferenciando os que estão situados próximos uns dos outros na escala de habilidades, e não somente entre os de maior habilidade em relação aos de menor habilidade. Um teste deve conter itens fáceis, médios e difíceis, distribuindo-se continuamente em toda a extensão da escala de habilidades. A sugestão é que os itens sejam distribuídos sobre a escala numa disposição que se assemelha à da curva normal: maior parte dos itens de dificuldade mediana e diminuindo progressivamente em direção às caudas (itens fáceis e itens difíceis em número menor). A razão desse critério encontra-se no fato de que a grande maioria da população possui nível de habilidade que se distribui mais ou menos dentro da curva normal, isto é, a maioria das pessoas possui magnitudes medianas de habilidade para o traço latente avaliado, e umas poucas possuem magnitudes grandes ou pequenas (PASQUALI, 2010).

No ML3, não são esperados valores do parâmetro *a* negativos, pois indicariam que a probabilidade de um indivíduo acertar o item diminui quanto maior for a sua habilidade. Valores pequenos de a indicam que o item é pouco discriminativo, isto é, pessoas com habilidades muito diferentes têm probabilidades aproximadas de acertar o item, e valores muito altos indicam que o item separa os indivíduos em dois grupos, aqueles que possuem tal habilidade (suas habilidades estão situadas acima do parâmetro *b*) e aquelas que não possuem (suas habilidades estão abaixo do parâmetro *b*) (ANDRADE; TAVARES; VALLE, 2000). Segundo Vieira (2017), para ser considerado bom em relação à discriminação, o item deve ter o parâmetro a entre 0,7 e 2,5.

Observa-se na segunda coluna da Tabela 3 que os valores do parâmetro *a* mais baixos referem-se aos mesmos itens problemáticos já apontados anteriormente, os de números 16, 19, 26 e 33 com valores do parâmetro *a* menores do que 0,1. Outros dois itens, 44 e 45, possuem o parâmetro a em torno de 0,25, ainda assim, considerados valores muito baixos. Esses índices indicam mais uma vez que os itens não desempenharam seus papéis nessa avaliação, serviram apenas para que os avaliandos perdessem tempo com a leitura e com a tentativa de resolução.

Uma ferramenta útil para uma avaliação rápida do comportamento dos itens é a visualização das suas curvas características. Por meio delas, é possível identificar de maneira rápida quais itens são eficientes e quais não cumprem muito bem o papel de discriminar, além de facilitar a obtenção de uma ideia geral sobre a dificuldade dos itens. A matriz com as curvas características dos itens exibida na Figura 4 está organizada de acordo com os valores da correlação bisserial (CB), da menor para a maior.

**Figura 4 –** Curvas características dos itens – Caderno azul




Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

As curvas traçadas nas duas primeiras linhas da matriz e a do item 30, na terceira linha, correspondem aos itens com CB menores do que 0,15. Observa-se claramente que

tais itens, na sua maioria, não são eficientes para discriminar os indivíduos por suas habilidades, isto é, a probabilidade de se responder corretamente ao item é quase a mesma em toda a escala de habilidades. Alguns, porém, possuem comportamento um pouco diferente por terem o parâmetro da discriminação muito alto, como é o caso dos itens 39 e 40, ambos com $a \approx 3{,}9$. Esses itens discriminam as pessoas em apenas dois grupos, os que estão acima dos que estão abaixo do parâmetro $b$ correspondente, que, para esses itens, também possuem valores elevados. Talvez fossem adequados para pessoas com nível de habilidade mais elevada, embora não seja possível afirmar isso sem a aplicação de um teste para uma amostra de população como essa.

Na verdade, as curvas que apresentam a forma mais parecida com a letra S correspondem aos itens mais eficientes, pois indicam que quanto maior a habilidade do examinando, maior a probabilidade de ele responder corretamente o item. Praticamente, somente a partir da quarta linha da matriz é que as curvas começam a tomar esta forma. O item 14 já possui a CB maior do que 0,24, mas o item número 5 é o primeiro entre os itens que apresenta a CB maior do que 0,3. Confirma-se, desse modo, a relação entre a eficiência do item e a sua CB relativamente alta (maior do que 0,3).

Na quarta coluna da Tabela 3, estão expostos os parâmetros que indicam a probabilidade de acerto ao acaso (chute). O esperado é, para valores desse parâmetro, em torno de 0,2, pois os itens possuem 5 alternativas. Os itens 9, 31 e 34 são os que possuem os maiores valores desse parâmetro, entre 0,26 e 0,31, aproximadamente. Isso indica que uma ou mais alternativas desses itens devem ser muito óbvias, podendo ser descartadas facilmente.

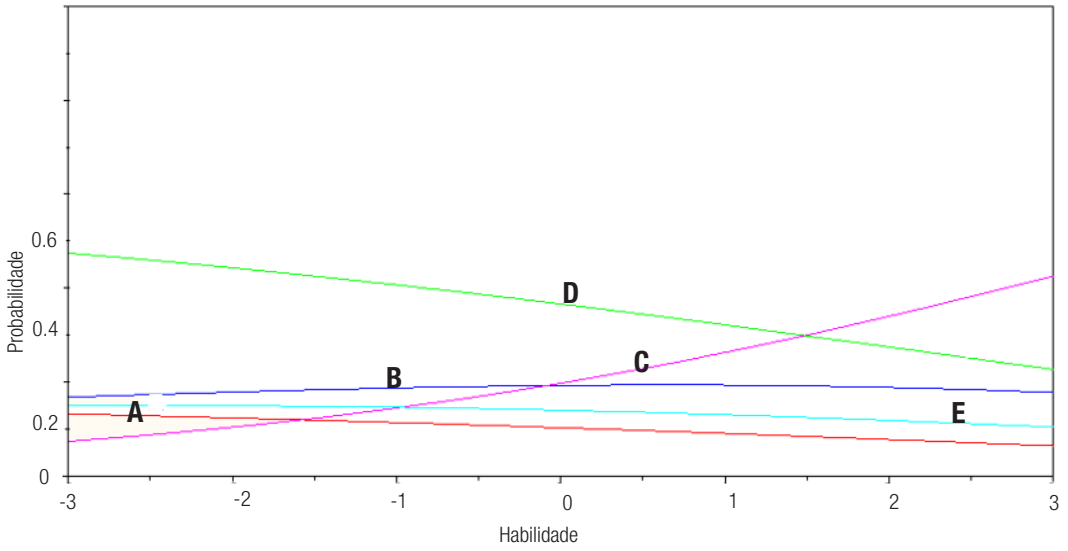
Outra informação interessante é que, quanto mais para a direita estiver a curva, mais difícil é o item. Observa-se que as melhores curvas para a eficiência dos itens estão situadas à direita do ponto médio da escala, confirmando que esses itens são para os que possuem habilidades altas.

## Análise da prova pelo MRN

A interpretação e análise dos dados provenientes do MRN devem ser feitas com base em dois parâmetros para cada opção de resposta, a discriminação ($a$) e a dificuldade ($b$). Inicialmente são analisados alguns dos itens que possuem a correlação bisserial menores do que 0,15, entre eles os de números 16 e 33 que possuem correlações bisseriais negativas. A maioria desses itens apresenta comportamentos estranhos aos esperados pelo MRN.

Para o item número 16 (Figura 5), a resposta indicada no gabarito da prova como correta é a alternativa A, mas, na Tabela 4, o maior valor do parâmetro de discriminação é o da alternativa C ($a = 0{,}29$), e o maior valor do parâmetro de dificuldade é o da alternativa D ($b = 0{,}70$). São esperados para a alternativa correta os maiores valores desses parâmetros. Observa-se que a alternativa supostamente correta (A) é praticamente a pior opção de resposta em toda a escala de habilidades, enquanto a alternativa D foi a melhor opção para pessoas de habilidade abaixo de $\theta \approx 1{,}5$, embora essa curva não seja muito discriminativa, além de ser decrescente. Apenas a partir da habilidade ($\theta \approx 1{,}5$) é que ocorre a maior probabilidade de a alternativa C ser escolhida.

**Figura 5 –** Curva característica do item 16 – MRN




Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

**Tabela 4 –** Parâmetros do item 16 – MRN

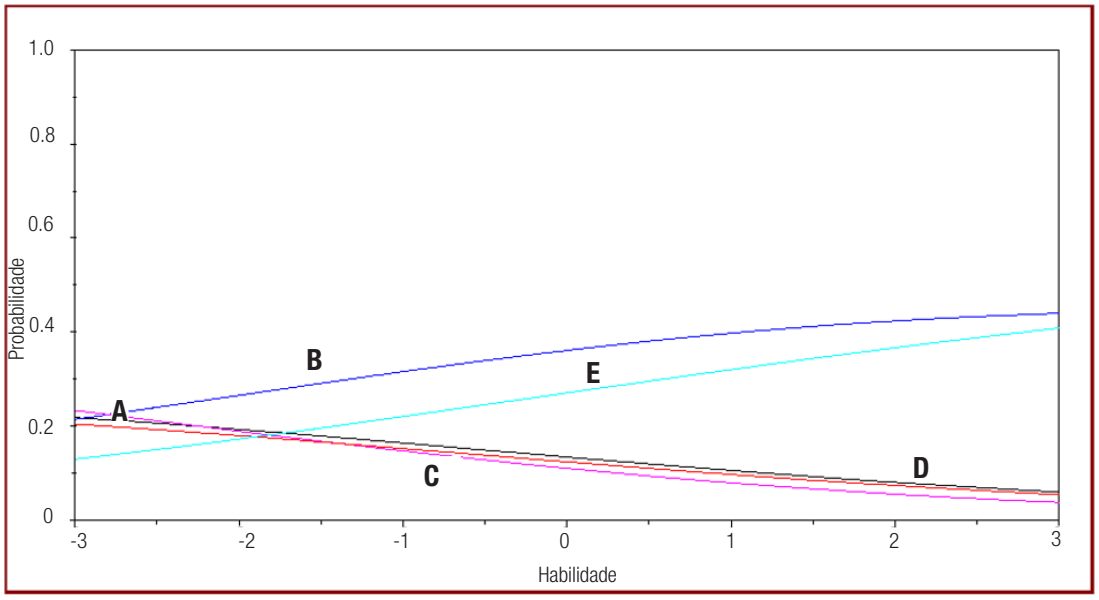
| | A | B | C | D | E |
|---|---|---|---|---|---|
| *a* | -0,12 | 0,01 | 0,29 | -0,12 | -0,06 |
| b | -0,57 | 0,06 | 0,08 | 0,70 | -0,26 |

Gabarito da prova: A

Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

Para o item 33 (Figura 6), a resposta indicada no gabarito como correta é a alternativa A, mas, na Tabela 5, o parâmetro da discriminação ($a$) dessa alternativa é negativo, indicando que a probabilidade de ser escolhida é menor quanto maior for a habilidade do indivíduo. O mesmo ocorre para as alternativas C e D, o que pode ser observado no gráfico por suas curvas estarem praticamente sobrepostas. O maior valor para o parâmetro da discriminação ocorre para a alternativa E ($a = 0,28$), e o maior valor para o parâmetro da dificuldade é o da alternativa B ($b$ = 0,71). A maior probabilidade de escolher entre as respostas foi para a alternativa B em praticamente todos os pontos da escala de habilidades. Mas, mesmo que a alternativa B fosse a correta, esse item não seria muito discriminativo. Outro problema relacionado com esse item é que as curvas das alternativas A, C e D são muito próximas entre si indicando que elas têm probabilidades aproximadas de serem escolhidas. Esses fatos confirmam a má qualidade dos itens de números 16 e 33 também sob o enfoque do MRN.

**Figura 6 –** Curva característica do item 33– MRN




Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

**Tabela 5 –** Parâmetros do item 33– MRN

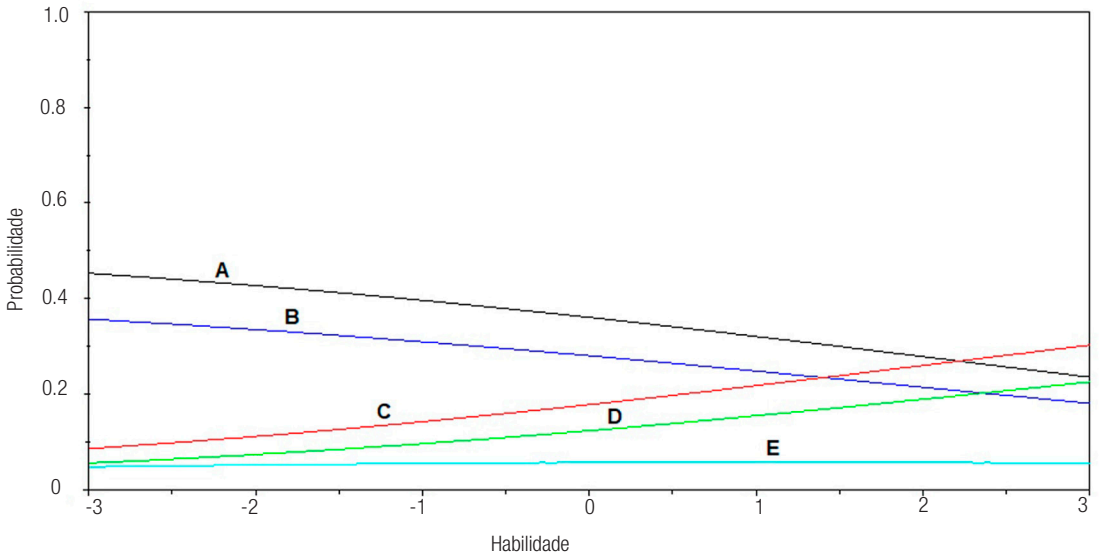
| | A | B | C | D | E |
|---|---|---|---|---|---|
| *a* | -0,13 | 0,21 | -0,22 | -0,13 | -0,28 |
| *b* | -0,28 | 0,71 | -0,48 | -0,36 | 0,42 |

Gabarito da prova: A

Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

A Figura 7 traz as curvas características das alternativas, e a Tabela 6, os parâmetros do item número 12. Para esse item, a alternativa correta indicada no gabarito oficial da prova é a de letra C, mas o maior valor do parâmetro a corresponde a alternativa D (*a = 0,18*), e o maior valor para o parâmetro *b* é para a alternativa A (*b* = 0,71). Observando-se o gráfico dessas CCAs, vê-se que os participantes da prova com habilidade $\theta < 2,2$ têm maior probabilidade de escolher a alternativa A, a partir daí a maior probabilidade de escolha é para a alternativa C. Observa-se também que a alternativa E é a pior opção para os indivíduos com habilidade em toda a escala.

**Figura 7 –** Curva característica do item 12– MRN




Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

**Tabela 6 –** Parâmetros do item 12– MRN

| | A | B | C | D | E |
|---|---|---|---|---|---|
| *a* | -0,16 | -0,16 | 0,16 | 0,18 | -0,03 |
| *b* | 0,77 | 0,52 | 0,07 | -0,30 | -1,07 |

Gabarito da prova: C

Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

As curvas características e os parâmetros do item 15 encontram-se na Tabela 7 e na Figura 8, respectivamente. Para esse item, a alternativa correta indicada no gabarito da prova é a de letra A, e os maiores valores dos parâmetros *a* e *b* ocorrem para a alternativa B, *a*=0,37 e *b*=0,61.

A alternativa indicada como correta não foi a melhor opção de escolha para os indivíduos no intervalo de habilidade estudado. Para indivíduos com habilidade acima de -0,8 aproximadamente, a maior probabilidade de escolha é a alternativa B.

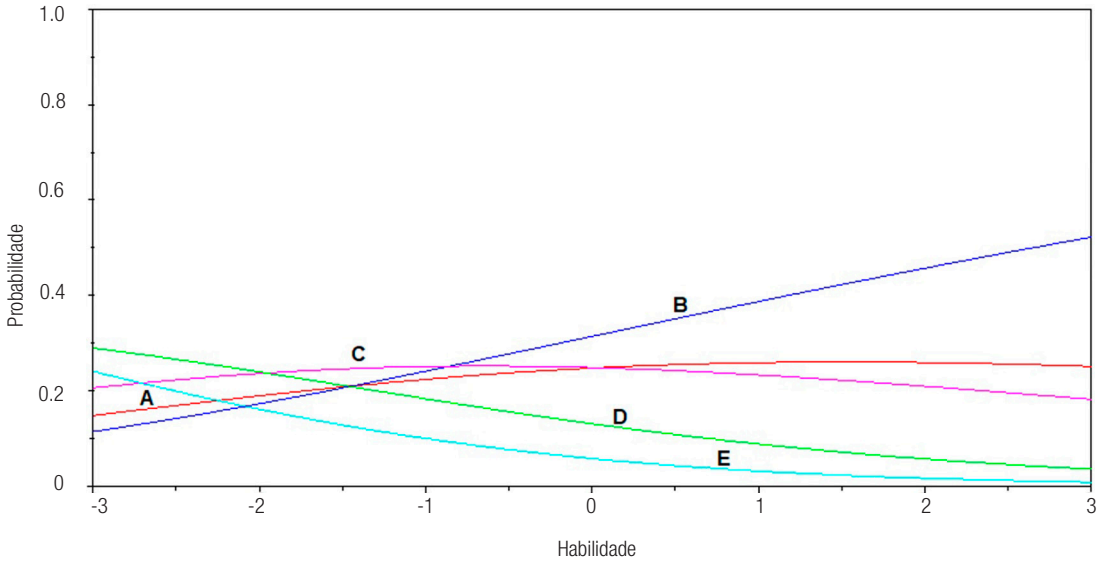
**Tabela 7 –** Parâmetros do item 15– MRN

| | A | B | C | D | E |
|---|---|---|---|---|---|
| *a* | 0,20 | 0,37 | 0,10 | -0,23 | -0,44 |
| b | 0,37 | 0,61 | 0,37 | -0,27 | -0,08 |

Gabarito da prova: A

Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

**Figura 8 –** Curva característica do item 15– MRN




Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

Confirma-se também, pelas análises segundo o MRN, que os itens que possuem as correlações bisseriais para a alternativa correta menores do que 0,15 possuem comportamentos inadequados. Um dos principais problemas desses itens é que a alternativa indicada como correta no gabarito oficial da prova não coincide com a alternativa supostamente correta modelada pelo MRN, ao menos para habilidades pertencentes ao intervalo (-3,3). A alternativa correta deve sempre ser escolhida pelos indivíduos com habilidade mais alta, e as demais alternativas, as incorretas, devem ser escolhidas pelos indivíduos com habilidade inferior. Quando ocorre a inversão dessa regra, é necessário verificar a existência de problemas em relação ao item, que podem ser conceituais, na formulação, quanto à alternativa correta, entre outros.

Observando a matriz das CCIs do ML3, (Figura 4) e lembrando que ela foi organizada de acordo com a medida da CB, em ordem crescente, pode-se constatar que as CCIs começam a melhorar mais ou menos a partir do item 2 (final da terceira linha da matriz), que possui CB da alternativa correta em torno de 0,22. Esse fato é repetido também com os parâmetros obtidos pelo MRN. Os itens que possuem CB < 0,22, quase na totalidade, são considerados ruins ou péssimos, com algumas poucas exceções. Isso significa que os itens realmente bons são aqueles que possuem CB maiores do que 0,3.

## Resumo sobre as análises

A Tabela 8 exibe os itens considerados ruins ou péssimos do ponto de vista dos métodos TCT, ML3 e MRN. Pela TCT, foram considerados ruins aqueles itens com CB<0,15; pelo ML3, foram considerados ruins aqueles com o parâmetro *a* fora do intervalo 0,7<*a*<2,5

e também aqueles com o parâmetro $b>3$; para MRN, foram escolhidos aqueles no qual a resposta correta apontada no gabarito da prova não coincidiu com a resposta considerada correta pelo MRN. Os itens destacados em cor cinza indicam que estes são comuns aos três métodos e correspondem a 20% do total dos itens da prova. Os outros itens foram considerados razoáveis ou bons.

**Tabela 8 –** Itens considerados ruins ou péssimos

| | TCT | ML3 | MRN |
|---|---|---|---|
| *Itens* | 12 15 16 19 26 27 28 29 30 33 36 39 40 44 45 | 6 9 11 12 13 15 16 17 19 22 25 26 28 29 30 33 34 35 39 40 42 44 45 | 12 15 16 19 22 24 26 27 29 33 34 35 39 44 45 |
| Porcentagem | 33,33% | 51,11% | 33,33% |

Fonte: elaborada pela autora com base nos dados da pesquisa (2017).

## Considerações finais

Este estudo se limitou a analisar a qualidade dos itens da prova de matemática do Enem 2015 por meio de alguns índices da TCT e pelos modelos ML3 e MRN da TRI, não tendo preocupação com análises sobre o conteúdo abordado, nem mesmo com a forma dos itens.

Segundo Messick (1996), as questões fundamentais para uma avaliação em larga escala eficiente consistem em validade, confiabilidade, comparabilidade e justiça. O autor afirma ainda que esses conceitos não se resumem a princípios de medição. São valores sociais com significado, não simples medidas, e devem ser considerados sempre que decisões de valores são tomadas com base nas avaliações (TOFFOLI et al., 2016).

Uma avaliação de qualidade deve permitir aos participantes condições de respostas que assegurem inferências corretas sobre seu desempenho em relação às habilidades que estão sendo medidas. Quando os testes são administrados para populações diversas, como no Enem, especificações que assegurem a validade do teste são mais difíceis de serem alcançadas, assim como é mais difícil obter medidas precisas dos conhecimentos e das competências dos respondentes. As questões sobre justiça estão relacionadas com a equidade do teste, ou a possibilidade de garantir oportunidades iguais a todos os participantes, e, para isso, é necessário que os testes sejam imparciais e apropriados para os vários grupos que serão testados.

A American Educational Research Association (AERA), a American Psychological Association (APA) e a National Councilon Measurement in Education (NCME) (AERA; APA; NCME, 2014) estabelecem que todos os examinandos devem ter oportunidade de demonstrar a sua posição na escala de habilidades em relação ao construto que o teste é concebido para medir. A validade do teste depende dessa oportunidade dada aos participantes da avaliação que, nesse contexto, está relacionada principalmente aos itens.

A TRI coloca todos os parâmetros relacionados aos itens (dificuldade, discriminação, probabilidade de acerto ao acaso) e os relacionados às pessoas (habilidade) em uma mesma métrica. Desse modo, é possível comparações, por exemplo, entre a dificuldade

dos itens e a habilidade das pessoas. Na prova de matemática do Enem 2015, o item mais fácil obteve o parâmetro da dificuldade em $b$ = -1,98, e aproximadamente 23,4% dos participantes possuem habilidades menores do que esse valor, indicando que esses indivíduos não souberam responder nenhum entre os 45 itens, só acertando itens ao acaso, ou popularmente falando, *no chute*.

Baseando-se no conceito de validade proposto por Borsboom et al. (2004, p. 1061) citado anteriormente, para esses participantes da avaliação, essa prova não produziu variações nos resultados da avaliação. Assim, como os próprios autores estabeleceram, "a medida não é eficiente ou está medindo algo completamente diferente". O mesmo pode ser dito sobre os itens considerados muito ruins em relação à discriminação: 16, 19, 26, 33 e também 44 e 45. Esses itens provavelmente não exercem efeito algum sobre a classificação dos participantes da avaliação.

A formação de um banco de itens calibrados com as ferramentas fornecidas pela TRI consiste em uma tecnologia bastante utilizada no mundo inteiro, portanto, com muitos estudos publicados nas melhores revistas científicas (MOREIRA JUNIOR, 2011; SOARES, 2005; VEY, 2011).

Desde 2009, quando o Enem passou a ter o formato atual, o seu objetivo passou de apenas avaliar o conhecimento dos jovens ao final do ensino médio para tentar uma aproximação da reforma desse nível de ensino e das tendências internacionais, destacando a importância de uma formação geral na educação básica (BRASIL, 1999). Além disso, o resultado do exame está sendo utilizado por instituições de ensino superior como critério de acesso à universidade e ao financiamento estudantil.

Desse modo, é inadmissível que em uma avaliação em larga escala com a dimensão e as propostas do Enem se conceba uma avaliação com qualidade tão aquém da desejada. Desde a sua reestruturação em 2009, a avaliação tem sofrido diversos tipos de problemas e muitos recursos públicos têm sido utilizados para a sua elaboração em cada edição. Estudos sobre a qualidade das avaliações são importantes para que os erros que já ocorreram sejam corrigidos e para que os índices de qualidade e os procedimentos sejam melhorados a cada edição.

Aliás, os procedimentos relacionados a todas as etapas – elaboração, aplicação e classificação dos participantes de avaliações em larga escala – são amplamente debatidos na literatura científica que trata do assunto (TOFFOLI et al., 2016; TOFFOLI, 2015). Pode-se, por meio da qualidade constatada nessa prova, intuir que procedimentos importantes estão sendo negligenciados.

Existem outros trabalhos que abordam a qualidade das provas do Enem em anos anteriores, como o do pesquisador Travitzki (2017), que realizou uma avaliação do Enem dos anos de 2009 e 2011 utilizando técnicas da TCT e também do modelo logístico de dois parâmetros da TRI. Constatou-se que a prova de matemática de 2009 apresentou 49% dos itens com correlação bisserial menor do que 0,3 indicando que esses itens podem ser considerados regulares ou ruins. A prova de 2011 apresentou 17% dos itens de matemática nessa categoria. Já a prova de matemática do Enem 2015 obteve 62,2% dos itens com a correlação bisserial menor do que 0,3, com 24,4% delas consideradas ruins, confirmando

que, em relação a esse índice, essa prova tem uma qualidade pior quando comparada com as edições de 2009 e de 2011.

Uma proposta para trabalhos futuros consiste em analisar os itens pedagogicamente, utilizando-se o MRN. Enquanto o ML3 é aplicado a testes com itens dicotômicos, nos quais apenas uma das opções de respostas é considerada correta e todas as outras alternativas incorretas. O MRN considera também a alternativa incorreta que foi assinalada pelo participante e, de certa forma, estabelece uma ordem entre as alternativas incorretas, desde a mais incorreta (mais distante da correta), até a mais próxima da correta. As análises dos itens podem se dar de forma muito mais completa, pois é possível detectar qual conhecimento a mais é útil para se deixar de escolher uma alternativa e escolher uma outra, considerada mais correta. Esses estudos podem ser de grande interesse para os professores e pesquisadores da área de ensino.

## Referências

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

ANDRADE, Dalton Francisco; KLEIN, Ruben. Aspectos quantitativos da análise dos itens da prova do Enem. In: BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Exame Nacional do Ensino Médio (Enem):** fundamentação teórico-metodológica. Brasília, DF: Inep, 2005. p. 107-112.

ANDRADE, Dalton Francisco; TAVARES, Héliton Ribeiro; VALLE, Raquel da Cunha. **Teoria da Resposta ao Item: conceitos e aplicações.** São Paulo: Associação Brasileira de Estatística, 2000.

ANDRADE, Josemberg Moura de; LAROS, Jacob Arie; GOUVEIA, Valdiney Veloso. O uso da Teoria de Resposta ao Item em avaliações educacionais: diretrizes para pesquisadores. **Avaliação Psicológica,** Porto Alegre, v. 9, n. 3, p. 421-435, dez. 2010.

BOCK, R. Darrel. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. **Psychometrika,** Switzerland, v. 37, n. 1, p. 29-51, 1972.

BORSBOOM, Denny; MELLENBERG, Gideon; VAN HEERDEN, Jaap. The concept of validity. **Psychological Review,** Washington, DC, v. 111, n. 4, p. 1061-1071, 2004.

BRASIL. Ministério da Educação. **ENEM:** uma avaliação inovadora. Brasília, DF: MEC, 1999. Disponível em: <http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/enem-uma-avaliacao-inovadora/21206>. Acesso em: 11 set. 2017.

BRASIL. Ministério da Educação. **Pré-testes seguem o rigor de segurança dos demais exames.** Brasília, DF: MEC, 2011. Disponível em: <http://portal.mec.gov.br/component/tags/tag/35239>. Acesso em: 11 set. 2017.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Exame Nacional do Ensino Médio. Brasília,** DF: MEC/INEP, [2017a]. Disponível em: <http://inep.gov.br/web/guest/enem>. Acesso: 11 set. 2017.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Saeb.** Brasília, DF: MEC/INEP, [2017b]. Disponível em: <http://portal.inep.gov.br/web/guest/educacao-basica/saeb>. Acesso: 11 set. 2017.

DE AYALA, Rafael Jaime. **The theory and practice of item response theory.** New York: The Guilford Press, 2009.

DE AYALA, Rafael Jaime; SAVA-BOLESTA, Monica. **Item parameter recovery for the nominal response model.** Applied Psychological Measurement, Thousand Oaks, v. 23, n. 1, p. 3-19, 1999.

KANE, Michael T. Validating the interpretations and uses of test scores. **Journal of Educational Measurement, Washington,** DC, v. 50, n. 1, p. 1-73, 2013.

KELLEY, Truman Lee. **Interpretation of educational measurements.** New York: MacMillan, 1927.

LORD, Frederic M. **Applications of item response theory to practical testing problems.** New Jersey: Lawrence Erlbaum Associates. 1980.

McNAMARA, Tim. **Language testing.** Oxford: Oxford University Press, 2000.

MESSICK, Samuel. Validity. In: LINN, Robert (Ed.). **Educational measurement.** 3. ed. New York: Macmillan, 1989. p. 13-103.

MESSICK, Samuel. Validity of performance assessments. In: PHILLIPS, Gary W. (Ed.). **Technical issues in large-scale performance assessment.** Washington, DC: National Center for Education Statistics, 1996. p. 1-18.

MOREIRA JUNIOR, Fernando de Jesus. **Sistemática para implantação de testes adaptativos informatizados baseados na teoria da Resposta ao Item.** 2011. (Tese de Doutorado) – Universidade Federal de Santa Catarina, Florianópolis, 2011.

PASQUALI, Luiz. Testes referentes a construtos: teoria e modelos de construção. In: PASQUALI, Luiz (Org.). **Instrumentação psicológica:** fundamentos e práticas. Porto Alegre: Artmed, 2010. p. 165-198.

SCARAMUCCI, Matilde Virginia Ricardi. Validade e consequências sociais das avaliações em contexto de ensino de línguas. **Lingvarvm Arena,** Porto, v. 2, p. 103-120, 2011.

SOARES, Tufi Machado; MENDONÇA, Márcia Cristina Meneghin. Construção de um modelo de regressão hierárquico para os dados do SIMAVE-2000. **Pesquisa Operacional,** Rio de Janeiro, v. 23, n. 3, p. 421-441, 2003.

SOARES, Tufi Machado. Utilização da Teoria da Resposta ao Item na produção de indicadores socioeconômicos. **Pesquisa Operacional,** Rio de Janeiro, v. 25, n. 1, p. 83-112, 2005.

THISSEN, David. **MULTILOG user's guide:** multiple, categorical item analysis and test scoring using item response theory (Version 6.0). Chicago: Scientific Software International, 1991. Computer program.

TOFFOLI, Sonia Ferreira Lopes. **Avaliações em larga escala com itens de respostas construídas no contexto do modelo Multifacetas de Rasch.** 2015. (Tese de Doutorado) – Universidade Federal de Santa Catarina, Florianópolis, 2015.

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

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

VEY, Ivan Henrique. **Avaliação de desempenho logístico no serviço ao cliente baseada na Teoria de Resposta ao Item.** 2011. (Tese de Doutorado) – Universidade Federal de Santa Catarina, Florianópolis, 2011.

VIEIRA, Nara Núbia. As provas das quatro áreas do Enem vistas como uma prova única na ótica de modelos da teoria da resposta ao item uni e multidimensional. **Boletim na Medida,** Brasília, DF, v. 5, n. 11, p. 29-40, fev. 2017.

ZIMOWSKI, Michelle F. et al. **Bilog-MG.** Lincolnwood: Scientific Software International, 2002.

*Recebido em: 26.10.2017*
*Revisado em: 24.08.2018*
*Aprovado em: 06.11.2018*

**Sônia Ferreira Lopes Toffoli** é doutora em engenharia de produção pela Universidade Federal de Santa Catarina. É professora adjunta no Departamento de Matemática da Universidade Estadual de Londrina (UEL).