André Luiz Santos de Oliveira

AVALIAÇÃO PSICOMÉTRICA DA MEDIDA DO COMPONENTE DE FORMAÇÃO GERAL DA PROVA DO EXAME NACIONAL DE DESEMPENHO DE ESTUDANTES (ENADE) DE 2010, 2011 E 2012.

Dissertação submetida ao Programa de Pós-Graduação em Métodos e Gestão em Avaliação da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Métodos e Gestão em Avaliação.
Orientador: Prof. Dr. Dalton Francisco de Andrade

Florianópolis
2017

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Oliveira, André Luiz Santos de
Avaliação psicométrica da medida do componente de formação geral da prova do Exame Nacional de Desempenho de Estudantes (ENADE) de 2010, 2011 e 2012. / André Luiz Santos de Oliveira ; orientador, Dalton Francisco de Andrade, 2017.
107 p.

Dissertação (mestrado profissional) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Métodos e Gestão em Avaliação, Florianópolis, 2017.

Inclui referências.

1. Métodos e Gestão em Avaliação. 2. Métodos e Gestão em Avaliação. 3. Psicometria. . 4. Teoria da Resposta ao Item. . 5. Exame Nacional de Desempenho dos Estudantes. I. Andrade, Dalton Francisco de . II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Métodos e Gestão em Avaliação. III. Título.

André Luiz Santos de Oliveira

AVALIAÇÃO PSICOMÉTRICA DA MEDIDA DO COMPONENTE DE FORMAÇÃO GERAL DA PROVA DO EXAME NACIONAL DE DESEMPENHO DE ESTUDANTES (ENADE) DE 2010, 2011 E 2012.

Esta Dissertação foi julgada adequada para obtenção do Título de mestre e aprovada em sua forma final pelo Programa de Pós-Graduação em Métodos e Gestão em Avaliação.

Florianópolis, 07 de março de 2017.

_____
Prof. Renato Cislaghi, Dr.
Coordenador do Curso

Banca Examinadora:

_____
Prof. Dalton Francisco de Andrade, Dr.
Orientador
Universidade Federal de Santa Catarina

_____
Prof. Mauro Luiz Rabelo, Dr.
Universidade de Brasília

_____
Prof. Marcelo Menezes Reis, Dr.
Universidade Federal de Santa Catarina

_____
Prof. Pedro Alberto Barbetta, Dr.
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Agradeço a todas as pessoas que colaboraram direta ou indiretamente na construção deste trabalho, em especial àquelas do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) e da Universidade Federal de Santa Catarina (UFSC).

"*Strictly speaking, any test reported in a single score should consist of items drawn from one-dimensional universe*." (BOCK, GIBBONNS & MURAKI, 1988, p. 261) [1]

[1] Tradução livre: Estritamente falando, qualquer prova resumida em uma única nota deveria consistir de itens tirados de um universo unidimensional.

RESUMO

O Exame Nacional de Desempenho de Estudantes (ENADE) é parte do Sistema Nacional de Avaliação da Educação Superior (SINAES) e consiste em uma prova composta pelos componentes da Formação Geral, comum a todas as áreas de conhecimento, e do Componente Específico de cada curso. Dada a existência de poucos estudos específicos sobre o componente da Formação Geral, que é aplicado anualmente e sua utilização na regulação da Educação Superior brasileira é imprescindível avaliar a qualidade desse componente com base na teoria da medida. Para tanto, avaliou-se os três anos correspondentes ao terceiro ciclo do ENADE – 2010, 2011 e 2012. Entre as questões objetivas do exame, identificou-se apenas uma dimensão predominante na prova de cada ano por meio da análise fatorial de correlações tetracóricas com o método dos resíduos mínimos. Verificou-se baixa consistência interna desses itens, com Alfa Cronbach variando entre 0,28 e 0,40 e médias de Correlação Ponto-Bisserial dos itens variando entre 0,14 e 0,23. Desta forma, sob a ótica da Teoria Clássica dos Testes, conclui-se que a nota referente às questões objetivas desse componente não deveria ser utilizada para a tomada de decisões. Adicionalmente, por meio da Teoria de Resposta ao Item, se realizou a análise de uma hipotética escala unidimensional extraída dessas questões e, foi observado que alguns itens não se ajustavam ao um modelo unidimensional de cada ano. Comparou-se o ajuste dos modelos logísticos de três e quatro parâmetros e verificou-se que o último teve melhor ajuste aos itens, com ganho de variância explicada em todos os anos. Por fim, propõe-se a utilização de um modelo de itens de formato misto, permitindo calibrar simultaneamente os itens objetivos e discursivos em uma única escala, aumentando a precisão da medida e integrando as notas desses dois tipos de questões, atualmente calculadas separadamente.

Palavras-chave: 1. Psicometria. 2. Teoria da Resposta ao Item. 3. Educação Superior. 4. Exame Nacional de Desempenho dos Estudantes

ABSTRACT

The National Student Performance Examination (ENADE) is part of the National System for the Evaluation of Higher Education (SINAES) and consists of a test composed of the components of General Formation, common to all areas of knowledge, and of the Specific Component of each course. Given the existence of few specific studies on the General Formation component, which is applied annually and its use in the regulation of Brazilian Higher Education, it is essential to evaluate the quality of this component based on the measurement theory. In order to do so, the three years corresponding to the third cycle of ENADE - 2010, 2011 and 2012 were analyzed. Among the objective questions of the examination, only one predominant dimension was identified in the test of each year by means of the factorial analysis of tetracorical correlations with the method of the minimum residues. There was a low internal consistency of these items, with Cronbach's Alpha varying between 0.28 and 0.40 and means of Point-Bisserial Correlation of items varying between 0.14 and 0.23. Thus, from the standpoint of the Classical Theory of Tests, it is concluded that the note regarding the objective questions of this component should not be used for decision making. Additionally, through the Item Response Theory, an analysis of a hypothetical one-dimensional scale extracted from these questions was performed, and it was observed that some items did not fit the one-dimensional model of each year. The adjustment of the logistic models of three and four parameters was compared and it was verified that the last had better adjustment to the items, with gain of variance explained in all the years. Finally, it is proposed to use a mixed-format item model, allowing simultaneous calibration of objective and discursive items on a single scale, increasing the precision of the measurement and integrating the notes of these two types of questions, currently calculated separately.

Keywords: 1. Psychometrics. 2. Item Response Theory. 3. Higher education. 4. National Exam for the Assessment of Student Performance.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AF – Análise Fatorial
CE – Componente específico do ENADE
CPC – Conceito Preliminar de Curso
ENADE - Exame Nacional de Desempenho dos Estudantes
ENC – Exame Nacional de Cursos
FG – Componente da Formação Geral do ENADE
INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC – Ministério da Educação
ML3 – Modelo logístico de três parâmetros
ML4 - Modelo logístico de quatro parâmetros
MRG – Modelo de Resposta Gradual
TCT – Teoria Clássica dos Testes
TRI – Teoria de Resposta ao Item

Sumário

1 INTRODUÇÃO

1.1 Contextualização

1.1.1 A Avaliação da Educação Superior no Brasil

A Lei de Diretrizes e Bases (lei 9.394 de 20 de dezembro de 1996) disciplina a educação escolar, estabelecendo, entre outras coisas, seus princípios e fins e sua organização. Entre as finalidades da Educação Superior, três se destacam por não refletirem diretamente uma competência técnica específica de cada área de conhecimento:

> "I – estimular a criação cultural e o desenvolvimento do espírito científico e do pensamento reflexivo; [...]
> V – suscitar o desejo permanente de aperfeiçoamento cultural e profissional e possibilitar a correspondente concretização, integrando os conhecimentos que vão sendo adquiridos numa estrutura intelectual sistematizadora do conhecimento de cada geração;
> VI – estimular o conhecimento dos problemas do mundo presente, em particular os nacionais e regionais, prestar serviços especializados à comunidade e estabelecer com esta uma relação de reciprocidade; [...]".

Assim, para verificar se a Educação Superior reflete essas finalidades, que vão além de uma prova de proficiência em uma área de conhecimento isolada, sua avaliação deverá realizada de uma forma específica. TENÓRIO & ANDRADE (2009) explicam a avaliação é um princípio básico de regulação pelo Estado e contribui para a melhoria dos programas e influencia a definição de políticas, práticas e decisões. Diante disse, vistas ao seu papel de regulador, o estado brasileiro deve empregar instrumentos e técnicas efetivos diante da grande diversidade de Instituições de Ensino Superior no Brasil, para o desenvolvimento de políticas mais acertadas.

A Avaliação da Educação Superior brasileira atualmente é regida pela lei 10.861 de 2004 e é feita por meio do Sistema Nacional de Avaliação da Educação Superior – SINAES. As finalidades desse sistema são: a melhoria da qualidade da educação superior, a orientação da expansão de sua oferta, o aumento permanente de sua eficácia

institucional e efetividade acadêmica e social e, especialmente, a promoção do aprofundamento dos compromissos e responsabilidades sociais das instituições de educação superior, por meio da valorização de sua missão pública, da promoção dos valores democráticos, do respeito à diferença e à diversidade, da afirmação da autonomia e da identidade institucional (BRASIL, 2004). O responsável pela operacionalização desse sistema é o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), incumbido da avaliação dos três tipos de organização acadêmica existentes no Brasil – Universidades, Centros Universitários e Faculdades.

Assim, dentro da lógica de um sistema de avaliação, o INEP utiliza diversos instrumentos e procedimentos para a avaliação dos cursos de graduação, entre os quais estão: a autoavaliação, realizada pelas próprias instituições; a avaliação *in loco*, realizada pelas comissões de especialistas; o Censo da Educação Superior, realizado anualmente; o questionário do estudante, respondido pelos concluintes dos cursos de graduação avaliados a cada ano; e a Prova do Exame Nacional de Desempenho dos Estudantes (ENADE). Assim com o objetivo de regulação, há integração entre esses instrumentos e procedimentos, a exemplo do indicador Conceito Preliminar de Curso (CPC) composto de dados do Censo da Educação Superior, do desempenho dos estudantes na prova do ENADE e de respostas do Questionário do Estudante. Caso um curso seja considerado insatisfatório nesse indicador, torna-se necessária a avaliação in loco para a renovação de seu reconhecimento.

1.1.2 O ENADE

O ENADE sucedeu o Exame Nacional de Cursos (ENC) que teve início com a Lei 9.131 (BRASIL, 1995). O ENC era aplicado anualmente e todos os concluintes de cada área avaliada respondiam a prova. Inicialmente apenas três áreas eram submetidas à prova, porém, em sua última edição, em 2003, esse número chegou a vinte e uma. A participação nesse exame era obrigatória para a obtenção do diploma por parte dos alunos. O ENC foi alvo de diversas críticas - sua concentração no desempenho na prova, o distanciamento das instituições de ensino do processo avaliativo, a ênfase apenas nos resultados de aprendizagem, os problemas de comparabilidade, a impossibilidade de comparação das notas de diversos anos, seus custos crescentes e a confusão entre regulação e avaliação (VERHINE; DANTAS & SOARES, 2006).

Essas críticas levaram ao desenvolvimento do ENADE, que conforme o parágrafo 1º do artigo 5º da Lei 10.861/2004 deverá aferir.

> "§1º O ENADE aferirá o desempenho dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação, suas habilidades para ajustamento às exigências decorrentes da evolução do conhecimento e suas competências para compreender temas exteriores ao âmbito específico de sua profissão, ligados à realidade brasileira e mundial e a outras áreas do conhecimento."
> (BRASIL, 2004)

A esses objetivos se relacionam, respectivamente, o Componente Específico da área (CE) e o componente da Formação Geral (FG). O componente da FG é avaliado por meio de 10 itens, sendo dois discursivos e oito de múltipla escolha e o CE, por trinta itens, sendo três discursivos e vinte e sete de múltipla escolha, totalizando quarenta questões para serem respondidas no dia do exame que tem duração máxima de quatro horas.

GAUDIO (2014) explica que a Formação geral tem como "... objetivo de investigar a aquisição de competências, habilidades e conhecimentos considerados essenciais na formação de qualquer estudante de qualquer área da educação superior." (p. 54) Essa "é constituída por questões comuns a todas as áreas do conhecimento". "São questões de conhecimento geral ou sobre o mundo em que vivemos e questões de ética e cidadania, consideradas por especialistas necessárias ou importantes para a educação de todos os universitários, independentemente de suas áreas de especialização." (p. 55). Evidentemente esse componente se relaciona a aferição das finalidades I, V e VI da LDB (BRASIL, 1996).

Já o CE tem seu objetivo explicitamente definido pela Portaria N. 40, de 12 de dezembro de 2007, do Ministério da Educação, "aferir as competências, habilidades e conteúdos agregados durante a formação" (BRASIL, 2007), sendo essas competências, habilidades e conteúdos especificados nas diretrizes curriculares de cada curso e aprovadas pelo Conselho Nacional de Educação, diferentemente do que ocorre com o componente da FG.

Além disso, convém destacar que assim como o ENC, o ENADE é componente curricular obrigatório, sendo prevista a participação dos

estudantes ingressantes e concluintes. Contudo, desde 2011, o INEP dispensa do ENADE os ingressantes dos cursos avaliados em cada edição do exame em função da alteração da metodologia de cálculo do Conceito ENADE e do Conceito Preliminar de Curso, que passaram a utilizar o desempenho dos ingressantes no ENEM (BRASIL, 2012) em vez de seu desempenho nessa prova.

O SINAES se organiza em ciclos trienais uma vez que a lei estabelece essa periodicidade para a avaliação de cada área do conhecimento e, outras ações o âmbito do SINAES, como a avaliação *in loco*, são realizadas durantes os outros dois anos do ciclo. A vantagem desse espaçamento entre as provas é permitir a avaliação de mais áreas em cada ciclo e viabilizar que as instituições com avaliações negativas no ENADE realizem alterações em seus cursos com o objetivo de seu desenvolvimento.

Apesar da primeira edição do ENADE, em 2004 ter avaliado apenas 13 áreas, poucas quando comparadas as 21 áreas avaliadas na última edição do ENC, deve-se considerar a sua estrutura em ciclo trienal. Assim, dado que em 2005 e 2006 foram avaliadas, respectivamente, 20 e 15 áreas, tem-se que o primeiro ciclo do ENADE avaliou 48 áreas, mais que o dobro de áreas em relação à última edição do ENC. Esse quantitativo vem aumentando e, no terceiro ciclo do ENADE, foram avaliadas 69 áreas, respectivamente 19, 33 e 17 áreas nos anos de 2010, 2011 e 2012.

Ressalta-se que o ENADE é parte do SINAES, diferindo do ENC que só considerava o desempenho do concluinte – ao integrar no calculo de seu indicador outras medidas, como as competências transversais medidas na FG, os dados do Censo da Educação Superior e outros indicadores de características do curso coletados por meio de questionários, ele permite representar melhor a qualidade geral do curso (VERHINE; DANTAS & SOARES, 2006). Todavia, para permitir a análise das inúmeras instituições e cursos, tanto o desempenho no ENADE quanto no ENC são resumidos em um único valor. No caso do ENADE, esse valor utilizado em um indicador conforme a NOTA TÉCNICA Nº 029 DE 15 DE OUTUBRO DE 2012.

Por fim, as características do ENADE e suas diferenças com relação ao ENC podem ser resumidas conforme o Quadro 1.

Quadro 1 Semelhanças e diferenças entre ENC e ENADE

Aspecto	ENC	ENADE
Integração com outras fontes de informação	Não há.	Integra desempenho individual com: • Diferença entre ingressantes e concluintes; • Censo da Educação Superior e • Questionários de percepção sobre o curso.
Avalia competências transversais	Não.	Sim, no componente da Formação Geral.
Periodicidade da avaliação de cada área	Anual.	Trienal.

Percebe-se que o ENADE pode ser encarado como uma evolução do ENC, com objetivos mais amplos e integrando mais fontes de informação quando comparado com seu antecessor. Contudo, ressalta-se que o ENADE é parte do SINAES, que é composto por diversos outros instrumentos e técnicas que visam o desenvolvimento e aperfeiçoamento da Educação Superior brasileira.

1.2 Objetivos

Nessa seção serão apresentados o objetivo geral e os objetivos específicos do trabalho.

1.2.1 Objetivo Geral

O objetivo geral desse trabalho é avaliar a qualidade da medida do componente da Formação Geral de cada prova do terceiro ciclo do ENADE - 2010 a 2012 - primeiro ciclo no qual o INEP realizou a elaboração e revisão dos itens de prova.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Investigar a dimensionalidade da Formação Geral para avaliar se o desempenho nessa prova deveria ser resumido em um único valor;
- Investigar parâmetros psicométricos de fidedignidade (precisão) de uma escala unidimensional da Formação Geral, conforme sua utilização atual, para esclarecer o quanto uma nota na prova é confiável do ponto de vista da teoria da medida;
- Avaliar a adequação do modelo logístico de três parâmetros da Teoria de Resposta ao Item, já utilizado pelo Inep no ENEM, e do modelo logístico de quatro parâmetros, ainda pouco investigado no Brasil para a análise das questões objetivas de Formação Geral;
- Propor um modelo de itens de formato misto da Teoria de Resposta ao Item para integrar as questões objetivas e as discursivas da Formação Geral em uma só escala para produzir uma medida com maior informação.

1.2 Justificativa

A prova de Formação Geral é aplicada anualmente a todos os concluintes dos cursos superiores avaliados em cada ano pelo ENADE, totalizando 1105944 estudantes ao longo do terceiro ciclo do ENADE (anos de 2010, 2011 e 2012). Em função de sua participação no cálculo do Conceito Preliminar de Curso, utilizado na regulação da Educação Superior brasileira, é imprescindível avaliar se esse instrumento tem fornecido uma medida coerente e confiável do ponto de vista psicométrico. A escolha desses anos se deve a indisponibilidade dos dados referentes ao Enade 2015 durante a produção desse estudo e à tentativa de minimizar vieses pessoais, uma vez que o autor trabalha com a elaboração e revisão das provas do ENADE desde 2013.

1.3 Estrutura do Trabalho

O trabalho tem cinco capítulos.

O capítulo 1 contém a contextualização do trabalho e seus objetivos.

O capítulo 2 contém a revisão de literatura e os referenciais teóricos utilizados.

O capítulo 3 contém a descrição do método empregado para a análise de dados.

O capítulo 4 contém os resultados obtidos.

O capítulo 5 contém as conclusões do trabalho e algumas considerações sobre o tema estudado.

2 Referencial teórico

Nesse capítulo serão expostas as bases teóricas que levaram as análises realizadas nesse trabalho e uma breve revisão de estudos que contém análises quantitativas de resultados do componente da Formação Geral do ENADE.

2.1 Psicometria

PASQUALI (2009) explica que a Psicometria se fundamenta na teoria da medida em ciências em geral envolvendo tanto a teoria quanto as técnicas de medida dos processos mentais. Esse conceito parte do clássico conceito de mensuração recomendado por STEVENS (1946, p. 680) "a atribuição de numerais para determinadas coisas de tal forma que estes representem fatos e convenções sobre eles." [2] Assim, sob a perspectiva da Teoria Clássica dos Testes, um teste propõe-se a medir traços latentes, entidades hipotéticas inferidas sobre características das pessoas, através de uma amostra de seus comportamentos – as respostas no teste (URBINA, 2007). De forma similar, a teoria de resposta ao item busca inferir a quantidade de traço latente do respondente por meio da análise da resposta do sujeito para cada item (ANDRADE *et al*, 2000). Assim, é possível dividir as diversas técnicas e conceitos da Psicometria em Teoria Clássica dos Testes (TCT) e, a mais moderna, Teoria de Resposta ao Item (TRI), ressaltando-se que os últimos não são incompatíveis, nem os tornaram obsoletos (ANDRIOLA, 2009) (PASQUALI & PRIMI, 2003).

Ressalta-se que é fundamental primeiramente se analisar e "justificar a legitimidade de se converter procedimentos e operações empíricas, como a resposta a testes e exames em uma representação numérica, ou o score" (p. 22 PASQUALI, *et al*, 1996). Esse problema é trabalhado na psicologia por meio da avaliação da validade de um teste, entendida como um julgamento sobre se este de fato mede aquilo que é proposto. URBINA (2007) destaca três tipos de fontes de validade - Validade de Critério, Validade de Construto e Validade de Conteúdo. Essa última consiste na verificação da adequação da representação do construto em um teste e pode envolver técnicas como a análise fatorial e

[2] Tradução livre de "… *the assignment of numerals to things so as to represent facts and conventions about them*.".

a avaliação de correlação de um teste com outros que medem o(s) mesmo(s) construto(s) (CHIODI & WECHSLER, 2008).

2.1.1 Análise Fatorial

LAROS (2005) explica que a análise fatorial (AF) é um dos procedimentos psicométricos mais utilizados na construção, revisão e avaliação de instrumentos psicológicos. O ponto central da AF é a parcimônia, ou seja, explicar uma grande quantidade de variáveis[3] com um número menor de variáveis hipotéticas – os fatores (p. 164). "A análise fatorial tem como lógica precisamente verificar quantos construtos comuns são necessários para explicar as covariâncias (as intercorrelações) dos itens. As correlações entre os itens são explicadas, pela análise fatorial, como resultantes de variáveis-fonte que seriam as causas destas covariâncias". (PASQUALI, *et al*, 1996, p. 96).

Para analisar dados binários, como os vetores de acerto ou erro em cada resposta de uma prova é necessário utilizar técnicas específicas como a análise fatorial feita a partir de matrizes tetracóricas ou a Análise Fatorial de Informação Plena (ANDRADE *et al*, 2010, p. 16), pois o uso de técnicas padrões produzem resultados enviesados com esse tipo de variável (LAROS, 2005). A Análise Fatorial de Informação Plena é vantajosa em relação as matrizes tetracóricas por evitar coeficientes tetracóricos indeterminados para itens com dificuldades muito altas ou baixas, acomodar melhor o efeito de itens omitidos ou não alcançados e possibilitar o teste da significância estatística de fatores adicionais (p. 278, BOCK, GIBBONNS & MURAKI, 1988). Realizada a análise apropriada é preciso decidir sobre a quantidade de fatores a serem extraídos. LAROS (2005) recomenda que se sigam critérios como considerar apenas fatores cujo autovalor é superior a (1,0) e a análise do teste *scree* de Cattell, que consiste na análise gráfica de um gráfico com os autovalores encontrados, entre outros.

Aspectos técnicos adicionais sobre a análise fatorial podem ser verificados em MANLY (2008), LAROS (2005) e DAMÁSIO (2012).

[3] A informação sobre acerto ou erro em cada informação é uma variável, tipicamente representada como 0 ou 1. Uma sequência de valores sobre várias questões pode ser entendida como um vetor.

2.1.1.1 Unidimensionalidade

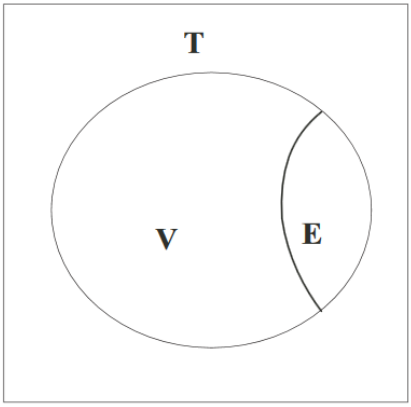
Conceitualmente "a unidimensionalidade é uma proposição teórica parcimoniosa e elegante, segundo a qual toda a complexidade intrínseca ao ato de resolução de um problema – de natureza cognitiva ou não – deve ter como causa uma única estrutura latente, denominada θ" (ANDRIOLA, 2009, p. 327). Essa exigência nunca pode ser plenamente atendida, uma vez que é claro que "qualquer desempenho humano é sempre multideterminado ou multimotivado." (ANDRADE; TAVARES; VALLE, 2000, p. 16). Assim, "a unidimensionalidade torna-se então uma questão de grau" (VITORIA; ALMEIDA; PRIMI, 2006, p. 6).

Como ela é pressuposto de vários modelos da Teoria de Resposta ao Item, é necessário adotar algum critério para considerá-la satisfeita. Para RECKEASE (1979) a unidimensionalidade pode ser considerada satisfeita caso o primeiro fator extraído na análise fatorial explique pelo menos 20% da variância do teste e os demais não o façam. Caso haja outros fatores de interesse, esses devem preferencialmente ser produzidos e calibrados separadamente.

2.1.2 Teoria Clássica dos Testes

PASQUALI (2009, p. 993) explica que a Teoria Clássica dos Testes "se preocupa em explicar o resultado final total, isto é, a soma das respostas dadas a uma série de itens, expressa no chamado escore total". Assim essa teoria se preocupa com o desempenho no teste como um todo. O modelo geral da TCT considera que o escore que o respondente obtém em um teste é função tanto de seu escore verdadeiro, que se deseja medir, quanto dos diversos erros de medida associados (PASQUALI, 2009). Esses últimos são referentes a diversas fontes como o contexto de aplicação, o próprio teste e o respondente (URBINA, 2007). Assim o escore bruto (T) obtido por uma pessoa num teste é uma união de seu escore verdadeiro (V) e o erro de medida (E), conforme ilustra a Figura 1.

Figura 1 Os componentes do escore verdadeiro



Fonte (PASQUALI, 2009, P. 994)

O conceito de Fidedignidade (ou Precisão) diz respeito à quão bem a nota em o teste é uma boa representação do que se pretende medir. Assim, a Fidedignidade é o grau de confiança em que o escore obtido pelo respondente representa seu escore real e por isso está fortemente relacionado ao conceito de erro de medida (URBINA, 2007). Entre as principais formas de se avaliar a precisão de um teste estão à análise da correlação entre suas metades e da consistência interna[4] (PASQUALI, 2009). Considerando-se os custos operacionais de se utilizar o método das metades, índices de consistência interna como o Alfa de Cronbach são amplamente empregados. Esse índice é afetado pela quantidade de itens do instrumento, pela variância de cada item e a correlação entre eles, e pode ser calculado da seguinte forma:

$$\propto = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_x^2}\right)$$

Onde:
$\propto$ é o alfa de Cronbach
n é a quantidade de itens;
σ_i^2 a variância do item i;
σ_x^2 a variância dos escores observados.

4

O alfa de Cronbach usualmente assume um valor entre 0 e 1. São considerados valores de α aceitáveis aqueles superiores a (0,5) ou, mesmo, (0,7), dependendo da aplicação e contexto do teste de tal forma que não existe um consenso. (MAROCO, J.; GARCIA-MARQUES, 2006). Desta forma, para homogeneizar a interpretação dos dados nesse trabalho, será utilizada a classificação de MURPHY & DAVIDSHOLDER (1988, apud MAROCO & GARCIA-MARQUES, 2006):

Tabela 1 Interpretação do Alfa de Cronbach

Confiabilidade	Valor α
Inaceitável	0,0 a 0,6
Baixa	0,6 a 0,7
Moderada	0,7 a 0,9
Elevada	0,9 a 1,0

Também é possível analisar a discriminação individual de cada item, por meio do índice de correlação ponto bisserial, cuja fórmula é a que se segue:

$$r_{pb} = \frac{\bar{C}_A - \bar{C}_T}{DP_t}\sqrt{\frac{p}{q}}$$

Onde:
$\bar{C}_A$ é o score médio dos alunos que acertaram a questão;
$\bar{C}_T$ é o score médio de todo os alunos;
DP_t é o desvio-padrão dos escores;
p é a proporção de alunos que acertaram a questão;
q é a proporção de alunos que erraram a questão.

O INEP (2012) adota a seguinte tabela de referência para interpretação do índice correlação ponto bisserial:

Tabela 2 Interpretação do índice correlação ponto bisserial

Índice de Discriminação	Classificação
≥0,40	Muito Bom
0,30 a 0,39	Bom
0,20 a 0,29	Médio
≤0,19	Fraco

Para que uma nota em uma prova seja coerente, a maioria dos seus itens precisa ter bons índices de discriminação. Nesse sentido, PRICE (2016, p. 187) recomenda que seja feita uma correção do r_{pb} caso um teste seja composto de menos de 25 itens - o item analisado deve ser retirado do computo do escore total no teste, evitando um efeito espúrio devido a grande proporção de variância explicada pelo desempenho no próprio item sobre o escore total. Disso decorre que o coeficiente divulgado nos relatórios síntese de área com os resultados do ENADE, precisa dessa correção é superestimado[5] uma vez que não há menção dessa correção na metodologia de cálculo empregada pelo INEP.

PASQUALI (1996) esclarece que essas técnicas de avaliação de precisão da medida se baseiam no pressuposto que o teste é unidimensional, e que a correlação ponto bisserial parece particularmente incongruente, dado que

> "... a adequação dos demais itens também está por ser demonstrada, inclusive a esta altura das análises do teste ainda não se sabe se os itens do teste são homogêneos, isto é, se o teste é unidimensional, suposição necessária para se poder obter um escore total. Tenta-se resolver este problema procedendo-se a uma análise fatorial dos itens antes da própria análise individual dos mesmos." (PASQUALI, 1996, p. 86).

[5] O Apêndice 3 contém esses índices já corrigidos. A comparação dos resultados pode ser feito acessando os relatórios síntese de área, disponíveis no sítio do INEP.

Ou seja, antes mesmo de se considerar fazer essas análises é indispensável analisar a dimensionalidade do teste. Conforme visto no capítulo 2.1.1 Análise Fatorial.

De forma geral, são feitas diversas críticas às medidas obtidas pela TCT, entre as quais se destacam: os parâmetros de um teste são relativos apenas à população na qual esse foi aplicado, a discriminação dos itens é calculada em referência ao desempenho total no teste, a dificuldade de se comparar testes, a presunção de homogeneidade na variância dos erros de medida e a reduzida confiança nos escores que se afastam do escore mediano (PASQUALI & PRIMI, 2003). Tais críticas, não invalidam as contribuições dessa teoria para a identificação das falhas das provas por meio de índices como o Alfa de Cronbach e o índice de correlação ponto bisserial, o que permite identificar falhas e direcionar a investigação de problemas nos testes, como foi visto ao longo desse capítulo.

2.1.3 Teoria de Resposta ao Item

A Teoria de Resposta ao Item busca analisar a relação entre o nível do traço latente do respondente (θ) e cada item, cujas características são representadas matematicamente pelos seus respectivos parâmetros. Assim, a TRI se interessa por "produzir tarefas (itens) de qualidade." (PASQUALI, 2009, p. 993). Essa teoria tem dois axiomas fundamentais:

> "1º O desempenho em um item é causado por um conjunto de traços latentes;
>
> 2º A relação entre o desempenho na tarefa e o conjunto dos traços latentes pode ser descrita por uma equação monotônica crescente, chamada de CCI (Curva Característica do Item)". (PASQUALI, 2009, p. 994)

De acordo com ANDRADE, TAVARES & VALLE (2000), a escolha de qual modelo se adotar depende da natureza dos itens (dicotômico[6] ou não), do número de populações envolvidas e da quantidade de traços latentes (ANDRADE, TAVARES & VALLE,

[6] Item que é corrigido como certo ou errado – um item de múltipla escolha pode ser considerado dicotômico desde que sua correção resulte em 0 ou 1.

2000, p. 7). Para itens dicotomizados é comum a utilização dos modelos logísticos de 1, 2 e 3 parâmetros (ANDRADE, TAVARES & VALLE, 2000, p. 9). ANDRIOLA (2009, p. 326) aponta que existe um modelo logístico de quatro parâmetros, porém esse ainda é pouco utilizado. As características dos modelos de 3 e 4 parâmetros serão tratadas ao longe desse capítulo.

Já os itens com respostas não dicotômicas, como itens discursivos e escalas tipo Likert, precisam de modelos mais complexos, como o Modelo de Resposta Gradual de Samejima, que é particularmente útil por não presumir a distância (em termos de θ) entre as categorias de resposta.

Consulte MOREIRA JUNIOR (2011) e ANDRADE, TAVARES & VALLE (2000) para se aprofundar nesses e outros modelos.

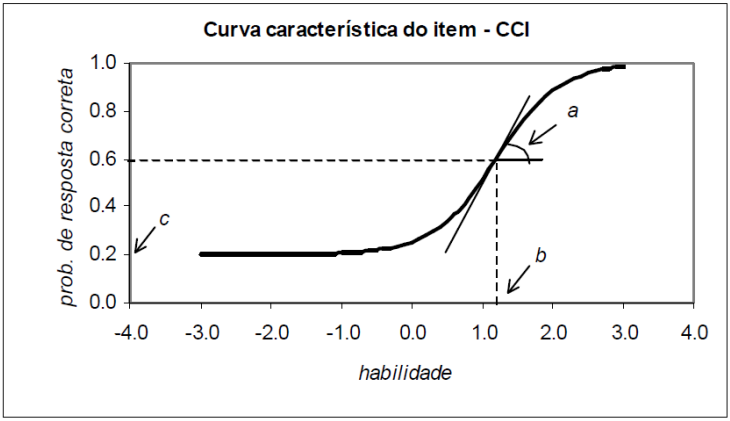
2.1.3.1 Modelo logístico de 3 parâmetros

O modelo logístico de três parâmetros é adequado para um conjunto de itens relativos a uma única dimensão. De forma geral, considerando que um item i meça um traço latente, a magnitude desse traço θ de um sujeito j influencia sua probabilidade P de acerto neste. Essa probabilidade é modelada em função de três parâmetros do item: sua discriminação a, sua dificuldade b e a probabilidade mínima de acerto por indivíduos com baixa habilidade c. Um sujeito tem a probabilidade de responder corretamente o item conforme a seguinte equação:

$$P_j(U_{ij} = 1|\theta j) = c_j + (1 - c_j)\frac{1}{1 + e^{-Da_j(\theta - b_j)}}$$

A forma gráfica dessa equação é conhecida como Curva Característica do Item – CCI e está representada na Figura 2.

Figura 2 Exemplo de Curva Característica do Item do modelo logístico de três parâmetros



Fonte: (ANDRADE; TAVARES; VALLE, 2000, p. 11) ·.

Essa curva permite visualizar rapidamente os parâmetros do item e avaliar suas características. Percebe-se que essa função é monotônica e crescente, ou seja, um aumento de θ nunca corresponde a um decréscimo de P. Além disso, o parâmetro b utiliza a mesma escala da habilidade (θ) e sua posição corresponde a ponto onde $P = (1 - c)/2$ e, apesar de b poder assumir valores entre -∞ e +∞, somente são úteis valores contemplados por magnitudes de θ encontrados na população[7]. Já o parâmetro a é proporcional a inclinação da reta no ponto b, valores baixos de a implicam em a probabilidade de acerto de pessoas com habilidade em niveis muito diferentes sejam parecidas[8]. Por fim, o parâmetro c, assume valores entre 0 e 1, usualmente próximo de $1/n$ onde n é o numero de alternativas que a questão apresenta. O D geralmente assume o valor de 1, mas pode ser substituído por 1,7 quando se deseja que a função forneça resultados semelhantes ao da função ogiva normal[9]. (ANDRADE, TAVARES & VALLE, 2010).

[7] Como θ tem distribuição normal, valores afastados mais que 3 desvios da média devem se examinados com cautela.

[8] MOREIRA JUNIOR (2011, p. 53-54) explica que apesar de não existir um valor exato a para decidir se um item discrimina bem ou não, são aceitáveis a maiores que 0,7, e ideais àqueles maiores ou iguais a 1,0.

[9] Observe que o coeficiente D geralmente assume o valor de 1 e é omitido em muitas formulas. Não se deve confundi-lo com o parâmetro adicional do ML4 que é representado por d.

Além da Curva Característica do Item é importante analisar a Curva de Informação do Item (também conhecida como Função de Informação do Item), que permite identificar quanto um item i contribui para estimar a habilidade de um sujeito em cada nível de habilidade θ, Segundo Andrade, Tavares e Valle (2000), a FII de um ML3 é dada por:

$$I_i(\theta) = D^2 a_i^2 \frac{1 - P_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i} \right]$$

Ou seja, um item fornece mais informação num nível θ quanto maior for o parâmetro a, mais próximo b for de θ e menor for o parâmetro c. O somatório de todas as Funções de Informação do Item de um teste é a Função de Informação do Teste. O erro padrão de estimação é o inverso da raiz da informação naquele ponto. Decorre disso que o erro da estimativa de habilidade de um teste é diferente em níveis diferentes de habilidade em função da informação que seus itens fornecem naquela faixa.

2.1.3.2 Modelo logístico de 4 parâmetros

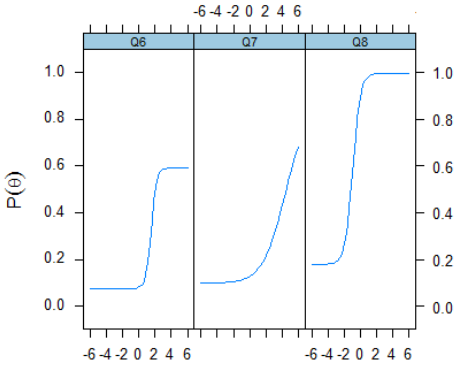
A diferença entre os modelos logísticos de três e quatro parâmetros é a inclusão do parâmetro d, que corresponde à probabilidade máxima de acerto (ou endosso[10]) de um item onde o aumento desse traço não corresponde a um aumento na probabilidade. Em função da falta de estudos nacionais com esse modelo, segue uma breve revisão de trabalhos com esse modelo. A proposição desse modelo foi realizada por BARTON & LORD (1981) e nesse trabalho o parâmetro foi empregado com o propósito de lidar com o problema de respondentes de alta habilidade errar acidentalmente itens abaixo de suas habilidades e consequentemente terem a estimativa de suas habilidades drasticamente reduzidas. A equação proposta para esse modelo foi:

$$P_j(\theta) = c_j + (d_j - c_j) \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}$$

[10] No caso de itens de respostas que não são categorizadas como certa ou errada, a escolha de uma alternativa compatível com níveis maiores de θ é denominada endosso.

Para ilustrar as diferenças entre o ML3 e o ML4, seguem três exemplos de CCI nesse modelo, cada uma com seu próprio valor do parâmetro d, em contraste com o estudo supracitado:

Figura 3 Exemplo de três de curvas característica de item no ML4



Percebe-se que a principal diferença com relação ao ML3 é que a assíntota superior pode assumir valores diferentes de um. MAGIS (2013) define a função da curva de informação de um item j no ML4 como:

$$I_j(\hat{\theta}) = \frac{a_j{}^2\left[P_j(\hat{\theta}) - c_j\right]^2\left[d_j - P_j(\hat{\theta})\right]^2}{\left(d_j - c_j\right)^2 P_j(\hat{\theta})Q_j(\hat{\theta})}$$

É evidente que caso o valor de d seja 1, essa equação se converte naquela do ML3.

Em seu estudo piloto, BARTON & LORD (1981) fixaram o quarto parâmetro, d, igualmente para todos os itens utilizando três valores (0,98), (0,99) e (1,00). Além disso, utilizaram sub-amostras de 1000 pessoas de quatro testes educacionais estadunidenses, metade relativa a habilidades matemáticas e a outra metade linguísticas, os parâmetros a, b e c foram calibrados normalmente utilizando um ML3. Para dois desses testes, o ajuste do modelo foi melhor com parâmetro d = (1,00) e que para os outros dois com d = (0,99). Apesar disso, o ajuste dos modelos não diferiu por mais que 0,02%. As estimativas de habilidades dos modelos diferiram entre (-0,36) e (0,64) desvios-padrões.

Ainda, contrariando a hipótese de que erros são mais prováveis com testes com pressão de tempo, o ML4 não se ajustou melhor que o ML3 para o único desses testes com limitação considerável de tempo. Deste modo, os autores concluíram que o ML4 não resultou em ajuste melhor aos dados ou em mudança significativa para a estimação das habilidades, além de não recomendar seu uso em função do aumento do tempo computacional. Além disso, a pequena diferença entre os modelos e os escores contrariaria a tese que ML3 superestima as habilidades dos estudantes de altas habilidades.

LOKEN & RULISON (2010) utilizaram dados simulados para verificar a consequência da utilização de modelos logísticos de 2 e 3 parâmetros com testes programados para terem os 4 parâmetros com distintas quantidades de itens. De forma geral, observou-se que a discriminação dos itens era menor e sua dificuldade maior quando se utilizava os modelos de 2 e 3 parâmetros para itens simulados para terem 4 parâmetros. De forma semelhante, observou-se que as curvas de informação dos testes com menos parâmetros tem diferenças sistemáticas quando comparada àquelas do ML4 – no ML3 ela é deslocada para direita e superestimada, pois se atribui mais informação que o apropriado à cauda superior e difere daquela do modelo de 2 parâmetros apenas em relação à informação, que seria superestimada, ao se atribuir mais informação em ambas às caudas, apesar de menor discriminação.

LOKEN & RULISON (2010) questionam se é realmente apropriado se fixar o parâmetro d, uma vez que os demais parâmetros são relativos ao item e não aos respondentes. Esse desajuste pode ser ilustrado em estudos sobre escalas atitudinais, onde sujeitos com θ elevado não endossam universalmente itens relativamente fáceis – o que pode ser atribuído a fatores como desejabilidade social. Além de analisar dados simulados, os autores aplicaram o ML4 nos dados coletados em uma pesquisa da mensuração autorrelatada de delinquência[11] de 2005 – nesses dados os parâmetros tiveram padrão similar ao observado na simulação. Os autores apontaram a necessidade de se tratar e aperfeiçoar questões computacionais e conceituais relativas à estimação dos parâmetros c e d.

MAGIS (2013) num trabalho de cunho algébrico busca encontrar o nível de habilidade latente que maximiza a informação que um item fornece no ML4. O autor propõe que a utilização de suas equações

[11] *Self-report measure of delinquency*

melhoraria o cálculo e permitiria, por conseguinte uma maior robustez na estimação do traço latente.

YEN *et al* (2012) investigaram a utilização do modelo de 4 parâmetros no contexto da testagem adaptativa computadorizada (*computerized adaptive test*) utilizando dados não simulados, utilizando um d fixo de (0,98). A investigação envolveu uma condição experimental na qual os estudantes tinham erros automáticos nos dois primeiros itens do teste e verificou-se que a utilização do ML4 serve tanto para reduzir o efeito de erros de estimativa causados por desatenção do respondente, quanto para aumentar a precisão da medida - pois eram necessários menos itens no ML4 que no ML3 para atingir o mesmo erro de medida durante a testagem adaptativa. Para estudos futuros os autores recomendam a utilização de itens calibrados com todos os parâmetros individualizados, a realização de experimentos modelos com erros forçados no meio e final do teste e a aplicação de testagem adaptativa dos dois sistemas aos mesmos estudantes.

ŚWIST (2015) investigou se o ML4 poderia ser utilizado para detectar erros na elaboração de item - comparando os parâmetros dos itens nos modelos logísticos de 2, 3 e 4 parâmetros com uma análise qualitativa. Para esse fim, valores não usuais ou expressivos dos parâmetros c e d, valores pequenos do parâmetro a ou ainda valores extremos do parâmetro b foram considerados como indicativos de problemas nos itens. Constatou-se que itens com problemas de elaboração geralmente tinham o parâmetro a ruim, mesmo em modelos com 2 e 3 parâmetros, não havendo vantagem na utilização do quarto parâmetro para a detecção de problemas de elaboração. Além disso, alguns itens que não violaram as regras gerais de elaboração tiveram parâmetro d afastado de zero, o que teria sido ocasionado devido à ambiguidade de seus conteúdos. A autora recomenda que futuras pesquisas envolvendo ML4 utilizem dados simulados e se concentrem na estimação de parâmetros d individuais.

CHENG & LIU (2015) analisaram analiticamente o efeito dos parâmetros c e d na testagem adaptativa e exploraram o problema da escolha do próximo item para apresentar ao respondente. Os autores defendem que o método de escolha do próximo item baseado no parâmetro b[12] fornece mais informação que o método de máxima informação de Fisher, devido ao ajuste mais conservador desse método, o que permite compensar o efeito de erros ou acertos acidentais. Também foram contextualizadas três razões que reacenderam o interesse

[12] Tradução livre de *"b-matching method"*

nos ML4: Seu valor para o melhor ajuste em modelos psicológicos, a compensação de erros iniciais no score final estimado e os avanços em computação estatística que tornaram mais simples calcular o parâmetro d individualmente para cada item.

2.1.3.3 Modelo de resposta gradual

ANDRADE, TAVARES & VALLE (2010) explicam que a TRI possui modelos adequados para itens abertos ou que possuam diversas possibilidades de resposta. O MRG de SAMEJIMA assume que há uma hierarquia entre as respostas e que cada opção de resposta corresponde a uma curva do modelo logístico de dois parâmetros. Segue um exemplo CCI e CI de um item calibrado nesse modelo, com três classificações de resposta.

Figura 4 Exemplo de CCI de item calibrado pelo Modelo de Resposta Gradual

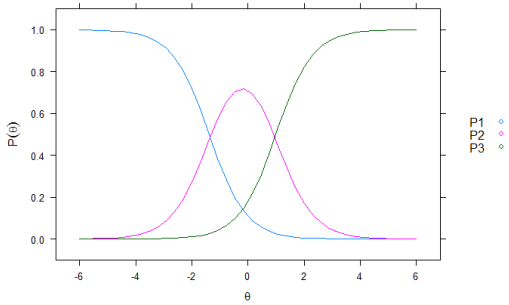
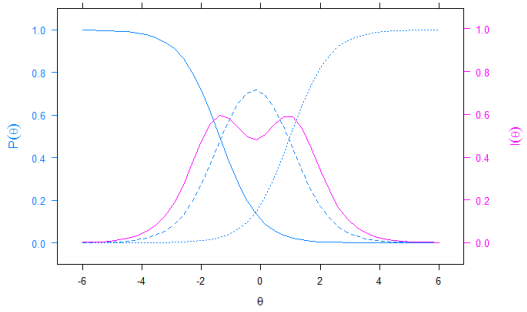


Figura 5 Exemplo de curva de informação de item calibrado pelo Modelo de Resposta Gradual



A curva vermelha evidencia que esse modelo fornece informação sobre a habilidade do respondente ao longo dos diversos níveis de desempenho equivalentes a cada nível ou opção de reposta (representadas pelas curvas azuis). Essa vantagem permite que a resposta do item seja útil para determinar a habilidade (θ) ao longo de uma ampla faixa de respostas, diferente dos modelos para itens dicotômicos, cuja informação é maximizada no nível de dificuldade correspondente ao parâmetro b. Em contrapartida a essa vantagem, o erro é acentuando se houver a interposição de múltiplos níveis de desempenho, o que torna necessário trabalhar com uma quantidade limitada de categorias de nível de desempenho.

Esse modelo não funciona bem se houver bastante interposição entre vários níveis de desempenho, o que torna necessário agrupar algumas faixas de desempenho. Em 2002, o INEP (2002, p. 16), agrupou as notas das redações do ENEM em três faixas de desempenho - Insuficiente a Regular [0, 40], Regular a Bom (40, 70] e Bom a Excelente (70, 100], padrão similar ao utilizado para as questões objetivas do mesmo exame, porém o propósito deste procedimento era unicamente facilitar a interpretação. KLEIN & FONTANIVE (2009) na perspectiva de empregar a Teoria de Resposta ao Item no novo modelo de correção das questões discursivas do ENEM, no qual cinco competências são avaliadas com quatro níveis de desempenho já definidos, recomendaram que o julgamento dos dois corretores independentes fosse convertido em uma escala de sete pontos, com três pontos extras correspondendo aos níveis intermediários entre os quatro pontos originais. BRAGA (2015) faz uma proposta diferente para a correção do ENEM, atribuindo seis pontos que configurariam a concordância com o quanto o desempenho do respondente se aproxima do desempenho ideal referente àquela competência. Já NOGUEIRA (2008, p. 71) optou por converter os itens discursivos em itens dicotômicos, atribuindo acerto (1) àqueles no qual a nota do estudante foi superior à média de todos os estudantes e errado (0) aos demais.

Detalhes sobre o modelo de resposta gradual podem ser consultados em ANDRADE, TAVARES & VALLE (2010) e SAMEJIMA (1972).

2.1.3.4 Modelos de itens de formato misto

ERCIKAN et al (1998, p. 137) explicam que a combinação de itens de múltipla escolha e discursivos tem tido acentuado interesse, dado que os dois tipos de itens possuem vantagens complementares. De

um lado itens objetivos são eficientes em termos de tempo de aplicação e permitem o aumento da validade de um teste ao melhorar sua representatividade dos conteúdos. Já os discursivos são mais apropriados para alguns tipos de tarefas como a resolução de problemas, e, portanto aumentariam a validade da avaliação. Os autores defendem que a calibração simultânea de diversos tipos de itens utiliza a informação a respeito do desempenho dos respondentes de uma forma estatisticamente ótima (IBID, p. 138). Além disso, os itens discursivos, em função da ausência de acerto (e erro) casual, fornecem informação sobre o desempenho de estudantes com valores extremos de θ de uma forma que os itens objetivos não poderiam (IBID, p. 152).

WAINER & THISSEN (1993) explicam que a teoria e a tecnologia daquele ano já permitiam a calibração simultânea de itens em modelos distintos da Teoria de Resposta ao Item desde que uma larga quantidade de respondentes estivesse disponível. Os autores esclarecem que a calibração simultânea evita problemas dos ajustes das escalas dos diferentes modelos.

2.2 Revisão Bibliográfica

Nessa sessão será apresentada uma breve revisão de trabalhos com análises quantitativas dos resultados do componente da Formação Geral do ENADE.

2.2.1 Artigos

Acessou-se a Biblioteca Eletrônica *Scientific Electronic Library Online* - SciELO pelo sítio http://www.scielo.br/, onde se realizou a pesquisa utilizando os seguintes critérios:

1. Trabalhos publicados entre 2010 e 2014;
2. Que contivessem o termo ENADE em qualquer campo;
3. Contivessem uma análise quantitativa de resultados do ENADE incluindo o componente da Formação Geral.

A partir do critério 1 e 2, foram levantados 21 trabalhos, após leitura, concluiu-se que apenas 6 atendiam o critério 3.

As diferenças de desempenho de ingressantes e concluintes foram investigadas em alguns trabalhos. GURGEL (2010) verificou que as notas de FG de ingressantes e concluintes diferiam pouco em magnitude, tanto para cima, quanto para baixo, nos anos de 2004 e de

2007 entre os estudantes de universidades do Piauí, já LANZILLOTTI & LANZILLOTTI (2014), ao analisarem os resultados dos estudantes de estatística no ENADE 2009, verificaram que em 2009 a média nacional na FG dos estudantes ingressantes de estatística foi superior em 0,99 pontos aos dos concluintes. Esses autores afirmam que isso sinalizaria "... que os conteúdos que contribuem para a atualização dos fatos relativos à formação cultural estão sendo negligenciados ao longo do curso" (LANZILLOTTI & LANZILLOTTI, 2014, p. 164). Em nível regional, os autores observaram esse padrão nas regiões Sul e Sudoeste, porém o resultado foi oposto nas regiões Norte e Nordeste, todavia não foi possível analisar os dados da região Centro-Oeste.

Outros estudos comparam o desempenho do CE e da FG, CORTELAZZO & RIBEIRO (2013) analisaram as médias e a variabilidade do desempenho de alunos de biologia de escolas estaduais e municipais de São Paulo nos exames de 2005 e de 2008 e verificaram que as instituições com piores desempenhos foram às faculdades e as Universidades Federais. Ao verificar a proximidade do desempenho de ingressantes e concluintes na FG, os autores questionam se os objetivos da FG não seriam mais pertinentes ao Exame Nacional do Ensino Médio. GONTIJO *et al* (2011) analisou o desempenho de estudantes de medicina no ENADE de 2004 e de 2007 e verificaram que 75% das questões de FG foram fáceis ou muito fáceis para estudantes desses cursos[13]. Também verificou uma relação linear entre o desempenho dos concluintes desses cursos na prova da Formação Geral e do Componente Específico – para cada aumento de 1 ponto na FG correspondia a 0,3 no CE entre os concluintes do curso. Contudo os mesmo autores apontam que "as notas muito próximas de ingressantes e concluintes suscitam questionamentos sobre a qualidade das questões." (GONTIJO *et al*, 2011, p. 212).

Por fim, alguns estudos investigaram os fatores associados ao desempenho no ENADE e na FG. SANTOS (2011) com o objetivo de observar o efeito da participação em uma metodologia pedagógica implantada por uma das instituições comparou o desempenho de ingressantes de quatro cursos na FG de instituições de um mesmo município e encontrou uma correlação entre a autoavaliação de participação nessas atividades e a respectiva nota na FG. Já SILVA, VENDRAMINI & LOPES, (2010) estudaram diversos fatores socioeconômicos associados ao desempenho no ENADE 2005 e

[13] Uma questão fácil tem taxa de acerto entre 61 e 85%, já uma questão muito fácil tem taxa de acerto igual ou superior a 86%

verificaram diferenças de desempenho por sexo em quinze dos vinte cursos avaliados, oito favorecendo as mulheres. Além disso, verificaram uma diferença de desempenho nesse componente favorável aos brancos em 12 cursos.

Dessa forma se observa que a maioria dos estudos se concentra seja nos resultados de uma área de conhecimento ou nos resultados de uma instituição, tipicamente considerando os fatores associados ao desempenho, conforme se observa no Quadro 2.

Quadro 2 Abrangência da temática dos estudos selecionados na revisão de literatura

Autor (es)	Estudo restrito a uma área do conhecimento	Estudo restrito a resultados de uma instituição ou estado
GURGEL (2010)	Não	Sim
SANTOS *et al* (2011)	Não	Sim
GONTIJO (2011)	Sim	Não
SILVA; VENDRAMINI; LOPES (2010)	Não	Não
CORTELAZZO; RIBEIRO (2013)	Sim	Sim
LANZILLOTTI & LANZILLOTTI (2014)	Sim	Não

2.2.2 Teses e Dissertações

De forma geral existem poucas Teses e Dissertações com análises quantitativas sobre os resultados da prova de Formação Geral. Nesse sentido, MOLCK & CALDERÓN (2014) realizaram uma revisão de literatura de trabalhos de mestrado e doutorado de 2004 -2010 com o tema ENADE, e identificaram sete eixos temáticos predominantes em 61 trabalhos, conforme a tabela 3.

Tabela 3 Número de trabalhos por eixo temático no estudo de MOLCK & CALDERÓN (2014)

Eixo Temático	Número de trabalhos
Melhorias do desempenho institucional	6
Melhorias dos cursos de graduação	5
Fatores de desempenho do aluno	5
Política pública avaliativa	5
Aspectos técnicos e operacionais	4
Estratégias didáticas para a melhoria do desempenho do aluno	4
Formação de professores	3

FONTE: MOLK & CALDERÓN (2014)

Dentre os quatro estudos que lidaram com aspectos técnicos e operacionais, três utilizaram a Teoria de Resposta ao Item em suas análises, porém cada um abordou o resultado de um curso. Achado semelhante é feito em relação aos trabalhos ligados aos fatores de desempenho do aluno – eles também se concentraram no desempenho dos alunos em seus respectivos cursos (MOLCK & CALDERÓN, 2014).

A fim de complementar essa revisão, acessou-se a Biblioteca Digital Brasileira de Teses e Dissertações pelo sítio http://bdtd.ibict.br/, onde se realizou a pesquisa utilizando os seguintes critérios:

1. Trabalhos de 2003 ou mais recentes
2. Conter análises quantitativas de resultados da FG

A partir do primeiro critério, foram levantados 51 trabalhos, cujos resumos foram analisados, elegendo-se apenas doze trabalhos que atendiam o segundo critério. Leram-se os resumos de cada um destes, e apenas se efetuou a leitura integral daqueles que contivessem análises quantitativas de resultados do componente da formação geral do ENADE, com o total de cinco trabalhos de mestrado e doutorado.

SILVA (2013), ao investigar a influência de disposições socioculturais nos ENADE dos anos de 2005, 2008 e 2011, encontrou que de "forma geral três fatores chamaram a atenção na análise realizada: a renda, o tipo de escola e a etnia dos estudantes avaliados" (p. 115). Os dados desse estudo também revelam uma disparidade entre o desempenho médio dos alunos de cada curso na Formação Geral nos

três anos. As notas médias mínimas e máximas na Formação Geral entre os cursos de cada ano foram respectivamente, 40,2 e 59,0 em 2011, 49,6 e 50,4 em 2008 e 45,7 e 62,5 em 2005. Esses valores sugerem um possível efeito da natureza do curso no desempenho na formação geral, todavia as diferenças dentro de cada curso decorrentes das diferenças sociodemográficas dos egressos são notáveis para a maioria das áreas em todos os anos.

GAUDIO (2014) investigou as diferenças de desempenho na Formação Geral do ENADE dos bolsistas do Programa Universidade para Todos (ProUni) com não bolsistas, verificou que esses tiveram um desempenho inferior na FG de até cinco pontos nos anos de 2006, 2007 e 2008. Ocorreram exceções nesse padrão em algumas áreas, como a Medicina, em que a média da nota geral dos bolsistas foi inferior em 14,49 pontos à média da nota geral verificada para a área. Os autores consideram que essas diferenças podem ser atribuídas ao perfil socioeconômico desses cursos e à falta de respondentes bolsistas em algumas áreas.

SANTOS (2012), ao estudar determinantes do desempenho de alunos de cursos de ciências contábeis nos anos de 2002, 2003 e 2006, tanto no ENADE quanto no ENC, concluiu que o desempenho "... é afetado pela interação entre características próprias dos discentes, como aspectos pessoais, socioeconômicos e os insumos das instituições de ensino." (p. 193). Essa conclusão é apoiada por um modelo hierárquico linear que explicou menos de 10% da variância total nos anos de 2002, 2003 e 2006.

MOREIRA (2010) estudou a influência de fatores institucionais sobre o rendimento de estudantes concluintes de quatro cursos distintos no ENADE 2005. Essa análise é feita por meio do controle das variáveis individuais. Foi elaborado um modelo para explicar a nota bruta do estudante do ENADE em relação a três conjuntos de variáveis – características socioeconômicas, características acadêmicas e características individuais. Destaca-se a utilização de dois índices para resumir a qualidade de fatores institucionais, qualidade de infraestrutura física e equipamentos e qualidade da biblioteca. O efeito dessas variáveis e das demais era distinto para cada um dos cursos.

PAIVA (2010) comparou os resultados dos cursos por modalidade – Presencial ou Educação a Distância - no ano de 2008 e observa que os estudantes da educação presencial tem média superior na FG - 52,1 - em comparação ao desempenho médio de concluintes EAD na FG - 48,16. Com relação ao componente específico foram feitos testes em cada área e não foi encontrado um padrão geral para as áreas

avaliadas, ora favorecendo os estudantes de curso presencial, ora favorecendo os cursos à distância.

Assim, de forma similar aos artigos, os trabalhos de mestrado e doutorado se concentraram nos resultados de uma área de conhecimento ou nos resultados de uma instituição e os fatores associados a eles.

2.3 Síntese do capítulo

A teoria da medida oferece muitas ferramentas para analisar os resultados de provas com itens de múltipla escolha e discursivos, além de orientar a análise desses resultados para a determinação se uma nota de prova, calculada a partir da taxa de acerto de diversos itens é apropriada ou não. É imprescindível, portanto se apropriar dessas técnicas e realizar a crítica sobre a qualidade dos instrumentos empregados na avaliação da educação.

Observou-se que poucos estudos fazem análises quantitativas da Formação Geral, sendo, portanto contribuir com esse tipo de análise. Além disso, a maioria dos trabalhos não faz critica com relação à qualidade da medida de Formação Geral que é produzida, investigando-se os fatores associados ao desempenho dela, sem estabelecer sua confiabilidade. Assim desenvolver um estudo com essa temática especifica se faz necessário.

Além disso, a Teoria de Resposta ao Item oferece uma série de técnicas que podem ser utilizadas para produzir medidas mais precisas e até mesmo integrar as notas das questões objetivas e discursivas. A grande quantidade de participantes do ENADE facilita a calibração de modelos – o que torna oportuno a utilização desses modelos, desde que atendidos seus pressupostos.

3 Método

Nesse capítulo serão apresentadas as provas cuja aplicação gerou os dados utilizados nesse estudo e os procedimentos de análise empregados nesse estudo.

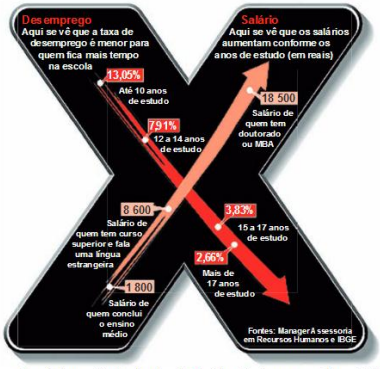
3.1 Instrumento

A prova da Formação Geral é concebida em função da lei 10.861 (BRASIL, 2004) e tem sua estrutura baseada nas diretrizes de prova publicadas anualmente pelo INEP. Nos três anos considerados nesse estudo (2010, 2011 e 2012) o componente da FG foi avaliado por meio de 10 itens, sendo dois discursivos e oito de múltipla escolha, cada um com cinco alternativas de resposta. A Figura 6 é um exemplo de item de múltipla escolha utilizado no ENADE.

Figura 6 Questão 6 utilizada no ENADE 2011

QUESTÃO 6

A educação é o Xis da questão



Disponível em: <http://ead.uepb.edu.br/noticias,82>. Acesso em: 24 ago. 2011.

A expressão "o Xis da questão" usada no título do infográfico diz respeito

A) à quantidade de anos de estudos necessários para garantir um emprego estável com salário digno.
B) às oportunidades de melhoria salarial que surgem à medida que aumenta o nível de escolaridade dos indivíduos.
C) à influência que o ensino de língua estrangeira nas escolas tem exercido na vida profissional dos indivíduos.
D) aos questionamentos que são feitos acerca da quantidade mínima de anos de estudo que os indivíduos precisam para ter boa educação.
E) à redução da taxa de desemprego em razão da política atual de controle da evasão escolar e de aprovação automática de ano de acordo com a idade.

Os itens de provas de Formação Geral tratam de temas diversos e visam avaliar se os estudantes têm capacidades e competências específicas conforme as diretrizes estabelecidas nas portarias de prova publicadas anualmente pelo INEP. Apesar de em 2010 não haver portaria específica para esse componente, cada portaria das áreas avaliadas pelo ENADE 2010 possui artigos especificando como será a avaliação desse componente, a exemplo da Portaria 214 que versou sobre a avaliação da Agronomia (MINISTÉRIO DA EDUCAÇÃO, 2010). Desde 2011, a avaliação desse componente é definida em portaria específica para àquele ano - a Portaria 188 em 2011 e a Portaria 207 em 2012 (MINISTÉRIO DA EDUCAÇÃO, 2011) (MINISTÉRIO DA EDUCAÇÃO, 2012).

De forma geral se observa muita semelhança entre as portarias sendo as de 2011 e 2012 iguais. Todavia, a de 2010 difere das demais por não avaliar a competência "VI - atuar segundo princípios éticos" e possuir 21 temas em vez dos 13 temas observados em 2011 e 2012.

3.2 Fontes de dados

Os dados relativos ao ENADE dos anos de 2010, 2011 e 2012 estão disponíveis no sítio eletrônico do INEP, no endereço http://portal.inep.gov.br/microdados. Nesses arquivos as bases de cada ano são disponibilizadas seus respectivos dicionários de variáveis e sintaxes para manipulação em alguns programas comerciais.

O procedimento de limpeza encontra-se descrito no Apêndice 1 – Roteiro de limpeza dos bancos de dados.

Foi necessário realizar o procedimento descrito no Apêndice 4 – Sintaxe de alteração do banco do ENADE 2011 no R para importar a base de dados de 2011 para o Excel. Na Tabela 4 está listada a quantidade de pessoas nos bancos a cada etapa do procedimento de limpeza.

Tabela 4 Quantidade de pessoas presentes no banco em cada etapa dos procedimentos de limpeza do banco

Tratamento do banco	Ano 2010	2011	2012
Nenhum tratamento	422896	376219	587351
Etapa 1: Remoção de ingressantes no curso	161151	376219	587352
Etapa 2: Remoção de pessoas ausentes na parte objetiva da FG	147737	296223	466469
Etapa 3: Remoção de pessoas com qualquer resposta em branco ou desconsiderada na parte objetiva[14]	145211	288698	456506
Etapa 4: Remoção de pessoas com ausência em algum item discursivo	103429	227121	349076

Como 2010 foi o último ano no qual os ingressantes dos cursos participaram do ENADE, estes foram removidos da análise para evitar a possível influência dessa condição no desempenho. Além disso, em função da parte objetiva da FG ser composta por apenas oito itens e uma parcela relativamente pequena ter tido alguma resposta ausente ou desconsiderada, foi decidido remover do banco qualquer estudante que tivesse alguma resposta ausente ou anulada em alguma das questões de múltipla escolha. Adicionalmente, dado o objetivo de se calibrar simultaneamente as questões discursivas, foi necessário remover qualquer estudante que deixou qualquer uma das duas questões discursivas em branco, o que resultou em uma redução de até 29% dos estudantes em comparação à etapa anterior.

Como o INEP remove do cômputo das notas de uma área as questões cuja correlação ponto-bisserial é inferior a 0,20, foi necessário recalcular a nota da parte objetiva da formação geral antes de se extrair

[14] Foram encontrados participantes que constavam como presentes na parte objetiva da formação geral, porém sem o vetor de escolha das questões, essas pessoas foram removidas na etapa 3.

as estatísticas descritivas. Assim, todas as questões foram consideradas para o cálculo das médias nacionais com seu gabarito original. Ilustra esse fenômeno a anulação da questão 6 em 2010 para dez áreas de conhecimento - Enfermagem, Fonoaudiologia, Nutrição, Fisioterapia, Serviço social, Terapia ocupacional, Tecnologia em Radiologia, Tecnologia em Agroindústria, Tecnologia em Gestão Hospitalar e Tecnologia em Gestão Ambiental. Já em 2011, foi feita a anulação da questão 8 apenas para o curso de Tecnologia em Alimentos.

3.3 Procedimento

A análise de dados seguiu os seguintes passos:

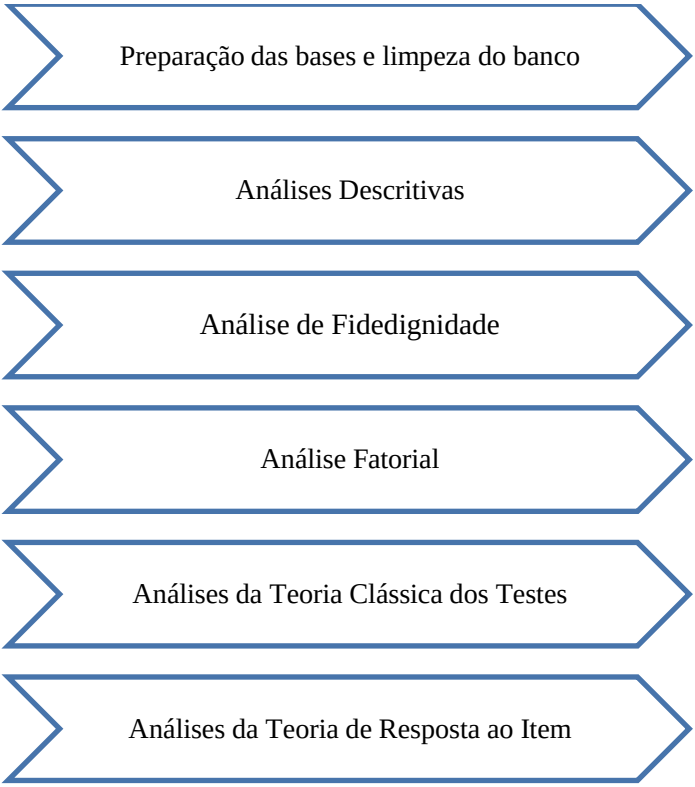


Figura 7 Etapas do procedimento de tratamento dos dados

Os arquivos com os bancos de dados foram inicialmente manipulados em um editor de planilha eletrônica, o Excel 2010©, conforme Apêndice 1 – Roteiro de limpeza dos bancos de dados.

As análises descritivas[15] foram realizadas utilizando o suplemento de "Análise de dados" do Excel 2010©. Calculou-se a média, o desvio padrão, a moda, a mediana, o mínimo, o máximo, o 1º Quartil e o 3º quartil da média geral na parte objetiva. Ressalta-se que a nota da parte objetiva é calculada numa escala de 0 a 100, e, consequentemente, cada questão objetiva vale 12,5 pontos.

Os itens discursivos são originalmente pontuados em uma escala de 0 a 100 pontos, com intervalos de cinco em cinco pontos. Assim, em função da falta de praticidade de se trabalhar potencialmente com 21 parâmetros de dificuldade no Modelo de Resposta Gradual foi necessário agrupar esses escores em menos pontos. Como não foi encontrado consenso na literatura sobre o melhor procedimento de agrupamento de faixas de desempenho e visando à aproximação com a análise da facilidade das questões objetivas do ENADE realizada pelo INEP no ENADE, e considerando que não há descrição comportamental do significado das faixas de desempenho de cada questão se realizou a seguinte recodificação:

[15] As análises descritivas são um conjunto de técnicas que buscam resumir uma distribuição de dados e suas características. Detalhes sobre cada índice e suas respectivas metodologias de cálculo podem ser observados em livros de estatística básica como HEALEY (2010).

Quadro 3 Categorização de escores de itens discursivos e sua correspondência às categorias do índice de facilidade utilizadas no ENADE

Categoria na questão discursiva	Nota na questão discursiva	Categoria correspondente do índice de facilidade
0	Até 15 pontos	Item muito difícil
1	Entre 16 e 40 pontos	Item difícil
2	Entre 41 e 60 pontos	Item médio
3	Entre 61 e 85 pontos	Item fácil
4	Mais de 85 pontos	Item muito fácil

As análises de correlação ponto-bisserial e alfa de Cronbach foram realizadas utilizando o programa ITEMAN for Windows 3.50.

Para o restante das análises foi utilizado o ambiente e linguagem R (R Core Team, 2016) e os pacotes Readxl (WICKHAM, 2016), Psych (REVELLE, 2016) e MIRT (CHALMERS, 2012).

A análise fatorial teve como objetivo investigar a pertinência de se utilizar técnicas apropriadas para uma dimensão e foram gerados os seus resultados e gráficos com os autovalores (*scree plot*). A programação dessa análise encontra-se disponível no Apêndice 2 – Programação no R O objetivo dessa análise foi verificar se o pressuposto de Unidimensionalidade da do capítulo 2.1.1.1 Unidimensionalidade foi atingido.

A calibração dos itens nos modelos da teoria de resposta ao item também foi realizada no R, gerando estatísticas que resumem os modelos, os parâmetros de cada item, comparações entre os modelos por meio da ANOVA e AIC e BIC, além de serem produzidos os gráficos com as Curvas Características do Item, Curvas de Informação do Item, Função de Informação Total do Teste.

4 Resultados

Nessa seção são apresentados os resultados de cada tipo de análise realizada e dos modelos calibrados pela Teoria de Resposta ao Item.

4.1. Análises descritivas

As análises descritivas dos itens objetivos e discursivos serão apresentadas a seguir. A análise dos itens objetivos tem o intuito de verificar se o procedimento de limpeza empregado alterou significativamente tipo de distribuição dos escores. Já a análise dos itens discursivos tem como objetivo verificar a adequação do procedimento de recodificação de variáveis empregado.

4.1.1. Itens objetivos

A distribuição das notas na parte objetiva de cada prova será analisada na Tabela 5. Também serão comparadas as análises descritivas relativas aos estudantes presentes no dia da prova, considerando-se dois segmentos - antes e depois da remoção de estudantes que deixaram alguma das questões discursivas em branco para verificar se esse procedimento não alterou a distribuição dos escores nessa parte da prova.

Tabela 5 Análises descritivas dos escores nas questões objetivas da Formação Geral das edições 2010, 2011 e 2012 do ENADE, antes e depois da remoção dos ausentes nas questões discursivas.

Ano	2010		2011		2012	
Estatística	Sem remoção	Com remoção	Sem remoção	Com remoção	Sem remoção	Com remoção
N	145211	103429	288698	227121	456506	349076
Média	47,14	48,75	50,59	51,67	46,45	47,56
Desvio padrão	19,13	18,70	19,28	18,81	20,25	19,98
Moda	50	50	50	50	50	50
Mínimo	0	0	0	0	0	0
1 Quartil	37,5	37,5	37,5	37,5	37,5	37,5
Mediana	50	50	50	50	50	50
3 Quartil	62,5	62,5	62,5	62,5	62,5	62,5
Máximo	100	100	100	100	100	100

Observa-se que as estatísticas descritivas se mantiveram estáveis mesmo com a remoção dos estudantes ausentes nas questões discursivas o reflete a adequação do procedimento de limpeza. As medidas de tendência central – média, mediana e moda – estão muito próximas antes e depois da limpeza, concentradas no meio dos valores possíveis dessa variável (0 a 100), que as medidas de dispersão Mínimo, 1º quartil, mediana, 3º quartil e máximo indicam a distribuição regular desses valores.

4.1.2. Itens discursivos

Nessa seção serão exibidas as distribuições de frequência de cada item discursivo das provas do ENADE antes e depois de procedimento de recodificação descrito na seção 3.3 Procedimento, para se avaliar se esse modificou significativamente o formato da distribuição. Para facilitar a comparação das questões discursivas de cada prova, serão exibidos em cada gráfico.

A Figura 8 contém a distribuição de pessoas em cada faixa desempenho das questões discursivas D1 e D2 do ENADE 2010 e a Figura 9 contém a distribuição após a conversão.

Figura 8 Gráfico de frequência de notas nas Questões Discursivas 1 e 2 do ENADE 2010

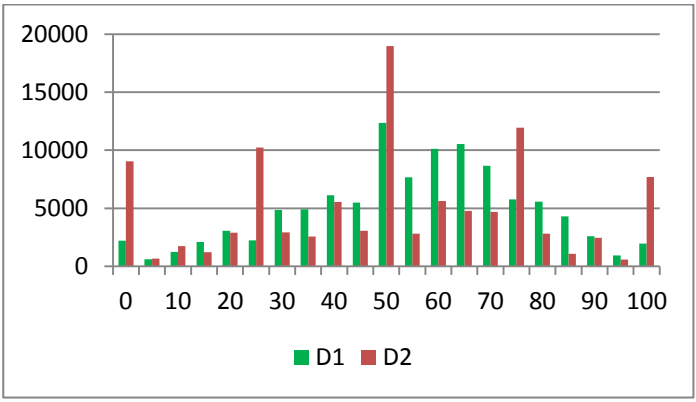
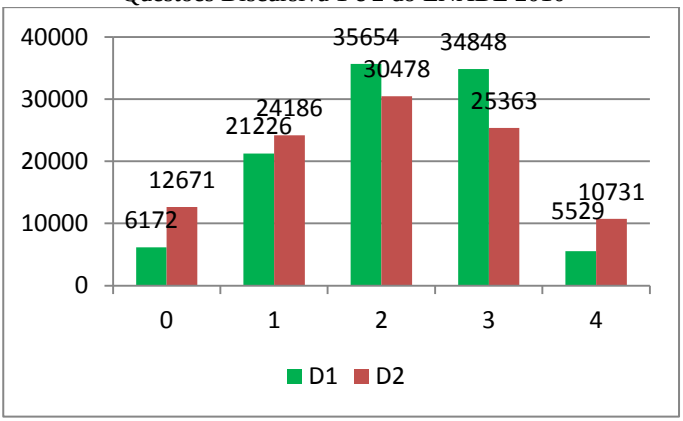


Figura 9 Gráfico de frequência de cada faixa de desempenho recodificado nas Questões Discursiva 1 e 2 do ENADE 2010



Verifica-se que a recodificação do desempenho dos itens discursivos do ENADE 2010 tem um padrão similar a distribuição original e tem formato mais semelhante à distribuição normal.

A Figura 10 contém a distribuição de pessoas em cada faixa desempenho das questões discursivas 1 e 2 do ENADE 2011 e a Figura 11 contém a distribuição após a conversão.

Figura 10 Gráfico de frequência de notas nas Questões Discursivas 1 e 2 do ENADE 2011

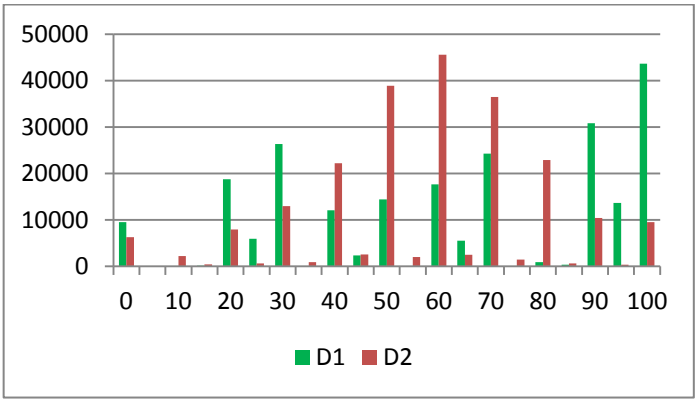
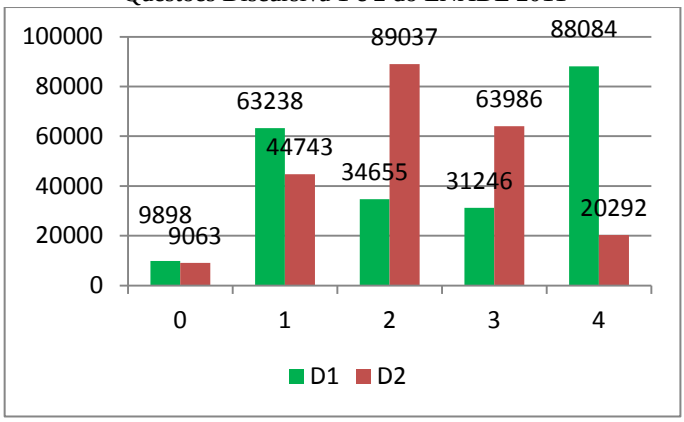


Figura 11 Gráfico de frequência de cada faixa de desempenho recodificado nas Questões Discursiva 1 e 2 do ENADE 2011



Verifica-se que a questão D1 de 2011 teve quantidade menos expressiva de escores nos níveis intermediários e sua distribuição não se assemelha a distribuição normal, diferente do que ocorreu com a questão D2. Após a recodificação, o formato de ambas as distribuições sem manteve.

A Figura 12 Figura 8contém a distribuição de pessoas em cada faixa desempenho das questões discursivas 1 e 2 do ENADE 2012 e a Figura 13 contém a distribuição após a conversão.

Figura 12 Gráfico de frequência de notas nas Questões Discursivas 1 e 2 do ENADE 2012

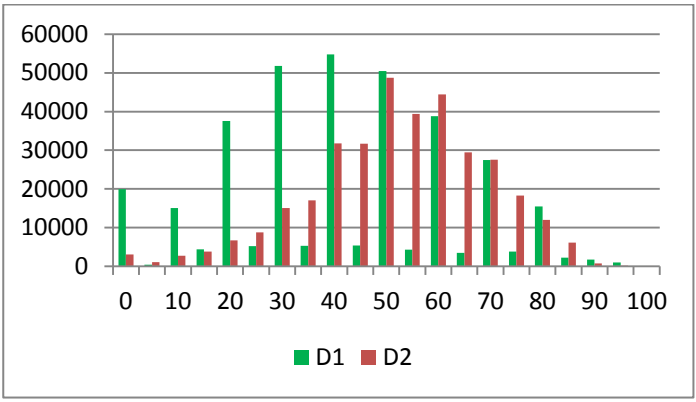
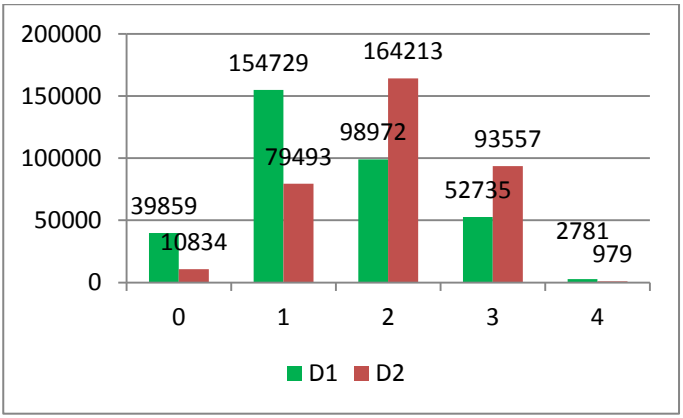


Figura 13 Gráfico de frequência de cada faixa de desempenho recodificado nas Questões Discursiva 1 e 2 do ENADE 2012



Verifica-se que a recodificação do desempenho dos itens discursivos do ENADE 2012 tem um padrão similar a distribuição original e tem formato mais semelhante à distribuição normal. Ressalta-se que a distribuições é mais concentrada nos valores mais baixos,

indicando que os estudantes tiveram desempenho pior nesses itens que nos anos anteriores.

Percebe-se que a recodificação dos itens manteve a forma da distribuição em todos os anos o que reflete a adequação desse procedimento.

4.2 Análise Fatorial

Nessa seção serão apresentados os índices de bondade de ajuste da análise fatorial dos itens objetivos e os gráficos de autovalor relativos a cada ano. Não foi possível analisar os resultados dos itens discursivos, mesmo após a recodificação no irt.fa. A análise fatorial utilizou correlações tetracóricas e o método dos resíduos mínimos. A seguir são apresentados os gráficos scree plot da prova de cada ano.

Figura 14 Scree plot das questões objetivas da FG de 2010

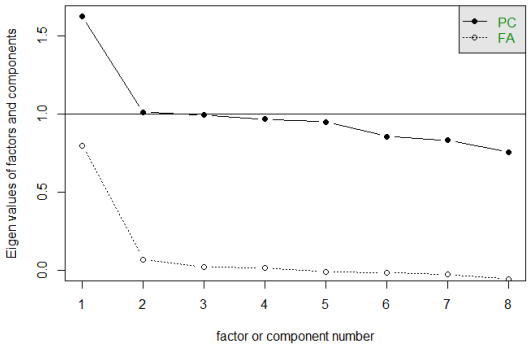


Figura 15 Scree plot das questões objetivas da FG de 2011

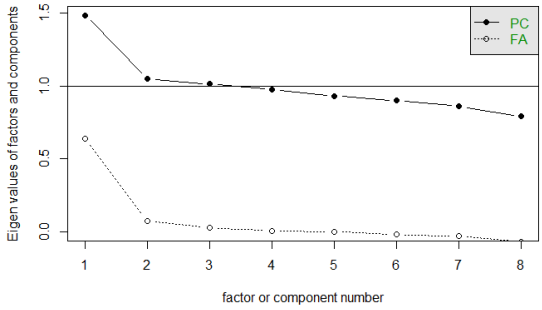
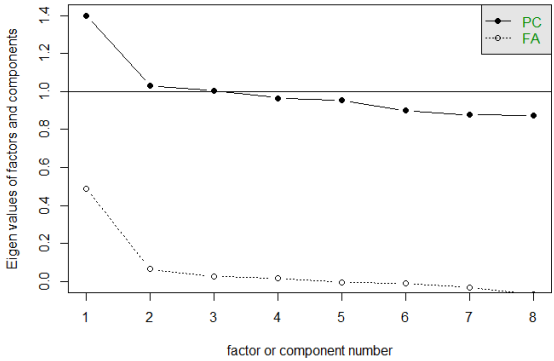


Figura 16 Scree plot das questões objetivas da FG de 2012



Apenas o primeiro fator se destaca em cada um dos s*cree plots*, o que não permite inferir a presença de mais que uma dimensão em cada ano pelo critério do "braço". Ressalta-se que um autovalor (*eingen value*) de 1 corresponde à variância de uma única variável sendo assim, em função de cada prova ter apenas oito itens objetivos, é plausível assumir que existam outros fatores nessas provas, que não foram esclarecidos pela ausência de variáveis (itens) suficientes para evidenciá-los.

Na Tabela 6 são apresentados os índices de ajuste do modelo. Ressalta-se que o programa indicou adequação de todos os modelos para uma estrutura de um fator.

Tabela 6 Índices de ajuste da Análise Fatorial

	2010	2011	2012
Qui-quadrado[16]	p < 0	p < 0	p < 0
Gl[17]	20	20	20
TLI (índice de Tucker Lewis):[18]	0,942	0,921	0,888
RMSEA[19]	0,036	0,032	0,029
SRMR[20]	0,03	0,03	0,03

Esses índices indicam uma relativa adequação do modelo aos dados, tendo apenas o TLI indicado um ajuste insatisfatório, considerando os índices mais rigorosos apontados por HU & BENTLER (1999).

4.3 Testes de Fidedignidade dos itens objetivos

Nessa seção estão comparados os índices de consistência dos itens objetivos. Esses índices servem para avaliar o quanto os itens de um teste se relacionam entre si, e, indiretamente, o quanto o desempenho em cada item se relaciona com o desempenho global nos itens. Na Tabela 7, estão exibidos o Alfa de Cronbach antes e depois da remoção dos ausentes nas questões discursivas para a prova de cada ano.

[16] GATIGNON (2010) esclarece que dado um elevado numero de respondentes é impossível rejeitar esse teste.

[17] Os graus de liberdade não dizem respeito diretamente ao ajuste da análise fatorial, mas são utilizados no computo de outros índices e refletem a complexidade do modelo.

[18] HU & BENTLER (1999) recomendam índices ≥ 0,95 para o TLI, mas ressaltam que anteriormente era aceito ≥ 0,90.

[19] HU & BENTLER (1999) recomendam um RMSEA abaixo de 0,06 para um ajuste aceitável ao modelo.

[20] HU & BENTLER (1999) recomendam um SRMR abaixo de 0,08 para um ajuste aceitável ao modelo.

Tabela 7 Alfa de Cronbach da escala dos itens objetivos da Formação Geral de 2010, 2011 e 2012 antes e depois da remoção dos ausentes nas questões discursivas.

Ano	Alfa de Cronbach	
	Sem remoção de ausentes	Com remoção de ausentes
2010	0,402	0,390
2011	0,317	0,286
2012	0,300	0,277

Conforme a Tabela 1 Interpretação do Alfa de Cronbach se verifica que a prova de nenhum dos anos atingiu um valor de alfa satisfatório para um exame desse tipo. Apesar da variação no índice após o procedimento de limpeza, não há alteração no sentido da interpretação nos dois casos. Os conjuntos de itens não fornecem uma medida confiável.

Para permitir a análise integral da prova por meio da correlação ponto bisserial, estão exibidas na Tabela 8 as médias de correlação ponto bisserial antes e depois da remoção dos ausentes nas questões discursivas, realizando-se o cálculo com e sem a inclusão do desempenho de cada item avaliado no cálculo do escore a partir do qual se faz a correlação. No Apêndice 3 – Correlação ponto bisserial de cada item estão especificadas as correlações corrigidas de cada item com relação aos alunos que responderam todos os itens objetivos e discursivos.

Tabela 8 Média da correlação ponto bisserial da escala dos itens objetivos da Formação Geral de 2010, 2011 e 2012 com e sem correção antes e depois da remoção dos ausentes nas questões discursivas.

Ano	Sem remoção de ausentes		Com remoção de ausentes	
	Sem correção	Com correção	Sem correção	Com correção
2010	0,590	0,230	0,596	0,227
2011	0,542	0,166	0,534	0,148
2012	0,519	0,148	0,512	0,135

É importante ressaltar que o INEP desconsidera do computo das notas de uma área todas as questões cuja correlação ponto-bisserial é

inferior a 0,20. Isso significa que caso esse índice fosse calculado com a correção para testes com poucos itens apontada por PRICE (2016), seriam considerados válidos apenas quatro, um e zero itens, respectivamente nas provas de Formação Geral de 2010, 2011 e 2012, conforme se observa no Apêndice 3.

4.3. Modelos da TRI

4.3.1 Comparação de ML3 e de ML4

Nesta seção são comparados o ML3 e o ML4 considerando apenas os itens objetivos. A tabela a seguir contém os índices conhecidos como AIC (*Akaike Information Criterion*) e BIC (*Bayesian Information Criterion*) [21], modelos com índices menores são preferíveis.

Tabela 9 Comparação da complexidade de ML3 e ML4 nos ENADE de 2010, 2011 e 2012.

	AIC		BIC	
Ano	ML3	ML4	ML3	ML4
2010	889605	889599	889835	889904
2011	2205705	2204582	2205953	2204913
2012	3750053	3750034	3750311	3750378

Verifica-se que o ML4 possui um AIC menor em todos os anos e um BIC maior nos anos de 2010 e 2012. A redução do AIC indica que o ML4 se ajusta melhor aos dados. Já o aumento do BIC era esperado, uma vez que o modelo possui um parâmetro adicional em cada variável e aumentos na complexidade de modelos são acompanhados de aumento desse índice (BURNHAM &ANDERSON, 2004).

A Tabela 10 contém o percentual de variância explicada por cada modelo considerando apenas os itens objetivos.

[21] Consultar BURNHAM &ANDERSON (2004) para mais informações.

Tabela 10 Comparação de percentual de variância explicada pelo ML3 e pelo ML4 nos ENADE de 2010, 2011 e 2012.

	Variância explicada	
Ano	ML3	ML4
2010	35,3	38,7
2011	15,5	51,4
2012	10,3	14,8

Verifica-se que com a inclusão do parâmetro d, a diferença entre os ML3 e ML4, aumentou a variância explicada em todos os anos. Observa-se também um comportamento anômalo no ano de 2011. Esse fenômeno pode ser compreendido por meio da Tabela 11, que explicita a carga fatorial de cada questão e sua variância explicada.

Tabela 11 Comparação da análise fatorial do modelo logístico de três e quatro parâmetros dos itens objetivos da prova de Formação Geral do ENADE 2011.

	ML3		ML4	
Item	$F1$	$h2$	$F1$	$h2$
Q1	0,495	0,245	0,507	0,257
Q2	0,560	0,314	0,739	0,546
Q3	0,664	0,441	0,849	0,721
Q4	-0,058	0,003	-0,930	0,864
Q5	0,390	0,152	0,405	0,164
Q6	0,180	0,032	0,932	0,870
Q7	0,221	0,049	0,848	0,719
Q8	0,046	0,002	0,055	0,003

Evidentemente a Q4 não se ajustou ao modelo de forma coerente, pois violaria o segundo axioma da TRI[22] uma vez que sua carga é negativa. A inclusão dessa questão no modelo não é apropriada.

As tabelas Tabela 12, Tabela 13 e Tabela 14 contêm os parâmetros dos itens calibrados comparativamente nos modelos de três e quatro parâmetros, nos anos de 2010, 2011 e 2012.

[22] É necessário que a relação entre o desempenho e a habilidade seja representada por uma curva monotônica crescente.

Tabela 12 Parâmetros dos itens objetivos no modelo logístico de três fatores do ENADE 2010

	ML3			ML4			
Item	a	b	c	a	b	c	d
Q1	0,40	0,97	0,07	0,67	-0,28	0,14	0,72
Q2	0,32	4,06	0,12	0,31	3,20	0,10	0,88
Q3	1,58	-0,80	0,23	1,39	-1,11	0,08	1,00
Q4	0,86	-0,97	0,05	1,25	-0,72	0,20	0,93
Q5	1,73	0,28	0,32	1,81	0,29	0,32	1,00
Q6	3,38	1,81	0,07	4,05	1,75	0,07	0,98
Q7	0,85	4,34	0,10	0,82	4,36	0,10	0,98
Q8	1,96	-1,00	0,13	2,13	-0,87	0,22	1,00

Verifica-se o aumento do parâmetro a em cinco dos oito itens (Q1, Q4, Q5, Q6, Q8), sinalizando que a inclusão de d permite ajuste melhor entre p (probabilidade de acerto) e θ(quantidade de traço latente do respondente). Observa-se também a expressiva diminuição de b em Q1 e Q2, com melhora significativa do parâmetro a para Q1 comparando ambos os modelos, o que pode ser atribuído à melhoria do ajuste do modelo aos dados, quando se considera o desempenho dos estudantes de altas habilidades, isso é, no ML3, tem-se um problema de convergência quando o aumento de θ não é acompanhado de um aumento de p.

Tabela 13 Parâmetros dos itens objetivos no modelo logístico de três fatores do ENADE 2011

	ML3			ML4			
Item	a	b	c	a	b	c	d
Q1	0,97	-0,71	0,06	1,00	-0,73	0,07	0,99
Q2	1,15	0,91	0,10	1,80	0,76	0,16	0,90
Q3	1,50	-0,75	0,06	2,78	-0,50	0,25	0,95
Q4	-0,10	-8,10	0,32	-4,17	1,83	0,11	0,55
Q5	0,72	-0,55	0,05	0,74	-0,76	0,02	0,97
Q6	0,31	-3,50	0,04	4,34	-1,82	0,07	0,79
Q7	0,39	3,02	0,08	2,71	0,17	0,23	0,41
Q8	0,08	21,90	0,04	0,09	12,95	0,06	0,59

Em 2011 se observa que o parâmetro a é superior no ML4 quando comparado ao ML3 em todos os itens, sinalizando que o melhor ajuste daquele modelo. Além disso, novamente observam-se problemas em relação a Q4 – há claro problema convergência desse item com

relação aos demais – seu parâmetro a é negativo – o que implica na violação do axioma da relação entre p e θ ser monotônica crescente. Evidentemente esse problema se deve a forte carga negativa desse item em relação ao fator predominante no conjunto de itens. Apesar da expressiva diminuição de b no Q8, este ainda não possui um bom ajuste no ML4, o que evidencia seu relacionamento fraco com os demais itens.

Tabela 14 Parâmetros dos itens objetivos no modelo logístico de três fatores do ENADE 2012

	ML3			ML4			
Item	a	b	c	a	b	c	d
Q1	0,75	-0,31	0,10	0,85	-0,37	0,13	0,95
Q2	0,81	0,24	0,04	0,97	-0,04	0,06	0,89
Q3	0,51	0,99	0,09	0,65	0,16	0,10	0,81
Q4	0,50	0,11	0,06	0,83	-0,89	0,08	0,75
Q5	-0,03	-35,07	0,06	-0,02	-38,50	0,05	0,96
Q6	0,55	-0,08	0,08	0,69	-0,30	0,13	0,90
Q7	0,89	0,04	0,07	1,05	0,01	0,12	0,94
Q8	0,16	4,32	0,08	0,16	3,43	0,06	0,96

Da forma semelhante aos outros anos, constatou-se que em 2012, o ML4 teve um ajuste melhor que o ML3 nos seis itens que tinham um ajuste aceitável no ML3, como se observa no aumento de a. Para os demais itens, Q5 e Q8, o ML4 não forneceu ajuste significativamente diferente. Destaca-se que, Q5 teve o parâmetro a é negativo em ambos os modelos, indicando seu desajuste a um modelo cuja unidimensionalidade é pressuposto, de forma semelhante observa-se que a correlação ponto bisserial desse item foi -(0,03), como se observa no Apêndice 3. Já Q8, que teve correlação ponto bisserial de (0,04), manteve seu parâmetro a praticamente inalterado em ambos os modelos.

De forma geral verificou-se um bom ajuste do ML4 aos dados, sendo inclusive mais apropriado que o ML3, dado os melhores parâmetros observados e aumento da variância explicada. Desta forma fica justificada a utilização desse modelo no modelo de itens de formato misto.

4.3.2 Modelo de itens de formato misto

Nesta sessão estão dispostos e analisados os modelos mistos com as oito questões objetivas e as duas discursivas para cada edição do

ENADE analisada. Para o modelo de cada ano será apresentada uma tabela com os parâmetros de cada item, sendo utilizado o ML4 para as questões discursivas e o modelo de resposta gradual para os itens discursivos. Além dos parâmetros, serão exibidas as curvas características e de informação de cada item e a curva de informação total do teste com se respectivo erro de estimação em cada nível de habilidade para permitir uma análise mais aprofundada da escala construída.

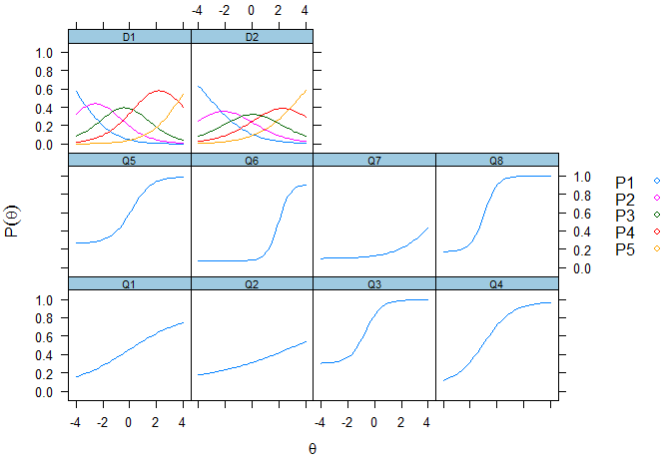
Na Tabela 15 estão dispostos os parâmetros dos itens relativos à escala do ENADE 2010:

Tabela 15 Parâmetros do Modelo de itens de formato misto do ENADE 2010

	Parâmetros							
Item	a	b	c	d	b_1	b_2	b_3	b_4
Q1	0,46	-0,11	0,04	0,85	-	-	-	-
Q2	0,29	3,81	0,10	0,96	-	-	-	-
Q3	1,71	-0,68	0,30	0,99	-	-	-	-
Q4	0,93	-0,99	0,07	0,97	-	-	-	-
Q5	1,35	0,13	0,26	0,99	-	-	-	-
Q6	2,35	1,95	0,07	0,90	-	-	-	-
Q7	0,72	4,68	0,10	0,99	-	-	-	-
Q8	2,03	-0,95	0,17	1,00	-	-	-	-
D1	0,93	-	-	-	-3,66	-1,39	0,62	3,80
D2	0,67	-	-	-	-0,32	-0,97	1,04	3,50

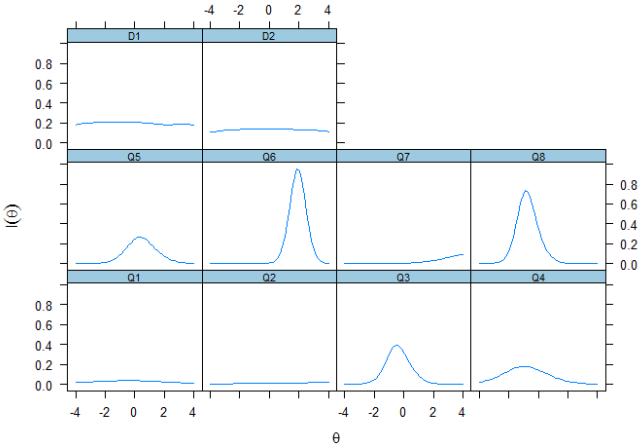
Verifica-se que Q1 e Q2 tem baixa discriminação (a) nesse modelo, o que torna recomendável sua remoção do cálculo da nota. Já o item Q7 apresenta um problema na estimação de sua dificuldade (b), o valor de 4,68 é irrealista, considerando uma população normalmente distribuída.

Figura 17 CCI do modelo de itens de formato misto do ENADE 2010



É evidente pela Figura 17 os parâmetros de Q1, Q2 e Q7 não são tão bons quanto aqueles da demais questões, como já tinha sido verificado na Tabela 15. Além disso, essa figura evidencia um bom ajuste dos itens D1 e D2 ao modelo, com uma clara distinção dos níveis de desempenho.

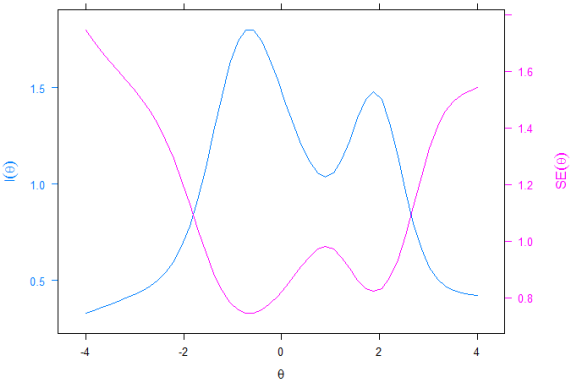
Figura 18 CII do modelo de itens de formato misto do ENADE 2010



É importante observar que as questões D1 e D2 fornecem consistentemente informação ao longo de toda a faixa de desempenho situada entre -4 e +4, o que as torna bastantes úteis para diferenciar

níveis diversos de desempenho. É evidente que D2 é um item com menor utilidade que D1 para discriminar diferentes faixas de desempenho, o que é refletido em seu parâmetro (a) menor, e consequentemente sua menor quantidade de informação. Já as questões Q1, Q2 e Q7, conforme já demonstrado, fornecem pouca informação.

Figura 19 Curva de Informação Total do Teste e Erro de Estimativa ENADE 2010



Percebe-se que o erro de estimativa é relativamente elevado na faixa de (-2) a (+2), situando-se entre 0,8 e 1,0 unidades isso implica em dificuldade de diferenciar de forma precisa o desempenho de pessoas com habilidade em níveis relativamente comuns. Além disso, se observa menor informação na região próxima a (+1) unidade do traço medido, o que significa à inclusão de questões desse nível de dificuldade no instrumento melhoria a confiabilidade de sua medida. De forma geral o recomendável é a inclusão de mais itens para a diminuição do erro de medida nas faixas de interesse.

Como o propósito desse estudo é analisar o instrumento aplicado não foi realizada a remoção de itens com ajuste ruim (Q1, Q2 e Q7) ao modelo proposto para as questões de formação geral do ENADE 2010. A remoção de itens com ajuste ruim tornaria a escala mais confiável sob o ponto de vista da medida.

Seguem análises relativas ao ENADE 2011

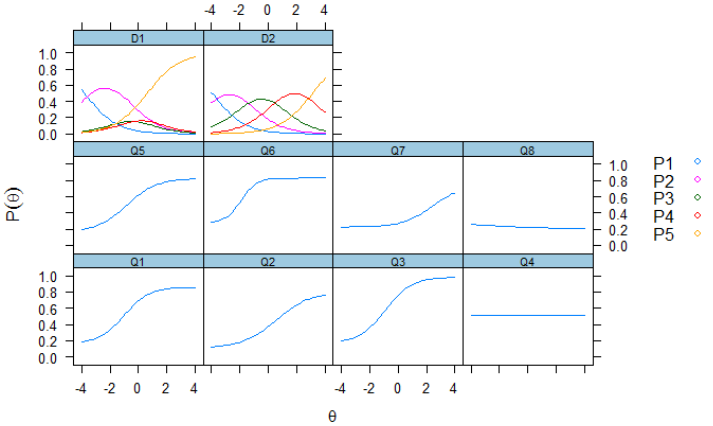
Tabela 16 Parâmetros do Modelo de itens de formato misto do ENADE 2011

	Parâmetros 4PL							
Item	a	b	c	d	b_1	b_2	b_3	b_4
Q1	1,14	-1,02	0,16	0,86	-	-	-	-
Q2	0,84	0,56	0,11	0,80	-	-	-	-
Q3	1,09	-0,89	0,17	0,98	-	-	-	-
Q4	0,97	0,04	0,52	0,52	-	-	-	-
Q5	0,95	-0,85	0,16	0,83	-	-	-	-
Q6	1,87	-1,91	0,28	0,83	-	-	-	-
Q7	0,99	2,38	0,23	0,73	-	-	-	-
Q8	-0,24	-11,4	0,20	0,64	-	-	-	-
D1	0,87	-	-	-	-3,80	-0,56	-0,15	0,60
D2	0,87	-	-	-	-3,94	-1,52	0,59	3,10

Observa-se que dois itens têm parâmetros problemáticos nesse modelo. O item Q4 possui parâmetros (c) e (d) muito próximos, o que, nesse caso, significa que o desempenho no item não tem relação com o nível de (θ). Já o item Q8 possui um valor baixo e negativo de (a) o que significa que ele viola o axioma da monotocidade e é uma questão que não deve ser incluída na mesma escala que as demais. Em relação às questões discursivas, observação um bom índice de discriminação com valores razoáveis de (a), porém D1 possui parâmetros b_2 e b_3 próximos o que torna à distinção desses níveis pouco relevante.

A curva característica desses itens está ilustrada na Figura 20.

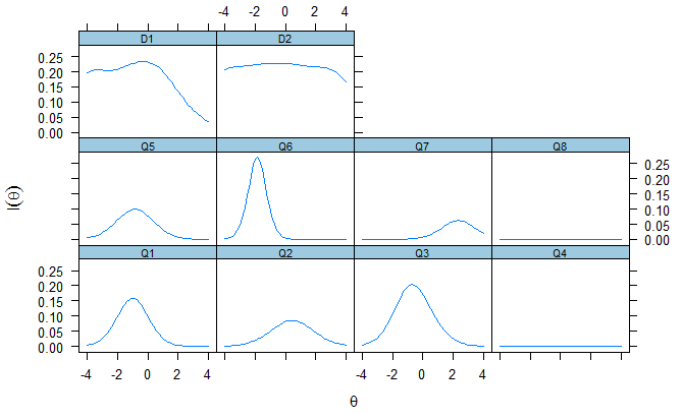
Figura 20 CCI do modelo de itens de formato misto do ENADE 2011



A Figura 20 evidencia os problemas dos parâmetros dos Q4, Q8 e D1 já observados na análise numérica. Observa-se que D1 ainda pode ser considerado um bom item apesar do problema observado em dois de seus níveis de dificuldade.

A curva de informação desses itens está ilustrada na Figura 21.

Figura 21 CII do modelo de itens de formato misto do ENADE 2011

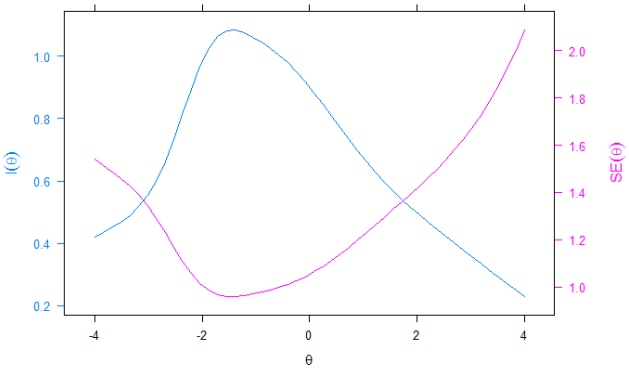


Verifica-se novamente que Q4 e Q8 não são úteis para medir o traço latente subjacente ao restante dos itens como já tinha sido observado na análise dos parâmetros estimados. Em particular Q4 evidencia a influência de diferença entre os parâmetros (c) e (d) no cálculo da quantidade de informação, como essa diferença é mínima a

informação é drasticamente reduzida. Além disso, a consequência da interposição entre alguns dos níveis desempenho em D1 tem como consequência a redução da informação fornecida para respondentes com habilidade próximas a essas faixas.

A Figura 22 contém a Curva de Informação Total do Teste e Erro de Estimativa ENADE 2011.

Figura 22 Curva de Informação Total do Teste e Erro de Estimativa ENADE 2011



Em comparação com 2010, observa-se por meio da linha vermelha que o erro de medida é maior na escala de 2011. O gráfico evidencia a necessidade mais questões na faixa entre (0) e (+2) e, de forma geral a necessidade de incluir mais questões por todas as faixas de desempenho. A inclusão de algumas questões discursivas com capacidade discriminação ao longo de uma faixa ampla de desempenho seria uma medida recomendável nesse caso.

A Tabela 17 contém os parâmetros de cada item relativos ao modelo de itens de formato misto do ENADE 2012.

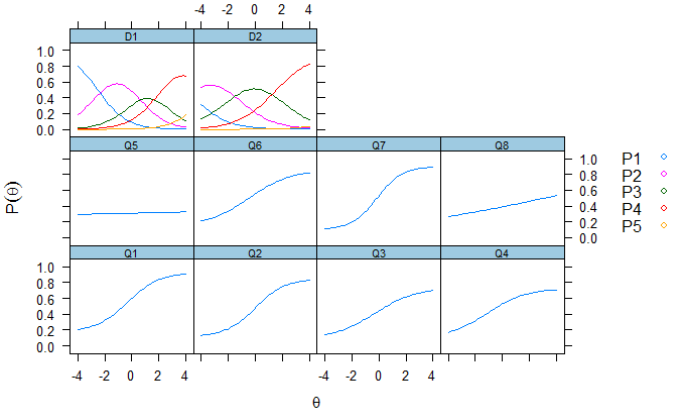
Tabela 17 Parâmetros do Modelo de itens de formato misto do ENADE 2012

	Parâmetros 4PL							
Item	a	b	c	d	b_1	b_2	b_3	b_4
Q1	0,86	-0,31	0,18	0,93	-	-	-	-
Q2	0,93	0,00	0,12	0,85	-	-	-	-
Q3	0,65	-0,29	0,08	0,74	-	-	-	-
Q4	0,71	-1,08	0,10	0,73	-	-	-	-
Q5	0,02	44,34	0,06	0,96	-	-	-	-
Q6	0,66	-0,46	0,15	0,86	-	-	-	-
Q7	1,09	-0,15	0,10	0,90	-	-	-	-
Q8	0,18	3,03	0,08	0,91	-	-	-	-
D1	0,94	-	-	-	-2,52	0,30	2,06	5,61
D2	0,73	-	-	-	-5,04	-1,59	1,51	8,40

No ENADE 2012 observaram-se valores médios de (a) menores quando comparado aos demais anos, o que significa que cada item individualmente tem menor poder de discriminação entre níveis distintos de habilidade. Contudo, apesar desse padrão, apenas dois itens objetivos tiveram parâmetros inaceitáveis (Q5 e Q8) e o último nível de desempenho dos dois itens discursivos. Destaca-se que os parâmetros (b) dos itens discursivos estão bastante espaçados.

A curva característica desses itens está ilustrada na Figura 23.

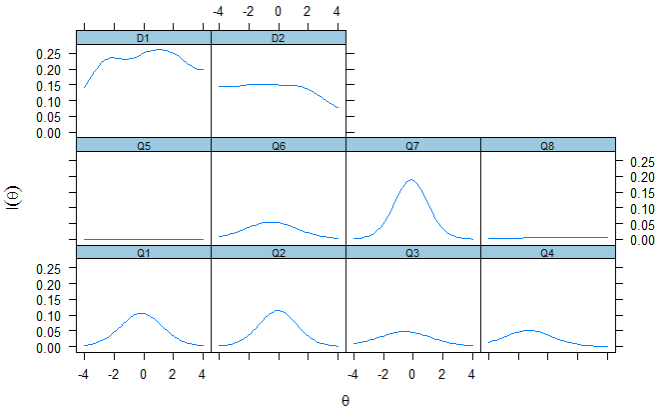
Figura 23 CCI do modelo de itens de formato misto do ENADE 2012



É visível que em 2012, os itens Q5 e Q8 não possuem muito poder discriminativo, como se observa pela pequena inclinação de suas curvas. Além disso, observa-se que o nível mais alto de desempenho de

D1 e D2 não é visível no gráfico, o que provavelmente se deve a influência de Q5 no modelo. A utilidade de cada item no instrumento é mais facilmente visualizada nas curvas de informação do item individuais, exibidas na figura a seguir.

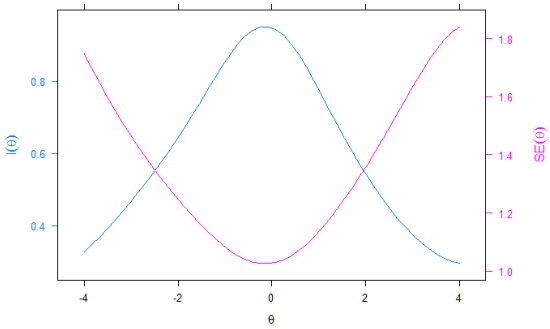
Figura 24 CII do modelo de itens de formato misto do ENADE 2012



Como visto, a quantidade de informação é uma função da discriminação, dificuldade e dos parâmetros c e d. Dito isso, nesse modelo as questões abertas forneceram substancial contribuição para a mensuração do traço latente, em função dos péssimos parâmetros das questões objetivas.

Por fim, a figura a seguir mostra o efeito cumulativo de cada curva de informação, percebe-se que a quantidade de informação é contrária ao erro.

Figura 25 Curva de Informação Total do Teste e Erro de Estimativa ENADE 2012



Em função da capacidade discriminativa dos itens desse ano com relação a um traça latente observa-se que o menor erro de estimativa é da ordem de uma unidade. Isso é reflexo da baixa variância explicada pelo conjunto de itens utilizados nesse ano.

De forma geral a utilização de uma nota baseada na quantidade de acertos não é confiável sob a luz da TCT, conforme observado pelo Alfa de Cronbach e a técnica da Correlação Bisserial com o ajuste para poucas variáveis. Apesar disso, sob a ótica da análise fatorial é possível considerar-se um modelo unidimensional, com pouca variância explicada. Estabelecido isso, a Teoria de Resposta ao Item permitiu o emprego de modelos que permitem identificar os itens com que não se ajustam bem a um modelo unidimensional, com ganho extra de variância explica utilizando-se o ML4, comparado ao ML3. Além disso, a utilização de um modelo de item de formato misto possibilitou o aumento na quantidade de informação da escala, um melhor ajuste dos demais itens e a comparação direta do desempenho nas questões objetivas e discursivas.

5 Considerações e conclusões

5.1 Contribuições do estudo

Conforme exposto na revisão de literatura, a maioria dos estudos quantitativos sobre o ENADE não se aprofunda na análise sobre o componente da Formação Geral e sim no componente especifico, assim esse trabalho promoveu uma reflexão sobre a medida, que se obtém na aplicação desse componente.

Por meio da análise psicométrica clássica identificou-se que as questões da formação geral não possuem muito poder discriminativo. Uma possível razão para a baixa capacidade discriminativa de um teste é a quantidade insuficiente de questões, como se observa no termo $(K/K-1)$ da fórmula do Alfa de Cronbach. Dito de outra forma, uma quantidade pequena de questões não fornece uma medida confiável sobre o que se pretende medir, mesmo que esses itens sejam bastante correlacionados entre si. Além dos índices medíocres, entre 0,28 e 0,40 do alfa, a correlação ponto bisserial evidencia a baixa consistência dos itens, realizada correção do seu cômputo.

Os modelos da TRI conseguem desenvolver uma medida melhor ao reduzir o efeito dos itens mais problemáticos, mas percebeu-se que todos os modelos produzidos na seção 4, tiveram pelo menos 2 itens inadequados[23].

De um ponto de vista estrito, conforme apontado por BOCK, GIBBONNS & MURAKI (1988), caso se deseje resumir o desempenho de um conjunto de itens em uma nota simples, esses itens devem se referir a uma única dimensão, assim é inadequado produzir uma nota de Formação Geral se os itens dessa prova não são claramente relativos à mesma dimensão. Assim, o procedimento recomendável é a identificação e produção de medidas separadas para cada dimensão de forma mais confiável e independente entre si.

Feita a mensuração das diversas dimensões, é possível resumir essas medidas em um indicador de Formação Geral, calculado por meio da manipulação numérica dos escores de cada dimensão. Ressalta-se que indicadores não possuem propriedades psicométricas, não fazendo sentido falar-se em consistência interna ou dimensionalidade do indicador. Todavia, a teoria da medida deve ser aplicada rigorosamente

[23] Parâmetro a abaixo de 0,70 ou $|b|$ maior que 3. No caso de itens no ML4, também podem ser considerados inadequados itens com c e d muito próximos.

para cada dimensão identificada, e as propriedades psicométricas de cada uma dessas deve ser individualmente estabelecida independentemente.

PASQUALI (1996, p.79) explica que responder um item é um comportamento multimotivado, como qualquer outro comportamento humano, daí, a rigor é impossível falar em unidimensionalidade estrita. Todavia a análise da covariação entre o item e o traço latente que se deseja medir é possível. Essa covariação é expressa pela carga fatorial do item, o que significa que itens com pouca carga são representações inadequadas do traço e devem ser removidos do teste. Assim a exclusão de itens tomando como critério a análise fatorial é mais razoável que baseá-la no índice de correlação ponto bisserial, que esse pressupõe tanto a unidimensionalidade, quanto a adequação dos demais itens do instrumento.

É importante ressaltar que as técnicas de análise de fidedignidade não diferenciam problemas de elaboração de um item de problemas de dimensionalidade das competências avaliadas pelo conjunto de itens. Assim é imprescindível avaliar a dimensionalidade da Formação Geral em um instrumento composto de mais itens e evitar problemas de elaboração como o que ocorreu com o item 4 de 2011, cuja carga negativa pode ser resultado de problema de elaboração desse item do tipo asserção-razão.

Um modelo para itens de formatos mistos da Teoria de Resposta ao Item possibilita colocar os itens discursivos na mesma escala dos itens objetivos, como se observou na seção 4.3.2. Essa integração é vantajosa, pois aumenta a quantidade de itens no instrumento e resolve o problema da atribuição de peso as questões. Além disso, itens discursivos fornecerem informação em diversas faixas de habilidade diferentemente das questões dicotômicas.

Assim, recomenda-se ajustar a correção dos itens discursivos, para ser pautada em níveis de competência, para aumentar a aderência ao objetivo de se avaliar as competências dos egressos dos cursos de graduação, estabelecido pela lei 10.861. Assim, com o intuito de melhorar a precisão da medida de FG, é recomendável a utilização de um modelo da teoria de resposta ao item que integre o desempenho das questões objetivas e discursivas na mesma escala como o modelo de itens de formato misto proposto.

5.2 Perspectivas para futuros estudos

Ao revisar a literatura sobre o ML4 observou-se a recomendação de se esclarecer a interpretação dos parâmetros c e d e o seu processo de calibração, especialmente quando se consideram que vários estudos fixaram um ou ambos os parâmetros, em vez de calibrá-los. Nesse sentido, RECKEASE (2009, p. 31) explica que apesar das descrições iniciais da Teoria de Resposta ao Item chamarem o parâmetro c de parâmetro do "chute" e que seu valor fosse próximo de $1/\mathrm{m}$, onde m seria a quantidade de alternativas, ele representa apenas a probabilidade de acerto das pessoas cuja habilidade é tão baixa em relação ao item, que sua redução não implica em alteração de p. A mesma crítica precisa ser feita para a interpretação do parâmetro d. Será que ele é relativo apenas aos erros de desatenção, como foi proposto por BARTON & LORD (1981)?

A comparação dos ML3 e ML4 revelou uma aplicação para este último. Ele forneceu melhor ajuste para alguns itens de baixa discriminação. É evidente que o parâmetro d pode ser interpretado como limite da influência de θ na probabilidade de acerto de um item. Essa interpretação é relevante com relação o problema apontado por LOKEN & RULISON (2010) sobre esse parâmetro não ser uma propriedade do item. Faz-se pertinente a realização de estudos adicionais com itens em múltiplas dimensões bem esclarecidas para verificar a pertinência dessa interpretação.

Além disso, como o ENADE é um teste de baixo impacto para os estudantes (*low stakes*), é possível que o parâmetro d esteja refletindo a falta de empenho consistente dos estudantes de altas habilidades em maximizar seu desempenho. Recomenda-se assim a aplicação do modelo a conjuntos de dados referentes à mensuração de habilidades semelhantes nos quais haja alto impacto (*high stakes*) para o respondente de seu desempenho.

Outra questão que merece investigação é a pequena diferença observada entre os ingressantes e concluintes dos diversos cursos na Formação Geral dos ciclos anteriores, conforme observado em GURGEL (2010), SILVA, VENDRAMINI E LOPES (2010), GONTIJO *et al* (2011) e LANZILLOTTI & LANZILLOTTI (2014), chegando inclusive a observar desempenho superior por parte dos ingressantes. Como é possível que não exista ganho de "Formação Geral" durante o curso se essa prova deveria evidenciar:

> “‘‘... a compreensão de temas que transcendam ao seu ambiente próprio de formação e sejam importantes para a realidade contemporânea. Essa compreensão vincula-se a perspectivas críticas, integradoras e à construção de sínteses contextualizadas.
> (MINISTÉRIO DA EDUCAÇÃO, 2010) (MINISTÉRIO DA EDUCAÇÃO, 2011) e (MINISTÉRIO DA EDUCAÇÃO, 2012).

Há de se questionar se existe realmente uma FG decorrente da participação na educação superior? Ou será que o problema é os objetivos desse componente não se aproximam mais de um conceito de cidadania mais pertinente ao Ensino Médio como apontam CORTELAZZO & RIBEIRO (2013)? Nesse sentido VERHINE; DANTAS & SOARES (2006, p. 300) destacam:

> “A partir da leitura dos relatórios e do Resumo Técnico (BRASIL, 2005), não é possível identificar uma lógica coerente embasando o componente de Formação Geral. Registra-se ainda que não houve relato de estudos que considerassem se o número de questões foi realmente adequado. Como será visto adiante, a inadequação do componente de formação geral do teste é evidenciada pelo fato de que os alunos ingressantes e concluintes atingiram, em média, escores muito próximos.”

Ou seja, a validade da prova de formação geral ainda está por ser demonstrada. Viu-se na seção 4.2 que apesar de ser possível se identificar apenas uma dimensão nas provas, ela explica pouca variância e não se destaca muito sobre os demais fatores. Ou seja, é possível que existam dimensões subjacentes a FG que não estão esclarecidas em função da baixa quantidade de itens no instrumento. Assim é recomendável a construção e aplicação de uma prova de Formação Geral com mais itens para avaliar a quantidade de dimensões apropriadas.

Como visto em KLEIN & FONTANIVE (2009), a avaliação do desempenho nas questões discursivas do ENEM teve uma alteração no procedimento de correção para pautá-las em competências. É possível aplicar essa metodologia ao ENADE que já possui as competências da FG descritas em suas portarias de prova. Ademais, como visto em

KLEIN & FONTANIVE (2009) e BRAGA (2015) é possível pautar a correção dessas questões em níveis de desempenho, categorias de desempenho menos arbitrárias que aquelas utilizadas nesse estudo que foram baseadas na interpretação dos escores.

Por fim, as sintaxes e roteiros disponibilizadas integralmente nos Apêndices 1 e 2, possibilitam a realização das análises desse estudo nos dados de outros ciclos do ENADE para verificar se conclusões similares podem ser levantadas sobre as características desse exame.

REFERÊNCIAS

ANDRADE, D. F.; TAVARES, H. R. e VALLE, R. C. **Teoria da Resposta ao Item: Conceitos e Aplicações.** In: SINAPE 2000, São Paulo, 2000, ABE-Associação Brasileira de Estatística. p. 01-164.

ANDRADE, J. M., LAROS, J. A. e GOUVEIA, V. V. (2010). O uso da teoria de resposta ao item em avaliações educacionais: diretrizes para pesquisadores. **Avaliação Psicológica**, v. 9(3), pp. 421-435

ANDRIOLA, W. B. A. Psicometria Moderna: características e tendências. **Est. Aval. Educ**., São Paulo, v. 20, n. 43, maio/ago. 2009.

BALANDA, K. P.; MACGILLIVRAY, H.L. (1988). Kurtosis: A Critical Review. **The American Statistician.** V. 42 (2): pp. 111–119.

BARTON, M. A., & LORD, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. Princeton, NJ: **Educational Testing Service**.

BOCK, R. D., GIBBONS., R. & MURAKI, E. (1988) Full-information item factor analysis. **Applied Psychological Measurement**, v. 12 n. 3, set 1988, pp. 261-280.

BRAGA, B. M. A. **Teoria da resposta ao item: o uso do modelo de Samejima como proposta de correção para itens discursivos.** 2015. vii, 60 f., il. Dissertação (Mestrado Profissional em Matemática)—Universidade de Brasília, Brasília, 2015.

BRASIL. **Lei 10.861, de 14/4/2004**. Institui o Sistema Nacional de Avaliação da Educação Superior – SINAES e dá outras Providências. Diário Oficial da União, Brasília, DF, 15 abr. 2004. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm>

_____. Decreto nº 3.860, de 9 de julho de 2001. Dispõe sobre a organização do ensino superior, a avaliação de cursos e instituições e dá outras providências. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 10 jul. 2001.

BRASIL. **Lei de Diretrizes e Bases da Educação Nacional**. Lei número 9394, 20 de dezembro de 1996.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Cálculo do Conceito Preliminar de Curso (CPC) referente ao ano de 2011**. NOTA TÉCNICA Nº 029 DE 15 DE OUTUBRO DE 2012.

BRASIL. Ministério da Educação. **PORTARIA NORMATIVA Nº 40, DE 12 DE DEZEMBRO DE 2007**, do Ministério da Educação.

_____. Lei nº 9.131, de 24 de novembro de 1995. Altera dispositivos da Lei nº 4.024, de 20 de dezembro de 1961 e dá outras providências. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 24 nov. 1995. Edição Extra, p. 19257. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/L9131.htm>. Acesso em: 24 set. 2006.

BURNHAM, K. P.; ANDERSON, D. R. (2004). Multimodel inference: understanding AIC and BIC in Model Selection. Sociological Methods & Research, v. 33, n. 2, p. 261–304.

CHALMERS, R. P. (2012). **mirt**: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29. Disponível em: <http://www.jstatsoft.org/v48/i06/>.

CHENG Y. & LIU C. The Effect of Upper and Lower Asymptotes of IRT Models on Computerized Adaptive Testing. **Applied Psychological Measurement.** v. 39(7) p. 551–565, 2015.

CHIODI, M. G, & WECHSLER, S. M. Avaliação psicológica: contribuições brasileiras. **Bol. - Acad. Paul. Psicol**., São Paulo , v. 28, n. 2, p. 197-210, dez. 2008 . Disponível em <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1415-711X2008000200008&lng=pt&nrm=iso>. Acesso em: 20 nov. 2016.

CORTELAZZO, A. L.; RIBEIRO, V. K. ENADE 2005 e 2008: desempenho dos estudantes de biologia de instituições de Educação Superior estaduais e municipais de São Paulo. **Ciênc. educ**. (Bauru), Bauru , v. 19, n. 2, 2013 . Disponível em

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-73132013000200012&lng=pt&nrm=iso . Acessado em 26 maio 2014.

DAMASIO, B. F. Uso da análise fatorial exploratória em psicologia. **Aval. psicol.,** Itatiba , v. 11, n. 2, p. 213-228, ago. 2012 . Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712012000200007&lng=pt&nrm=iso>. Acesso em 01 fev. 2017.

ERCIKAN, K.; SCHWARZ, R. D.; JULIAN, M. W.; BURKET G. R. WEBER, M. M. & LINK, V. Calibration and Scoring of Tests With Multiple-Choice and Constructed-Response Item Types. **Journal of Educational Measurement.** V. 35 n. 2, p. 137–154, jun. 1998.

GAUDIO, A. P. S. **O PROUNI como política de inclusão social: uma avaliação por meio do ENADE**. Universidade Católica de Brasília. Brasília. 2014

GATIGNON, H. (2010). **Statistical Analysis of Management Data.** Springer, Nova Iorque, NY, 2a ed.

GONTIJO, E. D. *et al* . Cursos de graduação em medicina: uma análise a partir do sinaes. **Rev. bras. educ. med**., Rio de Janeiro , v. 35, n. 2, jun. 2011 . Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-55022011000200010&lng=pt&nrm=iso. Acessado em 26 maio 2014.

GURGEL, C. R.. Análise do Exame Nacional de Desempenho dos Estudantes após o primeiro ciclo avaliativo das áreas de agrárias, saúde e serviço social do Estado do Piauí. **Ensaio: aval. pol. públ. Educ**., Rio de Janeiro , v. 18, n. 66, mar. 2010 . Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-40362010000100006&lng=pt&nrm=iso>. Acessado em 26 maio 2014.

HEALEY, J. F. **The Essentials of Statistics**: A Tool for Social Research. 2010, 2007 Wadsworth, Cengage Learning.

HU, L., & BENTLER, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. **Structural Equation Modeling**, v. 6,n. 1, p. 1-55.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Enem:** Exame Nacional do Ensino Médio -Documento Básico. Brasília: INEP, 2002. Disponível em: <http://www.publicacoes.inep.gov.br/portal/download/265>.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Exame Nacional de Desempenho de Estudantes (ENADE 2012). **Relatório Síntese: Administração**. 2012. Disponível em: < http://download.inep.gov.br/educacao_superior/enade/relatorio_sintese/2012/2012_rel_administracao.pdf> Acesso em: 29 jul. 2017

KLEIN, R.; FONTANIVE, N.. Uma nova maneira de avaliar as competências escritoras na redação do ENEM. **Ensaio: aval. pol. públ. Educ.,** Rio de Janeiro , v. 17, n. 65, p. 585-598, Dec. 2009 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-40362009000400002&lng=en&nrm=iso>. Acessado em 08 fev. 2017.

LANZILLOTTI, R. S.; LANZILLOTTI, H. S. Resultados do ENADE/2009 dos cursos de estatísticas em discussão. **Avaliação** (Campinas), Sorocaba , v. 19, n. 1, mar. 2014 . Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772014000100008&lng=pt&nrm=iso > . Acessado em 26 maio 2014.

LAROS, J. A. (2005). **O uso da análise fatorial: algumas diretrizes para pesquisadores**. In L. Pasquali (Org.), **Análise fatorial para pesquisadores** (pp. 163-184). Brasília, DF: LabPAM.

LOKEN, E. & RULISON, K. L. (2010). Estimation of a 4-parameter Item Response Theory model. **The British Journal of Mathematical and Statistical Psychology**, 63(3), 509-525.

MAGIS, D. (2013). A Note on the Item Information Function of the Four-Parameter Logistic Model. **Applied Psychological Measurement** 37(4) 304–315.

MANLY, B.J.F. **Métodos estatísticos multivariados**: uma introdução. 3.ed. Porto Alegre: Bookman, 2008. 229p.

MAROCO, J.; GARCIA-MARQUES, T. Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? **Laboratório de Psicologia**, Lisboa, v. 4, n. 1, p. 65-90, 2006.

MINISTÉRIO DA EDUCAÇÃO. **Portaria Nº 188 de 12 de julho de 2011**. Publicada no Diário Oficial de 13 de julho de 2011, Seção 1, p. 11.

MINISTÉRIO DA EDUCAÇÃO. **Portaria Nº 207 de 22 de junho de 2012**. Publicada no Diário Oficial de 25 de junho de 2012, Seção 1, p. 16.

MINISTÉRIO DA EDUCAÇÃO. **Portaria Nº 214 de 23 de julho de 2010.** Publicada no Diário Oficial de 14 de julho de 2010, Seção 1, p. 828.

MOLCK, ADAUTO. MARIN. CALDERÓN, ADOLFO IGNACIO. (2014) Exame Nacional de Desempenho de Estudantes: mapeamento e tendências temáticas da produção científica brasileira (2004 -2010). **Revista Educação Online**, n. 15, jan./abr. 2014, p. 57-77.

MOREIRA JUNIOR F. J. (2011). **Sistemática para implantação de testes adaptativos informatizados baseados na teoria da resposta ao item** [Tese]. Florianópolis: Faculdade de Engenharia Produção. Universidade Federal de Santa Catarina.

MOREIRA, A. M. A. (2010). **Fatores institucionais e desempenho acadêmico no ENADE:** um estudo sobre os cursos de biologia, engenharia civil, história e pedagogia. Tese de doutorado, Programa de Pós Graduação em Educação, Universidade de Brasília, Brasília, DF.

NOGUEIRA, S. O. (2008). **ENADE**: Análise de Itens de Formação Geral e de Estatística pela TRI. Dissertação de Mestrado, Programa de Pós Graduação Stricto Sensu em Psicologia, da Universidade São Francisco, Itatiba, SP.

PAIVA, G. S. **Exame Nacional de Desempenho dos Estudantes – ENADE**: Recortes da Educação Superior Presencial e a Distância. Universidade Católica de Brasília. Brasília, DF. 2010. 147 p.

PASQUALI, *et al*, 1996. **Teoria e Métodos de Medida em ciências do comportamento**. Brasília: Laboratório de Pesquisa me Avaliação e Medida. Instituto de Psicologia, Universidade de Brasília. INEP. 432p.

PASQUALI, L. Psicometria. **Rev. esc. enferm**. USP, São Paulo , v. 43, n. spe, p. 992-999, Dec. 2009 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-62342009000500002&lng=en&nrm=iso>. Acesso em 01 nov. 2016.

PASQUALI, L; PRIMI, R. Fundamentos da teoria da resposta ao item: TRI. **Aval. psicol**., Porto Alegre , v. 2, n. 2, p. 99-110, dez. 2003 . Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712003000200002&lng=pt&nrm=iso>. Acesso em 20 nov. 2016.

PRICE, L. R. **Psychometric Methods**: Theory into Practice. New York, NY: Guilford Publications, Inc. 2016.

R CORE TEAM (2016). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.R-project.org/.>

RECKEASE, M. D. Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. **Journal of Educational Statistics**, Vol. 4, No. 3 (Outono, 1979), pp. 207-230 1979.

RECKASE, M. D. **Multidimensional item response theory**. New York: Springer, 2009.

REVELLE, W. (2016) **psych**: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. Disponível em: <https://CRAN.R-project.org/package=psych Version = 1.6.12.>.

SAMEJIMA, F. A general model for free response data. **Psychometrika Monograph Supplement**, n. 18, 1972.

SANTOS, N. A. **Determinantes do Desempenho Acadêmico dos alunos de ciências contábeis**. Universidade de são Paulo. São Paulo 2012. 248 p.

SANTOS, M. A. P. *et al* . Avaliar projeto metodológico: isto é possível?. **Avaliação** (Campinas), Sorocaba , v. 16, n. 2, jul. 2011 . Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772011000200011&lng=pt&nrm=iso>. Acesso em 26 maio 2014.

SILVA, M. **A influência das disposições culturais no Exame Nacional de Desempenho dos Estudantes do Ensino Superior (ENADE).** Universidade Federal de São Carlos. São Carlos, SP. 2013.

SILVA, M. C. R.; VENDRAMINI, C. M. M.; LOPES, F. L.. Diferenças entre gênero e perfil sócio-econômico no exame nacional de desempenho do estudante. **Avaliação** (Campinas), Sorocaba , v. 15, n. 3, 2010 . Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772010000300010&lng=pt&nrm=iso>. Acessado em 26 maio 2014.

STEVENS, S. S. On the Theory of Scales of Measurement. Science, **New Series,** Vol. 103, No. 2684. (Jun. 7, 1946), pp. 677-680. Disponível em: < http://marces.org/EDMS623/Stevens%20SS%20(1946)%20On%20the%20Theory%20of%20Scales%20of%20Measurement.pdf>

SWIST K. (2015). Item analysis and evaluation using a four-parameter logistic model. **Edukacia** v. 3(134) p. 77–97.

TENÓRIO, R. M. & ANDRADE, M. A. B. (2009) A avaliação da educação superior no Brasil: desafios e perspectivas. Em: LÔRDELO, J. A. C & DAZZANI, M. V. (Org.). **AVALIAÇÃO EDUCACIONAL Desatando nós**. 1ª ed. Salvador: EDUFBA, 2009, p. 31-55

URBINA, S. **Fundamentos da testagem psicológica**. Porto Alegre: Artmed, 2007. 320 p.

VERHINE, R. E.; DANTAS, L. M. V.; SOARES, J. F. S. (2006). Do Provão ao ENADE: uma análise comparativa dos exames nacionais utilizados no Ensino Superior Brasileiro. **Ensaio**: aval. pol. públ. Educ., Rio de Janeiro, v.14, n.52, p. 291-310, jul./set. 2006

VITORIA, F.; ALMEIDA, L. S.; PRIMI, R. (2006). Unidimensionalidade em testes psicológicos: conceito, estratégias e dificuldades na sua avaliação. **Psic**, São Paulo , v. 7, n. 1, p. 01-07, jun. 2006 . Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1676-73142006000100002&lng=pt&nrm=iso>. acesso em 01 fev. 2017.

WAINER, H. & THISSEN, D. (1993) Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction, **Applied Measurement in Education**, v. 6 n. (2), p. 103-118.

WICKHAM, H. (2016). **readxl: Read Excel Files.** R package version 0.1.1. https://CRAN.R-project.org/package=readxl

YEN, Y.-C., HO, R.-G., LAIO, W.-W., CHEN, L.-J., & KUO, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. **Applied Psychological Measurement**, 36, 75-87.

Apêndice 1 – Roteiro de limpeza dos bancos de dados

1. Realizar download de arquivos em http://portal.inep.gov.br/microdados Importar arquivo no formato".txt" no Microsoft Excel – salvar campos numéricos como texto para evitar corrupção de informações salvas e preservar campos com algarismo zero a esquerda
2. Remover todos os sujeitos com "in_grad = 1" (ingressantes).
3. Remover todos os sujeitos cujo "tp_pr_ob_fg<>555", removendo os alunos que não responderam as questões objetivas da formação geral e também remover todos os sujeitos com ausência em qualquer uma das questões discursivas "tp_sfg_d1<>555" e "tp_sfg_d2<>555".
4. Apagar qualquer linha que "vt_esc_ofg" seja vazio.
5. Quebrar a célula "vt_esc_ofg" em oito – "E1", "E2", "E3", "E4", "E5", "E6", "E7", "E8" - cada uma correspondendo à escolha em cada questão da parte objetiva da prova da formação geral.
6. Realizar a contagem das seguintes marcações nas variáveis anteriores"." (resposta ausente) e de "*" (seleção de múltiplas opções de resposta) – calcular quantidade de respostas escolhidas por cada respondente. Remover linhas nas quais houve ausência de qualquer resposta objetiva.
7. Realizar nova correção verificando correspondência exata com o gabarito de cada ano, salvar 8 variáveis "Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8" com os valores 0 e 1. Substituir os valores da coluna "NT_OBJ_FG" pelo produto da soma das oito células anteriores com (12,5).
   a. Os gabaritos finais das questões estão disponíveis no endereço http://portal.inep.gov.br/web/guest/provas-e-gabaritos3 e são, respectivamente:
      i. ENADE 2010: CACEEADB
      ii. ENADE 2011: DAEACBBE
      iii. ENADE 2012: DEDECABD
8. Recalcular os valores de "NT_FG_D1" e "NT_FG_D2" fazendo a correspondência com as categorias apresentadas na sessão método e salvar o novo valor nas variáveis "D1" e "D2".
9. Remover todos os sujeitos cuja contagem de respostas escolhidas seja menor que oito. Salvar arquivo.

Apêndice 2 – Programação no R

```
####Programa para analisar o ano de 2010##########
##Para realizar outros anos substitua "2010" por outro valor


###### Pacotes necessários ######
library("mirt")
library("psych")
library(readxl)
###################################

####### Definição de vetores necessários para as análises a seguir###
OBJ<- c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8")
DIS<- c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "D1", "D2")
OBJ3PL<- c("3PL", "3PL", "3PL", "3PL", "3PL", "3PL", "3PL", "3PL")
DIS3PL<- c("3PL", "3PL", "3PL", "3PL", "3PL", "3PL", "3PL", "3PL",
"graded", "graded")
OBJ4PL<- c("4PL", "4PL", "4PL", "4PL", "4PL", "4PL", "4PL", "4PL")
DIS4PL<- c("4PL", "4PL", "4PL", "4PL", "4PL", "4PL", "4PL", "4PL",
"graded", "graded")



DIS2010 <- read_excel("C:/Users/andre.oliveira/Desktop/R 2010.xlsx",
 col_types = c("text", "numeric", "numeric",
 "numeric", "numeric", "numeric",
 "numeric", "numeric", "numeric",
 "numeric", "numeric"))



####Análise fatorial para verificar se um fator é adequado
irt.fa(DIS2010[OBJ])

### Scree plot ####
scree(DIS2010[OBJ], main ="Scree plot FG 2010")


###Definição de modelos
```

```
OBJ2010_3PL_1 = mirt(DIS2010[OBJ],1, TOL=.001,
itemtype=OBJ3PL)
OBJ2010_4PL_1 = mirt(DIS2010[OBJ],1, TOL=.001,
itemtype=OBJ4PL)
DIS2010_4PL_1 = mirt(DIS2010[DIS],1, TOL=.001,
itemtype=DIS4PL)

##Resumos dos modelos
summary(OBJ2010_3PL_1)
summary(OBJ2010_4PL_1)

summary(DIS2010_4PL_1)

##Comparação de modelos para as questões objetivas
anova(OBJ2010_3PL_1, OBJ2010_4PL_1)
coef(OBJ2010_3PL_1, IRTpars=TRUE)
coef(OBJ2010_4PL_1, IRTpars=TRUE)


##Parâmetros do modelo de itens de formato misto
coef(DIS2010_4PL_1, IRTpars=TRUE)


#Curva característica de cada item
plot(DIS2010_4PL_1, type = 'trace', theta_lim=c(-4,4) )

#Curva de informação de cada item
plot(DIS2010_4PL_1, type = 'infotrace', theta_lim=c(-4,4) )


#Curva de informação do teste e erro de medida
plot(DIS2010_4PL_1, type = 'infoSE', theta_lim=c(-4,4) )
```

Apêndice 3 – Correlação ponto bisserial de cada item

ENADE 2010 – Correlação ponto bisserial sem efeito espúrio

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	0-1	46	.49	.12	A	8	15	2	-.39	
					B	11	18	5	-.38	
					C	46	23	72	.12	*
					D	23	31	13	-.44	
					E	12	12	9	-.26	
2	0-2	31	.36	.06	A	31	16	52	.06	*
					B	25	31	17	-.41	
					C	36	43	27	-.44	
					D	5	7	3	-.23	
					E	2	3	1	-.16	
3	0-3	78	.43	.25	A	4	8	1	-.30	
					B	1	3	0	-.23	
					C	78	54	96	.25	*
					D	10	21	2	-.46	
					E	7	14	1	-.41	
4	0-4	69	.49	.21	A	5	9	1	-.32	
					B	4	9	1	-.33	
					C	13	21	5	-.41	
					D	10	18	2	-.42	
					E	69	43	92	.21	*
5	0-5	60	.58	.23	A	4	8	1	-.29	
					B	6	12	1	-.35	
					C	12	21	3	-.44	
					D	18	28	6	-.45	
					E	60	31	88	.23	*
6	0-6	12	.22	.13	A	12	4	26	.13	*
					B	5	7	3	-.24	
					C	26	26	23	-.32	
					D	23	26	19	-.35	
					E	34	36	29	-.36	
7	0-7	14	.18	.05	A	43	48	35	-.41	
					B	35	33	35	-.29	
					C	6	9	3	-.27	
					D	14	6	24	.05	*
					E	3	5	2	-.20	
8	0-8	80	.40	.28	A	2	4	0	-.25	
					B	80	57	98	.28	*
					C	7	15	1	-.43	
					D	2	5	0	-.25	
					E	9	19	1	-.46	

ENADE 2011 – Correlação ponto bisserial sem efeito espúrio

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	0-1	66	.50	.18	A	4	8	2	-.26	
					B	3	7	1	-.28	
					C	12	22	4	-.42	
					D	66	37	87	.18	*
					E	14	26	5	-.46	
2	0-2	38	.48	.15	A	38	13	61	.15	*
					B	19	27	12	-.41	
					C	7	13	3	-.34	
					D	21	25	15	-.36	
					E	15	22	9	-.37	
3	0-3	71	.51	.24	A	20	39	6	-.55	
					B	2	5	0	-.26	
					C	2	4	0	-.24	
					D	5	11	1	-.34	
					E	71	41	92	.24	*
4	0-4	53	.32	-.01	A	53	36	67	-.01	*
					B	31	38	23	-.43	
					C	7	11	4	-.30	
					D	7	11	4	-.29	
					E	2	4	1	-.20	
5	0-5	61	.49	.15	A	16	27	7	-.46	
					B	13	20	6	-.39	
					C	61	33	83	.15	*
					D	2	5	1	-.22	
					E	8	14	4	-.34	
6	0-6	75	.32	.08	A	15	24	9	-.41	
					B	75	57	89	.08	*
					C	2	4	0	-.24	
					D	4	7	2	-.27	
					E	3	7	1	-.29	
7	0-7	31	.35	.08	A	45	47	40	-.39	
					B	31	13	48	.08	*
					C	17	27	9	-.43	
					D	4	7	2	-.26	
					E	2	5	1	-.22	
8	0-8	19	.21	.03	A	16	18	13	-.30	
					B	7	11	4	-.28	
					C	36	35	36	-.31	
					D	22	27	17	-.37	
					E	19	9	29	.03	*

ENADE 2012 – Correlação ponto bisserial sem efeito espúrio

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	0-1	59	.45	.15	A	5	8	1	-.30	
					B	9	14	3	-.34	
					C	6	11	2	-.33	
					D	59	38	83	.15	*
					E	21	29	11	-.43	
2	0-2	48	.48	.16	A	12	18	6	-.35	
					B	4	7	1	-.27	
					C	25	35	14	-.46	
					D	10	15	5	-.32	
					E	48	26	74	.16	*
3	0-3	44	.42	.11	A	6	9	2	-.28	
					B	21	27	13	-.40	
					C	28	36	18	-.43	
					D	44	25	67	.11	*
					E	2	3	0	-.19	
4	0-4	52	.42	.11	A	2	4	1	-.23	
					B	17	25	8	-.42	
					C	4	6	1	-.26	
					D	25	32	16	-.41	
					E	52	32	74	.11	*
5	0-5	31	.26	-.01	A	25	29	21	-.35	
					B	36	40	30	-.37	
					C	31	19	45	-.01	*
					D	4	6	2	-.23	
					E	4	6	2	-.22	
6	0-6	55	.45	.13	A	55	34	79	.13	*
					B	19	26	11	-.40	
					C	5	9	2	-.29	
					D	17	26	8	-.43	
					E	3	5	1	-.21	
7	0-7	53	.49	.17	A	14	20	7	-.36	
					B	53	29	79	.17	*
					C	18	27	8	-.44	
					D	11	16	4	-.36	
					E	4	7	2	-.25	
8	0-8	39	.34	.04	A	5	8	2	-.27	
					B	6	9	3	-.27	
					C	15	21	9	-.38	
					D	39	24	57	.04	*
					E	35	38	29	-.36	

Apêndice 4 – Sintaxe de alteração do banco do ENADE 2011 no R

```
#Processo necessário apenas em 2011

x <-
readLines("C:/Users/andre.oliveira/Desktop/2011/microdados_ENADE
_2011/2.DADOS/ENADE_2011.TXT")


ENADE2011 <- data.frame(var1 = substr(x,1,4),
var2 = substr(x,5,8),
var3 = substr(x,9,11),
var4 = substr(x,12,14),
var5 = substr(x,15,18),
var6 = substr(x,19,21),
var7 = substr(x,22,24),
var8 = substr(x,25,32),
var9 = substr(x,33,35),
var10 = substr(x,36,36),
var11 = substr(x,37,68),
var12 = substr(x,69,70),
var13 = substr(x,71,74),
var14 = substr(x,75,78),
var15 = substr(x,79,81),
var16 = substr(x,82,84),
var17 = substr(x,85,87),
var18 = substr(x,88,90),
var19 = substr(x,91,93),
var20 = substr(x,94,96),
var21 = substr(x,97,99),
var22 = substr(x,100,102),
var23 = substr(x,103,105),
var24 = substr(x,106,108),
var25 = substr(x,109,111),
var26 = substr(x,112,114),
var27 = substr(x,115,118),
var28 = substr(x,119,122),
var29 = substr(x,123,126),
var30 = substr(x,127,130),
var31 = substr(x,131,134),
```

```
var32 = substr(x,135,138),
var33 = substr(x,139,142),
var34 = substr(x,143,146),
var35 = substr(x,147,150),
var36 = substr(x,151,154),
var37 = substr(x,155,158),
var38 = substr(x,159,161),
var39 = substr(x,162,164),
var40 = substr(x,165,167),
var41 = substr(x,168,170),
var42 = substr(x,171,178),
var43 = substr(x,179,220),
var44 = substr(x,221,228),
var45 = substr(x,229,236),
var46 = substr(x,237,278),
var47 = substr(x,279,320),
var48 = substr(x,321,328),
var49 = substr(x,329,331),
var50 = substr(x,332,334),
var51 = substr(x,335,342),
var52 = substr(x,343,350),
var53 = substr(x,351,358),
var54 = substr(x,359,361),
var55 = substr(x,362,364),
var56 = substr(x,365,367),
var57 = substr(x,368,375),
var58 = substr(x,376,383),
var59 = substr(x,384,391),
var60 = substr(x,392,394),
var61 = substr(x,395,397),
var62 = substr(x,398,398),
var63 = substr(x,399,399),
var64 = substr(x,400,400),
var65 = substr(x,401,401),
var66 = substr(x,402,402),
var67 = substr(x,403,403),
var68 = substr(x,404,404),
var69 = substr(x,405,405),
var70 = substr(x,406,406),
var71 = substr(x,407,409),
var72 = substr(x,410,412),
```

```
var73 = substr(x,413,413),
var74 = substr(x,414,414),
var75 = substr(x,415,415),
var76 = substr(x,416,416),
var77 = substr(x,417,417),
var78 = substr(x,418,418),
var79 = substr(x,419,419),
var80 = substr(x,420,420),
var81 = substr(x,421,421),
var82 = substr(x,422,422),
var83 = substr(x,423,423),
var84 = substr(x,424,424),
var85 = substr(x,425,425),
var86 = substr(x,426,426),
var87 = substr(x,427,428),
var88 = substr(x,429,429),
var89 = substr(x,430,430),
var90 = substr(x,431,431),
var91 = substr(x,432,432),
var92 = substr(x,433,433),
var93 = substr(x,434,434),
var94 = substr(x,435,435),
var95 = substr(x,436,436),
var96 = substr(x,437,437),
var97 = substr(x,438,438),
var98 = substr(x,439,439),
var99 = substr(x,440,440),
var100 = substr(x,441,441),
var101 = substr(x,442,442),
var102 = substr(x,443,443),
var103 = substr(x,444,444),
var104 = substr(x,445,445),
var105 = substr(x,446,446),
var106 = substr(x,447,447),
var107 = substr(x,448,448),
var108 = substr(x,449,449),
var109 = substr(x,450,450),
var110 = substr(x,451,451),
var111 = substr(x,452,452),
var112 = substr(x,453,453),
var113 = substr(x,454,454),
```

```
var114 = substr(x,455,455),
var115 = substr(x,456,456),
var116 = substr(x,457,457),
var117 = substr(x,458,458),
var118 = substr(x,459,459),
var119 = substr(x,460,460),
var120 = substr(x,461,461),
var121 = substr(x,462,462),
var122 = substr(x,463,463),
var123 = substr(x,464,464),
var124 = substr(x,465,465),
var125 = substr(x,466,466),
var126 = substr(x,467,467))


names2011 = c("NU_ANO" , "CO_IES" , "CD_CATAD" ,
"CD_ORGAC" , "CO_GRUPO" , "CO_REGIAO_CURSO" ,
"CO_UF_CURSO" , "CO_MUNIC_CURSO" , "NU_IDADE" ,
"TP_SEXO" , "NO_MUNIC" , "SG_UF" , "ANO_FIM_2G" ,
"ANO_IN_GRA" , "TP_SEMEST" , "IN_MATUT" , "IN_VESPER" ,
"IN_NOTURNO" , "AMOSTRA" , "PESO_AMOST" , "IN_GRAD" ,
"STATUS" , "IN_INSCR" , "TP_DEF_FIS" , "TP_DEF_VIS" ,
"TP_DEF_AUD" , "TP_PRES" , "TP_PR_GER" , "TP_PR_OB_FG" ,
"TP_PR_DI_FG" , "TP_PR_OB_CE" , "TP_PR_DI_CE" ,
"TP_SFG_D1" , "TP_SFG_D2" , "TP_SCE_D1" , "TP_SCE_D2" ,
"TP_SCE_D3" , "NU_QUE_OFG" , "NU_QUE_DFG" ,
"NU_QUE_OCE" , "NU_QUE_DCE" , "VT_GAB_FG" ,
"VT_GAB_CE" , "VT_ESC_OFG" , "VT_ACE_OFG" ,
"VT_ESC_OCE" , "VT_ACE_OCE" , "NT_OBJ_FG" , "NT_FG_D1" ,
"NT_FG_D2" , "NT_DIS_FG" , "NT_FG" , "NT_OBJ_CE" ,
"NT_CE_D1" , "NT_CE_D2" , "NT_CE_D3" , "NT_DIS_CE" ,
"NT_CE" , "NT_GER" , "NU_QUE_QIP" , "TP_SIT_QIP" ,
"CO_QPP_I1" , "CO_QPP_I2" , "CO_QPP_I3" , "CO_QPP_I4" ,
"CO_QPP_I5" , "CO_QPP_I6" , "CO_QPP_I7" , "CO_QPP_I8" ,
"CO_QPP_I9" , "NU_QUE_SOC" , "TP_SIT_SOC" , "CO_RS_S1" ,
"CO_RS_S2" , "CO_RS_S3" , "CO_RS_S4" , "CO_RS_S5" ,
"CO_RS_S6" , "CO_RS_S7" , "CO_RS_S8" , "CO_RS_S9" ,
"CO_RS_S10" , "CO_RS_S11" , "CO_RS_S12" , "CO_RS_S13" ,
"CO_RS_S14" , "CO_RS_S15" , "CO_RS_S16" , "CO_RS_S17" ,
"CO_RS_S18" , "CO_RS_S19" , "CO_RS_S20" , "CO_RS_S21" ,
"CO_RS_S22" , "CO_RS_S23" , "CO_RS_S24" , "CO_RS_S25" ,
```

```
"CO_RS_S26" , "CO_RS_S27" , "CO_RS_S28" , "CO_RS_S29" ,
"CO_RS_S30" , "CO_RS_S31" , "CO_RS_S32" , "CO_RS_S33" ,
"CO_RS_S34" , "CO_RS_S35" , "CO_RS_S36" , "CO_RS_S37" ,
"CO_RS_S38" , "CO_RS_S39" , "CO_RS_S40" , "CO_RS_S41" ,
"CO_RS_S42" , "CO_RS_S43" , "CO_RS_S44" , "CO_RS_S45" ,
"CO_RS_S46" , "CO_RS_S47" , "CO_RS_S48" , "CO_RS_S49" ,
"CO_RS_S50" , "CO_RS_S51" , "CO_RS_S52" , "CO_RS_S53" ,
"CO_RS_S54")

#Atribuição de nomes nas colunas e criação de arquivo . txt ajustado
colnames(ENADE2011) = names2011
write.table(ENADE2011,
"C:/Users/andre.oliveira/Desktop/2011/microdados_ENADE_2011/2.DADOS/ENADE_2011_OK.txt", sep="\t")

#Remover elementos utilizados
rm(x)
rm(names2011)
rm(ENADE2011)

#Importe arquivo no EXCEL como texto e insira uma célula em A1 movendo as demais para a direita
```