http://dx.doi.org/10.18222/eae.v28i67.3910

# AVALIAÇÃO DA QUALIDADE DO ENEM 2009 E 2011 COM TÉCNICAS PSICOMÉTRICAS

**RODRIGO TRAVITZKI**

**RESUMO**

*Foi realizada uma meta-avaliação do Exame Nacional do Ensino Médio (Enem) de 2009 e 2011, aplicando-se técnicas psicométricas em 2.025.268 provas. Algumas técnicas foram exemplificadas com a apresentação da análise do conteúdo de itens, mostrando sua utilidade aos educadores. Comparou-se indicadores relacionados à prova (coeficiente α e correlação interitem média), mas principalmente relativos aos itens (correlação item-total, coeficiente bisserial – CB –, discriminação e ajuste do modelo logístico de dois parâmetros). A prova de Matemática de 2009 apresentou confiabilidade insuficiente (α < 0,6). As provas de Linguagens e Códigos e Ciências Humanas apresentaram, nos dois anos, melhores indicadores de qualidade. Ao todo, foram encontrados 25% de itens com comportamento empírico fora do esperado em 2009, e 17% em 2011.*

**PALAVRAS-CHAVE** META-AVALIAÇÃO • PSICOMETRIA • AVALIAÇÃO EM LARGA ESCALA • ENEM.

* Este trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Brasil.

# EVALUACIÓN DE LA CALIDAD DEL ENEM 2009 Y 2011 CON TÉCNICAS PSICOMÉTRICAS

**RESUMEN**

*Se llevó a cabo una metaevaluación del* Exame Nacional do Ensino Médio *(Enem) de 2009 y 2011 por medio de la aplicación de técnicas psicométricas en 2.025.268 pruebas. Algunas de estas técnicas se ejemplificaron con la presentación del análisis del contenido de ítems, lo que puso de manifiesto su utilidad a los educadores. Se compararon indicadores relacionados con la prueba (coeficiente α y correlación inter-ítem media), pero sobre todo los relativos a los ítems (correlación ítem-total, coeficiente biserial – CB –, discriminación y ajuste del modelo logístico de dos parámetros). La prueba de Matemáticas de 2009 presentó confiabilidad insuficiente (α < 0,6). Las pruebas de Lenguajes y Códigos y Ciencias Humanas presentaron, en los dos años de estudio, mejores indicadores de calidad. En total se encontró un 25% de ítems con comportamiento empírico fuera de lo esperado en 2009, y un 17% en 2011.*

**PALABRAS CLAVE** METAEVALUACIÓN • PSICOMETRÍA • EVALUACIÓN EN GRAN ESCALA • ENEM.

# EVALUATION OF ENEM (2009 AND 2011)QUALITY WITH PSYCHOMETRIC TECHNIQUES

**ABSTRACT**

*This study is a meta evaluation of the* Exame Nacional do Ensino Médio *[National Secondary Education Examination] (Enem) of 2009 and 2011, applying psychometric techniques to 2.025.268 tests. Some techniques were illustrated by analysis of item content, showing their usefulness to educators. We compared indicators related to the tests (α coefficient and average inter-item correlation), but mainly related to the items (item-whole correlation, biserial coefficient, discrimination and fit of two-parameter logistic model). The 2009 Mathematics test showed insufficient reliability (α <0.6). The Language and Human Sciences tests showed better quality indicators in both years. In total, we found 25% of items with anomalous empirical behavior in 2009 and 17% in 2011.*

**KEYWORDS** META EVALUATION • PSYCHOMETRICS • LARGE-SCALE EVALUATION • ENEM.

## INTRODUÇÃO

O Exame Nacional do Ensino Médio (Enem) encontra-se hoje, juntamente com o exame chinês para admissão na universidade (*gao kao*), entre os maiores testes do mundo quanto ao número de candidatos, contando oito milhões de inscritos em 2015. Sendo um exame voluntário, sua alta atratividade decorre das diversas recompensas a ele atreladas, tais como: admissão na maioria das faculdades públicas do país; obtenção de bolsa ou financiamento para cursar uma faculdade privada; obtenção de certificação de ensino médio (EM), mesmo sem ter frequentado a escola (BRASIL, 2010).

Por tais recompensas, pode-se considerar o Enem o exame de maior valor econômico do Brasil, visto que um resultado positivo traz grandes benefícios ao candidato. Segundo Gomes (2010), o Enem tem gerado efeitos gradativos no trabalho docente e forte impacto mercadológico. O alto valor, por sua vez, acaba criando dificuldades adicionais, como o desenvolvimento de um sistema capaz de garantir segurança para uma logística colossal e também uma importância maior da qualidade do instrumento de avaliação. Mas como garantir essa qualidade?

Uma das principais referências internacionais para a elaboração de testes educacionais e psicológicos é o *Standards for educational and psychological testing* (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION – AERA; AMERICAN PSYCHOLOGICAL ASSOCIATION – APA; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION – NCME, 2014), renovado com certa periodicidade por tradicionais instituições americanas. O documento estabelece que os exames devem possuir especificações psicométricas que

> [...] indicam as propriedades estatísticas desejadas para os itens (como, por exemplo, dificuldade, discriminação e correlações interitem), assim como as propriedades estatísticas desejadas para todo o teste. (AERA; APA; NCME, 2014, p. 79)

Nesse sentido, cabe perguntar: quais seriam as especificações do Enem?

Diferente do Sistema de Avaliação da Educação Básica (Saeb), parece não haver uma farta documentação detalhando procedimentos para criação, pré-testagem e seleção de itens, ou simplesmente definindo características psicométricas desejáveis para os itens do Enem. Foi encontrado apenas um documento com informações desse tipo – além de informações genéricas sobre o pré-teste de itens presentes em um relatório pedagógico (BRASIL, 2007). O documento intitula-se *Exame Nacional do Ensino Médio (Enem): fundamentação teórico-metodológica* (BRASIL, 2005). Num dos textos, Fini (2005) explica que os itens do Enem são elaborados em torno de uma situação-problema, e devem conter em seu enunciado as informações necessárias para a tomada de decisão. As alternativas devem ser coerentes com o problema proposto, expressando diferentes graus de entendimento. A prova, por sua vez, deve contar com 20% de itens fáceis, 40% de itens médios e 40% de itens difíceis. Essas proporções podem ter sido alteradas nas versões recentes do Enem, mais focadas na finalidade de seleção para o Ensino Superior. Enfim, em termos psicométricos, Fini esclarece que, após o pré-teste, eram selecionados para a prova os itens com coeficiente bisserial maior do que 0,30.

Neste trabalho, buscou-se contribuir para o debate a respeito da qualidade das avaliações utilizando, principalmente, técnicas psicométricas, das clássicas às modernas. O foco no Enem é decorrente de sua extrema importância hoje: é fundamental saber se as provas são suficientemente boas. Adicionalmente, para contribuir com a qualificação dos professores para o debate, buscou-se também, na medida do possível, explicar de forma mais didática alguns conceitos e técnicas utilizados.

Na primeira seção, resgatam-se alguns conceitos da Teoria Clássica dos Testes (TCT) e da Teoria da Resposta ao Item (TRI) que podem ser úteis para a meta-avaliação e também para educadores em geral. Em seguida, detalha-se os métodos e dados utilizados, incluindo os limites estabelecidos para cada indicador. A seção de resultados inicia com um estudo qualitativo de alguns itens, buscando ilustrar o uso das técnicas psicométricas na prática. Seguem os resultados principais, relativos à meta-avaliação das oito provas do Enem (edições de 2009 e 2011). Por fim, há uma síntese dos resultados principais e algumas considerações sobre a metodologia utilizada e suas possíveis utilidades.

## AVALIAR A AVALIAÇÃO

Em educação, há diversas formas e níveis para se conceituar qualidade (TRAVITZKI, 2013a). Os testes padronizados se tornaram uma solução bastante difundida para avaliar a qualidade na educação, talvez por serem objetivos e de baixo custo (MARTÍNEZ ARIAS, 2009). Para garantir a qualidade desses testes – tanto mais necessária quanto maiores as recompensas que eles proporcionam –, há diversos conceitos e técnicas disponíveis para cada contexto e finalidade.

Pode-se analisar, por exemplo, a *validade* do teste, que seria sua capacidade de mensurar, de fato, a proficiência que se deseja mensurar. Embora haja procedimentos reconhecidos para verificação da validade, esse permanece um conceito polêmico em termos teóricos e empíricos. Segundo Pasquali (2009, p. 995), a validação dos testes

> [...] apresenta dificuldades importantes que se situam em três níveis ou momentos do processo de elaboração do ins-

> trumento, a saber, ao nível da teoria, da coleta empírica da informação e da própria análise estatística da informação.

Dentre as dificuldades, o autor destaca a irredutibilidade dos princípios de diversas teorias psicológicas,[1] além de postulados da análise fatorial que não necessariamente correspondem aos modelos teóricos ou dados empíricos.

**1** Como a validade de um construto deve ser verificada tomando-se como referência algo distinto, o processo de validação de testes frequentemente utiliza-se de mais de uma teoria.

Outra propriedade importante é a *confiabilidade* do teste, que pode ser estimada a partir de um conjunto de respostas utilizando-se diferentes metodologias. De acordo com Kuder e Richardson (1937), as diferentes formulações de confiabilidade se fundamentam nas características de um exame em virtude das correlações positivas entre seus itens. A correlação interitem média pode ser utilizada como indicador de confiabilidade, porém não é sensível ao número de itens do exame, uma variável importante da confiabilidade. Um dos indicadores de confiabilidade mais utilizados é o coeficiente α (CRONBACH, 1951), que articula os dois tipos de informação (correlação interitem e número de itens).

Em geral, considera-se que um valor acima de 0,9 é excelente; acima de 0,7 é aceitável; e valores abaixo de 0,5, inaceitáveis – mas há críticas a essa interpretação (SCHMITT, 1996). Segundo Morris (2011), os testes padronizados em larga escala costumam ter alta confiabilidade, mas podem apresentar baixa validade. Isso pode acontecer, por exemplo, se um teste que tenha objetivo de avaliar a competência em raciocínio científico for composto por itens bem elaborados para avaliar interpretação de texto. Os indicadores de confiabilidade não identificarão qualquer problema no teste, pois os itens se comportam de maneira coerente entre si. No entanto, o teste não mensura aquilo a que se propõe.

Ambas as propriedades se referem ao teste como um todo. Há também propriedades empiricamente observáveis que se referem aos itens. Tais propriedades costumam ser mais informativas ao educador, pois permitem a compreensão não apenas do conjunto de itens, mas também de cada um deles individualmente (maiores detalhes a seguir).

Neste trabalho, investigou-se a qualidade técnica do Enem em termos psicométricos, no nível do teste, mas es-

pecialmente no nível dos itens. Trata-se, portanto, de uma meta-avaliação do exame. O termo "meta-avaliação" foi introduzido por Michael Scriven, em 1969,[2] referindo-se ao ato de se avaliar uma avaliação, com objetivo de evitar distorções e imprecisões que possam prejudicar o processo educativo de crianças e jovens (TUFFLEBEAM, 2001).

**2** Scriven, M. S. An introduction to meta-evaluation. *Educational Products Report*, v. 2, p. 36-38, 1969.

A psicometria pode fornecer instrumentos bastante úteis para a meta-avaliação. Destacam-se a seguir três exemplos de estudos brasileiros em que se buscou, como aqui, utilizar técnicas psicométricas para a meta-avaliação em educação.

O primeiro estudo se refere a um exame de medicina, na área de anatomia clínica, em que se avaliou o comportamento empírico de 100 itens de múltipla escolha (SEVERO; TAVARES, 2010). O estudo utilizou algumas medidas de comportamento dos itens, como a carga fatorial padronizada (CFP) e a qualidade de ajuste do modelo (p-valor) dos itens. Os autores consideram de baixa qualidade os itens com CFP menor que 0,30 – o que corresponderia a uma discriminação menor que 0,314 no modelo logístico da TRI. Nos resultados globais, destaca-se a confiabilidade alta ($\alpha$ = 0,85). Embora o estudo tenha confirmado a qualidade do exame, foram encontrados 15 itens de baixa qualidade (15% do total), ou seja, com CFP menor do que 0,3. Além disso, três itens apresentaram p-valor maior ou igual a 0,05 indicando baixa qualidade de ajuste do modelo TRI.

O segundo estudo refere-se ao Exame Nacional de Desempenho dos Estudantes (Enade) na área de psicologia (PRIMI; HUTZ; SILVA, 2011). Diversos aspectos do exame são descritos e analisados, como o perfil da comissão elaboradora de provas, dimensionalidade, matriz de competências e habilidades, relação das habilidades gerais com conteúdos específicos, além de algumas propriedades psicométricas do exame e dos itens. Além de útil para os elaboradores do exame e interessados no tema, o estudo procurou validar a prova, comparando o desempenho de estudantes ingressantes e concluintes do curso – que se mostraram significativamente diferentes. Dentre os resultados do estudo, cabe aqui destacar dois pontos de corte utilizados pelos autores na identifi-

cação de itens com comportamento psicométrico abaixo do esperado: a) correlação item-total menor que 0,10; e b) carga fatorial menor que 0,30.

O terceiro estudo, também sobre o Enade (área de pedagogia), tem maior foco nos aspectos psicométricos dos itens, além de apresentar uma interessante revisão dos estudos de meta-avaliação no Brasil (LOPES; VENDRAMINI, 2015). Os autores utilizaram quatro critérios para verificar a qualidade dos itens: a) correlação item-total idealmente acima de 0,3, sendo aceitável acima de 0,2, dependendo dos outros critérios e do conteúdo do item; b) ajuste do modelo (*infit* e *outfit*) com valor entre 0,5 e 1,4, e idealmente abaixo de 1,2; c) correspondência entre a distribuição da dificuldade dos itens (limitada ao intervalo de -4 a 4) e a distribuição da proficiência na população; d) carga residual menor que 0,40. A prova foi composta de uma parte geral e outra específica. Dos sete itens relativos ao componente geral, todos apresentaram valores aceitáveis na correlação item-total, no ajuste do modelo e na dificuldade. Em relação à carga residual, dois itens apresentaram valor maior que 0,40 (ou menor que -0,40). No componente específico da prova, foram anulados dois itens pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), restando 26 itens válidos. Desses, todos apresentaram ajuste dentro dos limites estabelecidos, e a distribuição da dificuldade acompanhou, de modo geral, a distribuição da proficiência na população. Por outro lado, um item obteve carga residual excessiva, enquanto oito itens (31%) apresentaram correlação item-total abaixo de 0,3.

Na próxima seção, há uma explicação mais detalhada de algumas técnicas e seus usos em avaliações educacionais.

## O COMPORTAMENTO EMPÍRICO DOS ITENS

Segundo o documento *Standards for educational and psychological testing* (AERA; APA; NCME, 2014, itens 4.2, 4.10), os exames devem apresentar especificações psicométricas. Precisam ser documentados diversos aspectos técnicos do teste, tais como o modelo de análise, o ajuste do modelo (no caso da TRI), assim como os critérios para triagem de itens,

tais como dificuldade, discriminação, correlação interitem, correlação item-total, além do Funcionamento Diferencial do Item (DIF) entre os maiores grupos. Os métodos de triagem de itens – fundamentais durante o pré-teste – também são úteis *a posteriori* para meta-avaliação, pois permitem avaliar a qualidade dos itens por seu comportamento empírico.

Há diversos métodos para analisar o comportamento empírico dos itens. Eles podem ser utilizados como critério para avaliar a qualidade dos testes, assim como para outras finalidades, e muitas vezes os resultados convergem entre os métodos (HUDSON, 1991). Há dois grandes grupos de técnicas para analisar os resultados de testes padronizados: aqueles reunidos sob a alcunha da Teoria Clássica dos Testes – que se baseia em conceitos estatísticos mais genéricos, como correlação e porcentagem –, e os da Teoria da Resposta ao Item – com modelos próprios, desenhados especificamente para analisar testes de inteligência, conhecimento ou outras propriedades que não podem ser diretamente observadas (SOARES, 2005). Há diversas formas de se diferenciar esses dois grandes conjuntos de modelos da psicometria, mas é importante ter em mente que "a TRI não é, como se poderia ingenuamente acreditar, um novo enfoque psicométrico, ainda que alcance solucionar certos problemas da TCT" (ANDRIOLA, 2009, p. 322). Ou seja, a TRI não viola os princípios básicos da TCT, apenas adiciona novos princípios que permitem responder novas questões.[3]

**3** O leitor poderá verificar essa convergência, por exemplo, na Tabela 5 e no Anexo deste artigo.

O Enem, em sua primeira década de existência, baseava-se na TCT, sendo a pontuação de cada aluno correspondente ao percentual de acertos. A partir de 2009, a TRI começou a ser aplicada no exame (ANDRADE; KARINO, 2011), sendo utilizado o modelo logístico de três parâmetros (ML3P). Esse modelo descreve a probabilidade de uma pessoa acertar um item (com parâmetros A, B e C conhecidos), dependendo da sua proficiência.[4] A proficiência seria um traço unidimensional, porém não observável: tipicamente, uma característica psicológica. O parâmetro B corresponde ao nível de dificuldade do item, enquanto A define sua discriminação, ou seja, sua capacidade de diferenciar as pessoas com proficiência acima de B das pessoas com proficiência abaixo de B.

**4** Essa relação entre probabilidade de certo e proficiência é representada pela Curva Característica do Item.

O parâmetro C representa a probabilidade de um indivíduo com baixa proficiência responder corretamente a questão, ou seja, a chance de acerto ao acaso (ANDRADE; TAVARES; VALLE, 2000). Nos itens de múltipla escolha com cinco alternativas, esse parâmetro varia em torno de 20%. No modelo logístico de dois parâmetros (ML2P), por sua vez, são estimados apenas os parâmetros A e B – discriminação e dificuldade –, sendo o parâmetro C sempre igual a zero.

Seja qual for o modelo, os parâmetros são estimados para cada item, o que pode ajudar a verificar a qualidade psicométrica de uma prova. Em um exame de seleção, por exemplo, é desejável que o parâmetro A seja alto, especialmente na porção superior da escala de proficiência, permitindo selecionar com acuidade os candidatos mais habilidosos. Também é desejável haver, no conjunto dos itens, uma distribuição equilibrada do parâmetro B ao longo da escala de proficiência. Além disso, a combinação desses dois parâmetros também é importante, pois define o quanto cada item pode ser informativo em cada ponto da escala de proficiência, o que pode ser representado na Curva de Informação do Item (ANDRADE; TAVARES; VALLE, 2000). Por fim, além dos parâmetros e da Curva de Informação do Item, também é possível utilizar como indicador de qualidade a consistência do ajuste dos modelos, a saber, o valor de prova dos modelos de cada item (SEVERO; TAVARES, 2010).

A seguir, detalham-se as técnicas utilizadas como critério de avaliação dos itens neste trabalho.

### CORRELAÇÃO ITEM-TOTAL

A correlação item-total é a correlação entre o escore do item (no caso, 0 ou 1), e a média do escore no conjunto dos itens (ou nos outros itens). É uma técnica bastante utilizada para verificação da qualidade psicométrica dos itens, servindo por vezes como critério para eliminação de itens potencialmente problemáticos. Uma baixa correlação item-total sugere que o item não mede o traço latente captado pelo conjunto dos itens. Dependendo do objetivo do teste, do número de itens, da confiabilidade desejada, o mínimo estabelecido para um bom item pode variar. Um estudo para

avaliar proficiência em cursos de medicina considerou inadequados os itens cuja correlação item-total fosse inferior a 0,4 (BASS; WILSON; GRIFFITH, 2003). Em outro estudo, analisando um questionário sobre satisfação marital, foi estabelecido um limite mais severo, mínimo de 0,5 (ROACH; FRAZIER; BOWDEN, 1981). Enquanto um terceiro estudo, sobre demência, optou por eliminar apenas os itens com correlação item-total inferior a 0,25 (GILLEARD; GROOM, 1994). Por fim, em um estudo sobre validade e confiabilidade de escalas sobre percepção de saúde para diabéticos, foram excluídos três itens (de um total de 36) que apresentaram correlação item-total menor que 0,3 (KARTAL et al., 2007).

### COEFICIENTE BISSERIAL

O coeficiente bisserial do item, em termos técnicos, está relacionado à correlação entre o acerto no item (0 ou 1) e o escore do examinando na prova. Na verdade, é calculado para todas as alternativas do item (cinco, no caso do Enem), de forma que "um item tem bom desempenho quando esse coeficiente tem valor 'alto' positivo associado à alternativa correta e valores negativos associados aos distratores" (ANDRADE; KLEIN, 2005, p. 109). O coeficiente bisserial indica o quanto cada alternativa atraiu os alunos mais proficientes.

No Saeb de 2005, essa técnica foi incluída nos procedimentos para triagem de itens depois do pré-teste. Para o gabarito (alternativa correta), foram encaminhados para análise pedagógica os itens com coeficiente bisserial menor ou igual a 0,15. Além disso, também ficaram retidos

> [...] itens com dois coeficientes bisseriais de distratores (alternativas erradas) maiores que 0,10 ou coeficiente bisserial de um distrator maior que a bisserial da alternativa correta. (ANDRADE; LAROS; GOUVEIA, 2010, p. 426)

No Enem, por sua vez,

> [...] após análise dos resultados do pré-teste, são selecionadas aquelas que apresentam pertinência mais direta com a habilidade, originalidade e coeficiente bisserial maior de 30. (FINI, 2005, p. 103)

Ao menos esse foi o procedimento adotado na época, e não foram encontrados outros documentos mais recentes descrevendo as especificações psicométricas do Enem. Em 2009, cabe lembrar, o Enem passou por importantes transformações, tais como: reestruturação da Matriz de Referências, consequente mudança nos tipos de prova aplicados (inicia-se em 2009 a divisão do Enem em quatro provas distintas, além da redação), ampliação para dois dias de prova e uso da TRI para análise dos resultados (TRAVITZKI, 2013a).

## DISCRIMINAÇÃO E DIFICULDADE

Para garantir a precisão do teste, é desejável que a discriminação seja alta, para que o item seja mais informativo em certa porção da escala, embora haja evidências de que uma discriminação muito alta possa produzir distorções (MASTERS, 1988). A dificuldade do item, por sua vez, não deve ser muito alta ou muito baixa, caso contrário o item apresenta pouca informação sobre a população analisada.

Tais critérios foram utilizados, por exemplo, no estudo de Alnabhan e Harwell (2001), para avaliar a qualidade de um teste de admissão para faculdade. Os autores se basearam em limites utilizados anteriormente por Muthén, Kao e Burstein (1991). Ajustaram o modelo de dois parâmetros (ML2P) e consideraram adequados os itens que apresentaram: a) dificuldade entre -3 e 3 de desvio padrão; b) discriminação maior do que 0,5 – segundo os autores, um ponto de corte comum nos estudos iniciais –; c) um bom ajuste no modelo TRI ($\alpha < 0.05$); d) percentual de acerto menor do que 90% (ALNABHAN; HARWELL, 2001). Os resultados apontaram que 30% dos itens não cumpriam pelo menos uma dessas quatro condições. Tais resultados foram enviados a um conjunto de especialistas que decidiria entre modificar ou excluir tais itens futuramente.

Uma classificação do parâmetro A (discriminação) do modelo logístico é apresentada por Baker (2001, p. 34). A discriminação, segundo o autor, poderia ser: nenhuma (0), muito baixa (0,01 a 0,34), baixa (0,35 a 0,64), moderada (0,65 a 1,34), alta (1,35 a 1,69), muito alta (maior que 1,70) ou perfeita (+ infinito). Adicionalmente, Severo e Tavares (2010)

consideram que os itens são de baixa qualidade quando a discriminação é menor do que 0,314 (que corresponderia a uma carga fatorial padronizada de 0,30).

## MÉTODOS

### *SOFTWARE*

Neste trabalho foram utilizados apenas *softwares* livres, como o LibreOffice (escritório) e a plataforma Linux (sistema operacional), para usos mais gerais. Para realização das análises psicométricas com base na TCT e TRI, utilizou-se o R (estatística), que permite a programação e automação de procedimentos customizados.

Uma das vantagens do R é que, por ser um *software* de desenvolvimento livre e de código aberto, permite análises estatísticas complexas, transparentes e reprodutíveis, na medida em que os dados e o código fonte estejam disponíveis. Além disso, a arquitetura aberta e não proprietária proporciona condições ideais para o trabalho colaborativo, de forma que pessoas em todo mundo estão constantemente criando e testando novos pacotes do R, o que gera grande diversidade de funcionalidades. Os principais pacotes utilizados neste trabalho foram o "ltm" (RIZOPOULOS, 2006), para estimações via TRI, e o "psych" (REVELLE, 2016), para técnicas da TCT.

## DADOS

Foram analisados os microdados do Enem de 2009 e 2011. Inicialmente, foi analisado apenas o exame de 2009, por tratar-se da primeira edição do Enem no novo formato.[5] Em virtude do "vazamento" da prova em 2009, optou-se por analisar também o de 2011. O exame é composto por quatro provas objetivas, que correspondem às quatro áreas de conhecimento avaliadas: Ciências Humanas (CH), Ciências Naturais (CN), Linguagens e Códigos (LC) e Matemática (MT). A análise psicométrica foi realizada independentemente nas oito provas objetivas, quatro de cada ano. O processo de filtragem reduziu a amostra a menos de 10% do total, como se vê na Tabela 1, para a prova de Ciências Humanas.

**5** Os resultados desta primeira análise foram expostos na VII Reunião da Associação Brasileira de Avaliação Educacional (Abave) (TRAVITZKI, 2013b).

Não foram encontrados documentos do Inep explicando certos detalhes metodológicos sobre o cálculo das notas do Enem (por exemplo, se são incluídos todos os registros na calibração dos itens, ou apenas dos concluintes), mas sabe-se que é utilizado o ML3P, que a proficiência é estimada pelo método *Expected A Posteriori* (EAP),[6] e que os cálculos são feitos por três grupos independentes de especialistas (KARINO; BARBOSA, 2012).

Na tentativa de garantir a qualidade da análise, os dados foram filtrados segundo quatro critérios: 1) ter estudado em escola regular no EM; 2) ser concluinte do EM; 3) estar presente na prova; 4) ter recebido o caderno azul de prova. A Tabela 1 exemplifica esse recorte da amostra no caso da prova de Ciência Humanas.

**6** Um método bayesiano bastante utilizado na estimação de proficiências pela TRI. A estatística bayesiana tem a vantagem de produzir estimativas mesmo quando todos os itens estão certos, ou todos estão errados (KLEIN, 2013).

**TABELA 1 – Tamanho da amostra antes e depois do processo de filtragem (provas de Ciências Humanas)**

| | ENEM 2009 | ENEM 2011 |
|---|---|---|
| Total de inscritos no Exame | 4.148.721 | 5.380.856 |
| Amostra analisada | 218.180 | 302.993 |

Fonte: Elaboração do autor a partir dos dados do Enem (BRASIL, 2016).

Em relação às provas propriamente ditas, o total foi de 354 itens analisados. Isso porque seriam, a rigor, 180 itens em cada prova, ou seja, 360 itens no total, mas na prova de LC/2011 houve cinco itens relativos à língua estrangeira que foram retirados da análise. Além disso, em LC/2009 há um item anulado no próprio gabarito dos microdados.

Um fato importante que interfere nos resultados é que o ano de 2009 foi certamente peculiar para o Enem, pois houve "vazamento" da prova, e um novo exame precisou ser produzido em curto prazo. Além disso, as mudanças na matriz de referências haviam acabado de ser implementadas; e, portanto, talvez não houvesse itens "de reserva" suficientes no banco de itens para a confecção da segunda prova. Seria esperada, assim, uma qualidade inferior do exame desse ano em relação a outros. Uma hipótese que foi confirmada em nossos resultados.

## CRITÉRIOS PSICOMÉTRICOS DO INDICADOR DE QUALIDADE DO ITEM

Para verificar o comportamento empírico dos itens, foram utilizadas técnicas da TCT e da TRI. Para a definição dos limites de cada critério, foram utilizados os referenciais da literatura supracitada, levando também em conta que o Enem é um exame extremamente importante – e deve, portanto, ser um instrumento de alta qualidade.

**TABELA 2 – Critérios psicométricos utilizados para avaliar a qualidade dos itens das provas objetivas consideradas no estudo**

| CLASSIFICAÇÃO DO ITEM | TCT | | TRI | |
|---|---|---|---|---|
| | CORRELAÇÃO (CORR) | 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: Elaboração do autor com base nos dados do Enem (BRASIL, 2016).

Notas: CORR = correlação item-total; BISS = coeficiente bisserial da alternativa correta; o modelo da TRI ajustado foi o ML2P.

O significado de cada indicador pode ser traduzido, em linguagem mais acessível, nas perguntas:

1. (CORR)[7] *O item está suficientemente relacionado ao resto da prova?*
2. (BISS) *A alternativa correta atraiu os mais proficientes?*
3. (a) *O item discriminou bem no seu nível de proficiência?*
4. (FIT) *Foi possível ajustar um bom modelo da TRI para este item?*

**7** A correlação item-total foi calculada incluindo-se todos os itens no total.

Os dois primeiros indicadores de qualidade do item são provenientes da TCT, enquanto os dois últimos provém da TRI. Seria possível incluir outros indicadores de qualidade, como correlação interitem, amplitude do parâmetro de dificuldade, ou ainda a existência de DIF.

O ML2P, que fornece os indicadores 3 e 4, foi ajustado com as mesmas configurações para as oito provas,[8] possibilitando a comparação dos resultados. Seria possível, caso necessário, calibrar de forma mais confiável os itens mal ajustados, com outras configurações ou algoritmos – o que provavelmente ocorreu na análise oficial do Enem.

**8** Foram utilizadas configurações normalmente suficientes para uma boa calibração dos itens: amostra de 30 mil provas; 61 pontos de quadratura; 1.000 iterações quase-Newton; 400 iterações *Estimation-Maximization*. O código R correspondente a essas configurações é: ltm(respostas-z1, control = list(GHk = 61, iter.em = 400, iter.qN = 1000)).

Além da análise individual de cada indicador, buscou-se reunir todos em um indicador global simples. Como foi observada correlação entre os três primeiros indicadores, o critério utilizado para o indicador global foi:

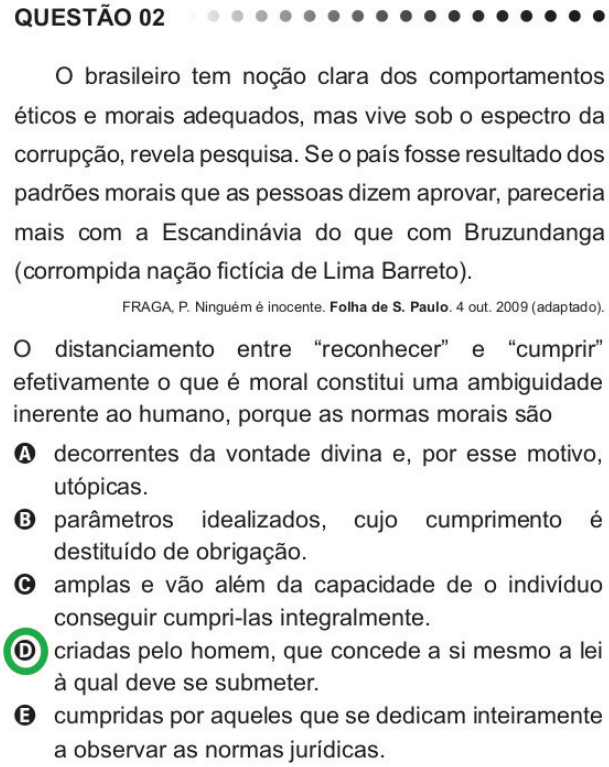
a. item duvidoso (indicador global): duvidoso em pelo menos três dos quatro indicadores, ou ruim em pelo menos um;
b. item ruim (indicador global): ruim em pelo menos três dos quatro indicadores, ou anulado no gabarito.

## RESULTADOS PSICOMÉTRICOS E CONTEÚDO DOS ITENS

Nesta seção analisa-se alguns itens para ilustrar a utilidade das técnicas psicométricas para educadores e formuladores de avaliações. Como o objetivo é investigar certos comportamentos empíricos anômalos, foram selecionados apenas itens com tal característica.

A questão 2 da prova de Ciências Humanas (Figura 1), por exemplo, foi classificada como "duvidosa" em três indicadores, sendo aceitável apenas em termos do FIT. A análise psicométrica revelou que, além disso, em um dos distratores (alternativa B) houve tanta atratividade quanto na alternativa correta (D), o que não seria esperado em um bom item. Ou seja, nas duas alternativas, o coeficiente bisserial ficou em torno de 0,2 (Figura 2A) – o que pode ser considerado alto para um distrator, e baixo para o gabarito.

FIGURA 1 – Item 2 da prova de Ciências Humanas do Enem 2011



**QUESTÃO 02**

O brasileiro tem noção clara dos comportamentos éticos e morais adequados, mas vive sob o espectro da corrupção, revela pesquisa. Se o país fosse resultado dos padrões morais que as pessoas dizem aprovar, pareceria mais com a Escandinávia do que com Bruzundanga (corrompida nação fictícia de Lima Barreto).

FRAGA, P. Ninguém é inocente. **Folha de S. Paulo**. 4 out. 2009 (adaptado).

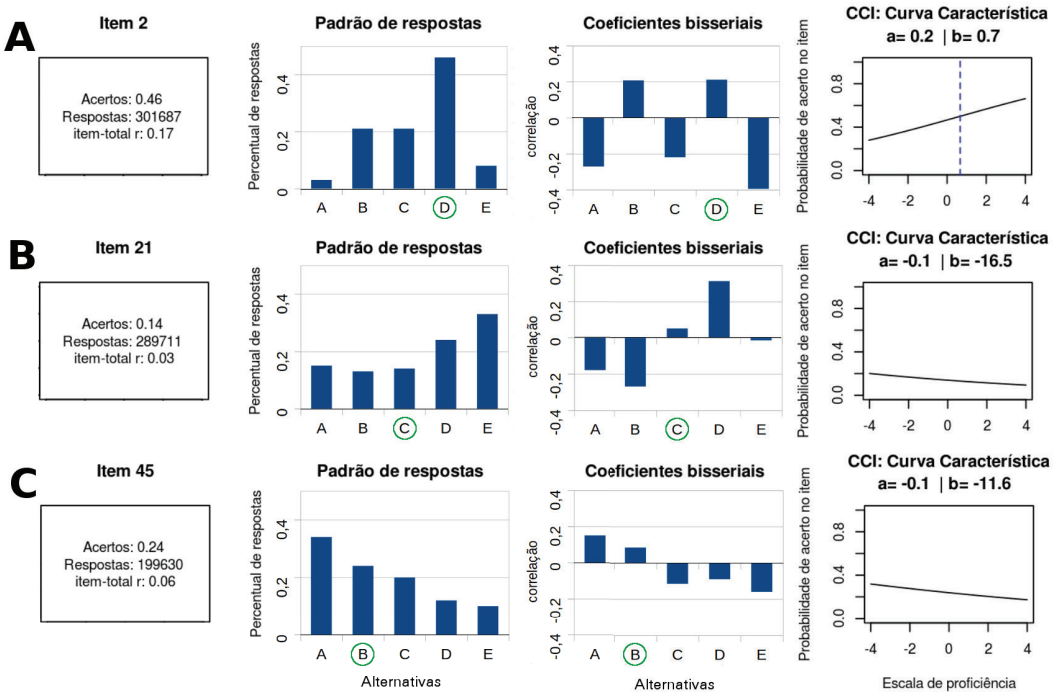
O distanciamento entre "reconhecer" e "cumprir" efetivamente o que é moral constitui uma ambiguidade inerente ao humano, porque as normas morais são

A) decorrentes da vontade divina e, por esse motivo, utópicas.
B) parâmetros idealizados, cujo cumprimento é destituído de obrigação.
C) amplas e vão além da capacidade de o indivíduo conseguir cumpri-las integralmente.
D) criadas pelo homem, que concede a si mesmo a lei à qual deve se submeter.
E) cumpridas por aqueles que se dedicam inteiramente a observar as normas jurídicas.

Fonte: Enem 2011 – prova de Ciências Humanas, caderno azul (BRASIL, 2016).

O enunciado correto (alternativa D) diz que "as normas morais são criadas pelo homem, que concede a si mesmo a lei à qual deve se submeter", o que parece bastante plausível, embora talvez um pouco idealizado. Contudo, ao notarmos que o enunciado trata inicialmente da "corrupção" e do "distanciamento entre reconhecer e cumprir" como "ambiguidade inerente ao humano", a hipótese anterior (alternativa D) é tão plausível quanto considerar que "as normas morais são parâmetros idealizados, cujo cumprimento é destituído de obrigação" (alternativa B). É justamente isso que informa o coeficiente bisserial: que as duas alternativas apresentaram a mesma capacidade de atrair os candidatos mais proficientes. Talvez tenha sido esse o motivo pelo qual o item foi classificado como "duvidoso" nos três indicadores.

FIGURA 2 – Resultados psicométricos para três itens avaliados no estudo




Fonte: Elaboração do autor com dados do Enem (BRASIL, 2016).

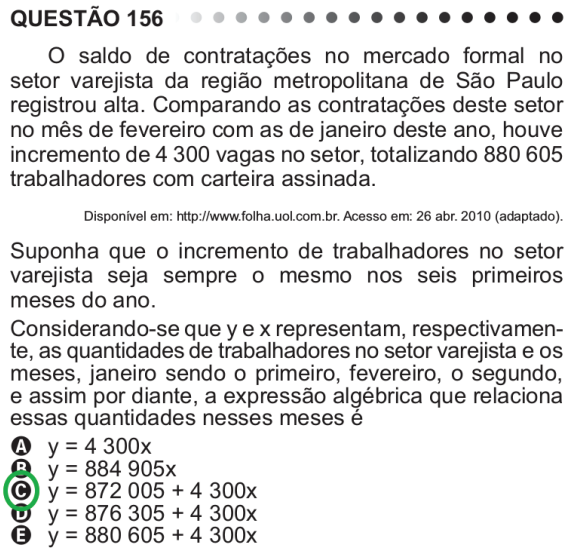
Nota: A é o item 2 da prova de Ciências Humanas do Enem 2011; B é o item 21 da prova de Matemática do Enem 2011; e C é o item 45 da prova de Matemática do Enem 2009.

Ou seja, a análise psicométrica e a leitura do enunciado convergem para o mesmo ponto: o item apresenta duas alternativas plausíveis. Embora não seja desejável em uma prova objetiva, tal duplicidade é bastante comum nas áreas humanas, em especial na filosofia. Diferente das ciências exatas e naturais – em que há grande esforço para se elaborar um único paradigma explicativo e, portanto, uma única resposta considerada correta –, nas ciências humanas e filosofia, é esperado que o aluno aprenda a interpretar a realidade de diversas formas, não necessariamente uma mais verdadeira do que outra. A temática da moral, tratada no item, apresenta diversas tensões internas, e pode ser compreendida sob diferentes perspectivas. O item, curiosamente, mostra duas perspectivas sobre a moral que nos parecem igualmente plausíveis – o que explicaria seu comportamento empírico anômalo. Tais considerações ilustram uma limitação importante das provas objetivas no que se refere a

avaliar certas habilidades. O problema proposto pelo item nos parece importante e interessante. Contudo, talvez ele fosse mais bem avaliado com outro tipo de instrumento, como um item dissertativo, uma redação ou mesmo uma prova de múltipla escolha que permitisse ao candidato escolher mais de uma alternativa.

No exemplo do item 2, a psicometria aponta a existência de um distrator demasiadamente atraente, e a análise qualitativa individual do item o confirma – uma evidência da consistência entre ambas as abordagens. A análise psicométrica pode ajudar a confirmar a opinião de educadores e avaliadores a respeito da qualidade dos itens e, no sentido inverso, a observação caso a caso dos itens pode ajudar a confirmar a eficácia da psicometria e interpretar seus resultados.

FIGURA 3 – Item 21 da prova de Matemática do Enem2011



**QUESTÃO 156**

O saldo de contratações no mercado formal no setor varejista da região metropolitana de São Paulo registrou alta. Comparando as contratações deste setor no mês de fevereiro com as de janeiro deste ano, houve incremento de 4 300 vagas no setor, totalizando 880 605 trabalhadores com carteira assinada.

Disponível em: http://www.folha.uol.com.br. Acesso em: 26 abr. 2010 (adaptado).

Suponha que o incremento de trabalhadores no setor varejista seja sempre o mesmo nos seis primeiros meses do ano.

Considerando-se que y e x representam, respectivamente, as quantidades de trabalhadores no setor varejista e os meses, janeiro sendo o primeiro, fevereiro, o segundo, e assim por diante, a expressão algébrica que relaciona essas quantidades nesses meses é

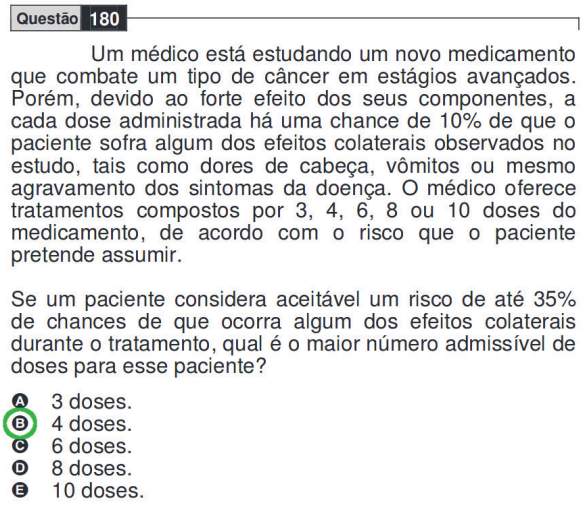
A) y = 4 300x
B) y = 884 905x
C) y = 872 005 + 4 300x
D) y = 876 305 + 4 300x
E) y = 880 605 + 4 300x

Fonte: Enem 2011 – prova de Matemática, caderno azul (BRASIL, 2016).

O item 21 da prova de Matemática de 2011 (Figura 3), por sua vez, foi avaliado como "ruim" nos três indicadores (exceto ajuste), sendo portanto considerado "ruim" no indicador global. Ao buscar respondê-lo, uma pessoa com habilidade de usar equações de primeiro grau chegaria, sem grandes dificuldades, à alternativa D. Contudo, para escolher

a alternativa correta C, é necessário também que ela seja habilidosa na linguagem escrita, além de ser atenta a pequenos detalhes: como, por exemplo, que o número citado no enunciado (880.605) se refere a fevereiro, e não a janeiro – e que, portanto, x é igual a 2, nesse caso, o que justifica a alternativa C como correta. Os coeficientes bisseriais confirmam a atratividade do distrator D, maior que a da alternativa correta C (Figura 2B).

**FIGURA 4 – Item 45 da prova de Matemática do Enem 2009**



Questão 180

Um médico está estudando um novo medicamento que combate um tipo de câncer em estágios avançados. Porém, devido ao forte efeito dos seus componentes, a cada dose administrada há uma chance de 10% de que o paciente sofra algum dos efeitos colaterais observados no estudo, tais como dores de cabeça, vômitos ou mesmo agravamento dos sintomas da doença. O médico oferece tratamentos compostos por 3, 4, 6, 8 ou 10 doses do medicamento, de acordo com o risco que o paciente pretende assumir.

Se um paciente considera aceitável um risco de até 35% de chances de que ocorra algum dos efeitos colaterais durante o tratamento, qual é o maior número admissível de doses para esse paciente?

A 3 doses.
B 4 doses.
C 6 doses.
D 8 doses.
E 10 doses.

Fonte: Enem 2009 – prova de Matemática, caderno azul (BRASIL, 2016).

Outro item de Matemática classificado como "ruim" nos três indicadores (exceto ajuste) foi o item 45 da prova de 2009 (Figura 4). Os resultados da análise psicométrica (Figura 2C) sugerem que, mais uma vez, o distrator A atraiu mais os candidatos proficientes do que a alternativa correta B. Embora o item esteja matematicamente correto, a formulação do problema – e as alternativas de resposta – parece ter confundido os candidatos. O item pede que se calcule probabilidades cumulativas partindo de um problema relacionado a remédios. Se o risco de se tomar uma dose é 10%, qual seria o risco de se tomar quatro doses? Aparentemente, seria 40%, como pensaram os candidatos que escolheram a alternativa A. Contudo, para resolver o problema, o candidato deveria

calcular a chance de as coisas darem certo, em vez do risco de darem errado. Ou seja, o cálculo correto seria multiplicar 90% por si mesmo quatro vezes ($0{,}9^4$), o que daria um risco em torno de 35%, em vez de 40%.

Em suma, a análise cuidadosa do enunciado dos itens confirma, na maioria dos casos, problemas que foram previamente detectados com técnicas psicométricas quantitativas, mostrando também como a articulação entre metodologias quantitativas e qualitativas pode ser frutífera. As técnicas psicométricas se mostram potencialmente úteis não apenas para gestores e sistemas de avaliação, mas também para os educadores que desejem compreender melhor seus educandos ou saber se suas provas estão bem elaboradas. Evidentemente, reconhecer a eficácia e utilidade das técnicas psicométricas não significa aderir acriticamente a elas nem tampouco que a avaliação educacional se deva reduzir à psicometria. Significa simplesmente que os educadores podem apropriar-se de algumas novas ferramentas que ajudem a resolver velhos e novos problemas.

## RESULTADOS PSICOMÉTRICOS E META-AVALIAÇÃO

Dentre as oito provas analisadas, apenas uma apresentou confiabilidade inadequada ($\alpha < 0{,}6$): a prova de Matemática de 2009, que também teve a mais alta dificuldade média (Tabela 3). Nota-se também que a prova de Ciências Naturais apresentou, nos dois anos, confiabilidade aceitável, porém próxima do limiar inferior. Se esse fato se confirma em outros anos, mereceria maiores investigações. As provas de Ciências Humanas e Linguagens e Códigos, por outro lado, mostraram-se excelentes em termos de confiabilidade, de acordo com os padrões esperados do coeficiente $\alpha$.

TABELA 3 – Características gerais das oito provas avaliadas no estudo

| | CIÊNCIAS HUMANAS | CIÊNCIAS DA NATUREZA | LINGUAGENS E CÓDIGOS | MATEMÁTICA |
|---|---|---|---|---|
| **ENEM 2009** | | | | |
| Número de itens | 45 | 45 | 44 | 45 |
| Número de respondentes | 218.180 | 218.180 | 200.242 | 200.242 |
| Média de acertos | 0,36 | 0,34 | 0,46 | 0,25 |
| Coeficiente α | 0,83 | 0,74 | 0,85 | 0,59 |
| Correlação interitem média | 0,09 | 0,06 | 0,11 | 0,03 |
| **ENEM 2011** | | | | |
| Número de itens | 45 | 45 | 40 | 45 |
| Número de respondentes | 302.993 | 302.993 | 291.219 | 291.219 |
| Média de acertos | 0,44 | 0,32 | 0,48 | 0,33 |
| Coeficiente α | 0,84 | 0,75 | 0,84 | 0,84 |
| Correlação interitem média | 0,10 | 0,06 | 0,11 | 0,10 |

Fonte: Elaboração do autor a partir dos dados do Enem (BRASIL, 2016).

Nota: Na prova de Linguagens e Códigos do Exame de 2009, houve uma questão anulada. E em 2011 essa prova começou a apresentar 5 itens de língua estrangeira, que foram excluídos da análise realizada no estudo.

Cabe notar também que, em todas as provas, a correlação interitem média foi maior do que zero, e não passou de 0,1, sugerindo que não há grande sobreposição entre os itens, de forma a garantir minimamente o pressuposto da independência local, necessário para os modelos da TRI utilizados neste trabalho e também pelo Inep, a partir de 2009. Uma correlação interitem muito alta pode indicar redundância de dois itens (ZAICHKOWSKY, 1994). Por outro lado, uma correlação interitem média de 0,03 confirma a baixa confiabilidade da prova de Matemática de 2009, apontada pelo coeficiente α.

O outro pressuposto, da unidimensionalidade, foi verificado por meio de análise fatorial para um fator, análise da estrutura muito simples (*Very Simple Structure*) e análise paralela sobre a matriz de correlações tetracóricas dos itens. Em todas as provas, a análise fatorial e a estrutura muito simples mostraram que um fator é suficiente para abarcar a variabilidade das respostas. Há, porém, frequentemente um segundo fator significativo, embora de menor importância, detectado pela análise paralela. De modo geral, esses resultados confirmam o cumprimento dos pressupostos básicos para o uso dos modelos logísticos da TRI.

Em relação ao grau de dificuldade, nota-se, de um modo geral, que nenhuma prova se mostrou fácil, o que seria esperado, de acordo com as especificações do Enem: 20% de itens fáceis; 40%, médios; e 40%, difíceis.[9] As provas de Lin-

**9** Para uma verificação mais precisa da distribuição do grau de dificuldade dos itens, seria necessária uma definição dos pontos de corte que distinguem os itens fáceis, médios e difíceis no Enem. Tal informação, contudo, não foi encontrada na documentação disponível.

guagens e Códigos apresentaram nível médio nos dois anos próximo a 50% de acertos. As provas de Ciências da Natureza e Matemática se mostraram consideravelmente difíceis, chegando a 25% de acertos, valor próximo à probabilidade de acerto ao acaso no Enem (20%). De acordo com os princípios da TRI, representados, por exemplo, na Curva de Informação do Teste (ANDRADE; TAVARES; VALLE, 2000), um exame que se concentra na porção superior da escala de proficiência pode ser pouco informativo sobre indivíduos de média e, principalmente, baixa proficiência. É possível, portanto, que o alto número de itens difíceis[10] tenha levado à baixa confiabilidade da prova MT/2009, tanto em α quanto na correlação interitem média.

**10** Nos dados em Anexo, pode-se conferir que apenas dois itens da prova MT/2009 apresentaram média de acerto superior a 50%, sendo que 38 deles (84% do total de itens) apresentou média de acerto inferior a 30%.

A Tabela 4 sintetiza os resultados gerais para cada indicador, em cada uma das oito provas. O critério que mais detectou itens possivelmente problemáticos foi, nos dois anos, a correlação item-total. Tal fato sugere que os limites adotados nesse critério podem ser demasiadamente rigorosos em relação aos outros; ou, alternativamente, também é possível que tais resultados indiquem certa dificuldade de se elaborar itens coerentes com uma suposta competência geral não observável. A discriminação do item também se mostrou sensível a comportamentos empíricos anômalos, sendo o segundo critério que mais identificou itens duvidosos ou ruins. Por fim, como seria de se esperar, dadas as condições privilegiadas de calibração – como uma grande amostra e número razoável de iterações –, há poucos itens com problema de ajuste. Mas destaca-se o fato de que, dos 45 itens da prova de Matemática de 2009, cinco apresentaram problemas de ajuste no ML2P.

**TABELA 4 – Porcentagem de itens classificados como *duvidosos ou ruins* segundo os critérios definidos no estudo, em cada uma das oito provas avaliadas**

| | CIÊNCIAS HUMANAS | CIÊNCIAS DA NATUREZA | LINGUAGENS E CÓDIGOS | MATEMÁTICA | TOTAL |
|---|---|---|---|---|---|
| **ENEM 2009** | | | | | |
| Correlação (CORR) | 29 % | 56 % | 23 % | 76 % | 46 % |
| Bisserial (BISS) | 11 % | 33 % | 7 % | 49 % | 25 % |
| Discriminação (a) | 18 % | 44 % | 14 % | 58 % | 34 % |
| Ajuste (FIT) | 2 % | 0 % | 0 % | 11 % | 3 % |
| **ENEM 2011** | | | | | |
| Correlação (CORR) | 24 % | 56 % | 20 % | 31 % | 33 % |
| Bisserial (BISS) | 13 % | 29 % | 8 % | 18 % | 17 % |
| Discriminação (a) | 16 % | 42 % | 15 % | 29 % | 26 % |
| Ajuste (FIT) | 0 % | 2 % | 0 % | 0 % | 1 % |

Fonte: Elaboração do autor a partir dos dados do Enem (BRASIL, 2016).

Nota-se também que, em 2009, houve 82 itens (46% do total) com correlação item-total menor do que 0,3 – condição importante para que um item possa de fato acrescentar algo à prova como um todo. Em 2011, foram identificados 58 itens (33%) com tal característica. Nesse sentido, embora a TRI permita que se produza resultados levando em conta essas diferentes "contribuições" de cada item,[11] uma frequência alta de itens pouco coerentes entre si (ou, em termos teóricos, com o construto subjacente à prova como um todo) pode ser considerada, no mínimo, certo desperdício de recursos públicos e privados.

Outra informação pertinente apresentada na Tabela 4 são os itens com bisseriais baixos. Isso porque o limite mínimo estabelecido para um bom item (0,3) corresponde ao limite estabelecido para triagem de itens na elaboração do "antigo Enem" (FINI, 2005). Com efeito, os resultados mostram que, segundo as especificações psicométricas do Enem, houve 45 itens (25%) inadequados na prova de 2009, enquanto em 2011 esse número foi reduzido a 30 (17%).

Observando cada indicador separadamente, cabe agora saber como eles se relacionam. A Tabela 5 mostra que há alta correlação entre eles, especialmente nas correlações que envolvem o parâmetro de discriminação – sugerindo que esse indicador poderia reunir, de forma sintética, os resultados do trio (correlação item-total, bisserial e discriminação). Cabe notar também que a correlação entre a correlação

**11** Como se pode notar na Tabela 5, há alta correlação entre a discriminação (de acordo com a TRI) dos itens e a correlação item-total (de acordo com a TCT). Tal discriminação, por sua vez, é fator predominante na determinação da quantidade máxima de informação que cada item pode prover a respeito do traço latente, seja qual for sua posição na escala.

item-total e a correlação bisserial se mostrou significativamente mais baixa do que as outras duas, que por sua vez se mostraram semelhantes entre si (Teste T, $p<0,01$). O ajuste (FIT) não foi incluído por ter menor variância, mas todos os sete itens com problema de ajuste apresentavam problemas nos outros indicadores.

**TABELA 5 – Correlações de Spearman entre os indicadores psicométricos adotados no estudo, considerando a classificação dos itens em cada um deles como "bom=0", "duvidoso=1" ou "ruim=2", para cada uma das oito provas avaliadas**

| ANO | PROVA | CORRELAÇÃO X BISSERIAL | CORRELAÇÃO X DISCRIMINAÇÃO | BISSERIAL X DISCRIMINAÇÃO |
|---|---|---|---|---|
| 2009 | Ciências Humanas | 0,66 | 0,79 | 0,81 |
| | Ciências da Natureza | 0,73 | 0,87 | 0,83 |
| | Linguagens e Códigos | 0,53 | 0,75 | 0,70 |
| | Matemática | 0,70 | 0,80 | 0,86 |
| 2011 | Ciências Humanas | 0,74 | 0,79 | 0,92 |
| | Ciências da Natureza | 0,74 | 0,84 | 0,81 |
| | Linguagens e Códigos | 0,65 | 0,87 | 0,73 |
| | Matemática | 0,71 | 0,95 | 0,74 |
| média | | 0,68 | 0,83 | 0,80 |

Fonte: Elaboração do autor a partir dos dados do Enem (BRASIL, 2016).

Algumas possíveis explicações, não excludentes entre si, para as altas correlações observadas seriam: os limites estabelecidos nos critérios deste trabalho são adequados e, em alguma medida, correspondentes entre si;[12] os três indicadores se referem a características semelhantes dos itens; as três características dos itens mensuradas pelos indicadores podem ser diferentes, mas estão intimamente ligadas na prática.

Seja qual for a explicação, a correlação encontrada entre os indicadores nos levou a certas opções no momento de articulá-los em um indicador global de qualidade do item. Foi considerado globalmente "duvidoso" o item avaliado como "duvidoso" em pelo menos três indicadores, ou "ruim" em um ou dois. De forma semelhante, foi considerado globalmente "ruim" o item avaliado como "ruim" em pelo menos três indicadores. Os resultados podem ser conferidos na Tabela 6.

**12** Na verdade, talvez os limites do indicador ATR estejam um pouco baixos em relação aos outros, dado que em muitos itens houve comportamento anômalo em COR e DCR, mas não em ATR.

**TABELA 6 – Porcentagem de itens classificados como *bons*, *duvidosos ou ruins*, segundo o indicador global definido no estudo, em cada uma das oito provas avaliadas**

| CLASSIFICAÇÃO DO ITEM | CIÊNCIAS HUMANAS | CIÊNCIAS DA NATUREZA | LINGUAGENS E CÓDIGOS | MATEMÁTICA | TOTAL |
|---|---|---|---|---|---|
| **ENEM 2009** | | | | | |
| Bom | 89 % | 67 % | 91 % | 51 % | 75 % |
| Duvidoso | 4 % | 29 % | 7 % | 33 % | 18 % |
| Ruim | 7 % | 4 % | 2 % * | 16 % | 7 % |
| **ENEM 2011** | | | | | |
| Bom | 87 % | 71 % | 92 % | 82 % | 83 % |
| Duvidoso | 11 % | 20 % | 3 % | 16 % | 13 % |
| Ruim | 2 % | 9 % | 5 % | 2 % | 4 % |

Fonte: Elaboração do autor a partir dos dados do Enem (BRASIL, 2016).
Nota: * Correspondente ao item anulado pelo Inep.

Considerando, portanto, o conjunto de itens "duvidosos" e "ruins", o indicador global apontou um total de 46 itens (25%) com comportamento empírico anômalo no ano de 2009, e 30 itens (17%) em 2011. Resultado praticamente idêntico ao obtido com apenas um dos três indicadores, o bisserial (BISS), segundo o critério estabelecido no "antigo Enem".

Cabe observar, por fim, que um elemento associado aos problemas de qualidade das provas de Matemática e Ciências da Natureza de 2009 é a posição do item na prova. Em geral, os itens com comportamento anômalo se encontram na parte final do teste (ver Anexo). Uma possível explicação é que os itens anteriores demandassem muito trabalho, levando os candidatos a uma maior exaustão mental (ou menor tempo disponível) ao final da prova. Outra possível explicação, não excludente, é que boa parte dos candidatos tenha deixado de fazer as questões por considerá-las excessivamente difíceis, sendo assim mais produtivo investir o tempo de prova nos outros itens. Ambas explicações estão de acordo com o alto grau de dificuldade observado nessas provas (ver Tabela 3).

## CONCLUSÃO

De modo geral, segundo os critérios adotados, o Enem apresentou qualidade adequada em 2011, porém duvidosa em 2009. A confiabilidade da prova de Matemática de 2009 se mostrou abaixo dos limites considerados aceitáveis ($\alpha < 0{,}6$).

Por outro lado, as outras sete provas não apresentaram esse problema. É importante levar em conta que, em 2009, um ano de grandes mudanças estruturais, a prova do Enem precisou ser refeita, devido a um "vazamento" de itens.

Um fato talvez relevante é que a prova de Ciências Naturais apresentou indicadores relativamente piores nos dois anos, o que pode sugerir necessidade de revisão de procedimentos para criação e revisão de itens dessa área do conhecimento. A área de Linguagens e Códigos, por sua vez, apresentou os melhores índices nos dois anos.

Para avaliar os itens, utilizou-se quatro indicadores: 1) CORR = correlação item-total; 2) BISS = coeficiente bisserial da alternativa correta; 3) a = Parâmetro a (discriminação) do ML2P; 4) FIT = ajuste ao modelo ML2P. Os três primeiros indicadores apresentaram alta correlação entre si, especialmente com a discriminação. Além disso, os resultados do BISS (coeficiente bisserial da alternativa correta) foram praticamente idênticos (quando vistos de forma geral) aos resultados do indicador global: 25% de itens com comportamento empírico anômalo em 2009, e 17% em 2011. Se considerados apenas os itens avaliados como "ruins" no indicador global, os resultados diminuem para 7% em 2009, e 4% em 2011.

Os resultados das diferentes técnicas utilizadas, de um modo geral, apresentaram considerável coerência entre si, confirmando a robustez das teorias e métodos adotados neste trabalho. É importante destacar, por outro lado, que foram consideradas apenas algumas propriedades psicométricas, como confiabilidade e discriminação, mas não se tratou a questão da validade, por exemplo. É possível, portanto, que itens com comportamento anômalo (segundo os critérios psicométricos aqui adotados) sejam ainda assim importantes devido ao seu conteúdo, caso este contribua significativamente para a validade do teste e sua coerência com a matriz de referência.

De modo geral, nossos resultados confirmam a qualidade técnica do Enem, levando em conta que em 2009 houve circunstâncias singulares que não mais se repetiram. Isso não significa que se deva deixar de aprimorar os procedimentos para criação, revisão e triagem de itens. Pelo contrário,

nossos resultados mostram que, mesmo em 2011, houve 17% de itens que, segundo o próprio critério psicométrico do Enem (coeficiente bisserial menor que 0,30), não deveriam ter sido incluídos na prova. Como se trata de um exame bastante importante para o país, qualquer imprecisão de medida pode gerar efeitos indesejáveis.

Vale destacar, ainda, a importância do trabalho realizado pelo Inep ao publicar as informações detalhadas sobre o exame (que incluem todas as respostas de cada aluno nas quatro provas e no questionário socioeconômico, além da documentação técnica), tornando possível grande diversidade de pesquisas nas áreas de educação, psicometria, sociologia, políticas públicas, dentre outras. Ao longo dos anos, o Inep tem primado pela transparência e aprimorado os procedimentos relativos ao Enem. Buscando contribuir nesse sentido, cabe-nos destacar algumas lacunas encontradas durante este estudo. Após uma intensa busca no portal do Inep, restaram algumas questões: 1) Todos os itens são sempre incluídos na estimação da proficiência? Se não, quais são os critérios para exclusão? Se sim, resta a questão dos itens mal ajustados ou que apresentaram correlação item-total negativa. 2) Como é feito o processo de calibração dos itens? Supõe-se que seja no pré-teste, de forma a permitir a comparabilidade das notas entre os anos. Neste caso, que tipo de amostra é utilizada? Ela representa os formandos de ensino médio de escolas regulares, ou outro tipo de população? 3) Quais foram os parâmetros de cada item utilizados na estimação das proficiências? Por que eles não são públicos como os microdados? 4) Quais são os procedimentos utilizados para garantir a qualidade dos itens e da prova, especialmente a partir de 2009?

Por fim, buscou-se neste artigo mostrar também algumas utilidades das técnicas psicométricas, tanto para educadores quanto para a meta-avaliação. Ao ressaltar a importância da psicometria para a educação, não se pressupõe aqui que ela produza resultados inequívocos, ou que a medida numérica seja sempre desejável na avaliação educacional. Tampouco se pressupõe que a educação deva, seja em termos epistemológicos ou práticos, ser reduzida ou submetida à psicometria.

Trata-se, aqui, apenas de esclarecer alguns conceitos da psicometria clássica e moderna, exemplificá-los e aplicá-los em um importante exame nacional, mostrando que há uma "caixa de ferramentas" interessante e útil aos educadores, que parece ainda subaproveitada no cotidiano escolar.

## REFERÊNCIAS


ALNABHAN, M.; HARWELL, M. Psychometric challenges in developing a college admission test for Jordan. *Social Behavior and Personality*: an international journal, Palmerston North, NZ, v. 29, n. 5, p. 445-458, jan. 2001.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. *Standards for educational and psychological testing*. Washington: American Educational Research Association, 2014.

ANDRADE, D. F.; KARINO, C. A. *Nota técnica*: Teoria de Resposta ao Item (TRI). Brasília, DF: Inep, 2011. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf>. Acesso em: 9 maio 2013.

ANDRADE, D. F.; KLEIN, R. 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, J. M.; LAROS, J. A.; GOUVEIA, V. V. 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.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da Resposta ao Item*: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

ANDRIOLA, W. B. Psicometria moderna: características e tendências. *Estudos em Avaliação Educacional*, São Paulo, v. 20, n. 43, p. 319-340, maio/ago. 2009.

BAKER, F. B. *The Basics of Item Response Theory*. 2. ed. Washington: ERIC Clearinghouse on Assessment and Evaluation, 2001.

BASS, P. F.; WILSON, J. F.; GRIFFITH, C. H. A shortened instrument for literacy screening. *Journal of General Internal Medicine*, Alexandria, VA, v. 18, n. 12, p. 1036-1038, Dec. 2003.

BRASIL. Portaria n. 807, de 18 de junho de 2010. Instituir o Exame Nacional do Ensino Médio (Enem) como procedimento de avaliação cujo objetivo é aferir se o participante do Exame, ao final do ensino médio, demonstra domínio dos princípios científicos e tecnológicos que presidem a produção moderna e conhecimento das formas contemporâneas de linguagem. *Diário Oficial da União*, Brasília, DF, 21 jun. 2010. Seção 1, p. 71.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Exame Nacional do Ensino Médio (Enem)*: fundamentação teórico-metodológica. Brasília, DF: Inep, 2005.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Enem*: relatório pedagógico 2004. Brasília, DF: MEC/Inep, 2007.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Microdados*. Brasília, DF: Inep/MEC, 2016. Disponível em: <http://portal.inep.gov.br/microdados> Acesso em: 3 mar. 2016.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, Madison, LA, v. 16, n. 3, p. 297-334, Sep. 1951.

FINI, M. E. Erros e acertos na elaboração de itens para a 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. 101-105.

GILLEARD, C.; GROOM, F. A study of two dementia quizzes. *British Journal of Clinical Psychology*, New Jersey, v. 33, n. 4, p. 529-534, Nov. 1994.

GOMES, C. M. A. Avaliando a avaliação escolar: notas escolares e inteligência fluida. *Psicologia em Estudo*, Maringá, PR, v. 15, n. 4, p. 841-849, out./dez. 2010.

HUDSON, T. Relationships among IRT item discrimination and item fit indices in criterion-referenced language testing. *Language Testing*, Londres, v. 8, n. 2, p. 160-181, 1 dez. 1991.

KARINO, C. A.; BARBOSA, M. T. S. *Nota técnica:* procedimento de cálculo das notas do Enem. Brasília, DF: Inep, 2012. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_procedimento_de_calculo_das_notas_enem_2.pdf>. Acesso em: 1 maio 2016.

KARTAL, A. et al. Validity and reliability study of the Turkish version of Health Belief Model Scale in diabetic patients. *International journal of nursing studies*, Amsterdam, v. 44, n. 8, p. 1447-1458, Nov. 2007.

KLEIN, R. Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 21, n. 78, p. 35-56, jan./mar. 2013.

KUDER, G. F.; RICHARDSON, M. W. The theory of the estimation of test reliability. *Psychometrika*, Madison, v. 2, n. 3, p. 151-160, Sep. 1937.

LOPES, F. L.; VENDRAMINI, C. M. M. Propriedades psicométricas das provas de pedagogia do Enade via TRI. *Avaliação*: Revista da Avaliação da Educação Superior, Campinas, SP, v. 20, n. 1, p. 27-47, mar. 2015.

MARTÍNEZ ARIAS, R. Usos, aplicaciones y problemas de los modelos de valor añadido en educación. *Revista de Educación*, Madrid, v. 348, p. 217-250, enero/abr. 2009.

MASTERS, G. N. Item discrimination: when more is worse. *Journal of Educational Measurement*, New Jersey, v. 25, n. 1, p. 15-29, Mar. 1988.

MORRIS, A. Student Standardised Testing: Current Practices in OECD Countries and a Literature Review. Paris: OECD, Out. 2011. 51 p. (*OECD Education Working Papers*, n. 65).

MUTHÉN, B. O.; KAO, C. F.; BURSTEIN, L. Instructionally Sensitive Psychometrics: Application of a New IRT – Based Detection Technique to Mathematics Achievement Test Items. *Journal of Educational Measurement*, New Jersey, v. 28, n. 1, p. 1-22, Mar. 1991.

PASQUALI, L. Psicometria. *Revista da Escola de Enfermagem da USP*, São Paulo, v. 43, n. Especial, p. 992-999, dez. 2009.

PRIMI, R.; HUTZ, C. S.; SILVA, M. C. R. da. A prova do Enade de Psicologia 2006: concepção, construção e análise psicométrica da prova. *Avaliação Psicológica*, Itatiba, SP, v. 10, n. 3, p. 271-294, dez. 2011.

REVELLE, W. Psych: Procedures for Personality and Psychological Research. 1.6.12. Evanston, IL: Northwestern University, 2016. Download. Disponível em: <http://cran.r-project.org/package=psych>. Acesso em: 2 abr. 2017.

RIZOPOULOS, D. ltm: an R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, Innsbruck, AT, v. 17, n. 5, p. 1-25, Nov. 2006.

ROACH, A. J.; FRAZIER, L. P.; BOWDEN, S. R. The marital satisfaction scale: development of a measure for intervention research. *Journal of Marriage and the Family*, New Jersey, v. 43, n. 3, p. 537-546, Aug. 1981.

SCHMITT, N. Uses and abuses of coefficient alpha. *Psychological Assessment*, Washington, v. 8, n. 4, p. 350-353, Dec. 1996.

SEVERO, M.; TAVARES, M. A. F. Metaevaluation in clinical anatomy: a practical application of Item Response Theory in multiple choice examinations. *Anatomical Sciences Education*, New Jersey, v. 3, n. 1, p. 17-24, Jan./Feb. 2010.

SOARES, T. M. Utilização da Teoria da Resposta ao Item na produção de indicadores sócio-econômicos. *Pesquisa operacional*, Rio de Janeiro, v. 25, n. 1, p. 83-112, jan./abr. 2005.

STUFFLEBEAM, D. L. The Metaevaluation Imperative. *American Journal of Evaluation*, Washington, DC, v. 22, n. 2, p. 183-209, jun. 2001.

TRAVITZKI, R. *Enem*: limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar. 2013. Tese (Doutorado em Educação) – Universidade de São Paulo, São Paulo, 2013a.

TRAVITZKI, R. Qualidade técnica dos itens do Enem 2009. In: REUNIÃO DA ASSOCIAÇÃO BRASILEIRA DE AVALIAÇÃO DA EDUCAÇÃO, 7., 2013, Brasília, DF. Avaliação e currículo: um diálogo necessário. *Anais...* Brasília, DF, 2013b.

ZAICHKOWSKY, J. L. The Personal Involvement Inventory: Reduction, Revision, and Application to Advertising. *Journal of Advertising*, Abingdon, v. 23, n. 4, p. 59-70, Dec. 1994.

RODRIGO TRAVITZKI
Doutor pela Faculdade de Educação da Universidade de São Paulo (FE/USP), São Paulo, São Paulo, Brasil
*travitzki@usp.br*

## ANEXOS

**TABELA A1 – Resultados item a item obtidos no estudo após a análise da prova de Ciências da Natureza do Enem 2009**

| ITEM | ACERTO | SEM RESPOSTA | CORRELAÇÃO ITEM-TOTAL | PARÂMETRO A | PARÂMETRO B | P-VALOR |
|---|---|---|---|---|---|---|
| 1 | 0,87 | 0,00 | 0,30 | 1,31 | -1,87 | 0,000 |
| 2 | 0,39 | 0,00 | 0,52 | 1,55 | 0,38 | 0,000 |
| 3 | 0,42 | 0,00 | 0,22 | 0,33 | 0,98 | 0,000 |
| 4 | 0,66 | 0,00 | 0,35 | 0,86 | -0,87 | 0,000 |
| 5 | 0,58 | 0,00 | 0,38 | 0,90 | -0,42 | 0,000 |
| 6 | 0,58 | 0,00 | 0,36 | 0,83 | -0,44 | 0,000 |
| 7 | 0,43 | 0,00 | 0,25 | 0,40 | 0,70 | 0,000 |
| 8 | 0,54 | 0,00 | 0,39 | 0,90 | -0,21 | 0,000 |
| 9 | 0,19 | 0,00 | 0,27 | 0,56 | 2,69 | 0,000 |
| 10 | 0,28 | 0,00 | 0,39 | 0,88 | 1,22 | 0,000 |
| 11 | 0,49 | 0,00 | 0,46 | 1,19 | 0,03 | 0,000 |
| 12 | 0,22 | 0,00 | 0,29 | 0,60 | 2,23 | 0,000 |
| 13 | 0,70 | 0,00 | 0,42 | 1,33 | -0,85 | 0,000 |
| 14 | 0,64 | 0,00 | 0,33 | 0,76 | -0,87 | 0,000 |
| 15 | 0,20 | 0,00 | 0,20 | 0,36 | 3,87 | 0,000 |
| 16 | 0,45 | 0,00 | 0,42 | 0,97 | 0,22 | 0,000 |
| 17 | 0,06 | 0,00 | 0,17 | 0,53 | 5,49 | 0,000 |
| 18 | 0,38 | 0,00 | 0,19 | 0,26 | 1,90 | 0,000 |
| 19 | 0,33 | 0,00 | 0,43 | 1,02 | 0,80 | 0,000 |
| 20 | 0,65 | 0,00 | 0,34 | 0,82 | -0,88 | 0,000 |
| 21 | 0,33 | 0,00 | 0,37 | 0,79 | 1,02 | 0,000 |
| 22 | 0,54 | 0,00 | 0,48 | 1,38 | -0,19 | 0,000 |
| 23 | 0,29 | 0,00 | 0,14 | 0,12 | 7,42 | 0,000 |
| 24 | 0,20 | 0,00 | 0,30 | 0,67 | 2,28 | 0,000 |
| 25 | 0,24 | 0,00 | 0,28 | 0,51 | 2,31 | 0,000 |
| 26 | 0,22 | 0,00 | 0,30 | 0,61 | 2,18 | 0,000 |
| 27 | 0,14 | 0,00 | 0,04 | -0,08 | -23,27 | 0,000 |
| 28 | 0,31 | 0,00 | 0,24 | 0,41 | 2,07 | 0,000 |
| 29 | 0,32 | 0,00 | 0,25 | 0,43 | 1,81 | 0,000 |
| 30 | 0,15 | 0,00 | 0,12 | 0,18 | 9,57 | 0,000 |
| 31 | 0,15 | 0,00 | -0,01 | -0,25 | -7,07 | 0,000 |
| 32 | 0,36 | 0,00 | 0,25 | 0,42 | 1,42 | 0,000 |
| 33 | 0,23 | 0,00 | 0,36 | 0,83 | 1,64 | 0,000 |
| 34 | 0,25 | 0,00 | 0,20 | 0,34 | 3,40 | 0,000 |
| 35 | 0,13 | 0,00 | 0,10 | 0,12 | 15,44 | 0,000 |
| 36 | 0,20 | 0,00 | 0,20 | 0,34 | 4,12 | 0,000 |
| 37 | 0,30 | 0,00 | 0,37 | 0,78 | 1,23 | 0,000 |
| 38 | 0,20 | 0,00 | 0,17 | 0,30 | 4,57 | 0,000 |
| 39 | 0,25 | 0,00 | 0,16 | 0,21 | 5,21 | 0,000 |
| 40 | 0,41 | 0,00 | 0,35 | 0,67 | 0,56 | 0,000 |
| 41 | 0,24 | 0,00 | 0,16 | 0,24 | 4,72 | 0,000 |
| 42 | 0,28 | 0,00 | 0,28 | 0,52 | 1,99 | 0,000 |
| 43 | 0,25 | 0,00 | 0,23 | 0,38 | 2,96 | 0,000 |
| 44 | 0,21 | 0,01 | 0,13 | 0,17 | 8,11 | 0,000 |
| 45 | 0,17 | 0,00 | 0,20 | 0,40 | 4,10 | 0,000 |

Fonte: Elaboração do autor a partir dos dados do Enem (BRASIL, 2016).

**TABELA A2 – Resultados item a item obtidos no estudo após a análise da prova de Matemática do Enem 2009**

| ITEM | ACERTO | SEM RESPOSTA | CORRELAÇÃO ITEM-TOTAL | PARÂMETRO A | PARÂMETRO B | P-VALOR |
|---|---|---|---|---|---|---|
| 1 | 0,13 | 0,00 | 0,27 | 0,62 | 3,23 | 0,000 |
| 2 | 0,26 | 0,00 | 0,41 | 1,12 | 1,13 | 0,000 |
| 3 | 0,11 | 0,00 | 0,26 | 0,66 | 3,39 | 0,000 |
| 4 | 0,27 | 0,00 | 0,22 | 0,33 | 3,10 | 0,000 |
| 5 | 0,24 | 0,00 | 0,28 | 0,54 | 2,28 | 0,000 |
| 6 | 0,41 | 0,00 | 0,39 | 1,03 | 0,43 | 0,000 |
| 7 | 0,36 | 0,00 | 0,33 | 0,74 | 0,88 | 0,000 |
| 8 | 0,17 | 0,00 | 0,19 | 0,33 | 4,89 | 0,000 |
| 9 | 0,25 | 0,00 | 0,17 | 0,17 | 6,39 | 0,000 |
| 10 | 0,11 | 0,00 | 0,16 | 0,25 | 8,45 | 0,000 |
| 11 | 0,37 | 0,00 | 0,36 | 0,87 | 0,69 | 0,000 |
| 12 | 0,52 | 0,00 | 0,36 | 0,92 | -0,09 | 0,000 |
| 13 | 0,33 | 0,00 | 0,17 | 0,22 | 3,25 | 0,000 |
| 14 | 0,70 | 0,00 | 0,29 | 0,81 | -1,19 | 0,000 |
| 15 | 0,24 | 0,00 | 0,31 | 0,63 | 2,01 | 0,000 |
| 16 | 0,27 | 0,00 | 0,32 | 0,70 | 1,58 | 0,000 |
| 17 | 0,11 | 0,00 | 0,22 | 0,50 | 4,35 | 0,000 |
| 18 | 0,27 | 0,00 | 0,24 | 0,40 | 2,61 | 0,000 |
| 19 | 0,27 | 0,00 | 0,39 | 0,96 | 1,24 | 0,000 |
| 20 | 0,24 | 0,00 | 0,24 | 0,42 | 2,79 | 0,000 |
| 21 | 0,21 | 0,00 | 0,21 | 0,38 | 3,61 | 0,000 |
| 22 | 0,16 | 0,00 | 0,13 | 0,11 | 15,15 | 0,000 |
| 23 | 0,25 | 0,00 | 0,19 | 0,28 | 4,04 | 0,000 |
| 24 | 0,19 | 0,00 | 0,41 | 1,11 | 1,59 | 0,000 |
| 25 | 0,28 | 0,00 | 0,28 | 0,56 | 1,82 | 0,000 |
| 26 | 0,23 | 0,00 | 0,18 | 0,22 | 5,76 | 0,000 |
| 27 | 0,24 | 0,00 | 0,22 | 0,37 | 3,22 | 0,000 |
| 28 | 0,26 | 0,01 | 0,31 | 0,61 | 1,86 | 0,000 |
| 29 | 0,26 | 0,01 | 0,32 | 0,69 | 1,66 | 0,000 |
| 30 | 0,21 | 0,00 | 0,18 | 0,27 | 5,10 | 0,000 |
| 31 | 0,27 | 0,00 | 0,20 | 0,28 | 3,67 | 0,000 |
| 32 | 0,11 | 0,00 | 0,23 | 0,51 | 4,45 | 0,000 |
| 33 | 0,23 | 0,00 | 0,17 | 0,19 | 6,28 | 0,000 |
| 34 | 0,25 | 0,01 | 0,29 | 0,58 | 2,06 | 0,000 |
| 35 | 0,17 | 0,01 | 0,17 | 0,26 | 6,11 | 0,000 |
| 36 | 0,33 | 0,01 | 0,10 | -0,05 | -13,22 | 0,100 |
| 37 | 0,23 | 0,01 | 0,13 | 0,09 | 12,88 | 0,240 |
| 38 | 0,23 | 0,01 | 0,12 | 0,06 | 21,07 | 0,110 |
| 39 | 0,28 | 0,01 | 0,12 | 0,01 | 138,46 | 0,200 |
| 40 | 0,22 | 0,01 | 0,11 | 0,00 | 795,67 | 0,020 |
| 41 | 0,23 | 0,01 | 0,17 | 0,18 | 6,64 | 0,000 |
| 42 | 0,23 | 0,01 | 0,18 | 0,22 | 5,49 | 0,000 |
| 43 | 0,14 | 0,01 | 0,11 | 0,11 | 16,45 | 0,120 |
| 44 | 0,25 | 0,01 | 0,13 | 0,09 | 13,05 | 0,000 |
| 45 | 0,24 | 0,00 | 0,06 | -0,10 | -11,60 | 0,000 |

Fonte: Elaboração do autor a partir dos dados do Enem (BRASIL, 2016).

**Recebido em:** MAIO 2016
**Aprovado para publicação em:** DEZEMBRO 2016