UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM EDUCAÇÃO

LEANDRO ARAUJO DE SOUSA

ANÁLISE COMPARATIVA DO EXAME NACIONAL DO ENSINO MÉDIO (ENEM) VIA TEORIA CLÁSSICA DOS TESTES E TEORIA DE RESPOSTA AO ITEM

FORTALEZA
2019

LEANDRO ARAUJO DE SOUSA

ANÁLISE COMPARATIVA DO EXAME NACIONAL DO ENSINO MÉDIO (ENEM) VIA TEORIA CLÁSSICA DOS TESTES E TEORIA DE RESPOSTA AO ITEM

Tese apresentada ao Programa de Pós-Graduação em Educação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Doutor em Educação. Área de concentração: Avaliação Educacional.

Orientadora: Profa. Dra. Adriana Eufrásio Braga

FORTALEZA

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S697a Sousa, Leandro Araujo de.
Análise comparativa do Exame Nacional do Ensino Médio (Enem) via teoria clássica dos testes e teoria de resposta ao item / Leandro Araujo de Sousa. – 2019.
105 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Faculdade de Educação, Programa de Pós-Graduação em Educação, Fortaleza, 2019.
Orientação: Profa. Dra. Adriana Eufrásio Braga.

1. Avaliação em larga escala. 2. Psicometria clássica. 3. Modelo de traço latente. I. Título.
CDD 370

LEANDRO ARAUJO DE SOUSA

ANÁLISE COMPARATIVA DO EXAME NACIONAL DO ENSINO MÉDIO (ENEM) VIA TEORIA CLÁSSICA DOS TESTES E TEORIA DE RESPOSTA AO ITEM

Tese apresentada ao Programa de Pós-Graduação em Educação da Universidade Federal do Ceará, como requisito parcial à aprovação no segundo exame de qualificação. Área de concentração: Avaliação Educacional.

Aprovada em: ___/___/_____.

BANCA EXAMINADORA

_____

Profa. Dra. Adriana Eufrásio Braga (Orientadora)

Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Wagner Bandeira Andriola

Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Edson Silva Soares

Universidade Federal do Ceará (UFC)

_____

Prof. Dr. João Batista Carvalho Nunes

Universidade Estadual do Ceará (UECE)

_____

Prof. Dr. José Airton de Freitas Pontes Junior

Universidade Estadual do Ceará (UECE)

A Deus.

Aos meus pais, Auzenir e Edmar.

A minha esposa, Grasiany.

AGRADECIMENTO

A Profa. Dra. Adriana Eufrásio Braga e ao Prof. Dr. Nicolino Trompieri Filho, pela excelente orientação.

Aos professores participantes da banca examinadora, Prof. Dr. José Airton de Freitas Pontes Junior, Prof. Dr. Wagner Bandeira Andriola, Prof. Dr. Edson Silva Soares e Prof. Dr. João Batista Carvalho Nunes pelo tempo, pelas valiosas colaborações e sugestões.

A minha esposa, Grasiany Almeida pelo apoio e compreensão das minhas ausências durante o doutorado.

A minha família, especialmente aos meus pais, Alzenir e Edmar, que me incentivaram e são os principais responsáveis pela minha carreira acadêmica e nunca deixaram de acreditar no estudo como mudança de vida.

Aos meus amigos de grupo de pesquisa pelas valiosas discussões proporcionadas e que enriqueceram minha carreira acadêmica.

“Teoria da mensuração: inconveniente necessário”

Nicolino Trompieri Filho e José Anchieta Esmeraldo Barreto

RESUMO

As avaliações educacionais em larga escala, entre elas o Exame Nacional do Ensino Médio (Enem), são importantes medidas de desempenho escolar realizadas atualmente no Brasil. Por muito tempo a Teoria Clássica dos Testes (TCT) foi dominantemente utilizada. Nesse contexto, o Enem desde a sua criação, em 1998, até 2008 utilizou-se desse paradigma de análise. No entanto, a partir da década de 1990 parte dos procedimentos da TCT têm sido substituídos pela Teoria de Resposta ao Item (TRI). Dessa forma, em 2009 esse exame passou a utilizar a TRI. Com isso, questionamos se há diferenças nas medidas a partir da TCT e TRI, portanto, se há comparabilidade entre os parâmetros dos itens e escores dos participantes. Dessa forma, esta pesquisa teve o objetivo de avaliar os resultados do Enem de 2017 a partir da TCT e TRI. Trata-se de uma pesquisa de abordagem quantitativa e com objetivo exploratório. Participaram da pesquisa uma amostra de 10.000 participantes selecionados por amostragem aleatória simples. A análise da dimensionalidade das provas do exame foi realizada com o teste de Análise Paralela e Análise Fatorial de Informação Plena. Posteriormente, foram estimados os parâmetros dos itens e dos participantes do exame a partir da TCT e TRI (1, 2 e 3 parâmetros) e comparados a partir do coeficiente de correlação de Pearson (r) e regressão linear simples. Para todas as análises foram utilizados pacotes estatísticos do *Software R*. Os resultados encontrados a partir da Análise Paralela indicaram evidências de haver uma dimensão dominante em cada prova do exame. A Análise Fatorial de Informação Plena indicou que a maioria dos itens apresentaram cargas fatoriais elevadas em um único fator. Mas alguns itens apresentaram cargas fatoriais baixas (<0,30). Após o ajuste de um modelo unidimensional e estimação dos parâmetros dos itens e dos participantes pela TCT e TRI, encontrou-se forte correlação entre os parâmetros de dificuldade e discriminação dos itens pela TCT e modelos de TRI de 1 e 2 parâmetros. Os valores de correlação decrescem com o modelo de 3 parâmetros. O mesmo ocorre na comparação entre os escores dos participantes. Em ambos os casos o modelo de regressão simples ajustado foi significativo. Conclui-se que há alta comparabilidade entre os parâmetros dos itens e escores dos participantes pela TCT e modelos de TRI. Entretanto, a comparabilidade torna-se mais frágil com o modelo de TRI de 3 parâmetros. Assim, questiona-se a necessidade do modelo de TRI para as finalidades desta avaliação.

Palavras-chave: Avaliação em larga escala. Psicometria clássica. Modelo de traço latente.

ABSTRACT

Large-scale educational assessments, including the National Exam of Upper Secondary Education (Enem), are important measures of school performance in Brazil today. For a long time the Classical Test Theory (CTT) has been predominantly used. In this context, Enem since its creation in 1998 until 2008 has used this analysis paradigm. However, since the 1990s, part of the CTT procedures have been replaced by Item Response Theory (IRT). Thus, in 2009 this exam started to use the IRT. Thus, we question whether there are differences in measurements from CTT and IRT, therefore, whether there is comparability between item parameters and participants' scores. Thus, this research aimed to evaluate the results of the 2017 Enem from CTT and IRT. It is a research with quantitative approach and exploratory objective. A sample of 10.000 participants selected by simple random sampling participated in the research. The analysis of the exams dimensionality was performed with the Parallel Analysis and Full Information Factor Analysis test. Subsequently, the parameters of the exam items and participants were estimated from CTT and IRT (1, 2 and 3 parameters) and compared using Pearson's correlation coefficient (r) and simple linear regression. Software *R* statistical packages were used for all analyzes. The results from the Parallel Analysis indicated evidence of a dominant dimension in each exam. Full Information Factor Analysis indicated that most items had high factor loadings in a single factor. But some items had low factor loadings (<0.30). After adjusting a one-dimensional model and estimating item and participant parameters by CTT and IRT, a strong correlation was found between the difficulty and item discrimination parameters by CTT and 1 and 2-parameter IRT models. Correlation values decrease with the 3-parameter model. The same occurs when comparing the participants' scores. In both cases the adjusted simple regression model was significant. It is concluded that there is high comparability between the parameters of the items and the participants' scores by CTT and IRT models. However, comparability becomes weaker with the 3-parameter IRT model. Thus, the need for the IRT model for the purposes of this evaluation is questioned.

Keywords: Large scale evaluation. Classical test theory. Latent trace model.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AERA	American Educational Research Association
AIC	Critério de Informação de Akaike
APA	American Psychological Association
BIC	Critério de Informação Baysiano
CCI	Curva Característica do Item
CFI	Comparative Fit Index
CML	Conditional Maximum Likelihood
EAP	Estimated a Posteriori
Enem	Exame Nacional do Ensino Médio
FIFA	Full Information Factor Analysis
ID	Índice de Dimensionalidade
JML	Joint Maximum Likelihood
MAP	Maximum a Posteriori
MEC	Ministério da Educação
ML	Maximum Likelihood
ML	Maximum Likelihood
MML	Marginal Maximum Likelihood
NCME	National Council on Measurement in Education
Prouni	Programa Universidade para Todos
RMSEA	Root-Mean-Square Error of Approximation
Saeb	Sistema de Avaliação da Educação Básica
Sinaes	Sistema Nacional de Avaliação da Educação Superior
Sisu	Sistema de Seleção Unificada
SRMSR	Standardised Root Mean Square Residual
TCT	Teoria Clássica dos Testes
TLI	Tucker-Lewis Index
TRI	Teoria de Resposta ao Item

Sumário

1 INTRODUÇÃO

As avaliações em larga escala são importantes para tomadas de decisão e direcionamento de políticas públicas educacionais (VIANNA, 2003). Entre essas avaliações, o Exame Nacional do Ensino Médio (ENEM) surge como parâmetro de avaliação para esse nível de ensino. Além disso, os resultados do exame têm sido usados pelas Instituições de Ensino Superior (IES) para selecionar candidatos. Por possibilitar tomadas de decisão importantes, esses exames precisam apresentar boa qualidade para avaliar (TOFFOLI et al., 2016). Nesse sentido, a psicometria auxilia na elaboração de testes que apresentem boa capacidade de realizar medidas. Entre as características necessárias a um teste educacional estão a validade e a fidedignidade, ou seja, deve apresentar boas evidências que realiza a medida do que se pretende e de forma precisa (TOFFOLI et al., 2016).

Por muito tempo, a Teoria Clássica dos Testes (TCT) foi e continua sendo utilizada na análise da qualidade métrica de instrumentos de medida nas avaliações educacionais (SARTES; SOUSA-FORMIGONI, 2013). No entanto, nos últimos anos, tem ganhado destaque a Teoria de Resposta ao Item (TRI) em avaliações em larga escala, sob a justificativa de oferecer vantagens como estabilidade e comparabilidade dos resultados, algo não oferecido pela TCT (ANDRADE; LAROS; GOUVEIA, 2010). A partir disso, considera-se que o modelo TRI não tem substituído totalmente a TCT, mas complementa as suas análises (SARTES; SOUSA-FORMIGONI, 2013).

A TCT, para alguns autores (ANDRADE; TAVARES; VALLE, 2000; KLEIN, 2013; SARTES; SOUSA-FORMIGONI, 2013), apresenta alguns problemas, entre outros: a dependência da amostra, ou seja, do particular conjunto de sujeitos avaliados, dessa forma o teste apresenta escores diferentes para grupos diferentes de avaliados; a dependência do teste e dos itens, pois escores distintos são obtidos se um grupo de sujeitos é avaliado com diferentes testes sobre o mesmo conhecimento e; em decorrência disso, os testes não permitem a comparabilidade dos resultados, sendo, dessa forma, instáveis.

Para propor alternativas a esses problemas surge a TRI, que alega proporcionar estabilidade dos resultados, ou seja, os sujeitos terão os mesmos escores, ou notas, mesmo que sejam utilizados testes com itens diferentes. Isso torna-se possível porque o parâmetro de análise é o item, em que, independentemente dos avaliados, terá sempre os mesmos parâmetros. Por apresentar essa invariabilidade, os resultados tornam-se comparáveis (PASQUALI, 2009; VALLE, 2000).

No entanto, com o objetivo de comparar a TCT e TRI nos resultados obtidos na análise de testes, alguns estudos foram realizados (ADEDOYIN; ADEDOYIN, 2013; ADEGOKE, 2013; COSTA; FERRÃO, 2015; GÜLER; UYANIK; TEKER, 2014), possibilitando alguns esclarecimentos acerca da problemática. Nesses estudos os autores indicaram alta correlação entre os parâmetros de dificuldade e discriminação dos itens estimados pela TCT e TRI, indicando compatibilidade dos resultados obtidos pelos dois modelos. Esses indícios levam a questionar as vantagens alegadas pela TRI.

Com isso, a partir de 2009 o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) passou a utilizar a TRI na validação dos itens e na análise dos resultados do Enem. O exame utiliza o modelo logístico de três parâmetros, em que considera a dificuldade, discriminação e probabilidade de acerto casual do item (BRASIL, 2011). Considera-se que a mudança de método de análise influencia na confiabilidade da medida realizada, pois o modelo mais convencional da TRI exige a unidimensionalidade dos itens, algo muito difícil de ser obtido. Algumas críticas questionam a coerência metodológica do exame (TAVARES, 2013). Entre as discussões, Tavares (2013) indaga a possibilidade de cumprir o pressuposto, já que o conhecimento humano é multideterminado, ou seja, depende de vários fatores. Questiona-se também a validade das medidas realizadas, já que o modelo não consegue explicar essa realidade, embora se advogue por uma "unidimensionalidade essencial" (STOUT, 1990).

Como possibilidade de um modelo de melhor ajuste a realidade surgem os modelos de TRI multidimensional (RECKASE, 2009). No entanto, esses modelos ainda são pouco implementados nas avaliações educacionais em larga escala.

A partir disso, considerando que o Enem é uma prova multidisciplinar, questiona-se a sua validade como um instrumento unidimensional. Embora o exame apresente áreas de avaliação bem determinadas (Linguagens e Códigos, Matemática, Ciências da Natureza e Ciências Humanas), cada área é composta por conhecimentos de diferentes disciplinas. Por exemplo, na prova de Linguagens e Códigos há itens referentes aos conhecimentos de Língua Portuguesa, Língua Estrangeira, Artes e Educação Física. Embora possamos compreender que todas essas áreas constituem uma linguagem, cada uma resguarda a sua especificidade que as diferenciam das outras.

Não obstante, o exame apresenta algumas inadequações quanto a sua proposta inicial (ANDRADE, 2012): i) para que os resultados sejam comparados é necessário que haja alguns itens em comum, no entanto, não é possível que se aplique itens já utilizados em outros

exames, já que é um teste de seleção e suas provas são divulgadas na íntegra; ii) para a validação dos itens é necessário que esses sejam pré-testados em uma amostra do universo de participantes, quando isso ocorre alguns possíveis candidatos têm acesso aos itens anteriormente; iii) uma das justificativas para a utilização da TRI é que esta tem foco na análise do item, assim é possível atribuir pesos distintos para os itens, algo que também é possível e legítimo com a TCT, com esta solucionaria o problemas da dificuldade de interpretação dos resultados do exame pelos candidatos, de modo que estes não têm como estimar o seu resultado.

Além desses problemas apresentados, há uma outra questão que precisa ser levantada. Segundo uma nota técnica publicada pelo Ministério da Educação (MEC)[1] a utilização da TRI no Enem tem duas finalidades principais: i) permitir a comparabilidade entre os anos; ii) permitir a aplicação do exame várias vezes ao ano.

Em relação ao primeiro objetivo apresentado cabe uma reflexão. A comparabilidade entre os resultados é particularmente importante quando se quer acompanhar a evolução do aprendizado de um determinado grupo (ANDRADE; LAROS; GOUVEIA, 2010) com vistas a tomada de decisão e redirecionamento de políticas e recursos. No entanto, o objetivo principal do Enem atualmente é ser parâmetro de seleção para os cursos de graduação nas universidades, institutos federais e instituições privadas de ensino superior. Dessa forma, os resultados têm como foco o desempenho individual dos candidatos, que não necessariamente são estudantes, para fins de classificação. Assim, resultados comparáveis para tomada de decisão e definição de política públicas educacionais já são oferecidos por outras provas, como as realizadas no âmbito do Sistema de Avaliação da Educação Básica (Saeb) (KLEIN, 2009).

Quanto ao segundo objetivo alegado pelo Inep, este não tem se concretizado na prática, pois o exame ainda é aplicado uma única vez ao ano. Além desses dois principais objetivos, aponta-se também a possibilidade de aplicar provas distintas, ou seja, com itens diferentes, sem, contudo, alterar o nível de dificuldade da prova. Esse procedimento também não tem sido realizado. Atualmente é aplicado cadernos de prova com cores diferentes, em que há apenas a alteração da posição dos itens em cada caderno.

Diante disso, esta tese está pautada no seguinte problema e questões decorrentes: Há comparabilidade dos parâmetros dos itens e dos escores dos participantes estimados a partir Teoria Clássica dos Testes e Teoria de Resposta ao Item? Além disso, o desenvolvimento da pesquisa foi norteado pelas seguintes questões: Os itens das provas do Enem 2017 são

[1] http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf

unidimensionais? Sendo possível ajustar um modelo unidimensional, há correlação entre os parâmetros dos itens e dos participantes estimados pela TCT e TRI?

Dessa forma, esta tese está pautada em duas hipóteses principais: i) Considerando que o conhecimento humano é multideterminado, ou seja, é influenciado por vários fatores ou dimensões latentes, supomos que os itens das provas do Enem 2017 não guardam o pressuposto da unidimensionalidade em cada prova; ii) Quando ajustado um modelo unidimensional para as provas do exame, os parâmetros dos itens e os escores dos participantes dados pela TRI são altamente comparáveis aos obtidos pela TCT.

Assim, a tese defendida neste trabalho de pesquisa é de que os itens do Enem 2017 não são unidimensionais e quando ajustado um modelo unidimensional para as provas do exame os parâmetros dos itens e o escore dos participantes são comparáveis. Para essa verificação foram estabelecidos os seguintes objetivos.

1.1Objetivos

1.1.1 *Objetivo geral*

Avaliar os resultados do Exame Nacional do Ensino Médio de 2017 a partir da Teoria Clássica dos Testes e Teoria de Resposta ao Item.

1.1.2 *Objetivos específicos*

Analisar a dimensionalidade das provas das quatro áreas do Exame Nacional do Ensino Médio, Linguagens e Códigos, Ciências Humanas, Ciências da Natureza e Matemática, a partir da análise paralela e análise fatorial de informação plena.

Estimar os parâmetros de dificuldade e discriminação dos itens e o escore dos participantes do Enem 2017 pela Teoria Clássica dos Testes.

Estimar os parâmetros de dificuldade, discriminação e acerto casual dos itens das provas do Enem 2017 e a habilidade dos participantes pela Teoria de Resposta ao Item.

Comparar e correlacionar os parâmetros dos itens da Teoria Clássica dos Testes com os parâmetros dos modelos logísticos de 1, 2 e 3 parâmetros da Teoria de Resposta ao Item.

Comparar e correlacionar o desempenho dos participantes estimados a partir da Teoria Clássica dos Testes com o desempenho estimado a partir dos modelos logísticos de 1, 2 e 3 parâmetros da Teoria de Resposta ao Item.

Nas seções posteriores está descrito o referencial teórico e metodológico de base para essa pesquisa. Em um primeiro momento realizou-se uma problematização dos aspectos políticos e pedagógicos implicados aos métodos de análise dos resultados do exame. Posteriormente está descrito brevemente as características do Enem. Por fim, é exposto o referencial metodológico de base utilizado neste estudo.

2 AVALIAÇÕES EM LARGA ESCALA: ASPECTOS POLÍTICOS E METODOLÓGICOS

As avaliações educacionais em grandes grupos populacionais surgem da necessidade de realização de prestação de contas de grandes investimentos financeiros aplicados nessa área, denominado de *accountability*, e por vezes criticado como instrumento de regulação da educação pelo Estado (FREITAS, 2004; SCHNEIDER; ROSTIROLA, 2015). Além disso, surgem também da importância de obter diagnósticos que permitissem o planejamento e direcionamento de políticas focadas em tomadas de decisões para a resolução de problemas específicos no campo educacional (VIANNA, 2003).

Para tanto, é necessário remeter à ideia de qualidade. E quando se trata de educação, o conceito de qualidade não é fixo, depende dos sujeitos envolvidos e de suas necessidades (HORTA NETO, 2010). Dessa forma, o que é qualidade para um determinado grupo pode não o ser para outro. Portanto, ao realizar uma avaliação é necessário estabelecer os critérios de qualidade, ou seja, o padrão de referência para comparação dos seus resultados. Embora critica-se as avaliações em larga escala por forçarem uma padrão único de "qualidade" para o sistema educacional ao utilizarem apenas os resultados das provas aplicadas (SOUSA, 2014). Quando se trada de qualidade do sistema, considera-se que este não é possível ser conhecido através de uma prova padronizada (CASASSUS, 2013), mas não podemos descartar que os indicadores revelados são aspectos necessários à ideia de qualidade em educação (MACHADO; ALAVARSE, 2014).

Diante disso, o Brasil implantou avaliações em larga escala em diferentes níveis administrativos a partir da década de 1990 na tentativa de solucionar problemas e melhorar a qualidade da educação ofertada que, nesse contexto, estava relacionado a proporcionar aquisição de competências e habilidades, ou seja, elevar o desempenho dos estudantes (VIANNA, 2003). Entre as avaliações criadas estão o Sistema de Avaliação da Educação Básica (Saeb), Prova Brasil, Exame Nacional do Ensino Médio (Enem) e Sistema Nacional de Avaliação da Educação Superior (Sinaes). Essas avaliações assumem duas funções distintas (MINHOTO, 2016):

> 1) levantar informações ou evidências necessárias à formulação de políticas educacionais, tendo como propósito ampliar e aprofundar o conhecimento sobre os sistemas de ensino para que as diferentes esferas de governo possam definir prioridades de intervenção; e 2) induzir mudanças ou consolidar reformas educacionais previamente estruturadas para os sistemas de ensino (p. 78).

No entanto, há uma contradição importante levantada por Vianna (2003 p. 44) que permanece atualmente: “são desenvolvidas competências e habilidades em nosso sistema educacional de uma forma sistemática, ou, explicitando, é o nosso ensino orientado para o desenvolvimento de competências?”. Esse questionamento aponta para uma possível fragilidade do sistema educacional e de suas avaliações ao se considerar a incoerência entre o ensinado e o avaliado. Além disso, ressalta que avaliações de grande abrangência populacional comprometem uma avaliação completa, ou seja, que contemple todas as dimensões dos avaliados, sendo limitada a uma métrica do que se supõe medir.

Para superar essa limitação, Dalben e Almeida (2015) propõem uma perspectiva multidimensional para as avaliações em larga escala, de modo que pudessem abranger os mais variados fatores possíveis associados a aprendizagem dos alunos. No entanto, os autores reconhecem a impossibilidade de se contemplar todo o currículo da escala, uma vez que esse sempre é mais abrangente que uma matriz de referência, por mais ampla que ela seja. Além disso, ressaltam que com uma avaliação multidimensional as escolas podem eleger suas prioridades e metas a serem alcançadas considerando os aspectos avaliados, com isso lateraliza-se o foco exclusivo no rendimento.

Nesse contexto, Bauer, Alavarse e Oliveira (2015) realizam uma sistematização do debate e ressaltam que não há consenso em relação as contribuições das medidas educacionais em larga escala, mas reconhecem sua importância. Diante disso, os autores apontam duas linhas de discussão: i) o papel das avaliações em larga escala nas reformas educacionais e; ii) avaliações educacionais como instrumento de gestão e política educacional.

Na primeira linha, Bauer, Alavarse e Oliveira (2015) reconhecem o papel das avaliações quando centrada no produto e não no processo. Nessa perspectiva, há argumentos favoráveis, uma vez que as escolas e professores são responsabilizados, estimulando o comprometimento destes com a aprendizagem dos estudantes. Além disso, instala-se uma cultura de avaliação, dando transparência aos resultados e processos do serviço público. Também possibilita aos gestores a comparação de resultados com outras escolas ou com diferentes níveis administrativos, o que permite tomada de decisões e direcionamento de políticas públicas. Em contraponto, há severas críticas ao papel político das avaliações em larga escala, principalmente quando têm grandes impactos (*high stakes testing*). Argumenta-se no sentido de que essas avaliações são instrumentos para imposição de reformas educacionais, aprofundando ainda

mais a desigualdade. Além disso, denunciam os interesses econômicos associados, tais como terceirização dos testes e venda de materiais pela iniciativa privada.

Na segunda linha encontra-se os usos e divulgação dos resultados da avaliação. Nessa perspectiva, considera-se a importância da divulgação dos resultados (BAUER; ALAVARSE; OLIVEIRA, 2015). Por exemplo, quando proporciona aos alunos e professores informações sobre a aprendizagem, possibilitando tomadas de decisões. Aos pais, quando proporciona informações de qualidade de ensino das escolas, auxiliando na escolha da instituição em que os filhos serão matriculados. No entanto, há críticas de como a divulgação é realizada. Entre as objeções apresentadas estão os ranqueamentos das instituições que não apresentam informações completas sobre a mesma. Outra questão colocada é em relação aos impactos causados às escolas e professores, quando é utilizada políticas de bonificação. Em decorrência disto, os professores acabam por adaptar seus métodos de ensino e limitam os conteúdos ensinados para atender as expectativas das avaliações.

Considerando esse viés político das avaliações em larga escala, Bonamino e Sousa (2012) apresenta a avaliação da educação básica no Brasil em três gerações, em que ressaltam principalmente o grau dos impactos sobre as escolas. Na primeira geração, as autoras apontam não haver impacto sobre as escolas e professores, tal como o Sistema de Avaliação da Educação Básica (Saeb), em que sua estrutura de avaliação visa diagnósticos gerais da educação no país, não permitindo comparações entre as escolas, deste modo apenas auxilia no direcionamento de políticas públicas para melhoria dos resultados. Em uma avaliação de segunda geração, já há alguma influência sobre as escolas, como é o caso da Prova Brasil, em que, por ser censitária, permite a comparação entre escolas e turmas, exercendo assim pressão sobre os gestores e professores para alcançar as metas. Por fim, em uma avaliação de terceira geração há grandes impactos sobre escolas e professores, principalmente com as políticas de bonificação realizadas a partir dos resultados das avaliações, como é o caso dos sistemas de avaliação de São Paulo e Pernambuco. Como consequência, pode haver alterações no processo de ensino e aprendizagem para melhorar os resultados.

Diante disso, colocamos o Enem como uma avaliação de terceira geração que, apesar de não ser propriamente um sistema de avaliação da Educação Básica, tem se estabelecido como referência de avaliação para o Ensino Médio e seus resultados acarretam consequências importantes, principalmente para os alunos, uma vez que são utilizados para ingresso em instituições de ensino superior gratuita e para acesso a programas de financiamento em instituições privadas. Além disso, tem influenciado nos métodos de ensino, currículos e gestão

das escolas (AMARO, 2013; FONTANIVE, 2013; LOPES; LÓPEZ, 2010; MACHADO, 2013; MESQUITA, 2012; SANTOS, 2011).

A partir disso, para atender a rigorosidade necessária nos processos de construção e análise dos resultados, muitos exames e sistemas de avaliações no Brasil e no mundo optaram por adotar a TRI como modelo de análise em detrimento da TCT, por considerarem permitir maior precisão e comparabilidade dos resultados. Diante disso, apresentamos o método (TCT ou TRI) de cálculo das proficiências (nota) dos participantes das avaliações estaduais e nacionais no Brasil.

Quadro 1 – Métodos de análise dos resultados nas avaliações nacionais e estaduais no Brasil.

Avaliações	**Ano de Criação**	**Entes da Federação**	**Análise dos Resultados**	
			Na criação	**Atualmente**
Avaliação Nacional da Educação Básica (Aneb)	1990	União	TCT	TRI
Avaliação Nacional do Rendimento Escolar (Aresc - Prova Brasil)	2005	União	TCT	TRI
Avaliação Nacional da Alfabetização (ANA)	2013	União	TRI	TRI
Exame Nacional do Ensino Médio (Enem)	1998	União	TCT	TRI
Sistema Estadual de Avaliação da Aprendizagem Escolar (Seape)	2009	Acre	TCT e TRI	TCT e TRI
Avaliação de Aprendizagem da Rede Estadual de Ensino de Alagoas (Areal)	2012	Alagoas	TCT e TRI	TCT e TRI
Sistema de Avaliação do Desempenho Educacional do Amazonas (Sadeam)	2008	Amazonas	TCT e TRI	TCT e TRI
Sistema de Avaliação Baiano da Educação (Sabe)	2011	Bahia	TCT e TRI	TCT e TRI
Sistema Permanente de Avaliação da Educação Básica do Ceará (Spaece)	1992	Ceará	TCT	TCT e TRI
Sistema Permanente de Avaliação Educacional do Distrito Federal (SipaeDF)	2014	Distrito Federal	TCT	TCT
Programa de Avaliação da Educação Básica do Espírito Santo (Paebes)	2009	Espírito Santo	TCT e TRI	TCT e TRI
Sistema de Avaliação Educacional do Estado de Goiás (Saego)	2011	Goiás	TRI	TRI
Sistema de Avaliação da Educação da Rede Pública de Mato Grosso do Sul (Saems)	2008	Mato Grosso do Sul	TCT e TRI	TCT e TRI
Avaliação Diagnóstica do Ensino Público do Estado de Mato Grosso (ADEPE-MT)	2016	Mato Grosso	TRI	TRI
Sistema Mineiro de Avaliação e Equidade da Educação Pública (SIMAVE)	2000	Minas Gerais	TCT	TCT e TRI
Sistema Paraense de Avaliação Educacional (SisPAE)	2013	Pará	TCT e TRI	TCT e TRI
Sistema Estadual de Avaliação da Educação da Paraíba (Avaliando IDEPB)	2012	Paraíba	TCT e TRI	TCT e TRI
Sistema de Avaliação da Educação Básica do Paraná (Saep)	2012	Paraná	TCT e TRI	TCT e TRI
Sistema de Avaliação Educacional de Pernambuco (Saepe)	2000	Pernambuco	TCT	TCT e TRI
Sistema de Avaliação Educacional do Piauí (Saepi)	2011	Piauí	TCT e TRI	TCT e TRI

Quadro 1 (Continuação)

Sistema de Avaliação da Educação do Estado do Rio de Janeiro (Saerj)	2008	Rio de Janeiro	TRI	TRI
Sistema Integrado de Monitoramento e Avaliação Institucional (Simais)	2016	Rio Grande do Norte	TCT e TRI	TCT e TRI
Sistema de Avaliação do Rendimento Escolar do Rio Grande do Sul (Saers)	1996	Rio Grande do Sul	TCT	TRI
Sistema de Avaliação Educacional de Rondônia (Saero)	2012	Rondônia	TCT e TRI	TCT e TRI
Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (Saresp)	1996	São Paulo	TCT	TCT e TRI
Sistema de Avaliação da Educação do Estado do Tocantins (Saeto)	2011	Tocantins	TCT	TCT

Fonte: Elaboração do autor.

Ao investigar os métodos de análise utilizados nas avaliações em larga escala no Brasil, atualmente, parte considerável dos testes adota a TRI na análise dos resultados, ou seja, no cálculo da proficiência dos participantes, embora em muitas dessas avaliações a TCT não tenha sido deixada totalmente de lado. Algumas dessas avaliações podem ser consideradas como de terceira geração, ou seja, têm muita influência ou impacto sobre os interessados, a saber, os alunos, professores, gestão e a própria escola, considerando que muitas administrações públicas utilizam os resultados como política de bonificação dos mesmos.

Nesse sentido, é importante que esses instrumentos apresentem informações confiáveis sobre o desempenho dos estudantes. Para tanto, é necessário dispor de métodos de análise dos resultados consistentes e que produzam boas informações. Diante disso, utilizar a TCT ou TRI pode influenciar nos resultados e em sua interpretação. Dessa forma, são importantes as evidências empíricas sobre a qualidade ou não desses métodos de análise dos resultados, principalmente da TRI, que tem sido cada vez mais adotada nos exames e avaliações educacionais.

Nesta seção foi discutido os aspectos políticos e pedagógicos implicados na metodologia de análise dos resultados do Enem. Na próxima seção será realizada uma breve caracterização do exame com o objetivo de compreender sua estrutura e como esta contradiz o pressuposto da unidimensionalidade das provas admitido na elaboração e análise dos resultados do exame.

3 O EXAME NACIONAL DO ENSINO MÉDIO (ENEM)

O Enem teve sua primeira edição em 1998, inicialmente com o objetivo de ser parâmetro de auto avaliação para os estudantes ao final do Ensino Médio, mas com a possibilidade de ter suas notas utilizadas como seleção para os cursos das instituições de Educação Superior públicas e privadas ou como parte delas (BRASIL, 1998).

Ao longo dos anos o exame tem ganhado maiores proporções tanto em número de inscrições como de impacto social, já que sua nota vem sendo utilizada por muitas instituições tanto públicas via Sistema de Seleção Unificada (Sisu) como privadas a partir do Programa Universidade para Todos (Prouni) para o ingresso nos cursos de graduação. Desde seu início, em 1998, até 2010 o número de inscritos saltou de 157.221 para 4.611.441 (CORTI, 2013), ou seja, quase 30 vezes mais inscritos.

A partir de 2009, quando o Inep inaugura o assim denominado "Novo Enem", há uma reestruturação da matriz de referência do exame que passa a ser composta por competências e habilidades divididas por área de conhecimento, a saber: Linguagens e Códigos, Matemática, Ciências Humanas e Ciências da Natureza e suas respectivas disciplinas, compondo um total de 180 itens mais uma redação (BRASIL, 2009). A estrutura do exame está disposta no Quadro 2.

Quadro 2 – Estrutura do Exame Nacional do Ensino Médio.

Área	Disciplinas	Competências e Habilidades	n de itens
Linguagens e Códigos	Português, Educação Física, Artes e Língua Extrangeira	9 competências 30 habilidades	45
Matemática	Matemática	7 competências 30 habilidades	45
Ciências Humanas	História, Geografia, Filosofia e Sociologia	6 competências 30 habilidades	45
Ciências da Natureza	Física, Química e Biologia	8 competências 30 habilidades	45

Fonte: Inep (2019).

Além das alterações realizadas na estrutura do exame, o Inep também passa a adotar a TRI como método de análise da prova e do desempenho dos candidatos. O modelo utilizado pelo exame é o unidimensional de três parâmetros, em que se considera a dificuldade, discriminação e a probabilidade de acerto casual. Segundo a nota técnica da instituição essas

alterações ocorreram com o objetivo de “(1) permitir a comparabilidade dos resultados entre os anos e (2) permitir a aplicação do Exame várias vezes ao ano”[2].

O modelo de TRI adotado no Enem é o unidimensional, em que considera que uma dimensão é responsável por um conjunto de itens. Na análise, os itens de cada área de conhecimento são considerados como de uma dimensão. No entanto, como observado no Quadro 2, cada área é composta por disciplinas diferentes. Nesse sentido, do ponto de vista de conteúdo dos itens, a prova de Ciências da Natureza, por exemplo, pode ser dividida em três disciplinas: física, química e biologia. Dessa forma, seria razoável pressupor pelo menos três dimensões.

Na breve descrição do Enem nesta seção foi realizado um esforço de expor as características do exame e a inadequação teórica deste ao pressuposto de unidmensionalidade exigido pelo modelo de TRI adotado para o exame. Na próxima seção é apresentado os modelos da TCT e TRI, modelos utilizados na comparação dos resultados do Enem nesta pesquisa.

[2] Disponível em http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf

4 PSICOMETRIA: TEORIA CLÁSSICA DOS TESTES E TEORIA DE RESPOSTA AO ITEM

A Psicometria iniciou o seu desenvolvimento no século XIX, junto com a consolidação da ciência com base no positivismo, com a necessidade de elaboração de instrumentos de avaliação psicológica que proporcionassem medidas válidas e precisas (SARTES; SOUSA-FORMIGONI, 2013). Tem fundamento em uma base epistemológica eminentemente quantitativista, assumindo pressupostos da teoria da mensuração, e é considerada como um ramo da psicologia, ou seja, das ciências empíricas (PASQUALI, 2009). Tem o objetivo de realizar medidas objetivas dos aspectos mentais. Pode ser definida como um conjunto de métodos e técnicas que utilizam de parâmetros métricos indispensáveis à medição das variáveis psicológicas independentes do campo de aplicação e dos instrumentos empregados (MUÑIZ, 1994).

Arias, Lloreda, Lloreda (2014) afirmam ser consenso, embora existam críticas, que a Psicometria como teoria da medição pode ser considerada como o processo de atribuição de números aos atributos dos sujeitos de modo a revelar seus diferentes graus. É importante ressaltar que a Psicometria é um termo bastante amplo e se refere a pelo menos cinco áreas especializadas, a saber: teoria da medição, teoria dos testes, escalas psicológicas, escalas psicofísicas e técnicas multivariadas (MUÑIZ, 1994). Neste texto será tradada, por especificidade, da Teoria dos Testes.

Inicialmente a Teoria dos Testes fundamentou-se unicamente no enfoque clássico, a TCT. No entanto, com os estudos que buscaram definir a estrutura dos traços latentes como causadores do comportamento observável surge a TRI (SARTES; SOUSA-FORMIGONI, 2013).

A TCT predominou por muito tempo, até meados da década de 1980. Muito por conta do seu modelo matemática relativamente simples. A TRI, por outro lado, possui um modelo matemático mais sofisticado e, embora tenha surgido bem antes, só começou a ser efetivamente utilizada na validação de instrumentos e análise dos resultados a partir dos anos de 1980 com o avanço tecnológico que permitiu a criação de programas de computadores que possibilitassem sua implementação (SARTES; SOUSA-FORMIGONI, 2013). Para melhor compreensão do desenvolvimento da psicometria apresenta-se a seguir um breve histórico dos modelos de TCT e TRI.

4.1 Breve história e desenvolvimento da teoria dos testes

O desenvolvimento dos testes, ao considerar uma perspectiva histórica, tem seu início em tempos bem remotos, desde a Antiguidade. Registra-se que já em 3.000 a.C o Império Chinês utilizava-se de testes para selecionar bons soldados para compor o exército (URBINA, 2004).

A teoria dos testes surgiu no início do século XX, inicialmente com os trabalhos de Spearman em 1904, ao propor o modelo linear clássico, na tentativa de fundamentação dos escores e da estimação dos erros de medidas associados em um teste (MUÑIZ, 1994; TRAUB, 1997), também teve contribuições de George Udny Yule, Truman Lee Kelley e outros (TRAUB, 1997). Esse modelo tornou-se o pressuposto fundamental da TCT, visto adiante com mais detalhes.

Alfred Binet e Théodore Simon tiveram papel de "pai fundador" da TRI (LINDEN, 2015). Binet foi solicitado a desenvolver na cidade de Paris um teste capaz de diferenciar estudantes com retardo mental e direcioná-los para a educação especial. A partir, disso, Binet pensou formas de medir uma variável não observável, do qual não se tem acesso direto, o que hoje denominamos de traço latente, variável latente ou teta. Em 1905, apenas um ano após a publicação de Spearmam em 1904, Binet publica um trabalho em que estão explícitas essas ideias, ou seja, os modelos de TCT e TRI sugiram no mesmo período.

Gulliksen, na década de quarenta apresentou uma sistematização da TCT, mostrando toda a sua estrutura, em que apresentou os postulados do modelo clássico detalhadamente (GULLIKSEN, 1943).

Também na mesma década Stevens (1946), ao apresentar as escalas de medidas, indica uma solução para os problemas referentes à mensuração de sensações humanas até então elencadas pela *Committee of the British Association for the Advancement of Scienc*. Entre os principais problemas era a própria definição de medição. Com isso, Stevens propôs a definição de medida sob uma variedade de formas, que estão associadas às propriedades da operação empírica com o objeto e às propriedades matemáticas das escalas. Dessa forma, as análises estatísticas empregadas deveriam levar em consideração a natureza da escala em que o objeto está sendo medido (STEVENS, 1946).

Thurstone também apresentou importantes contribuições para a teoria dos testes, principalmente ao publicar seu livro sobre análise fatorial, proporcionando grandes avanços na verificação da validade dos testes, sendo potencializados com a implementação de recursos computacionais (MUÑIZ, 1994). Thurstone também contribuiu para a TRI. Diferentemente de

Binet, ele desprendeu a inteligência da idade e construiu uma escala própria, impondo uma curva conhecida, a curva de distribuição cumulativa, em que os valores de localização estimados foram utilizados como valores de escala para os itens (LINDEN, 2015).

Embora Binet e Thurstone tenham dado importantes contribuições para o início da TRI, foi com o trabalho de Lord e Novick (1968), *Statistical Theories of Mental Tests Scores*, que é dado o início formal da TRI, pois esta obra marca o antes e depois da Teoria dos Testes, ou seja, marca o início da psicometria moderna, a Teoria de Resposta ao Item (MUÑIZ, 1994). Desde então a TRI tem sido muito pesquisada e expandida.

Após uma breve histórico do desenvolvimento da psicometria, buscou-se nas próximas subseções realizar uma breve descrição dos dois modelos, a TCT e TRI, apresentando o modelo matemático e seus pressupostos.

4.2 Teoria Clássica dos Testes (TCT)

O modelo da TCT, ou Modelo Linear Clássico, foi inicialmente desenvolvido por Spearman (MUÑIZ, 1994; PASQUALI, 2009) e axiomatizada por Gulliksen (GULLIKSEN, 1950). Alguns autores também apresentaram um resumo do modelo sistemático do modelo (LORD, 1959; NOVICK, 1966). Com base nesses autores será apresentado os fundamentos da TCT. São três os elementos que constituem o postulado fundamental da teoria:

$$T = V + E$$

Ou seja, o escore empírico é a soma do escore verdadeiro mais o erro, que se define como (GULLIKSEN, 1950):

T = escore bruto ou empírico do sujeito, que é a soma dos escores obtidos no teste;
V = escore verdadeiro, que seria a magnitude real daquilo que o teste quer medir no sujeito e que seria o próprio T se não houvesse erro de medida e;
E = o erro cometido nesta medida.

É importante ressaltar que o erro de medida está presente em qualquer operação empírica, assim, o objetivo da TCT é dispor de técnicas estatísticas que visem controlar ou predizer o tamanho do erro na aplicação dos testes (MUÑIZ, 1994; PASQUALI, 2009). Assim, é razoável assumir que erro é definido como a diferença entre o escore verdadeiro (a pontuação real do sujeito) e o escore observado (o escore do sujeito no teste), ou escore empírico (LORD, 1959).

É necessário também destacar que o erro é aleatório, assistemático, isto é, ao realizar várias vezes o teste, o sujeito obterá diferentes pontuações, e essas pontuações poderão ser maiores ou menores que a pontuação verdadeira. Portanto, como o erro é um evento casual, sabe-se que a média do erro é 0 (MUÑIZ, 1994).

São três os pressupostos fundamentais do modelo, dos quais não podem ser empiricamente comprovados de forma direta (MUÑIZ, 1994). O primeiro suposto diz que a pontuação verdadeira (V) é a esperança matemática da pontuação empírica (T): V = E(T). Isso significa que se o teste fosse aplicado infinitas vezes, a pontuação verdadeira seria a média da pontuação empírica. O segundo suposto diz que não há correlação entre a pontuação verdadeira e os erros de um teste, ou seja, a correlação é zero: p (V,E) = 0. Não há razões para supor que o tamanho da pontuação verdadeira está associado ao tamanho do erro, se assim fosse, o erro seria sistemático, portanto, controlável. O terceiro suposto assume que os erros em dois testes distintos não se correlacionam: p (E_j, E_k) = 0. Assim, se os testes são aplicados adequadamente, os erros serão aleatórios em cada teste.

Outro aspecto importante na TCT é a definição de testes paralelos. Assumindo que é possível elaborar, diz ser paralelos dois testes que medem a mesma coisa, mas com itens diferentes (MUÑIZ, 1994): são "tau equivalentes" quando possuem pontuações verdadeiras iguais e com variância de erro não necessariamente igual e "essencialmente tau equivalentes" quando as pontuações dos sujeitos são iguais em um e outro teste mais uma constante: $V_1 = V_2 + K$.

A partir desse modelo, algumas deduções imediatas são possíveis (MUÑIZ, 1994), conforme está apresentadas no Quadro 3.

Quadro 3 – Deduções imediatas do modelo clássico.

Deduções	Definição
$E = T - V$	O erro é igual o escore empírico menos o escore verdadeiro
$E(E) = 0$	A esperança matemática dos erros de média é zero
$\mu_T = \mu_V$	A média do escore empírico é igual a média do escore verdadeiro
Cov (V, E) = 0	A covariância entre escore verdadeiro e o erro é igual a zero
Cov (T, V) = Var (V)	A covariância entre escore empírico e verdadeiro é igual a variância do escore verdadeiro
Cov (Tj, Tk) = Cov (Vj, Vk)	
Var (T) = Var (V) + Var (E)	A variância do escore empírico é igual a variância do escore verdadeiro mais a variância do erro.
p (T, E) = σ_E/σ_T	A correlação entre o escore empírico e o verdadeiro é igual ao cociente dos desvios dos erros pelos desvios do escore empírico
$\mu_1 = \mu_2 = ... = \mu_k$ $\sigma^2 (T_1) = \sigma^2 (T_2) = ... \sigma^2 (T_k)$ p (T_1, T_2) = p (T_1, T_3) = ... p (Tj, Tk)	Para N testes paralelos as suas variâncias e sua correlações são iguais

Fonte: Adaptado de Muñiz (1994).

É importante ressaltar que esse modelo tem suas deduções com base nos parâmetros populacionais, dessa forma, é necessário amostras estritamente representativas, por tanto, suficientemente grandes para que os valores populacionais sejam adequadamente estimados (MUÑIZ, 1994). No modelo de TCT considera-se principalmente dois parâmetros dos itens, a dificuldade e a discriminação dos itens, detalhados a seguir.

4.2.1 *Parâmetros dos itens: dificuldade e discriminação*

A dificuldade dos itens segundo a TCT é definida como a proporção de sujeitos que respondem corretamente ao item (ANDRIOLA, 1998; VIANNA, 1976). Dessa forma, quanto mais sujeitos acertam determinado item, mais fácil ele é. Com isso, a dificuldade do item só é utilizada mais frequentemente em contexto de testes de aptidão, em que há apenas respostas certas e erradas (PASQUALI, 2009).

A discriminação na TCT é definida como a capacidade do item distinguir sujeitos de escores altos em relação àqueles de escores baixos ou diferenciar sujeitos de desempenho baixo e superior (PASQUALI, 2009; VIANNA, 1976). O índice de discriminação informa a coerência dos escores do item com os escores do teste, assim, quanto maior for o índice dos itens, maior será a homogeneidade do teste (SILVEIRA, 1983). O cálculo da discriminação do item pode ser obtido de duas formas (PASQUALI, 2009): 1) pelos grupos-critério, que em casos de testes de desempenho escolar utiliza-se como critério a diferença entre os 27% do grupo que obtiveram os melhores resultados e os 27% dos sujeitos com os resultados mais baixos e; 2) pela correlação do item com o escore total menos o escore do item (correlação item-total), que pode ser obtido através da correlação bisserial por ponto.

Além de considerar a dificuldade e a discriminação dos itens, a TCT considera duas características fundamentais na análise dos testes educacionais e psicológicos, a validade e a fidedignidade, apresentados com mais detalhes na próxima subseção.

4.2.2 *Validade e Fidedignidade dos testes*

Uma das grandes preocupações em relação aos instrumentos de avaliação é quanto a sua validade. É necessário que o teste seja adequado para o que se pretende medir. Um teste é válido quando mede aquilo que pretende medir (ANDRIOLA, 1998; URBINA, 2004; VIANNA, 1976). Mesmo assim, a questão da validade é necessária e generalizável a todos os testes, pois este deve realizar satisfatoriamente a medida para o qual foi construído (REQUENA, 1990). Alguns autores (PASQUALI, 2009; REQUENA, 1990; VIANNA, 1976)

falam especificamente de três tipos de validade dos testes: 1) validade de conteúdo, que refere à representatividade da amostra dos conteúdos e comportamentos de um determinado teste; 2) validade de critérios se refere a capacidade de um teste predizer um desempenho específico de um sujeito e; 3) validade de construto, que tem o objetivo de identificar em que medida as respostas de um teste tem um significado e determinar o grau de consistência na relação empírica do teste com esse significado (REQUENA, 1990).

No entanto, atualmente, essas categorias clássicas de validade têm sido fortemente criticadas. A partir disso, a American Educational Research Association (AERA), American Psychological Association (APA) e a National Council on Measurement in Education (NCME) lançam um novo entendimento sobre o conceito de validade. Assim, a validade passa a ser entendida como a concordância entre as evidências encontradas e a teoria de modo a oferecer suporte para a intepretação dos escores de um teste (AERA; APA; NCME, 2014). Desse modo, amplia-se as fontes de evidência para além das três categorias clássicas, validade de conteúdo, critério e construto. Diante disso, é possível elencar pelo menos 5 fontes de evidência, são os baseados: i) no conteúdo do teste; ii) nos processos de resposta; iii) na estrutura interna; iv) na relação com outras variáveis e; v) nas consequências da testagem (AERA; APA; NCME, 2014).

A fidedignidade ou precisão é um dos conceitos mais importantes da TCT (ANDRIOLA, 1998). Refere-se ao grau com que a medida é realizada com o mínimo de erro possível repetidas vezes com os mesmos sujeitos, produzindo resultados idênticos (URBINA, 2004; VIANNA, 1976). Segundo Vianna (1976) fidedignidade de um teste pode ser definido como o grau de estabilidade dos resultados, ou seja, a consistência interna dos escores, em que aplicando o instrumento diversas vezes nos mesmo sujeitos se produz os mesmo resultados. Uma questão importante na fidedignidade do teste é o erro, como aponta Maroco e Garcia-Marques (2006), pois quanto mais uma medida é ausente de erro mais consistente ela é, portanto, mais confiável. O erro tem basicamente duas fontes, uma interna, como suas condições física, psicológicas, e outra externa ao examinando, como as condições do ambiente de testagem (AERA; APA; NCME, 2014).Uma preocupação importante é com o impacto das tomadas de decisão decorrentes do processo de medida. Nesse sentido, quanto mais importante forem as decisões tomadas a partir da informação do teste mais precisão requer sua medida (AERA; APA; NCME, 2014).

Nesta seção foi apresentada as principais características da TCT, com ênfase nos parâmetros do itens, dificuldade e discriminação, e nas características dos testes, a validade e a

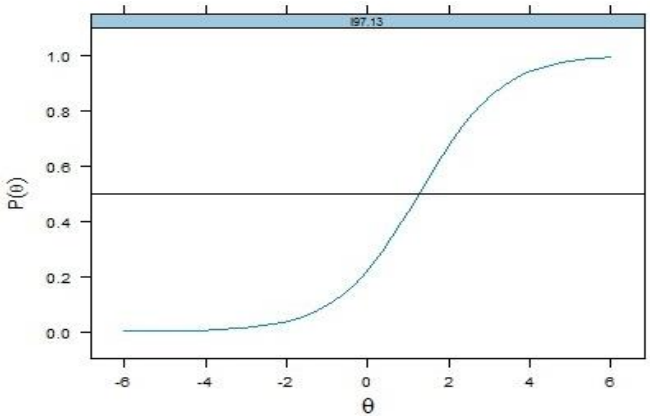
fidedignidade. Agora será apresentado o modelo de TRI e seu principal pressuposto, a unidimensionalidade.

4.3 Teoria de Resposta ao Item (TRI)

A TRI surge no âmbito da psicometria moderna como resposta aos problemas apresentados pela TCT e como complemento a esta e não como modelo substitutivo (ANDRIOLA, 2009). Esse modelo trabalha com o traço latente, que é o teta (θ), fenômeno psíquico, como critério a partir dos itens do teste (variáveis observáveis), assim, a qualidade do teste é determinada em função dos itens, com isso, ela objetiva construir itens de qualidade (PASQUALI, 2009), possui foco individualizado nos itens de um teste ou banco de itens (ANDRIOLA, 2009). A TRI sugere que a probabilidade de acertar o item decorra dos seus parâmetros e do traço latente (ou aptidão) do indivíduo exigida em um teste (VALLE, 2000). Dessa forma, esse modelo apresenta uma vantagem em relação ao clássico, a possibilidade de comparabilidade dos resultados de indivíduos de uma mesma população, mesmo que submetidos a itens diferentes (MUÑIZ, 1990; PASQUALI, 2009; VALLE, 2000).

Segundo Pasquali (2009), para essa teoria, a probabilidade de acerto a um item aumenta em indivíduos que apresentam maior aptidão e vice-versa. Essa probabilidade ou relação funcional é representada pelo que se denomina de Curva Característica do Item (CCI) (ANDRIOLA, 2009), como mostra a Figura 1.

Figura 1 - Curva Característica do Item.



Fonte: Elaborado pelo autor.

Como se pode observar no gráfico, na curva em formato de *S* ascendente, que é a característica do item, a probabilidade de acertar o item [P(θ), na ordenada] aumenta em indivíduos com maior aptidão (θ, na abscissa). Essa curva pode ser afetada por vários parâmetros dependendo do modelo utilizado (MUÑIZ, 1990). Nos modelos mais usuais de TRI

os itens apresentam três parâmetros, dificuldade, discriminação e acerto casual, descritos a seguir.

4.3.1 *Parâmetros da TRI: dificuldade, discriminação e acerto casual*

O primeiro parâmetro a ser considerado no item, talvez o mais importante, é o da dificuldade, em que é representado pela letra *b*. Diferentemente da TCT, a TRI o considera na mesma escala do traço latente, ou seja, do teta (θ). Dessa forma, a dificuldade está relacionada ao nível do teta necessário para responder o item (LAROS, 2009).

O índice de discriminação também é um parâmetro a ser considerado nos itens. Ele informa a capacidade deste em distinguir sujeitos com habilidades (aptidão) distintas, sendo representado pela letra *a*. Assim, quanto mais o item consegue diferenciar sujeitos com magnitudes próximas de habilidade, mais discriminativo ele é (LAROS, 2009).

Outro parâmetro do item considerado é o acerto ao acaso. Refere-se à probabilidade de acertar o item quando não se tem aptidão suficiente (MUÑIZ, 1990) e é representado pela letra *c*. Como este parâmetro se dá em termos de proporção, ele pode variar de 0 a 1. Quanto maior o valor, maior a probabilidade de os indivíduos acertarem o item ao acaso, dado uma aptidão inferior à dificuldade do item (MUÑIZ, 1990).

Antes de estimar os parâmetros dos itens e dos participantes é necessário verificar se os mesmos estão medindo um mesmo traço latente, ou seja, se são unidimensionais. Esse pressuposto será descrito com mais detalhes.

4.3.2 *Pressupostos da TRI: unidimensionalidade e independência local*

Ao analisar um conjunto de itens a partir da TRI, estes devem dispor de algumas características para que ocorra um adequado ajuste do modelo pretendido. São dois, a unidimensionalidade e independência local. Supõe-se que ao aplicar um teste, ou seja, um conjunto de itens, a probabilidade de acertá-los dependerá unicamente do traço latente do sujeito, do seu θ (MUÑIZ, 1990). Dessa forma, pressupõe-se que os itens estejam medindo um único traço latente, ou seja, que sejam unidimensionais. Este pressuposto "[...] é uma proposição teórica parcimoniosa e elegante, segundo o qual toda a complexidade intrínseca ao ato de resolução de um problema – de natureza cognitiva ao não – deve-se como causa uma única estrutura latente [...]" (ANDRIOLA, 2009 p. 327).

Outro pressuposto da TRI é o da independência local. Supõe-se que as respostas a um determinado item não sejam influenciadas por outros itens. Se existe independência local, a probabilidade de acertar um conjunto de itens é igual ao produto da probabilidade de acertar cada um destes (MUÑIZ, 1990; PASQUALI, 2009). Atendendo a esses pressupostos, três modelos de TRI comumente utilizados em testes educacionais são possíveis de serem aplicados aos dados. Esses modelos são melhor detalhados na seção posterior.

4.3.3 *Modelos logísticos da TRI*

Existem vários modelos de TRI na literatura e, como coloca Valle (2000), esses dependem fundamentalmente de três fatores: a natureza dos itens (dicotômicos ou não), o número de populações envolvidas e a quantidade de traços latentes medidos (quando mede mais de um denominam-se modelos multidimensionais). Ressalta a autora que os modelos unidimensionais para itens dicotômicos são os mais utilizados. Os modelos logísticos são classificados de acordo com os parâmetros utilizados (MUÑIZ, 1990; PASQUALI, 2009; VALLE, 2000): de um parâmetro: dificuldade; de dois parâmetros: dificuldade e discriminação; de três parâmetros: dificuldade, discriminação e probabilidade de acerto ao acaso.

O modelo de 1 parâmetro utiliza apenas o parâmetro de dificuldade dos itens, ou seja, a resposta ao item depende apenas deste e da aptidão do indivíduo, ou seja, da variável latente (MUÑIZ, 1990). O modelo de 2 parâmetros considera a dificuldade e a discriminação do item, em que a função logística deve considerar estes dois parâmetros (MUÑIZ, 1990; PASQUALI, 2009). O modelo de 3 parâmetros considera a dificuldade, discriminação e a probabilidade de acerto ao acaso (MUÑIZ, 1990; PASQUALI, 2009). Para Andrade, Tavares e Valle (2000) esse modelo é o mais utilizado atualmente.

A expressão matemática utilizada para esses modelos são:

Modelo Logístico de 1 parâmetro	$Pi\ (\theta) = \dfrac{e^{D^{(\theta - b_i)}}}{1 + e^{D^{(\theta - b_i)}}}$
Modelo Logístico de 2 parâmetros	$Pi\ (\theta) = \dfrac{e^{Da_i{}^{(\theta - b_i)}}}{1 + e^{Da_i{}^{(\theta - b_i)}}}$
Modelo Logístico de 3 parâmetros	$Pi\ (\theta) = c_i + (1 - c_i)\ \dfrac{e^{Da_i{}^{(\theta - b_i)}}}{1 + e^{Da_i{}^{(\theta - b_i)}}}$

em que,

Pi (θ) = é a probabilidade de um sujeito acertar o item dado um determinado teta;
θ = é o nível de habilidade do sujeito;
i = é o número do item no teste;
a = índice de discriminação;
b = índice de dificuldade do item;
c = índice de acerto casual;
e = uma base logarítmica de valor 2,72;
D = é uma constante de valor 1,7.

4.3.4 *Métodos de estimação dos parâmetros dos itens e da habilidade dos sujeitos*

Na prática de medidas educacionais um dos principais objetivos é obter os parâmetros dos itens e conhecer a habilidade ou proficiência dos sujeitos (BAKER, 1992). No entanto, de fato, a única informação que conhecemos são as respostas dos sujeitos aos itens. Com isto, temos um vetor de respostas de cada sujeito. No caso de testes de rendimento o vetor é constituído por zeros (0) e uns (1) que representam os itens dicotomizados em erros e acertos, respectivamente. Dessa forma, o que podemos obter são as probabilidades de ocorrência de cada vetor. A partir disso, podemos obter os parâmetros dos itens e a habilidade dos sujeitos (BAKER, 1992).

Entre os procedimentos mais utilizados para a estimação dos parâmetros dos itens estão os métodos de máxima verossimilhança (*maximum likelihood - ML*): i) máxima verossimilhança conjunta (*joint maximum likelihood – JML*), ii) máxima verossimilhança marginal (*marginal maximum likelihood – MML*), e iii) máxima verossimilhança condicional (*conditional maximum likelihood – CML*) (EMBRETSON; REISE, 2000).

Os métodos de ML apresentam vantagens com o aumento da amostra, tais como: i) convergência para o valor verdadeiro, ii) erro padrão minimizado, iii) normalidade na distribuição do erro de estimativa (EMBRETSON; REISE, 2000).

De modo geral os métodos de estimação dos parâmetros exigem pressupostos a serem atendidos. Os dois principais são a independência local e dimensionalidade apropriada (AYALA, 2009). Especificamente, apresenta independência local quando a resposta a um item ocorre em função de outro item e não do traço latente dominante. Apresenta dimensionalidade

apropriada quando o modelo se adequa bem aos dados e consegue prever o nível de traço latente dos sujeitos, ou seja, quando as respostas aos itens dependem unicamente de sua habilidade (EMBRETSON; REISE, 2000).

Os métodos de estimação dos parâmetros a partir dos três métodos de *ML* serão descritos a partir de Embretson e Reise (2000). O método *JML* é um processo de estimação dos parâmetros dos itens quando o nível de traço latente dos sujeitos é desconhecido. Parte-se de valores provisórios de traço latente para, a partir disso, estimar valores de parâmetros dos itens e, posteriormente, melhorar as estimativas do traço latente. Posteriormente, o processo é repetido. Dessa forma, aplica-se sucessivas iterações em busca de estimativas melhoradas dos parâmetros dos itens. Como coloca Ayala (2009) esse "ping-pong" de estágios é um processo contínuo que ocorre até que a melhoria obtida na estimação seja insignificativa.

Para Embretson e Reise (2000) o método de *JML* apresentam vantagens e desvantagens. Entre as vantagens estão: i) algoritmo facilmente programável, ii) é aplicável a muitos modelos de TRI, iii) computacionalmente eficiente. Entre as desvantagens estão: i) as estimativas dos parâmetros não apresentam propriedades adequadas, ii) o significado do erro padrão é questionável, iii) não produz estimativas para pontuações perfeitas (tudo 0 ou tudo 1), iv) as hipóteses sobre o modelo têm utilidade questionável.

No método *MML* os níveis de traços latentes desconhecidos são expressos em termos de probabilidade do padrão de resposta, considerando que estas são uma amostra aleatória de uma distribuição populacional (AYALA, 2009; EMBRETSON; REISE, 2000). Dessa forma, esse método modela a probabilidade de se observar determinado padrão de resposta na população. Diante disto, para estimar os parâmetros deve-se especificar a forma de distribuição populacional. Normalmente, assume-se que se distribui normalmente (EMBRETSON; REISE, 2000).

O modelo de *MML* tem como vantagens (EMBRETSON; REISE, 2000): i) é aplicável a todos os modelos de TRI; ii) mostra-se eficiente tanto para testes longos como curtos; iii) as estimativas de erro padrão do item são boas aproximações da variância das estimativas; iv) são possíveis as estimativas para padrões de respostas perfeitos (tudo zero ou tudo um); v) o modelo é útil para testes de ajuste, por exemplo, na comparação entre dois modelos. Todavia, as suas desvantagens são: i) é difícil programar o algoritmo desse modelo, mas vale ressaltar que atualmente com o avanço computacional, isso tem sido cada vez menos problemático; ii) a estimativa pressupõe assumir uma determinada distribuição populacional,

no entanto, isso não tem sido necessariamente um problema, já que se pode assumir que não há distribuição normal, além disso, esse modelo é robusto frente a pequenos desvios da normal.

Por fim, no método *CML*, quando há uma estatística suficiente, ou seja, quando não há necessidade de outra informação, as habilidades desconhecidas dos sujeitos são estimadas para o padrão de respostas sem ser necessário recorrer aos seus parâmetros (EMBRETSON; REISE, 2000). Neste caso, o escore total é uma estatística suficiente no modelo de Rasch e seus derivados, ao contrário de modelos mais complexos, como os de dois e três parâmetros, por exemplo. Deste modo, esse método de estimação é aplicado apenas a esses modelos de TRI.

Entre as vantagens do método de *CML* estão (EMBRETSON; REISE, 2000): i) não assume a distribuição populacional dos níveis de traço latente; ii) a partir disso, ao estimar os parâmetros dos itens, esse método atende ao princípio da invariância, já que não é influenciado pela amostra; iii) apresenta propriedades desejáveis em muitas condições, já que se mostra consistente e normalmente distribuído, mesmo quando alguns pressupostos não são atendidos. Por outro lado, as desvantagens são: i) é aplicável apenas aos modelos de Rach; ii) não estima os parâmetros de padrões de resposta perfeitos; iii) há problemas com estimativas de testes longos.

Quando se conhece os parâmetros dos itens, a habilidade dos sujeitos pode ser obtida via três métodos (EMBRETSON; REISE, 2000): i) máxima verossimilhança (*maximum likelihood – ML*), ii) máxima a posteriori (*maximum a posteriori – MAP*), e iii) estimativa a posteriori (*estimated a posteriori – EAP*). No processo de estimativa dos escores dos sujeitos, frequentemente se assume que estes proveem de uma população distribuída normalmente (EMBRETSON; REISE, 2000).

O método *ML*, como o nome indica, é um procedimento que encontra um escore que maximiza a probabilidade de ocorrência do padrão de resposta dos sujeitos (AYALA, 2009; HAMBLETON; SWAMINATHAN; ROGERS, 1991). Para tanto, é necessário calcular as probabilidades de resposta (acerto ou erro) individuais dos itens. Um dos problemas dessa estimação é que a probabilidade tende ao infinito quando se tem padrões de respostas com tudo 0 ou tudo 1 (EMBRETSON; REISE, 2000). No entanto, os autores ressaltam que esse procedimento e menos tendencioso ao erro.

Para evitar esses problemas, é possível incorporar no procedimento informações a priori sobre os itens. Isso é possível com o método *MAP*, que é uma estimativa *baysiana* em que o pesquisador faz uso de informações anteriores para se chegar ao nível de traço latente dos sujeitos (EMBRETSON; REISE, 2000). A informação prévia comumente usada é considerar

que a amostra provém de uma populacional normalmente distribuída, ou seja, com média 0 e desvio padrão 1. Utilizando essa informação prévia, o método MAP encontra o escore que maximiza a probabilidade de ocorrência do padrão de resposta do sujeito (EMBRETSON; REISE, 2000).

Já o método *EAP* é não-operatório, assim facilmente calculado pelos computadores. A diferença entre o *EAP* e *MAP*, é que este calcula com base na moda da distribuição, enquanto o outro o faz com base na média da distribuição a posteriori (EMBRETSON; REISE, 2000). Isso permite com que as probabilidades para padrões de respostas uniformes sejam possíveis de serem estimados (EMBRETSON; REISE, 2000).

Até esse momento foi descrito os modelos de TCT e TRI. Agora será apresentado algumas problematizações acerca da dimensionalidade dos testes educacionais, que tem uma característica peculiar, os itens são corrigidos em certo e errado, portanto são por natureza variáveis dicotômicas. Assim, é apresentado os métodos de análise da dimensionalidade disponíveis, ressaltando os mais indicados para essa finalidade.

4.4 Dimensionalidade de testes educacionais

As avaliações em larga escala brasileiras estão adotando cada vez mais os modelos baseados em TRI para a estimação das proficiências dos participantes (ver Quadro 1). Para tanto, os itens devem atender ao pressuposto da unidimensionalidade, o que garante a adequação dos modelos para itens dicotômicos frequentemente utilizados nessas avaliações.

A unidimensionalidade de um conjunto de itens pode ser entendida como um traço latente responsável pela variabilidade presente nesses itens (ZIEGLER; HAGEMANN, 2015), ou quando uma dimensão (traço latente) explica a performance no teste (BEJAR, 1980), quando os resíduos dos itens não se correlacionam, ou seja, os itens se ajustam a um modelo comum, no caso de itens dicotômicos, um traço latente (MCDONALD, 1981). Também há definições que apresentam a unidimensionalidade não como a presença de um fator dominante, mas no grau de interferência da multidimensionalidade nos itens (REISE; COOK; MOORE, 2015)

Em outras palavras, são unidimensionais os itens que mensuram o mesmo traço latente. Nos casos de avaliações em larga escala, é razoável considerar que itens de ciências mensurem esse construto, por exemplo. No entanto, o Enem apresenta dificuldades, pelo menos do ponto de vista teórico (a priori), de satisfazer a essa característica.

Como apresentado e discutido por alguns autores (CONDE; LAROS, 2007; TAVARES, 2013), esse pressuposto é difícil de ser atendido estritamente, considerando que os fenômenos humanos, principalmente os ligados ao processo de ensino e aprendizagem, são multideterminados. No entanto, consideram ser de bom senso que basta haver uma dimensão dominante, ou seja, uma "unidimensionalidade essencial" para que o pressuposto seja atendido (STOUT, 1990). Nesse sentido, a unidimensionalidade, seria considerada como uma questão de grau. Todavia, é necessário que critérios sejam estabelecidos para se admitir haver adequação a esse postulado.

É com esse objetivo que esta seção se apresenta, ou seja, apresentar os critérios presentes na literatura para se considerar um conjunto de itens como unidimensional. Além disso, buscar-se-á apresentar as principais técnicas estatísticas utilizadas para a análise desse pressuposto.

Na literatura, muitas técnicas estão disponíveis para a análise da dimensionalidade de um teste. Algumas com ressalvas na sua utilização e outras como maiores adequabilidades ou consistência nos resultados. Será brevemente apresentado essas técnicas e ao final indicar a que se julgou como mais adequada para os objetivos desta tese.

Estão dispostos na literatura cinco grupos de testes para determinar a unidimensionalidade (HATTIE, 1985): os baseados nos padrões de resposta, fidelidade, análise dos componentes principais, análise fatorial e índices baseados nos modelos de traços latentes ou TRI. Todos os índices apresentam problemas, no entanto, os dois últimos índices têm maior aceitação e são mais utilizados entre os pesquisadores. As características são postas a seguir, conforme é discutido por Hattie (1985).

Os índices baseados nos padrões de resposta partem da ideia de um teste perfeitamente unidimensional, identificado em função da quantidade de desvio que um conjunto de respostas dos itens apresentam em relação a uma escala ideal. Esses índices apresentam várias críticas. Uma delas é que a unidimensionalidade é satisfeita apenas quando os itens se aproximam de uma escala perfeita. Além disso, não apresenta métodos para se identificar a presença de outras dimensões ou fatores medidos pelo teste. Argumenta-se ainda, que uma escala perfeita não necessariamente indica que os itens estejam medindo uma única dimensão. Entre outros motivos, esses métodos não têm sido utilizados para a análise de unidimensionalidade (HATTIE, 1985).

Os índices pautados na fidelidade baseiam-se na ideia de que itens com alta consistência interna (o mais comum deles é o coeficiente alfa proposto por Cronbach)

apresentam grande parte da variância compartilhada, dessa forma, indica a presença de um único fator sendo medido por esses itens. No entanto, críticas são apresentadas indicando que o coeficiente alfa alta não indica unidimensionalidade, tendo em vista que o primeiro é afetado por outros fatores, tais como o número de itens no teste, quantidade de fatores medidos pelo mesmo item, pelas comunalidades dos itens, entre outros (HATTIE, 1985).

Outros dois índices apresentados baseiam-se no modelo fatorial, a análise de componentes principais e a análise fatorial. A diferença entre os dois modelos é que o primeiro extrai os componentes (dimensões) a partir da diagonal principal de uma matriz de correlações dos itens. Já a análise fatorial extrai os fatores a partir de uma matriz reduzida, de modo que a variância de cada item se decompõem em uma parte comum e uma parte própria de cada item (HATTIE, 1985).

As medidas de unidimensionalidade pautados na análise de componentes principais indicam quanto de variância é explicada pelo primeiro fator. Se for um percentual grande da variância, provavelmente os itens são unidimensionais. O problema tem sido no quão grande deve ser o percentual de variância explicada pelo primeiro componente para que seja considerado unidimensional. Alguns apontam 40%, outros 20%, mas sem justificativa plausível (HATTIE, 1985). Há também critérios baseados nos autovalores, considerando a diferença do autovalor do primeiro fator e o do segundo, mas ainda sem considerar o valor máximo fixo (HATTIE, 1985).

Por último, os índices baseados na teoria do traço latente, consideram que a probabilidade de resposta a um item está representada em uma função logística de três parâmetros. O pressuposto fundamental e mais crítico do modelo é o da independência local. Esse pressuposto indica que quaisquer pares de item não devem estar correlacionados. Supõe-se que o traço latente é o único fator importante para a resposta aos itens. Parece razoável admitir que quando um conjunto de itens não apresentam independência local, sejam unidimensionais. Todavia, um conjunto de itens podem apresentar independência local, mesmo que medindo duas dimensões (HATTIE, 1985).

Os índices para análise da unidimensionalidade pautados na teoria do traço latente têm sido os melhores e mais adequados (BARTHOLOMEW, 1980). Um modelo matemático baseado na teoria foi proposta e denominada de *Full Information Factor Analysis* - FIFA (BOCK; GIBBONS; MURAKI, 1988), ou Análise Fatorial de Informação Plena.

A FIFA foi recomendada para analisar a estrutura dimensional do Enem de 1999 (NOJOSA, 2002). A referida edição do exame foi construída sob uma matriz de 5 competências.

Atualmente, o Enem é constituído de quatro áreas de conhecimento, Linguagens e Códigos, Matemática, Ciências Humanas e Ciência da Natureza, além de uma redação, como pode ser observado no Quadro 2.

Outro método tem sido utilizado para a determinação da dimensionalidade, a Análise Paralela. No entanto, há evidência de problemas desse método em dados dicotômicos (TRAN; FORMANN, 2009). Uma forma de minimizar os problemas é aplicar a análise em uma matriz de correlação tetracórica. Apesar desse problema a Análise Paralela oferece um indicativo da dimensionalidade dos itens a partir do *scree plot* (DRASGOW; LISSAK, 1983; ZWICK; VELICER, 1986).

Dessa forma, será utilizado a Análise Paralela e a FIFA para analisar a dimensionalidade do exame. Mais detalhes serão apresentados a seguir.

4.4.1 *Análise Paralela*

O método da Análise Paralela foi desenvolvido por Horn baseado no modelo simples da regra de Kaiser (ZWICK; VELICER, 1986). Esse método tem sido citado na análise da dimensionalidade de testes com itens dicotômicos (DRASGOW; LISSAK, 1983; WENG; CHENG, 2005).

Este é um método relativamente simples e consiste em comparar os autovalores a partir de uma matriz de correlação, no caso de dados dicotômicos uma matriz de correlação tetracórica, e comparar com os autovalores obtidos para dados simulados com mesmo número de variáveis e tamanho amostral (LEDESMA; VALERO-MORA, 2007).

O critério para a determinação do número de fatores consiste no contraste entre os autovalores dos dados reais com os dos dados simulados, então observa-se o número de fatores em que os autovalores dos dados simulados são maiores que os dos dados reais, então esses são retidos (ZWICK; VELICER, 1986).

Hayton, Allen e Scarpello (2004) sugerem 4 etapas para a retenção do número de fatores a partir da Análise Paralela: i) gerar os dados aleatórios; ii) extrair os autovalores da matriz de correlação dos dados aleatórios; iii) Extrair os autovalores médios e o percentil 95 de todos os autovalores na matriz de dados aleatórios; iv) comparar com os autovalores dos dados reais.

4.4.2 *Análise Fatorial de Informação Completa*

Métodos de análises fatoriais têm sido amplamente adotadas na análise de itens. No entanto, essas análises não têm satisfeito plenamente e apresentado muitos problemas como os casos de Heywood, muito por conta das características dos itens que se apresentam de forma dicotômica ou politômica, comumente encontrados em testes educacionais, embora algumas soluções como a realização da análise sobre uma matriz tetracórica tenham sido implementadas (BOCK; GIBBONS; MURAKI, 1988). Como alternativa para a resolução desses problemas, foi apresentado um método baseado em TRI, a *Full Information Factor Analysis* (BOCK; AITKIN, 1981).

Atualmente, a Análise Fatorial de Informação Plena é comumente utilizada para a análise da dimensionalidade de variáveis dicotômicas ou politômicas, pois supera muitos dos problemas enfrentados com dados com essas características. O modelo pode ser interpretado como uma aproximação dos modelos convencionais de análise fatorial com o modelo de teoria de resposta ao item (NOJOSA, 2002).

Essa análise utiliza todos os vetores de resposta aos itens, por isso é denominada de "Full Information" (informação completa) ao contrário dos métodos baseados nas frequências de ocorrência conjunta das pontuações dos itens, denominados como de informação limitada (BOCK; GIBBONS; MURAKI, 1988) .

A Análise Fatorial de Informação Plena (FIFA) está implementada no pacote estatístico "mirt" (Multidimensional Item Response Theory) disponibilizado para o ambiente do *software R* (CHALMERS, 2012).

Apresentado o referencial teórico e metodológico que fundamentou a realização dessa pesquisa, a seguir é exposto os procedimentos metodológicos levados a cabo para atingir os objetivos e encontrar respostas para os problemas levantados anteriormente na introdução deste trabalho.

5 PROCEDIMENTOS METODOLÓGICOS

5.1 Tipologia da pesquisa

Esta pesquisa apresenta abordagem quantitativa e utiliza o método estatístico. A abordagem quantitativa caracteriza-se pelo emprego da quantificação e pela aplicação de testes estatísticos como forma de aproximação ao problema de pesquisa (RICHARDSON, 2012). Esta característica está presente neste estudo, considerando que tem como objetivo central a comparação de dois modelos matemático empregados na avaliação de testes psicológicos e educacionais assim como na interpretação das pontuações dos examinandos obtidas nesses testes. O método estatístico reduz os fenômenos a termos quantitativos de modo que permita a manipulação estatística e a comprovação da relação entre os fenômenos entre si (MARCONI; LAKATOS, 2003).

5.2 Fonte dos dados

Os microdados dos resultados das provas e as respostas de cada candidato nos itens do Enem 2017 estão disponíveis no site Inep em planilhas do *software* SPSS (*Statistical Package for Social Sciences*) e são de livre acesso ao público.

5.3 População e amostra

A população deste estudo é constituída 6.731.341 candidatos de todas as regiões e estados do Brasil que realizaram o Enem 2017. Foram excluídos os candidatos que não estiveram presentes no exame e os que não responderam a nenhum dos itens do exame permanecendo 4.426.755.

A amostra final desta pesquisa é formada por 10.000 participantes do Enem da edição de 2017 que estiveram presentes em todas as provas do exame. Esses participantes foram selecionados por amostragem aleatória simples. Esse número reduzido da amostra em relação ao quantitativo total dos participantes se deu pela capacidade limitada de processamento dos computadores utilizados nessa pesquisa. O volume muito grande de dados não foi suportado pela capacidade operacional, gerando instabilidade no processamento das máquinas. Considerando o tamanho populacional, o erro estimado foi menor que 1% para um intervalo de confiança de 95%.

Destaca-se ainda que se utilizou apenas um dos quatro cadernos de prova de cada área. Para Linguagens e Códigos e Ciências Humanas utilizou-se o caderno azul, para Ciências Humanas o caderno amarelo e para Matemática o caderno cinza.

A seleção da amostra foi realizada para cada caderno de prova separadamente, ou seja, a amostra de candidatos de um caderno não é a mesma de outra. Para o caderno de Linguagens e Códigos, o candidato opta por Língua Inglesa ou Espanhola. Neste estudo foi realizado com os candidatos que escolheram a segunda opção de Língua Estrangeira. Todas as escolhas foram realizadas por seleção aleatória simples, ou seja, por sorteio. As características sócias demográficas das amostras estão nas Tabelas 1 e 2.

Tabela 1 – Idade dos candidatos, Enem 2017.

Idade	LC	MT	CN	CH
Média	22,63	21,80	21,76	21,78
Desvio padrão	7,59	7,19	7,12	7,28
Mínimo	14	14	13	14
Máximo	68	78	70	68
Amplitude	54	64	67	54

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

A idade média dos participantes do Enem selecionados para este estudo está em torno de 22 anos de idade e os limites de idade variaram entre 13 e 78 anos. Isso indica uma variabilidade muito grande dos participantes em relação a idade. Isso se refletiu no desvio padrão que variou entre 7,12 a 7,59 anos de idade.

As características dos participantes para cada prova do exame em relação as variáveis sexo, e região de residência (Nordeste, Norte, Centro-Oeste, Sudeste e Sul) estão na Tabela 2, a seguir.

Tabela 2 – Características sócio demográficas dos candidatos, Enem 2017.

Variáveis		LC	MT	CN	CH
		%	%	%	%
SEXO	MAS	39,9	37,5	41,7	41,3
	FEM	60,1	62,5	58,3	58,7
REGIÃO	NE	34,0	40,2	34,2	34,1
	N	11,1	15,2	11,4	11,9
	SE	35,4	25,6	35,9	35,4
	CO	7,9	9,1	8,0	8,2
	S	10,6	10,0	10,5	10,3

Legenda: MAS: Masculino; FEM: Feminino; NE: Nordeste; N: Norte; SE: Sudeste; CO: Centro-Oeste; S: Sul. LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

5.4 Caracterização do exame

O Enem é uma prova constituída de itens de múltipla escolha e uma redação, na qual busca avaliar competências e habilidades desenvolvidas pelos alunos no decorrer da Educação Básica, sendo orientada por uma matriz de referência construída especificamente para o exame.

A matriz de referência atualmente é dividida em quatro grandes áreas do conhecimento: Linguagens, Códigos e suas Tecnologias, que contempla os conhecimentos de Português, Educação Física, Artes, Língua Estrangeira Moderna e Tecnologia da Informação e Comunicação; Ciências da Natureza e suas Tecnologias, que abrange a Biologia, Química e Física; Ciências Humanas e suas Tecnologias, envolvendo a História, Geografia, Sociologia e Filosofia e; Matemática e suas Tecnologias.

5.5 Tratamento para dados ausentes

Entre os participantes da amostra desse estudo ocorreram ausência de respostas a determinados itens das provas. No entanto, em nenhuma delas a quantidade de valores ausentes ultrapassou 1%. Para não eliminar os participantes das análises posteriores optou-se por substituir os valores ausentes pelo método de imputação múltipla.

O método de imputação múltipla consiste na substituição dos valores ausentes a partir de várias simulações até chegar a uma aproximação estatística ótima. Esse método é bastante robusto quando os valores ausentes são completamente aleatórios (NEWMAN, 2014) e em casos de dados binários ou dicotômicos (BÉLAND et al., 2018). Para a aplicação desse método utilizamos o *software* SPSS 20.0.

5.6 Análise estatística

Para todas as análises realizadas neste trabalho foi utilizado o *software* R, programa livre e amplamente utilizados pelos pesquisadores e psicometristas para implementação de análises estatísticas. Esse *software* incorpora uma ampla possibilidade e flexibilidade nas análises.

Inicialmente foram obtidas a discriminação dos itens a partir da TCT. Foram estimados os parâmetros de discriminação através da correlação ponto-bisserial, uma vez que os itens foram dicotomizados em certo e errado (PASQUALI, 2009). Para essas análises foi utilizado o pacote “ltm” (RIZOPOULOS, 2006). Itens com discriminação muito baixa (r_{pb} < 0,15) foram excluídos da análise, pois indica baixa correlação do item com o escore total.

Inicialmente a discriminação foi utilizada com finalidade de analisar sua adequação para a realização da análise fatorial exploratória. Posteriormente, foram utilizados para analisar os itens no modelo fatorial final.

Para uma análise exploratória da dimensionalidade submeteu-se os dados a um teste de Análise Paralela. Para aplicação deste teste utilizou-se o pacote "psych" (REVELLE, 2017). Essa análise consiste na comparação dos *eigen value* dos dados reais com os de um conjunto dados simulados gerados aleatoriamente com igual número de variáveis e de mesmo tamanho amostral (HAYTON; ALLEN; SCARPELLO, 2004). O critério para a determinação do número de fatores a serem retidos se baseia na comparação dos *eigen value* dos dados reais e dos dados gerados. Retém-se os fatores no momento em que o valor *eigen* dos dados reais é menor que o dos dados simulados. Para a análise da existência de uma dimensão dominante nos dados, realizou-se uma comparação do primeiro *eigen value* com o segundo (Eigen1/Eigen2).

Outra análise utilizada para verificar o pressuposto da unidimensionalidade foi a Análise Fatorial de Informação Plena, um modelo baseado na teoria do traço latente, considerada mais adequado em situações de itens dicotômicos (BARTHOLOMEW, 1980), típicos de testes educacionais, como é o caso deste trabalho que utilizara os dados da prova do Enem. Para essa análise foi utilizado o pacote estatístico "mirt" (CHALMERS, 2012).

O ajuste dos modelos foi realizado com base nas medidas do *Root-Mean-Square Error of Approximation* (RMSEA), *Standardised Root Mean Square Residual* (SRMSR), *Tucker-Lewis Index* (TLI) e *Comparative Fit Index* (CFI). Para essas análises foi utilizado o pacote estatístico "mirt" (CHALMERS, 2012). São considerados desejáveis para um bom ajuste dos modelos valores de RMSEA e SRMSR abaixo de 0,05. Já para os valores de TLI e CFI são considerados os valores acima de 0,95 como indicativos de bom ajuste.

Também foi analisado os valores dos índices de ajuste dos modelos com base no Critério de Informação de Akaike (AIC - Akaike's Information Criterion) e o Critério de Informação Baysiano (BIC - Bayesian Information Criterion) (NYLUND; ASPAROUHOV; MUTHÉN, 2007). O modelo que produz menores valores de ambos os critérios é o modelo de melhor ajuste. No entanto, o AIC tende a superestimar a quantidade de dimensões e o BIC a subestimar. O índice considerado neste estudo foi o BIC, pois considerado mais consistente que o índice AIC através de estudos de simulação Monte Carlo (NYLUND; ASPAROUHOV; MUTHÉN, 2007).

A análise do ajuste do modelo também foi realizado a partir do Índice de Dimensionalidade (ID) considerando as recomendações de Nojosa (2002). Consiste,

inicialmente, em obter os valores de ajuste do modelo com um fator (M1). Posteriormente estima-se os modelos com dois (M2), três fatores (M3) e assim sucessivamente, e então compará-los entre si.

Ao comparar os modelos obtêm-se um valor com uma distribuição qui-quadrado (X^2). No entanto, esse valor é superestimado. Recomenda-se então dividir esse valor por dois ou por três para um ajuste mais adequado (NOJOSA, 2002). Esse valor será denominado nesta tese de qui-quadrado corrigido (X^2corr). Esse valor então é dividido pelos graus de liberdade (gl). Nojosa (2002) ressalta que esse valor não é interpretável diretamente, pois só a diferença do valor X^2 entre os modelos deve ser considerado. Esse valor obtido será denominado de Índice de Dimensionalidade (ID). Na comparação de dois modelos, M1 x M2, um ID com valor positivo maior que 2,0, indica que o segundo modelo é melhor, se o valor for menor que 2,0, o primeiro modelo é preferível. Nesse sentido, espera-se que se os itens forem unidimensionais, a comparação entre os modelos 1 (M1) e 2 (M2) forneça um ID positivo menor que 2,0.

Após ajustar um modelo fatorial dos itens, foram estimados a estatísticas clássicas de discriminação e dificuldade dos itens e dos participantes. Foram estimados os parâmetros de discriminação através da correlação ponto-bisserial. Nesta fase da análise considerou-se como discriminativo os itens com valores acima de 0,20. Já a dificuldade do item varia de 0 a 1, em que quanto mais próximo de 0 mais difícil é o item. Esse parâmetro indica o percentual de acerto do item. Para essas análises foi utilizado o pacote "ltm" (RIZOPOULOS, 2006).

Para análise da fidedignidade das provas foi utilizado o coeficiente de Kuder-Richardson, mais recomendado para o caso de testes com itens dicotômicos como os do Enem. Esse coeficiente varia entre 0 e 1, em que quanto mais próximo a 1 maior é a fidedignidade. Para obter esse valor foi utilizado o pacote "*validateR"* (DESJARDINS, [s.d.]).

Também foram estimados os parâmetros dos itens e a habilidade dos participantes a partir da TRI. Essas análises foram realizadas para os modelos unidimensionais de 1, 2 e 3 parâmetros. Portanto, foi estimado os parâmetros de dificuldade, discriminação e acerto casual. Para Baker (2001) o índice de dificuldade é relativo à habilidade do respondente. Dessa forma, o conceito de item "fácil" ou "difícil" deve ser interpretado considerando também a habilidade de quem responde aquele item. A avaliação da dificuldade do item também depende do objetivo do teste (PASQUALI, 2009). Por exemplo, se é um teste realizado para selecionar indivíduos, a dificuldade dos itens deve refletir o critério de habilidade desejado. Por outro lado, se o objetivo é diferenciar indivíduos de baixa habilidade dos de alta habilidade uma melhor distribuição dos níveis de dificuldade dos itens é desejável.

A discriminação dos itens pode ser classificada da forma como se segue (BAKER, 2001):

Classificação	Valor
Não	0
Muito baixo	0,01 – 0,34
Baixo	0,35 – 0,64
Moderado	0,65 – 1,34
Alto	1,35 – 1,69
Muito Alto	> 1,70
Perfeito	+>1,70

O autor destaca que o parâmetro de acerto casual é interpretado diretamente. Ou seja, o valor indica a probabilidade de acertar o item ao acaso para todos os níveis de habilidade. Por exemplo, um item como o parâmetro de acerto ao acaso de c=0,10 indica uma probabilidade de 10% de acertar o item casualmente.

Para análise da comparabilidade entre os parâmetros dos itens e escores dos participantes pela TCT e TRI utilizamos da análise de correlação de Pearson (r). Os valores dessa análise variam entre 0 e 1, em que quanto mais próximo a 1 mais forte é a correlação entre as variáveis. Para obter em porcentagem o quanto que a variação de uma variável é explicada pela outra também foi calculado o coeficiente de determinação (R^2), obtido pelo quadrado do coeficiente de correlação. As correlações foram plotadas em um gráfico de dispersão de pontos para melhor visualização das correlações entre as medidas. O pacote utilizado foi o “ggplot2” (WICKHAM et al., 2019).

Para a comparação dos dois modelos realizou-se a classificação do parâmetro de dificuldade dos itens em ordem crescente com o objetivo de identificar alterações na magnitude da dificuldade dos itens.

Também foi implementado uma análise de regressão linear simples com o objetivo de identificar se os escores obtidos pela TCT conseguem prever os obtidos pela TRI. O ajuste do modelo foi obtido pelo R^2. A análise da significância do modelo foi realizada através de uma anova da regressão. Foi considerado significativos os valores de $p \leq 0,01$. Nesses casos os escores de TCT foram considerados bons previsores dos escores da TRI.

Nesta seção foi descrito o caminho metodológico percorrido para alcançar os objetivos propostos para o estudo. A seguir são apresentados os resultados encontrados a partir do método empregado. Optou-se nesta tese por realizar a discussão à medida que se expõe os achados com a finalidade de aproximar as explicações e problematizações aos resultados para facilitar a compreensão.

6 RESULTADOS E DISCUSSÃO

6.1 Análise descritivas dos itens pela teoria clássica

Para compreender melhor o comportamento dos itens foi estimado os parâmetros a partir da teoria clássica. O objetivo foi analisar a discriminação dos itens a partir do coeficiente de correlação ponto bisserial de modo a identificar o quanto o item se correlaciona com o escore total do conjunto de itens. Os valores estão dispostos na Tabela 3.

Tabela 3 – Parâmetros clássicos dos itens do Enem, 2017.

Área	LC		MT		CN		CH	
Item	r_{pb}	d	r_{pb}	d	r_{pb}	d	r_{pb}	d
1	0,34	0,31	0,31	0,29	0,29	0,25	0,48	0,30
2	0,30	0,46	0,33	0,64	0,22	0,25	0,47	0,55
3	0,36	0,45	0,22	0,27	0,23	0,55	0,16	0,23
4	0,15	0,29	0,27	0,17	0,35	0,49	0,37	0,53
5	0,30	0,33	0,33	0,43	0,15	0,16	0,40	0,49
6	0,36	0,65	**0,09**	0,11	0,29	0,16	0,39	0,59
7	0,34	0,37	0,31	0,12	0,40	0,46	0,47	0,35
8	0,24	0,41	0,22	0,22	0,16	0,19	**0,12**	0,23
9	0,36	0,50	0,44	0,40	0,23	0,60	0,51	0,35
10	0,27	0,26	0,25	0,24	**0,13**	0,12	0,26	0,49
11	0,45	0,61	**0,00**	0,12	**0,14**	0,17	0,39	0,44
12	0,26	0,20	0,31	0,30	0,19	0,32	0,22	0,23
13	0,20	0,19	0,21	0,25	0,30	0,42	0,50	0,58
14	0,44	0,37	0,30	0,25	0,18	0,16	0,21	0,36
15	0,23	0,50	0,29	0,28	0,21	0,35	0,46	0,41
16	0,27	0,33	0,29	0,31	0,18	0,19	0,29	0,54
17	0,36	0,51	0,34	0,30	0,40	0,25	0,41	0,31
18	0,26	0,23	**0,14**	0,16	0,21	0,28	0,28	0,31
19	0,50	0,58	**0,13**	0,19	0,43	0,37	0,19	0,36
20	0,23	0,13	0,29	0,27	0,33	0,37	0,39	0,36
21	0,31	0,66	**0,11**	0,13	**0,06**	0,47	0,19	0,23
22	0,33	0,37	0,22	0,29	0,38	0,23	0,37	0,43
23	0,39	0,57	0,35	0,43	**0,13**	0,21	**-0,05**	0,21
24	0,33	0,28	0,25	0,38	0,22	0,24	0,30	0,29
25	**0,10**	0,15	0,25	0,17	0,15	0,26	**0,04**	0,25
26	0,21	0,41	0,40	0,59	**0,13**	0,23	0,38	0,48
27	0,46	0,65	0,37	0,50	0,17	0,18	0,50	0,42
28	0,23	0,54	0,31	0,27	0,19	0,15	0,31	0,30
29	**0,05**	0,14	0,15	0,06	0,36	0,46	0,34	0,39
30	0,26	0,30	0,24	0,28	**0,10**	0,14	**0,04**	0,20
31	0,48	0,46	0,33	0,34	0,25	0,21	0,39	0,38
32	0,45	0,39	0,29	0,30	**0,09**	0,08	0,32	0,32
33	**0,00**	0,17	0,24	0,11	0,33	0,36	0,52	0,49
34	0,39	0,77	**0,13**	0,23	**0,10**	0,13	0,20	0,24
35	**0,13**	0,25	0,38	0,37	**0,07**	0,09	0,52	0,31
36	**0,05**	0,25	**0,05**	0,23	0,35	0,51	**0,06**	0,27
37	0,35	0,51	0,25	0,27	0,20	0,16	0,15	0,27

Tabela 3 (Continuação)

38	0,29	0,30	0,24	0,15	0,26	0,20	0,26	0,23
39	**0,08**	0,18	0,37	0,12	0,31	0,27	0,40	0,45
40	0,37	0,41	**0,11**	0,17	0,20	0,11	0,19	0,22
41	0,22	0,31	**0,14**	0,16	0,27	0,29	0,47	0,45
42	0,36	0,41	0,20	0,28	0,24	0,33	0,29	0,38
43	0,34	0,53	0,27	0,24	0,40	0,29	0,30	0,31
44	0,35	0,35	**0,12**	0,23	**0,12**	0,23	0,20	0,14
45	0,26	0,19	0,15	0,23	0,31	0,25	0,15	0,11

Legenda: r_{bp}: correlação ponto bisserial; *d*: dificuldade do item; LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

A partir disso, sete itens (4, 25, 29, 33, 35, 36 e 39) de Linguagens e Códigos, dez itens de Matemática (6, 11, 18, 19, 21, 34, 36, 40, 41 e 44), dez itens (10, 11, 21, 23, 26, 30, 32, 34, 34 e 44) de Ciências da Natureza e cinco itens (8, 23, 25, 30 e 36) da prova de Ciências Humanas apresentaram discriminação abaixo de um valor minimamente aceitável (>0,15). Portanto, esses itens foram excluídos das análises seguintes. Esse valor de discriminação também foi utilizado como parâmetro para exclusão de itens da análise fatorial no estudo de Ferreira (2009).

Diante disso, observa-se problemas em muitos itens das provas do Enem de 2017. Muitos itens apresentam discriminação muito baixa. Isso indica que esses itens não têm boa correlação com o escore total da prova (MUÑIZ, 1994). Do ponto de vista prático, isto indica que o item não está conseguindo diferenciar candidatos de escores total mais alto dos com escore total mais baixo (MUÑIZ, 1994; VIANNA, 1976), ou seja, ambos tiveram probabilidade de acertar o item bem próximo (ANDRADE; LAROS; GOUVEIA, 2010). Dito de outra forma, o item não está discriminando os examinandos eficazes e os ineficazes em um teste (ANDRIOLA, 1998; MUÑIZ, 1994). Itens com problema de discriminação devem ser rejeitados (VIANNA, 1976).

Esse tipo de problema é uma tarefa que pode ser resolvida pelos elaboradores de itens. Comumente, problemas de clareza, objetividade do item causam dificuldades de sua interpretação, o que pode induzir candidatos de bom desempenho ao erro. Em um estudo foi identificado que itens com formulação complexa e enunciado confuso apresentam baixa discriminação (CANÇADO; CASTRO; OLIVEIRA, 2013). Por outro lado, o mesmo estudo indicou que os itens simples, objetivos e claros apresentam boa qualidade discriminativa.

Diante disso, verifica-se um problema importante na estruturação da prova. Pois considerando que um exame que objetivo, atualmente, a seleção de candidatos para os cursos de graduação oferecidos pelas Instituições de Ensino Superior, é desejável que apresentem boa

qualidade discriminativa. Após essa análise inicial, foi realizada a análise da dimensionalidade a partir da análise paralela e a *Full Information Factor Analysis* com o objetivo de compreender a estrutura fatorial dos itens.

6.2 Análise da dimensionalidade das provas do Enem 2017

Nesta seção foi conduzida uma análise paralela com a análise de componentes principais a partir de uma matriz de correlações tetracóricas. Na Tabela 4 estão dispostos os valores próprios (*eigen value*) para as cinco primeiras dimensões de cada prova do Enem 2017.

Tabela 4 – *Eigen value* da análise de componentes principais, Enem 2017.

Fatores	LC		MT		CN		CH	
	c.p	d.s	c.p	d.s	c.p	d.s	c.p	d.s
1	6,80	1,12	5,04	1,11	4,52	1,11	8,09	1,12
2	1,52	1,10	1,81	1,10	1,61	1,1	1,48	1,11
3	1,23	1,10	1,17	1,09	1,17	1,09	1,11	1,10
4	1,11	1,09	1,15	1,08	1,14	1,08	1,09	1,09
5	1,06	1,08	1,11	1,07	1,13	1,07	1,07	1,08
6	1,05	1,07	1,07	1,07	1,10	1,07	1,05	1,08
7	1,03	1,07	1,03	1,06	1,06	1,06	1,04	1,07
8	1,00	1,06	1,02	1,05	1,05	1,05	1,03	1,06
9	0,98	1,06	1,01	1,05	1,03	1,05	0,99	1,06
10	0,98	1,05	0,99	1,04	1,02	1,04	0,96	1,05
11	0,96	1,04	0,97	1,04	1,00	1,04	0,95	1,05
12	0,94	1,04	0,95	1,03	0,98	1,03	0,93	1,04
Eigen1/Eigen2	4,47		2,78		2,80		5,47	

Legenda: c.p: *Eigen values* dos componentes principais; d.s: *Eigen values* dos dados simulados; LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

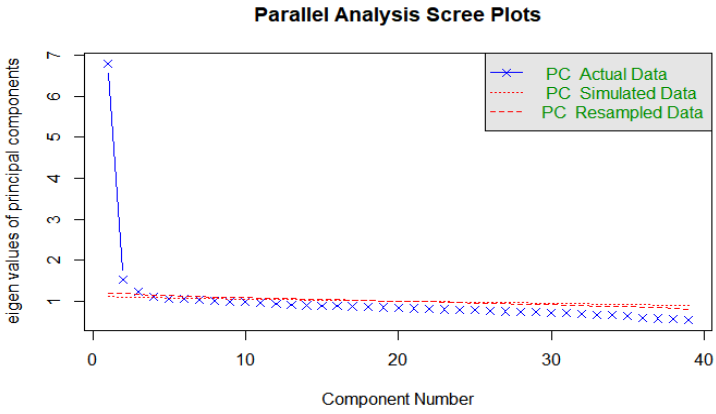
Ao submeter os dados de cada prova a análise paralela, esperava-se a existência ou predominância de um fator, ou seja, que os itens fossem unidimensionais, considerando que esse é o pressuposto do modelo de TRI utilizado no exame. No entanto, as comparações dos *eigen value* dos dados reais com os simulados, o teste indicou a existência de quatro fatores para a prova de Linguagens e Códigos, cinco fatores para as provas de Matemática, seis fatores para a prova de Ciências da Natureza e três fatores para a prova de Ciências Humanas.

Entretanto, há um certo consenso no entendimento de que é suficiente admitir a existência de uma dimensão dominante, uma "dimensionalidade essencial" (STOUT, 1990). Diante disso, apenas duas provas parecem apresentar uma dimensão dominante. A prova de Linguagens e Códigos apresentou a razão entre o eigen1 e eigen2 de 4,47, ou seja, cerca de quatro vezes maior que o segundo e a prova de Ciências Humanas com valor 5,47, ou seja, o

eigen1 é cinco vezes maior que o eigen2. Para as provas de Matemática e Ciências da Natureza o eigen1 é apenas duas vezes maior que o eigen2.

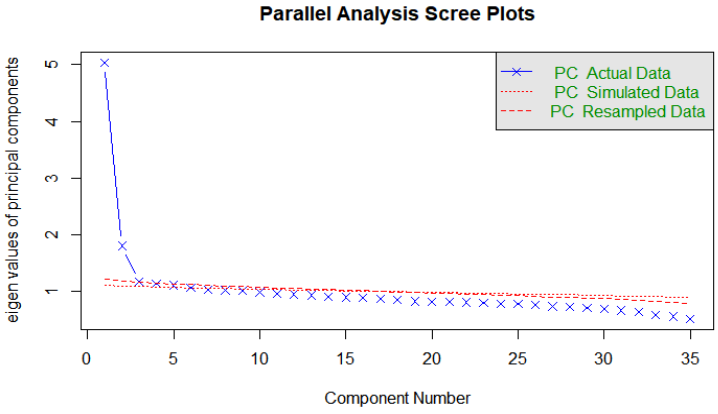
Para melhor visualização da magnitude dos *eigen value* das provas foi construído um *scree plot* dos valores de cada fator dos dados reais e dos dados simulados. Nas figuras 2, 3, 4 e 5 podem ser visualizados esses valores.

Figura 2 – *Scree Plot* da análise paralela da prova de Linguagens e Códigos, Enem 2017.



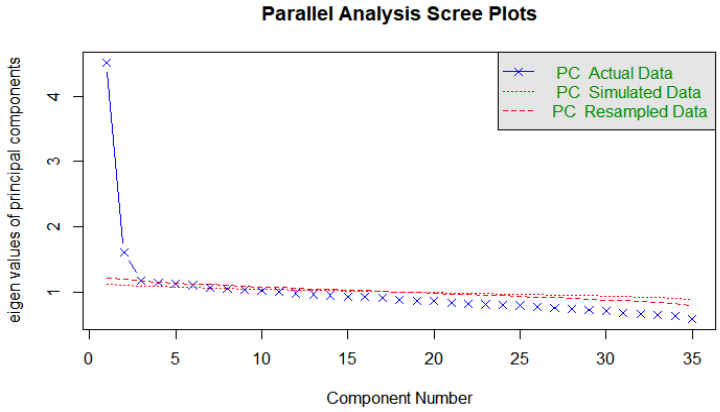
Fonte: Da pesquisa (2019).

Figura 3 – Scree Plot da análise paralela da prova de Matemática, Enem 2017.



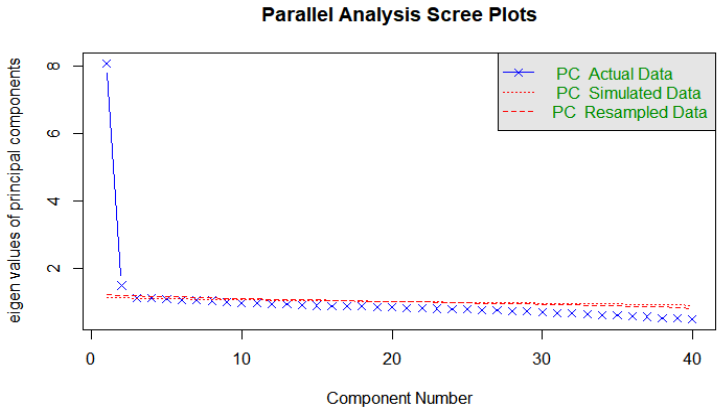
Fonte: Da pesquisa (2019).

Figura 4 – Scree Plot da análise paralela da prova de Ciências da Natureza, Enem 2017.



Fonte: Da pesquisa (2019).

Figura 5 – Scree Plot da análise paralela da prova de Ciências Humanas, Enem 2017.



Fonte: Da pesquisa (2019).

Nesses gráficos torna-se mais evidente a existência de uma dimensão dominante nas provas de Linguagens e Códigos e Ciências Humanas. No entanto, na prova de Matemática e Ciências da Natureza os valores dos dois primeiros fatores se destacam.

Após uma análise exploratória inicial da dimensionalidade dos dados a partir da correlação ponto bisserial e da análise paralela submetemos os itens a uma Análise Fatorial de Informação Plena (*Full Information Factor Analysis*). Uma primeira análise foi realizada com a partir dos valores de ajuste dos modelos com base nas medidas do *Root-Mean-Square Error of Approximation* (RMSEA), *Standardised Root Mean Square Residual* (SRMSR), *Tucker-Lewis Index* (TLI) e *Comparative Fit Index* (CFI). Os valores desses índices podem ser observados para cada prova na Tabela 5.

Tabela 5 – Índice de ajuste dos modelos estimados para as provas do Enem, 2017.

Provas	Modelos	RMSEA	SRMSR	TLI	CFI
LC	1	0,011	0,014	0,987	0,988
	2	0,009	0,013	0,991	0,992
	3	0,008	0,012	0,993	0,994
	4	0,007	0,011	0,993	0,995
MT	1	0,009	0,014	0,983	0,985
	2	0,009	0,013	0,984	0,987
	3	0,008	0,013	0,988	0,990
	4	0,007	0,012	0,989	0,992
CN	1	0,009	0,014	0,977	0,980
	2	0,009	0,013	0,980	0,983
	3	0,008	0,012	0,984	0,988
	4	0,008	0,012	0,984	0,988
CH	1	0,010	0,013	0,992	0,993
	2	0,009	0,012	0,994	0,995
	3	0,008	0,011	0,995	0,996
	4	0,007	0,010	0,996	0,997

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

Como foi constatado na tabela acima, os valores de ajuste para todas as provas estimados para os modelos de 1, 2, 3 e 4 parâmetros mostraram-se adequados. Os valores do RMSEA e SRMSR se mantiveram abaixo de 0,05 e os valores do TLI e CFI acima de 0,95. Esses índices não conseguiram identificar com precisão o modelo fatorial mais adequado para esses itens.

Também foi analisado os valores dos índices ajuste dos modelos com base nos valores de AIC e BIC. Os dados dessa análise estão dispostos na Tabela 6.

Tabela 6 – Critério de Informação Baysiano (BIC) e Critério de Informação de Akaike (AIC) dos modelos estimados para as provas do Enem, 2017.

Dimensões	LC		MT		CN		CH	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
1	469439,3	470282,9	380552,8	381309,9	387059,2	387816,3	468863,2	469728,4
2	469253,3	470370,9	380444,1	381446,3	387063,7	388065,9	468714,5	469861,0
3	469121,0	470505,4	380366,6	381606,8	386936,4	388176,6	468616,5	470037,0
4	469059,6	470703,5	380355,4	381826,4	386892,6	388363,5	468608,2	470295,4

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

O índice de ajuste do AIC, como apresentados na literatura tende a ser mais permissivo quanto ao número de fatores, o que acorreu nessa pesquisa. Os valores decrescem à medida que se acrescenta mais fatores ao modelo. Indicando, por esse índice, que 4 ou mais fatores se ajustam bem.

Ao verificar os índices de ajuste oferecidos pelo BIC, observa-se que o menor valor, portanto, o modelo melhor ajustado, foram os dos modelos unidimensionais das provas. Nota-

se também que a medida que são acrescentados mais fatores os valores do BIC aumentam. Isso indica que se perde qualidade no ajuste ao estimar modelos com mais de uma dimensão. No entanto, embora os BIC tenha sido considerado mais consistente na indicação do número de dimensões a serem retidas na análise fatorial, o teste é considerado com conservador, ou seja, valoriza modelos com menos dimensões.

Também foi analisado o Índice de Dimensionalidade a partir da comparação entre os modelos. Os ID das comparações dos modelos estão dispostos na Tabela 7.

Tabela 7 – Índice de dimensionalidade das provas do Enem 2017.

Provas	**Comparação**	**X^2**	**X^2corr**	**gl**	**ID**
LC	M1 - M2	262,002	87,334	38	2,298
	M2 - M3	206,238	68,746	37	1,858
	M3 - M4	133,464	44,488	36	1,235
MT	M1 - M2	176,698	58,899	34	1,732
	M2 - M3	143,442	47,814	33	1,499
	M3 - M4	75,188	25,063	32	0,783
CN	M1 - M2	63,559	21,186	34	0,623
	M2 - M3	193,233	64,411	33	1,952
	M3 - M4	107,810	35,937	32	1,123
CH	M1 - M2	226,643	75,547	39	1,937
	M2 - M3	173,993	57,998	38	1,526
	M3 - M4	82,315	27,438	37	0,741

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

Em todas as comparações realizadas para as provas de Matemática, Ciências da Natureza e Ciências Humanas os valores de ID foram positivos menores que 2,0. Dessa forma, O ID indica que essas provas são unidimensionais. No prosseguimento das comparações a prova de Linguagens e Códigos se ajustou melhor ao modelo de 2 fatores. As demais provas tiveram melhor ajuste ao modelo unidimensional. Diante disso, esses resultados corroboram com os apresentados pelo critério BIC.

Como forma de obter mais informações em relação a dimensionalidade das provas, também foi verificado o percentual de variância explicada por cada modelo de 1, 2, 3 e 4. Esses valores estão apresentados na Tabela 8.

Tabela 8 – Percentual de variância explicada nos modelos das provas do Enem, 2017.

Provas	Modelos			
	1	2	3	4
LC	29,5	31,9	32,3	32,7
MT	58,1	57,4	59,9	61,4
CN	52,4	50,8	53,8	57,0
CN	49,9	42,1	42,0	42,1

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

Ao analisar o percentual da variância total explicada por cada modelo, verifica-se um valor maior da variância explicada pelo modelo unidimensional. A medida que se acrescenta mais fatores ao modelo pouca explicação da variância é acrescentada. Por outro lado, algumas provas têm esse percentual diminuído ao acrescentar algum fator, como é o caso das provas de Ciências da Natureza e Ciências Humanas. Neste ponto da análise parece haver evidências suficientes de que o modelo unidimensional é o mais adequado para todas as provas do Enem 2017.

No prosseguimento das análises, as cargas fatoriais de cada item também são indicativos da dimensionalidade destes, já que consiste na correlação do item com o fator. Dessa forma, espera-se que se os itens das provas, especialmente da prova de Linguagens e Códigos, são unidimensionais, cargas fatoriais altas, 0,30 ou mais, sejam identificadas. Optou-se por realizar sucessivas análises das cargas fatoriais dos itens até se obter itens com cargas satisfatórias. As cargas fatoriais dos itens para o modelo unidimensional estão na Tabela 9.

Tabela 9 – Cargas fatoriais dos itens com 1 fator do Enem 2017.

	LC			MT	CN		CH	
Itens	1º Extração	2º Extração	3º Extração	1º Extração	1º Extração	2º Extração	1º Extração	2º Extração
i1	0,54	0,52	0,51	0,79	0,74	0,74	0,78	0,77
i2	0,69	0,69	0,69	0,63	0,38	0,36	0,68	0,67
i3	0,51	0,49	0,49	0,82	0,46	0,45	0,41	0,38
i4	0,40	0,41	0,41	0,86	0,48	0,48	0,44	0,44
i5	0,65	0,65	0,65	0,83	0,64	0,64	0,59	0,58
i6	0,45	0,45	0,45	*	0,82	0,81	0,48	0,48
i7	0,57	0,56	0,56	0,92	0,63	0,63	0,81	0,81
i8	**0,24**	**	**	0,57	0,87	0,87	*	*
i9	0,46	0,46	0,45	0,76	**0,26**	**	0,77	0,76
i10	0,50	0,49	0,48	0,88	*	*	**0,28**	**
i11	0,59	0,59	0,59	*	*	*	0,47	0,47
i12	0,44	0,43	0,43	0,87	0,60	0,60	0,68	0,67
i13	0,54	0,53	0,53	0,81	0,63	0,63	0,76	0,76
i14	0,77	0,78	0,77	0,78	0,95	0,95	**0,25**	**
i15	**0,23**	**	**	0,86	0,69	0,69	0,73	0,73
i16	0,44	0,44	0,44	0,79	0,86	0,86	0,32	0,32
i17	0,41	0,41	0,41	0,60	0,76	0,76	0,82	0,82
i18	0,60	0,59	0,59	*	0,67	0,67	0,77	0,76
i19	0,67	0,66	0,66	*	0,81	0,81	**0,18**	**
i20	0,77	0,77	0,77	0,89	0,43	0,43	0,50	0,49
i21	0,36	0,36	0,36	*	*	*	0,62	0,60
i22	0,58	0,58	0,57	0,67	0,87	0,87	0,53	0,53
i23	0,46	0,46	0,46	0,57	*	*	*	*
i24	0,59	0,58	0,58	0,34	0,78	0,78	0,35	0,35
i25	*	*	*	0,84	**0,12**	**	*	*
i26	**0,19**	**	**	0,72	*	*	0,50	0,50
i27	0,63	0,63	0,63	0,52	0,68	0,66	0,74	0,74

Tabela 9 (Continuação)

i28	0,23	**0,23**	***	0,81	0,81	0,80	0,78	0,78
i29	*	*	*	0,86	0,67	0,67	0,55	0,55
i30	0,44	0,44	0,45	0,76	*	*	*	*
i31	0,66	0,66	0,66	0,60	0,87	0,87	0,67	0,66
i32	0,71	0,71	0,70	0,85	*	*	0,81	0,81
i33	*	*	*	0,72	0,82	0,81	0,85	0,85
i34	0,59	0,59	0,61	*	*	*	0,48	0,48
i35	*	*	*	0,66	*	*	0,87	0,86
i36	*	*	*	*	0,73	0,75	*	*
i37	0,40	0,40	0,40	0,65	0,63	0,63	0,70	0,71
i38	0,59	0,59	0,58	0,91	0,86	0,85	0,55	0,54
i39	*	*	*	0,92	0,68	0,70	0,59	0,58
i40	0,67	0,68	0,68	*	0,89	0,89	0,71	0,71
i41	0,33	0,33	0,33	*	0,75	0,76	0,73	0,72
i42	0,47	0,48	0,48	0,78	0,86	0,86	0,66	0,65
i43	0,39	0,39	0,39	0,73	0,86	0,85	0,71	0,70
i44	0,74	0,74	0,73	*	*	*	0,86	0,86
i45	0,80	0,80	0,79	0,74	0,89	0,89	0,63	0,62
%Var	29,5	31,2	31,8	58,1	52,4	55,0	40,9	43,0

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
*Itens excluídos por baixa discriminação (<0,15).
**Itens excluídos por baixa carga fatorial na primeira extração (<0,30).
***Itens excluídos por baixa carga fatorial na segunda extração (<0,30).
Fonte: Da pesquisa (2019).

Ao analisar as cargas fatoriais dos itens das provas do Enem 2017, apenas os itens da prova de Matemática apresentaram adequação já na primeira extração, com cargas fatoriais variando entre 0,34 a 0,92. O modelo final apresentou um percentual de variância explicado de 58,1%. Durante o processo de ajuste de um modelo fatorial para essa prova, 10 de 45 itens foram excluídos por baixa discriminação, permanecendo 35 itens.

Os itens da prova de Linguagens e Códigos foram necessárias três extrações para a obtenção de um modelo fatorial satisfatório. No modelo final 10 itens também foram excluídos, inicialmente por baixa discriminação e ao aplicar o critério da carga fatorial durante o ajuste do modelo, permanecendo 35 itens. As cargas fatoriais desses itens variaram entre 0,33 a 0,79. O percentual de variância explicado foi de 31,8%.

Para os itens da prova de Ciências da Natureza foram necessárias duas extrações para a obtenção de um modelo fatorial adequado. No modelo final 12 itens foram excluídos por baixa discriminação e baixas cargas fatoriais, permanecendo 33 itens no modelo final. As cargas fatoriais variaram entre 0,36 e 0,95. O percentual de variância explicado foi de 55%.

Por último, os itens da prova de Ciência Humanas se ajustaram também na segunda extração. Inicialmente foram excluídos 5 itens por baixa discriminação e depois mais 3 itens por baixa carga fatorial na primeira extração. O modelo final resultou em 37 itens como cargas fatoriais entre 0,32 e 0,86 e percentual de variância explicado de 43%.

No quadro a seguir estão os itens excluídos durante o processo de análise fatorial para cada prova do exame.

Quadro 4 – Itens excluídos durante a análise fatorial exploratória, Enem 2017.

Provas	Total de itens	$r_{pb} < 0,15$	1º Extração >0,30	2º Extração >0,30	Total de itens excluídos
LC	45	25, 29, 33, 35, 36 e 39	8, 15 e 28	28	10
MT	45	6, 11, 18, 19, 21, 34, 36, 40, 41 e 44	_____	_____	10
CN	45	10, 11, 21, 23, 26, 30, 32, 34, 35, e 44	9 e 25	_____	12
CH	45	8, 23, 25, 30 e 36	10, 14 e 19		8

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

Durante a análise fatorial exploratória foram excluídos 10 itens da prova de Linguagem e Códigos e Matemática, 12 itens de Ciências da Natureza e 8 itens de Ciência Humanas. Esperava-se que as análises indicassem que todos os itens apresentassem unidimensionalidade. No entanto, não ocorreu. Isso demonstra problemas métricos dos itens das provas. Investigações mais aprofundadas são necessárias para analisar o impacto desses problemas na estimação das habilidades dos sujeitos. Supomos, inicialmente, que esses problemas podem prejudicar os candidatos no processo de seleção para os cursos de graduação das Instituições de Ensino Superior.

Para todas as provas do exame foi possível ajustar um modelo unidimensional, no entanto após a exclusão de muitos itens (ver Quadro 6). Embora seja parcimonioso um modelo unidimensional, considerando apenas uma dimensão dominante (STOUT, 1990), essa dimensão não conseguiu explicar a variabilidade dos outros itens. Provavelmente esses itens podem ser explicados pela existência de outras dimensões. Nesse caso, modelos de TRI multidimensionais são mais adequados (RECKASE, 2009).

Os pressupostos dos modelos de TRI unidimensionais, o mesmo utilizado no Enem, têm sido fortemente criticados. Tavares (2013 p. 69) faz o seguinte questionamento sobre este ponto: “Se pensarmos em um exame específico utilizado em larga escala em nosso país, como o Enem, podemos assegurar que a TRI baseada em um modelo logístico unidimensional de três parâmetros é a melhor opção metodológica para esse caso?”. Ainda acrescenta apontando a perspectiva interdisciplinar do exame como um de seus diferenciais, o que torna mais difícil ainda essa ideia.

Uma das saídas apontadas para esse problema da dimensionalidade do Enem e seu caráter interdisciplinar não comportado pelo modelo de TRI unidimensional é a utilização de

modelos multidimensionais (OLIVEIRA, 2015). Esse problema da estrutura fatorial do exame foi testado empiricamente em algumas pesquisas.

Em alguns estudos que realizam uma análise da estrutura fatorial dos itens para a compreensão dos construtos medidos por esse exame têm sido realizados. Em uma pesquisa se analisou o Enem de 2010 a partir de análise fatorial exploratória e confirmatória (MUNER, 2013). Nesse estudo a autora conseguiu ajustar um modelo unidimensional com explicação de 30% da variância apenas para a prova de Linguagens e Código, mas após a exclusão de alguns itens. A prova de Ciência Humanas e Matemática foi ajustado um modelo bidimensional, ambos com 40% de variância explicada e a prova de Ciências da Natureza um modelo com três dimensões e 50% de variância explicada. A estrutura fatorial dessas provas foi ratificada pela análise fatorial confirmatória com bons valores de ajuste do modelo.

Resultados similares são encontrados em outro estudo que analisa o Enem, mais especificamente a prova de matemática de 2015 (PRIMI; CICCHETTO, 2018). Nesta pesquisa os autores implementaram uma análise de componentes principais sobre os resíduos obtidos no ajuste do modelo de TRI de Rasch. Com essa análise foi identificado a existência de um fator secundário importante indicando sistematicidade no acerto de participantes considerados como de baixa habilidade em itens difíceis.

Em outro estudo analisou-se as quatro provas do Enem de 2012 tomadas com prova única com 180 itens sob a ótica da TRI uni e multidimensional (VIEIRA, 2016). As análises indicaram que um modelo unidimensional se ajustou bem. O modelo multidimensional com quatro fatores também apresentou bom ajuste, mas não sugeriu que os itens estavam relacionados com suas respectivas áreas de conhecimento. Isso dificulta a interpretação semântica dos fatores.

Embora Vieira (2016) tenha evidenciado que um modelo unidimensional se ajusta bem aos itens das quatro áreas, torna-se mais complicado ainda a interpretação desse fator e recai nas discussões apresentadas por Tavares (2013) de que é difícil compreender que apenas um fator latente é responsável pelas respostas aos itens de um teste educacional, ainda mais se tratando do Enem.

Por outro lado, uma outra pesquisa, ao analisar a estrutura fatorial de todos os itens do Enem de 2016 a partir do método de análise fatorial de informação plena, identificou a existência de pelo menos duas dimensões (PICCIRILLI; SOUZA, 2018). Ao realizar a interpretação dos itens os autores sugerem que os dois fatores medidos pelos itens são interpretação de texto e raciocínio lógico.

Também foram realizadas análises psicométricas da estrutura fatorial das provas do exame antes de 2009, quando o Enem ainda utilizava técnicas de psicometria clássica na análise dos resultados. Em estudo realizado com o Enem de 1999 (NOJOSA, 2002) e 2001 (COSTA, 2015) foi possível ajustar um modelo de TRI multidimensional com cinco fatores. Nojosa (2002) ressalta que esse modelo seria o mais adequado para a análise do exame considerando suas características de conteúdo.

A partir dessas evidências apontadas ainda não é possível realizar conclusões fortes sobre a dimensionalidade do Enem. Nesse ponto em particular, mais pesquisas são necessárias para a compreensão do que realmente este exame está a medir. Na próxima subseção é realizada a comparação dos resultados do exame a partir dos modelos de TCT e TRI.

6.3 Análise comparativa do Enem a partir da TCT e TRI

Uma vez ajustados os modelos unidimensionais para as provas do Enem 2017, realizou-se a análise dos itens a partir dos parâmetros da TCT e TRI. Para a TCT foi calculado as estatísticas de discriminação, dificuldade e o coeficiente de confiabilidade dos itens. Para a TRI foi calculado os parâmetros a partir dos modelos de 1 (1PL), 2 (2PL) e 3 (3PL) parâmetros. Para o modelo 1PL foi estimado a dificuldade (*b*). Para o modelo 2PL a dificuldade (*b*) e discriminação (*a*). Para o modelo 3PL a dificuldade (*b*) e discriminação (*a*) e o parâmetro de acerto ao acaso (*c*). Os valores estão padronizados (escore z) com média 0 e desvio padrão 1. Os valores dos parâmetros dos itens para cada prova do exame estão nas Tabelas de 10 a 13.

Na Tabela 10 estão os parâmetros dos itens para a prova de Linguagens e Códigos. Nessa prova apenas o item 4 apresentou dificuldade abaixo do esperado (0,16). A dificuldade dos itens dessa prova variou entre 0,13 e 0,66. O coeficiente KR quando o item é excluído ficou entre 0,77 e 0,79. Nenhum item influenciou significativamente o valor de KR da prova que foi de 0,78. Esse valor está dentro do aceitável.

Para o modelo 1PL a dificuldade do item (*b*) variou entre -2,04 e 1,33. No modelo 2PL o parâmetro de discriminação (*a*) varia entre 0,20 e 1,39 e a dificuldade (*b*) entre -1,97 e 1,53. Para o modelo 3PL o parâmetro de discriminação (*a*) varia entre 0,59 e 2,22, a dificuldade (*b*) entre -4,39 e 1,30 e o acerto casual entre 0,01 e 0,34.

Tabela 10 – Parâmetros dos itens pela TRI da prova de Linguagens e Códigos do Enem 2017.

	TCT		TRI					
			1 parâmetro	2 parâmetros		3 parâmetros		
Item	r_{bp}	*d*	*b*	*a*	*b*	*a*	*b*	*c*
i1	0,35	0,31	-0,91	0,72	-0,91	1,01	-1,44	0,10
i2	0,31	0,46	-0,17	**0,53**	-0,17	1,61	-2,13	0,34
i3	0,37	0,45	-0,21	0,70	-0,20	0,96	-0,74	0,16
i4	**0,16**	0,29	-1,01	**0,20**	-0,92	0,77	-2,74	0,23
i5	0,32	0,33	-0,79	**0,59**	-0,76	1,45	-2,26	0,20
i6	0,37	0,65	0,70	0,82	0,73	0,87	0,58	0,08
i7	0,35	0,37	-0,61	0,67	-0,60	1,14	-1,46	0,17
i9	0,37	0,50	0,02	0,75	0,02	0,86	-0,24	0,10
i10	0,28	0,26	-1,16	**0,54**	-1,11	0,93	-1,94	0,12
i11	0,46	0,61	0,50	1,19	0,59	1,24	0,48	0,04
i12	0,26	0,20	-1,50	**0,54**	-1,44	0,80	-1,98	0,07
i13	0,20	0,19	-1,59	**0,37**	-1,49	1,05	-3,00	0,13
i14	0,45	0,37	-0,60	1,03	-0,64	2,08	-1,94	0,17
i16	0,28	0,33	-0,76	**0,49**	-0,73	0,84	-1,53	0,16
i17	0,37	0,51	0,03	0,75	0,03	0,77	-0,02	0,02
i18	0,28	0,23	-1,34	**0,55**	-1,29	1,25	-2,58	0,13
i19	0,51	0,58	0,37	1,39	0,47	1,51	0,27	0,07
i20	0,23	0,13	-2,04	**0,55**	-1,97	2,02	-4,39	0,09
i21	0,31	0,66	0,73	**0,63**	0,72	0,65	0,71	0,01
i22	0,34	0,37	-0,61	0,65	-0,60	1,18	-1,57	0,18
i23	0,40	0,57	0,30	0,86	0,31	0,87	0,29	0,01
i24	0,35	0,28	-1,07	0,72	-1,07	1,20	-1,89	0,11
i27	0,47	0,65	0,71	1,31	0,86	1,36	0,74	0,05
i30	0,27	0,30	-0,95	**0,47**	-0,90	0,85	-1,76	0,15
i31	0,48	0,46	-0,17	1,20	-0,18	1,48	-0,55	0,09
i32	0,46	0,39	-0,48	1,06	-0,52	1,69	-1,33	0,14
i34	0,39	0,77	1,33	1,19	1,53	1,29	1,30	0,15
i37	0,36	0,51	0,04	0,70	0,04	0,74	-0,05	0,04
i38	0,30	0,30	-0,96	**0,57**	-0,92	1,22	-2,13	0,17
i40	0,38	0,41	-0,39	0,79	-0,39	1,57	-1,61	0,22
i41	0,23	0,31	-0,87	**0,37**	-0,81	**0,59**	-1,47	0,14
i42	0,37	0,41	-0,42	0,75	-0,42	0,92	-0,74	0,09
i43	0,35	0,53	0,14	0,70	0,14	0,71	0,12	0,01
i44	0,35	0,35	-0,71	0,70	-0,70	1,82	-2,43	0,21
i45	0,26	0,19	-1,57	**0,55**	-1,51	2,22	-4,24	0,14

Legenda: *b*: dificuldade; *a*: discriminação; *c*: acerto ao acaso.
Fonte: Da pesquisa (2019).

Em Matemática (Tabela 11) os itens 29 e 45 tiveram discriminação abaixo do considerado adequado. A dificuldade dos itens variou entre 0,11 e 0,64 e o coeficiente alfa quando o item é excluído entre 0,67 e 0,69. O coeficiente KR de todos os itens foi 0,68. Esse valor está abaixo do considerado aceitável. No modelo 1PL a dificuldade variou entre -2,81 e 0,60. No modelo 2PL a discriminação dos itens está entre 0,15 e 1,39 e a dificuldade entre -2,72 e 0,67. Para o modelo 3PL os valores dos parâmetros dos itens estão para a discriminação entre 0,61 e 4,01, a dificuldade entre -7,93 e 0,44 e o acerto casual entre 0,01 e 0,34.

Tabela 11 – Parâmetros dos itens pela TRI da prova de Matemática do Enem, 2017.

	TCT		TRI					
			1 parâmetro	2 parâmetros		3 parâmetros		
Item	r_{bp}	*d*	*b*	*a*	*b*	*a*	*b*	*c*
i1	0,32	0,29	-0,97	**0,61**	-0,97	2,20	-3,72	0,21
i2	0,35	0,64	0,60	0,92	0,67	1,37	-0,08	0,29
i3	0,22	0,27	-1,08	**0,29**	-1,03	2,45	-5,55	0,24
i4	0,27	0,17	-1,70	0,50	-1,68	2,91	-6,05	0,13
i5	0,34	0,43	-0,28	**0,64**	-0,28	2,49	-3,24	0,34
i7	0,31	0,12	-2,13	0,71	-2,20	3,99	-7,89	0,08
i8	0,23	0,22	-1,34	**0,39**	-1,30	1,18	-3,01	0,16
i9	0,46	0,40	-0,44	1,27	-0,49	1,98	-1,41	0,15
i10	0,25	0,24	-1,21	**0,39**	-1,18	3,10	-6,38	0,21
i12	0,32	0,30	-0,91	**0,56**	-0,91	2,99	-5,00	0,24
i13	0,21	0,25	-1,18	**0,29**	-1,13	2,34	-5,53	0,22
i14	0,31	0,25	-1,18	**0,56**	-1,17	2,09	-3,94	0,19
i15	0,30	0,28	-1,02	**0,51**	-1,01	2,90	-5,22	0,23
i16	0,29	0,31	-0,86	**0,49**	-0,85	2,18	-4,06	0,25
i17	0,35	0,30	-0,89	0,72	-0,92	1,28	-1,88	0,14
i20	0,29	0,27	-1,06	**0,49**	-1,04	3,30	-5,93	0,23
i22	0,23	0,29	-0,95	**0,32**	-0,91	1,55	-3,66	0,25
i23	0,37	0,43	-0,28	0,79	-0,29	1,19	-0,97	0,17
i24	0,26	0,38	-0,51	**0,41**	-0,50	**0,61**	-1,03	0,14
i25	0,25	0,17	-1,71	**0,45**	-1,68	2,61	-5,75	0,14
i26	0,43	0,59	0,38	1,39	0,51	1,74	0,44	0,03
i27	0,39	0,50	0,02	0,97	0,04	1,04	0,00	0,01
i28	0,32	0,27	-1,09	**0,60**	-1,09	2,35	-4,08	0,20
i29	**0,15**	0,06	-2,81	**0,35**	-2,72	2,88	-7,68	0,05
i30	0,24	0,28	-1,01	**0,37**	-0,98	2,01	-4,35	0,24
i31	0,35	0,34	-0,72	0,73	-0,75	1,27	-1,70	0,16
i32	0,29	0,30	-0,91	**0,50**	-0,90	2,69	-4,84	0,25
i33	0,24	0,11	-2,20	**0,58**	-2,21	1,79	-4,45	0,07
i35	0,40	0,37	-0,58	0,89	-0,62	1,50	-1,61	0,17
i37	0,26	0,27	-1,06	**0,41**	-1,03	1,45	-3,26	0,21
i38	0,24	0,15	-1,84	**0,44**	-1,80	3,79	-7,93	0,12
i39	0,36	0,12	-2,08	0,88	-2,24	4,01	-7,32	0,08
i42	0,20	0,28	-1,02	**0,27**	-0,97	2,12	-5,16	0,25
i43	0,28	0,24	-1,22	**0,49**	-1,20	1,80	-3,68	0,18
i45	**0,15**	0,23	-1,26	**0,15**	-1,19	1,90	-5,52	0,22

Legenda: *b*: dificuldade; *a*: discriminação; *c*: acerto ao acaso.
Fonte: Da pesquisa (2019).

Para a prova de Ciências da Natureza (Tabela 12) os itens 5, 8, 14, 16 e 72 apresentaram dificuldade inferior a 0,20 e variando entre 0,16 e 0,51. Os valores do coeficiente KR quando o item é excluído variou entre 0,64 e 0,66. O valor para todos os itens foi de 0,65. Para o modelo 1PL a dificuldade dos itens entre -2,19 a 0,22. No modelo 2PL a discriminação está entre 0,20 a 1,16 e a dificuldade entre -2,15 e 0,22. Para o modelo 3PL a discriminação variou entre 0,66 a 5,24, a dificuldade entre -12,24 e 0,15 e a probabilidade de acerto ao acaso entre 0,01 e 0,39.

Tabela 12 – Parâmetros dos itens pela TRI da prova de Ciências da Natureza do Enem, 2017.

	TCT		TRI					
			1 parâmetro	2 parâmetros		3 parâmetros		
Item	r_{bp}	*d*	*b*	*a*	*b*	*a*	*b*	*c*
i1	0,30	0,25	-1,18	**0,58**	-1,19	1,84	-3,49	0,18
i2	0,24	0,25	-1,15	**0,37**	-1,11	0,66	-1,92	0,13
i3	0,25	0,55	0,22	**0,37**	0,22	0,86	-1,21	0,39
i4	0,37	0,49	-0,05	0,80	-0,05	0,92	-0,15	0,03
i5	**0,15**	0,16	-1,74	**0,21**	-1,66	1,41	-4,74	0,14
i6	0,31	0,16	-1,78	0,67	-1,83	2,38	-4,91	0,11
i7	0,42	0,46	-0,19	1,07	-0,19	1,39	-0,58	0,10
i8	**0,17**	0,19	-1,56	**0,20**	-1,49	2,99	-7,42	0,17
i12	0,20	0,32	-0,81	**0,21**	-0,78	1,26	-3,76	0,29
i13	0,32	0,42	-0,33	**0,58**	-0,33	1,39	-1,88	0,28
i14	**0,18**	0,16	-1,73	**0,23**	-1,65	5,24	-12,24	0,15
i15	0,23	0,35	-0,67	**0,30**	-0,65	1,62	-3,73	0,30
i16	**0,19**	0,19	-1,57	**0,27**	-1,50	2,82	-6,91	0,17
i17	0,41	0,25	-1,18	1,01	-1,31	2,02	-2,82	0,12
i18	0,23	0,28	-1,01	**0,32**	-0,97	1,54	-3,79	0,24
i19	0,45	0,37	-0,54	1,16	-0,62	2,31	-2,08	0,18
i20	0,34	0,37	-0,56	0,70	-0,58	0,80	-0,63	0,01
i22	0,39	0,23	-1,30	0,90	-1,41	3,03	-4,63	0,15
i24	0,22	0,24	-1,24	**0,32**	-1,19	2,10	-5,09	0,21
i27	**0,18**	0,18	-1,59	**0,25**	-1,52	1,50	-4,54	0,16
i28	0,20	0,15	-1,82	**0,34**	-1,76	2,27	-5,79	0,13
i29	0,39	0,46	-0,19	0,83	-0,19	1,52	-1,27	0,23
i31	0,26	0,21	-1,38	**0,44**	-1,35	3,06	-6,45	0,19
i33	0,35	0,36	-0,59	**0,64**	-0,60	2,39	-3,49	0,28
i36	0,37	0,51	0,06	0,83	0,07	1,91	-1,50	0,32
i37	0,21	0,16	-1,75	**0,37**	-1,70	1,38	-3,95	0,12
i38	0,27	0,20	-1,47	**0,49**	-1,46	2,78	-5,88	0,17
i39	0,32	0,27	-1,08	**0,60**	-1,09	1,65	-3,02	0,18
i40	0,21	0,11	-2,19	**0,40**	-2,15	3,33	-7,81	0,09
i41	0,27	0,29	-0,94	**0,43**	-0,92	1,98	-3,99	0,24
i42	0,24	0,33	-0,75	**0,34**	-0,72	2,85	-5,71	0,30
i43	0,41	0,29	-0,95	0,95	-1,03	2,78	-3,66	0,18
i45	0,32	0,25	-1,16	**0,60**	-1,18	3,30	-5,74	0,20

Legenda: *b*: dificuldade; *a*: discriminação; *c*: acerto ao acaso.
Fonte: Da pesquisa (2019).

Em relação a prova de Ciências Humanas (Tabela 13) os itens 3, 21, 37, 40 e 45 tiveram índices de discriminação considerado inadequado. A dificuldade dos itens variou entre 0,11 e 0,55 e o coeficiente KR quando o item é excluído entre 0,81 e 0,82. O coeficiente da prova foi de 0,82. Esse valor é considerado bom, e indica alta consistência interna entre os itens. No modelo 1PL teve a dificuldade variando entre -2,26 e 0,42. No modelo 2PL a discriminação está entre 0,20 a 1,49 e a dificuldade entre -2,09 e 0,53. Já para o modelo 3PL a discriminação está entre 0,58 e 2,85, a dificuldade entre -6,09 e 0,42 e a probabilidade de acerto casual variou entre 0,01 e 0,26.

Tabela 13 – Parâmetros dos itens pela TRI da prova de Ciências Humanas do Enem, 2017.

	TCT		TRI					
			1 parâmetro	2 parâmetros		3 parâmetros		
Item	r_{bp}	*d*	*b*	*a*	*b*	*a*	*b*	*c*
i1	0,50	0,30	-0,97	1,24	-1,10	2,03	-2,09	0,10
i2	0,47	0,55	0,22	1,25	0,29	1,53	-0,13	0,13
i3	**0,17**	0,23	-1,37	**0,26**	-1,25	0,70	-2,62	0,16
i4	0,37	0,53	0,15	0,78	0,15	0,83	0,09	0,02
i5	0,41	0,49	-0,02	0,89	-0,01	1,21	-0,58	0,17
i6	0,39	0,59	0,42	0,90	0,45	0,93	0,42	0,01
i7	0,49	0,35	-0,67	1,17	-0,73	2,31	-2,13	0,16
i9	0,52	0,35	-0,71	1,33	-0,81	2,00	-1,61	0,10
i11	0,39	0,44	-0,26	0,83	-0,26	0,90	-0,36	0,03
i12	0,22	0,23	-1,33	**0,40**	-1,23	1,52	-3,47	0,18
i13	0,51	0,58	0,39	1,49	0,53	2,01	-0,07	0,18
i15	0,46	0,41	-0,42	1,06	-0,44	1,81	-1,44	0,17
i16	0,29	0,54	0,18	**0,53**	0,17	**0,58**	0,08	0,04
i17	0,42	0,31	-0,90	0,93	-0,93	2,47	-3,01	0,18
i18	0,29	0,31	-0,91	**0,52**	-0,86	2,02	-3,42	0,23
i20	0,40	0,36	-0,62	0,84	-0,63	0,96	-0,80	0,04
i21	**0,19**	0,23	-1,35	**0,32**	-1,23	1,28	-3,35	0,18
i22	0,37	0,43	-0,32	0,74	-0,32	1,05	-0,91	0,16
i24	0,30	0,29	-1,00	**0,58**	-0,96	**0,63**	-0,99	0,01
i26	0,38	0,48	-0,10	0,79	-0,10	0,97	-0,46	0,12
i27	0,51	0,42	-0,34	1,34	-0,37	1,85	-1,01	0,12
i28	0,32	0,30	-0,97	**0,60**	-0,94	2,09	-3,35	0,21
i29	0,35	0,39	-0,51	0,67	-0,50	1,11	-1,32	0,18
i31	0,40	0,38	-0,54	0,84	-0,55	1,50	-1,57	0,18
i32	0,32	0,32	-0,84	**0,62**	-0,81	2,31	-3,48	0,23
i33	0,53	0,49	-0,06	1,41	-0,03	2,69	-1,31	0,22
i34	0,20	0,24	-1,30	**0,34**	-1,20	0,92	-2,59	0,16
i35	0,53	0,31	-0,88	1,39	-1,02	2,85	-2,68	0,13
i37	**0,15**	0,27	-1,13	**0,20**	-1,02	1,71	-4,56	0,24
i38	0,26	0,23	-1,37	**0,51**	-1,30	1,08	-2,43	0,12
i39	0,41	0,45	-0,20	0,87	-0,20	1,23	-0,80	0,16
i40	**0,19**	0,22	-1,39	**0,32**	-1,27	1,69	-4,04	0,18
i41	0,48	0,45	-0,20	1,19	-0,20	1,78	-0,97	0,16
i42	0,30	0,38	-0,54	**0,55**	-0,52	1,44	-2,19	0,26
i43	0,31	0,31	-0,92	**0,58**	-0,88	1,66	-2,74	0,20
i44	0,21	0,14	-2,01	**0,45**	-1,88	2,82	-6,09	0,11
i45	**0,16**	0,11	-2,26	**0,34**	-2,09	1,34	-4,17	0,08

Legenda: *b*: dificuldade; *a*: discriminação; *c*: acerto ao acaso.
Fonte: Da pesquisa (2019).

Em todas as provas alguns itens apresentaram índices de discriminação pela TCT abaixo do valor considerado adequado (0,20). Alguns itens tiveram valores de discriminação entre 0,15 e 0,19. Em alguns trabalhos que analisam o Enem consideraram 0,15 como valor de discriminação aceitável (FERREIRA, 2009). O coeficiente KR das provas manteve-se acima de 0,65, mas só as provas de Linguagens e Códigos e Ciências Humanas tiveram valores acima de 0,70.

A partir da análise pela TCT mais alguns itens teriam que ser retirados da análise por apresentar baixa discriminação, ou seja, são ruins em diferenciar os participantes de menor

desempenho dos de maior desempenho. Para o caso do Enem, que é uma prova de seleção, isso não é interessante, pois poderiam prejudicar os participantes mais preparados.

Além disso, observa-se que os itens, em sua maioria, são mais difíceis em todas as provas. O item mais fácil apresentou índice de dificuldade de 0,66, ou seja, com 60% de acerto. Pela TCT é interessante que os itens apresentem índices entre todas as faixas de dificuldade, mas com a maioria dos itens entre 0,40 e 0,60. Isso é importante para que o teste tenha boa capacidade discriminativa.

Quando verificado pela TRI, os índices de discriminação dos itens para a prova de Linguagens e Códigos no modelo de 2PL, 14 itens apresentam valores baixos (<0,65). Para o modelo 3PL apenas um item. Na prova de Matemática 24 itens tiveram valores de discriminação baixos no modelo 2PL e apenas um no modelo 3PL. Na prova de Ciências da Natureza 23 itens apresentaram baixa discriminação e nenhum item no modelo 3PL. Na prova de Ciências Humanas 13 itens apresentaram baixos valores de discriminação e apenas dois itens no 3PL. Esses itens deveriam ser excluídos da prova.

Até aqui foi analisado o Enem 2017 a partir da TCT e TRI, em que foram estimados os parâmetros dos itens e o escore total (para a TCT) e a habilidade ou proficiência (para a TRI) dos participantes. Agora será realizado um comparativo entre as estatísticas dos itens e dos participantes a partir dos dois modelos.

Na Tabela 14 estão os coeficientes de correlação (r) e determinação (R^2) dos parâmetros dos itens para todas as provas. As análises foram realizadas entre a dificuldade e discriminação dos itens pela TCT e os parâmetros dos itens para os três modelos (1PL, 2PL e 3PL) na TRI.

Tabela 14 – Correlação entre os parâmetros dos itens da TCT e TRI do Enem 2017.

Provas	TCT	1PL		2PL		3PL	
		r	R^2	r	R^2	r	R^2
LC	d	0,997	99,40	0,997	99,40	0,931	86,68
	r_{pb}			0,944	89,11	0,251	6,30
MT	d	0,981	96,24	0,981	96,24	0,859	73,79
	r_{pb}			0,925	85,56	-0,162	2,68
CN	d	0,993	98,60	0,992	98,41	0,754	56,85
	r_{pb}			0,977	99,40	-0,129	1,66
CH	d	0,993	98,60	0,994	98,80	0,885	78,32
	r_{pb}			0,975	95,06	0,353	12,46

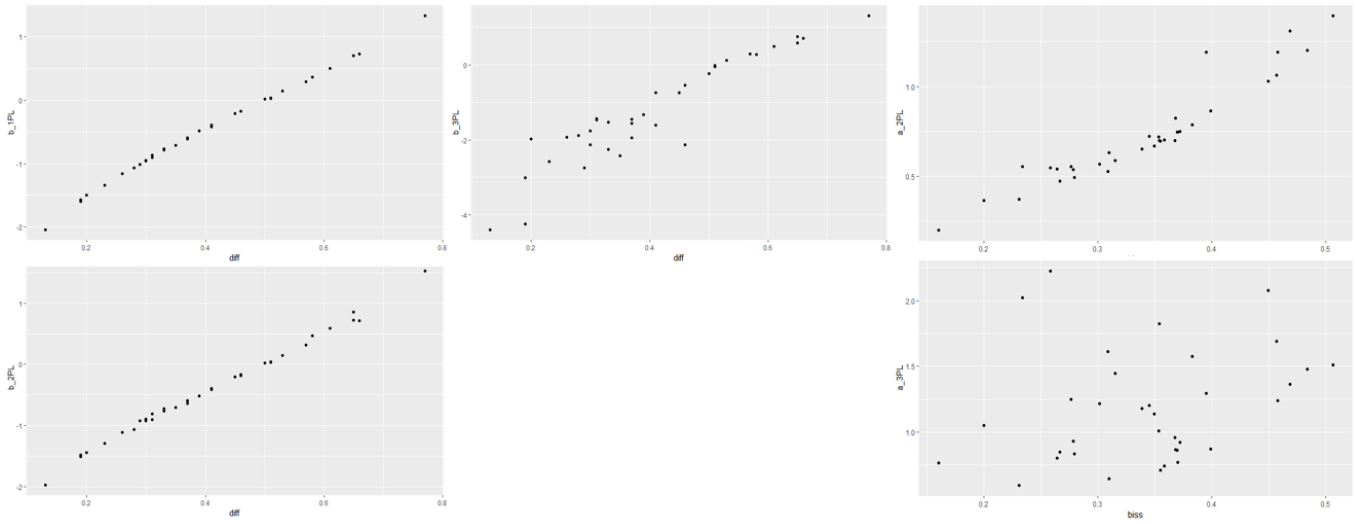
Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

Os coeficientes de correlação entre os parâmetros de dificuldade pela TCT e pela TRI nos modelos 1PL e 2PL para todas as provas estão acima de 0,99, o que é considerado uma correlação muito forte e quase perfeita. A correlação da dificuldade pela TCT e pela TRI modelo 3PL há diminuição dos valores que estão entre 0,75 e 0,93. Em todas as correlações entre os índices de discriminação pela TCT e TRI nos modelos de 1PL e 2PL os valores permanecem acima de 0,90. No entanto, quando são estimadas as correlações para a discriminação com o modelo 3PL os valores decrescem drasticamente para todas as provas (>0,40). Para as provas de Matemática e Ciências da Natureza as correlações são negativas.

As correlações analisadas são reforçadas pelo coeficiente de determinação. O percentual de explicação da variação existente entre o índice de dificuldade pela TCT e o modelo TRI de 1 parâmetro chegam a quase 100% para todas as provas do exame, em que o menor valor é de 96,24%. Os valores permanecem altos com o modelo de 2 parâmetros como um percentual mínimo de 95,06% de explicação. No entanto, caiu consideravelmente para o modelo de 3 parâmetros, com valor de 86,68% na prova de Linguagens e Códigos e apenas 56,85% na prova de Ciências da Natureza. Quando analisados esse valor para os índices de discriminação, o valor máximo indicado foi de apenas 12,46% de explicação.

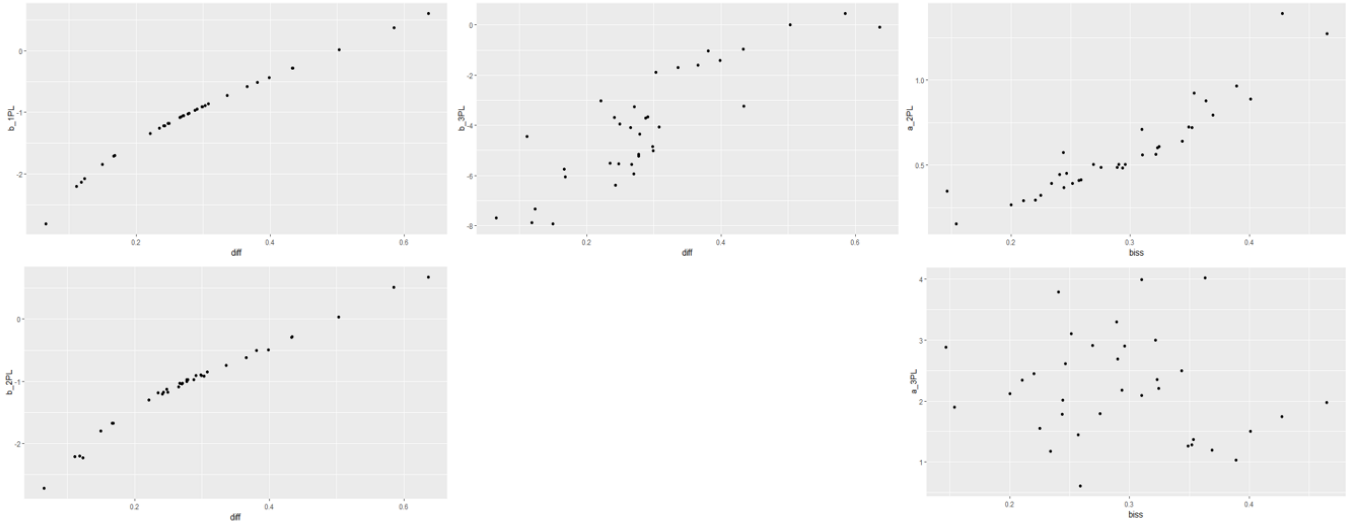
A representação da correlação a partir do gráfico de dispersão de pontos revela isso mais claramente. Essas ilustrações podem ser observadas nas Figuras de 6 a 9 a seguir. Os gráficos de dispersão entre a dificuldade pela TCT e TRI nos modelos de 1 e 2 parâmetros indicam uma correlação forte. No entanto, como se pode observar, os pontos dos gráficos que representam a correlação entre os parâmetros de dificuldade e discriminação pela TCT e TRI no modelo 3PL estão mais dispersos, ou sem padrão linear, o que indicam uma correlação fraca entre os parâmetros.

Figura 6 – Gráfico de dispersão de pontos entre os parâmetros dos itens pela TCT e TRI da prova de Linguagens e Códigos, Enem 2017.



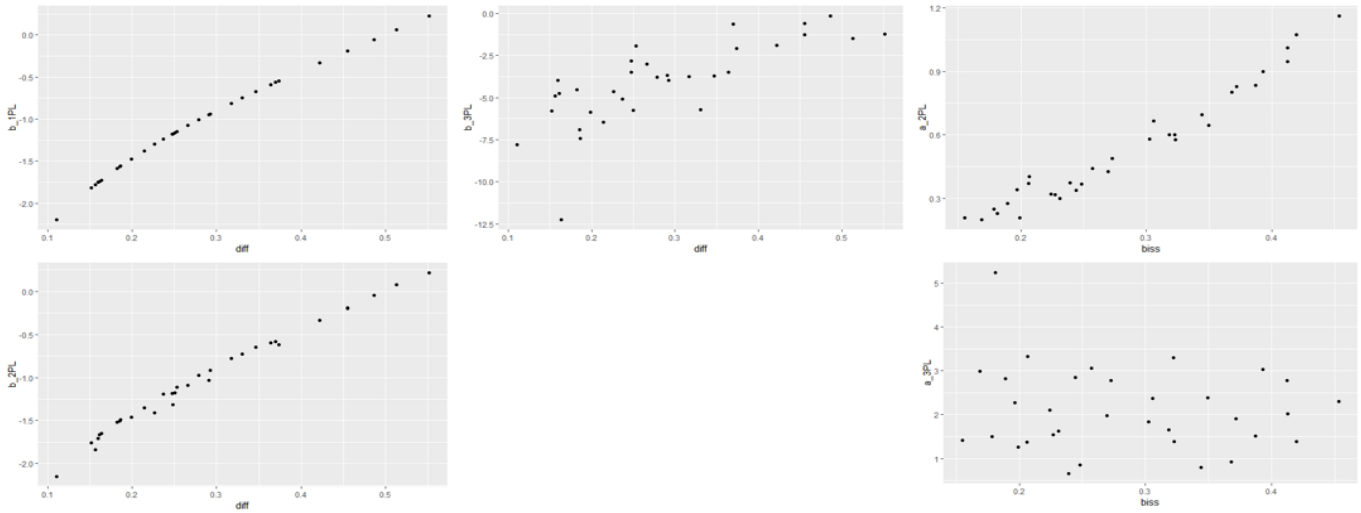
Fonte: Da pesquisa (2019).

Figura 7 – Gráfico de dispersão de pontos entre os parâmetros dos itens pela TCT e TRI da prova de Matemática, Enem 2017.



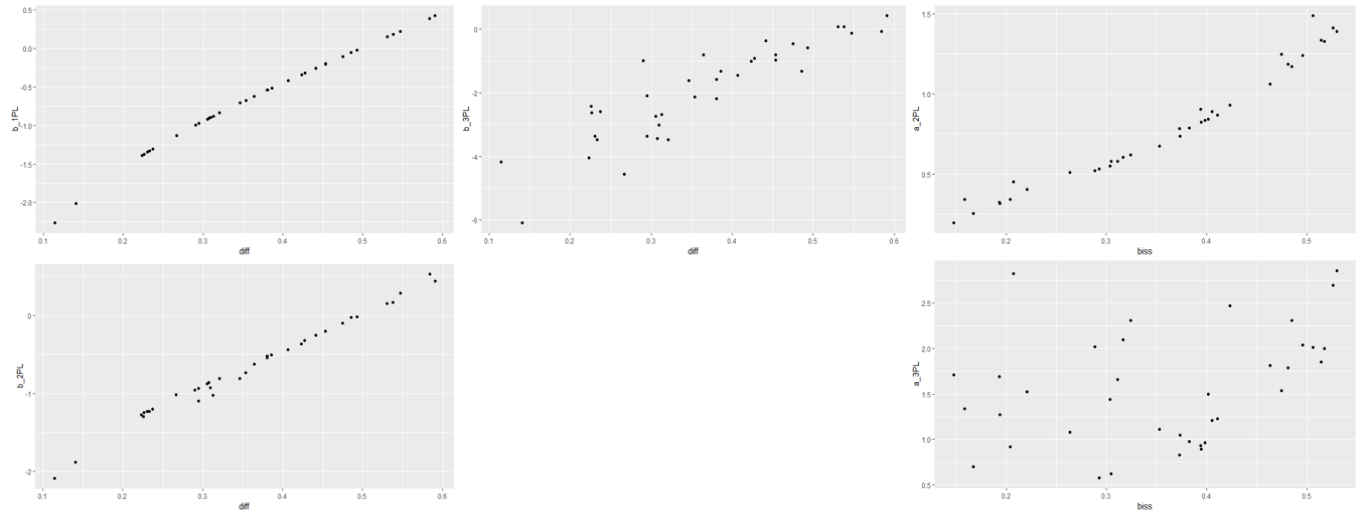
Fonte: Da pesquisa (2019).

Figura 8 – Gráfico de dispersão de pontos entre os parâmetros dos itens pela TCT e TRI da prova de Ciências da Natureza, Enem 2017.



Fonte: Da pesquisa (2019).

Figura 9 – Gráfico de dispersão de pontos entre os parâmetros dos itens pela TCT e TRI da prova de Ciências Humanas, Enem 2017.



Fonte: Da pesquisa (2019).

Para comparar os parâmetros dos itens estimados pela TCT e TRI foi classificado pelo nível de dificuldade em ordem decrescente. O objetivo foi identificar se ambos os modelos geram parâmetros de dificuldade diferentes para os itens. Os resultados estão dispostos na Tabela 15.

Tabela 15 – Classificação dos itens a partir do parâmetro de dificuldade.

	LC				MT				CN				CH			
	TCT	TRI			TCT	TRI			TCT	TRI			TCT	TRI		
Posição		1PL	2PL	3PL		1PL	2PL	3PL		1PL	2PL	3PL		1PL	2PL	3PL
1	i20	i20	i20	i20	i29	i29	i29	i38	i40	i40	i40	i14	i45	i45	i45	i44
2	i13	i13	i45	i45	i33	i33	i39	i7	i28	i28	i6	i40	i44	i44	i44	i37
3	i45	i45	i13	i13	i7	i7	i33	i29	i5	i6	i28	i8	i40	i40	i38	i45
4	i12	i12	i12	i4	i39	i39	i7	i39	i6	i37	i37	i16	i3	i3	i40	i40
5	i18	i18	i18	i18	i38	i38	i38	i10	i14	i5	i5	i31	i12	i38	i3	i32
6	i10	i10	i10	i44	i4	i25	i4	i4	i37	i14	i14	i38	i21	i21	i21	i12
7	i24	i24	i24	i5	i25	i4	i25	i20	i27	i27	i27	i28	i38	i12	i12	i18
8	i4	i4	i38	i2	i8	i8	i8	i25	i8	i16	i16	i45	i34	i34	i34	i21
9	i30	i38	i4	i38	i45	i45	i43	i3	i16	i8	i8	i42	i37	i37	i1	i28
10	i38	i30	i1	i12	i10	i43	i45	i13	i38	i38	i38	i24	i24	i24	i37	i17
11	i1	i1	i30	i14	i43	i10	i10	i45	i31	i31	i22	i6	i1	i1	i35	i43
12	i41	i41	i41	i10	i13	i13	i14	i15	i22	i22	i31	i5	i28	i28	i24	i35
13	i5	i5	i5	i24	i14	i14	i13	i42	i24	i24	i17	i22	i17	i43	i28	i3
14	i16	i16	i16	i30	i3	i28	i28	i12	i1	i1	i24	i27	i18	i18	i17	i34
15	i44	i44	i44	i40	i20	i3	i20	i32	i2	i17	i1	i41	i35	i17	i43	i38
16	i7	i22	i14	i22	i28	i20	i3	i33	i17	i45	i45	i37	i43	i35	i18	i42
17	i14	i7	i7	i16	i37	i37	i37	i30	i45	i2	i2	i18	i32	i32	i32	i7
18	i22	i14	i22	i41	i15	i15	i15	i28	i39	i39	i39	i12	i7	i9	i9	i1
19	i32	i32	i32	i7	i30	i42	i30	i16	i18	i18	i43	i15	i9	i7	i7	i9
20	i40	i42	i42	i1	i42	i30	i42	i14	i41	i43	i18	i43	i20	i20	i20	i31
21	i42	i40	i40	i32	i1	i1	i1	i1	i43	i41	i41	i1	i31	i31	i31	i15
22	i3	i3	i3	i3	i22	i22	i17	i43	i12	i12	i12	i33	i42	i42	i42	i29
23	i2	i31	i31	i42	i12	i12	i12	i22	i42	i42	i42	i39	i29	i29	i29	i33
24	i31	i2	i2	i31	i17	i32	i22	i37	i15	i15	i15	i17	i15	i15	i15	i27
25	i9	i9	i9	i9	i32	i17	i32	i5	i33	i33	i19	i19	i27	i27	i27	i24
26	i17	i17	i17	i37	i16	i16	i16	i8	i19	i20	i33	i2	i22	i22	i22	i41
27	i37	i37	i37	i17	i31	i31	i31	i17	i20	i19	i20	i13	i11	i11	i11	i22
28	i43	i43	i43	i43	i35	i35	i35	i31	i13	i13	i13	i36	i39	i39	i39	i20
29	i23	i23	i23	i19	i24	i24	i24	i35	i7	i7	i7	i29	i41	i41	i41	i39
30	i19	i19	i19	i23	i9	i9	i9	i9	i29	i29	i29	i3	i26	i26	i26	i5
31	i11	i11	i11	i11	i5	i5	i23	i24	i4	i4	i4	i20	i5	i33	i33	i26
32	i6	i6	i21	i6	i23	i23	i5	i23	i36	i36	i36	i7	i33	i5	i5	i11
33	i27	i27	i6	i21	i27	i27	i27	i2	i3	i3	i3	i4	i4	i4	i4	i2
34	i21	i21	i27	i27	i26	i26	i26	i27					i16	i16	i16	i13

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas;
Fonte: Da pesquisa (2019).

Quando se observa a classificação dos itens é possível identificar que alterações na dificuldade dos itens ocorrem no modelo de 3 parâmetros. Grandes alterações na classificação não ocorrem entre a dificuldade dos itens na TCT e TRI de 1 e 2 parâmetros.

Neste momento será realizado as análises a partir dos escores dos participantes, que atualmente são convertidos para uma escala de nota que vai de 0 a 1000 pontos. Como as notas são calculadas a partir da TRI, as notas apresentam média de 500 e desvio padrão de 100 pontos. Para fins de comparação, a nota dos participantes pela TCT também foi convertida para esta escala. Os escores médios dos participantes pela TCT, esses variam entre 362,00 e 406,41 nas

quatro provas do exame. No entanto, nos modelos de 1, 2 e 3 parâmetros observa-se que os valores se estabilizam na média de 500,0 pontos, o que era esperado, como está na Tabela 16.

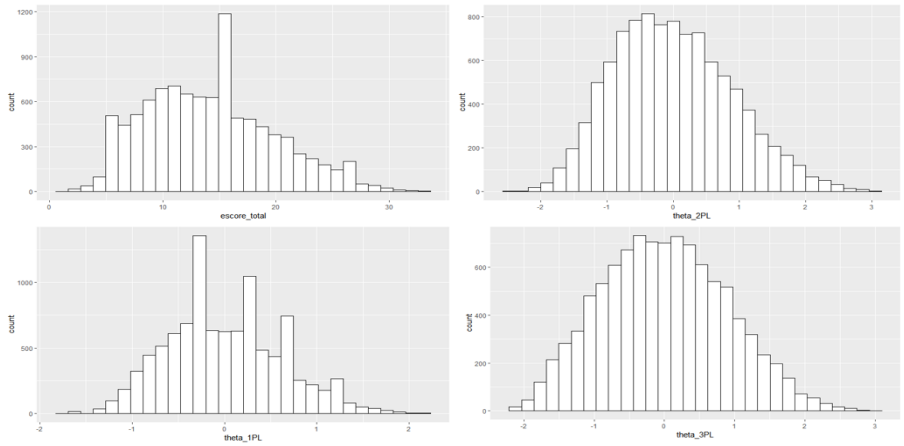
Tabela 16 – Desempenho médio dos participantes do Enem 2017.

Prova	TCT	TRI		
		1PL	2PL	3PL
LC	406,41	500,0	500,0	500,0
MT	288,70	500,0	500,0	500,0
CN	286,20	500,0	500,0	500,0
CH	362,00	500,0	500,0	500,0

Legenda: M: média; dp: desvio padrão; LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
Fonte: Da pesquisa (2019).

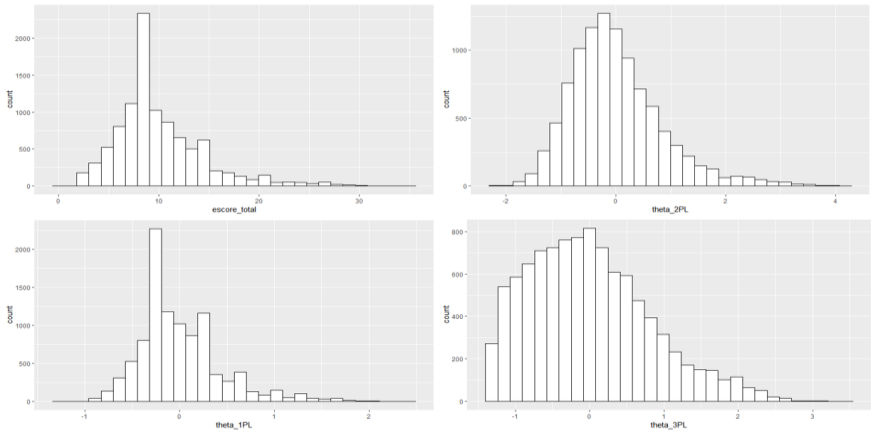
Um histograma também foi construído como forma de comparar a distribuição das pontuações dos participantes a partir da TCT e modelos de TRI de 1, 2 e três parâmetros. Esses histogramas estão nas Figuras de 10 a 13.

Figura 10 – Histograma do desempenho dos participantes pela TCT e TRI da prova de Linguagens e Códigos, Enem 2017.



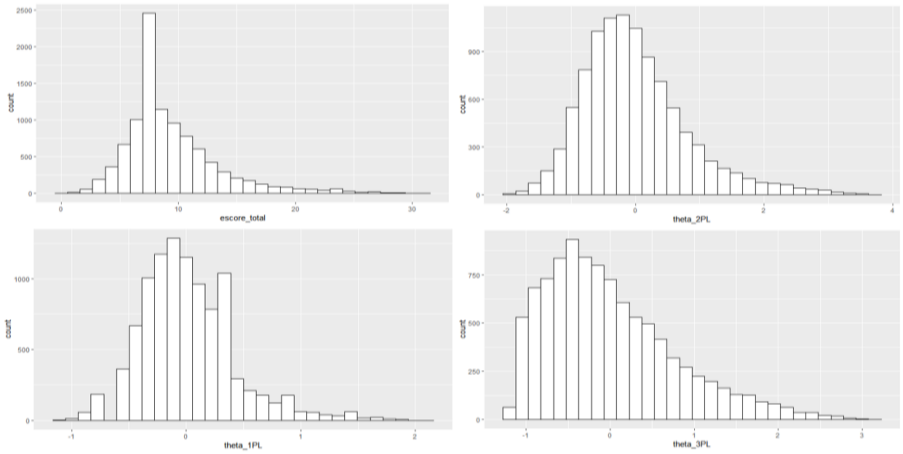
Fonte: Da pesquisa (2019).

Figura 11 – Histograma do desempenho dos participantes pela TCT e TRI da prova de Matemática, Enem 2017.



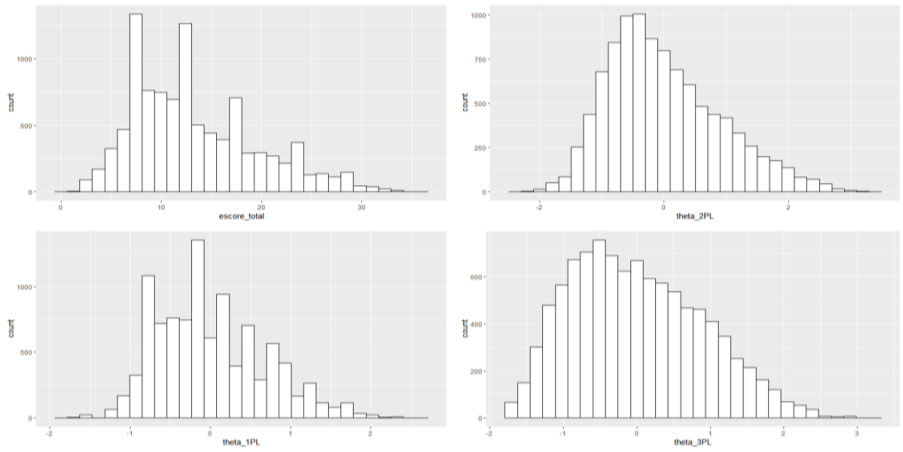
Fonte: Da pesquisa (2019).

Figura 12 – Histograma do desempenho dos participantes pela TCT e TRI da prova de Ciências da Natureza, Enem 2017.



Fonte: Da pesquisa (2019).

Figura 13 – Histograma do desempenho dos participantes pela TCT e TRI da prova de Ciências Humanas, Enem 2017.



Fonte: Da pesquisa (2019).

Nos histogramas apresentados é possível observar bastante semelhança entre as distribuições das pontuações dos participantes pela TCT e TRI no modelo de 1 parâmetro para todas as provas do exame. Também é possível observar que a distribuição das pontuações dos sujeitos obtidas pelo modelo de TRI de 3 parâmetros, com exceção na prova de Linguagens e Códigos, se desloca para a esquerda, ou seja, aumenta-se a frequência de participantes. Possivelmente isso ocorreu devido ao parâmetro de acerto casual.

Após essa análise descritiva das pontuações dos participantes a partir dos modelos analisados, o próximo passo foi correlacionar essas pontuações (*r*) e obter o coeficiente de determinação (R^2). Uma vez observada correlação entre os escores dos participantes foi implementado uma análise de regressão linear simples considerando o escore dos participantes pela TCT como uma variável independente e o escore do participante estimado pelo modelo de TRI como uma variável dependente. O objetivo da análise foi descobrir se os escores estimados pelo primeiro modelo são bons previsores do segundo. Esses valores estão dispostos na Tabela 17.

Tabela 17 – Correlação entre o escore total da TCT e a habilidade da TRI no Enem 2017.

Provas	1PL			2PL			3PL		
	r	R^2	F	*r*	R^2	F	*r*	R^2	F
LC	0,999	99,80	5302892*	0,986	97,23	364712,8*	0,975	96,06	191862,2*
MT	0,999	99,80	2406372*	0,966	93,31	140400,6*	0,885	78,32	36027,38*
CN	0.998	99,60	2717501*	0,961	92,35	120354,5*	0.903	81,54	44484,4*
CH	0,998	99,60	2716074*	0,984	96,83	306861,2*	0,969	93,90	156697,1*

Legenda: LC: Linguagens e Códigos; MT: Matemática; CN: Ciências da Natureza; CH: Ciências Humanas.
*Modelo de regressão estatisticamente significativo ($p<0,001$)
Fonte: Da pesquisa (2019).

Como se observa na tabela acima, todas as correlações realizadas foram acima de 0,80, consideradas correlações fortes. Para todas as provas a correlação entre o escore total e a pontuação dos participantes pela modelo de TRI com 1 parâmetro, ou seja, considerando apenas a dificuldade do item permanece com valores acima de 0,99, uma correlação quase perfeita. As correlações continuam fortes com os outros modelos, mas há um leve decréscimo. A correlação entre o escore total e a pontuação obtida pela TRI de 3 parâmetros para a prova de Matemática foi a menor com um valor de 0,885.

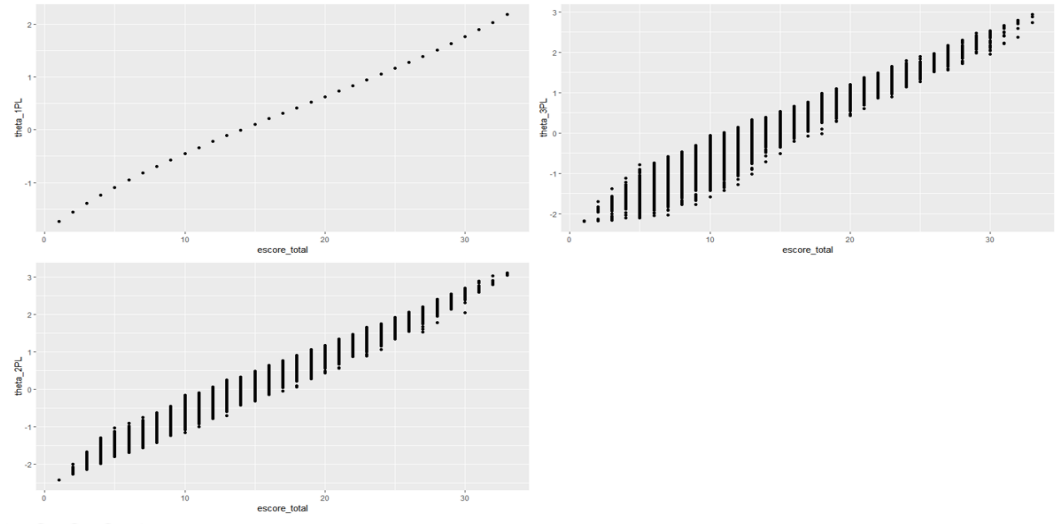
A análise do coeficiente de determinação indica que entre as pontuações dos participantes pela TCT e o modelo de TRI de 1 parâmetro explicam mais de 99% da variação uma da outra em todas as provas do exame. Esse valor tem leve diminuição quando comparado como o modelo de TRI de 2 parâmetros, mas permanecendo com um percentual de explicação acima de 90%. Quando comparado as pontuações dos participantes pela TCT e o modelo de

TRI de 3 parâmetros há uma diminuição considerável nas provas de Matemática e Ciências da Natureza, sendo 78,32% e 81,54%, respectivamente.

O modelo de regressão linear simples aplicado, em que considera o escore calculado pela TRI como variável dependente e pela TCT como variável independe apresentou significância em todas as provas. Isso indica que o escore total é um bom previsor dos escores da TRI nos três modelos.

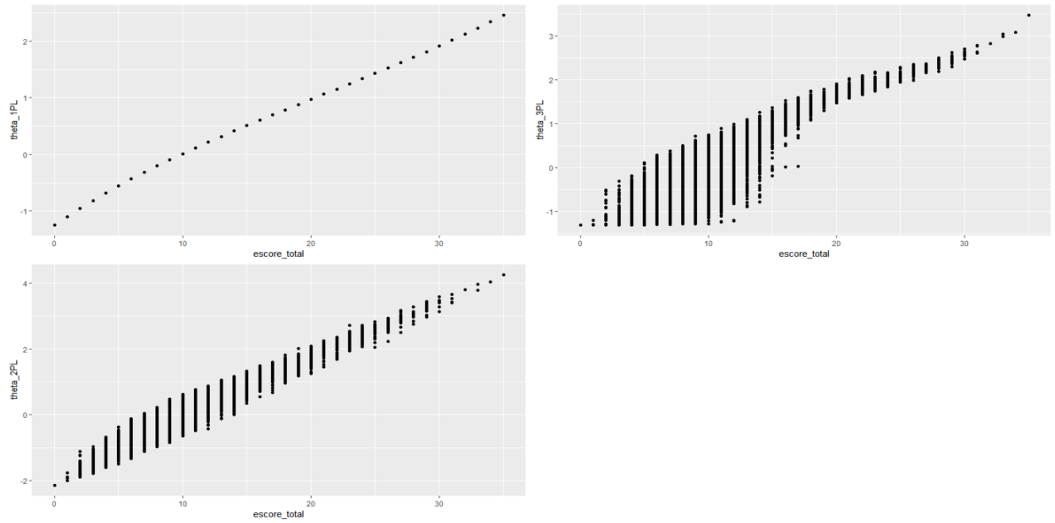
Como forma de ilustrar os valores de correlação apresentados acima, foi construído um gráfico de dispersão de pontos que estão nas Figuras de 14 a 17. Em todas as figuras pode-se observar uma correlação forte. Porém, como indicado nos valores apresentados anteriormente, a correlação entre o escore total e a pontuação do sujeito pela TRI decresce à medida que se acrescenta mais parâmetros ao modelo de TRI. Isso pode ser observado com os pontos dos gráficos mais dispersos para as correlações entre TCT e modelo de TRI de 2 e 3 parâmetros. Isso pode ser observado para todas as provas do Enem 2017.

Figura 14 – Dispersão de pontos entre o escore total dos participantes pela TCT e a habilidade pela TRI da prova de Linguagens e Códigos, Enem 2017.



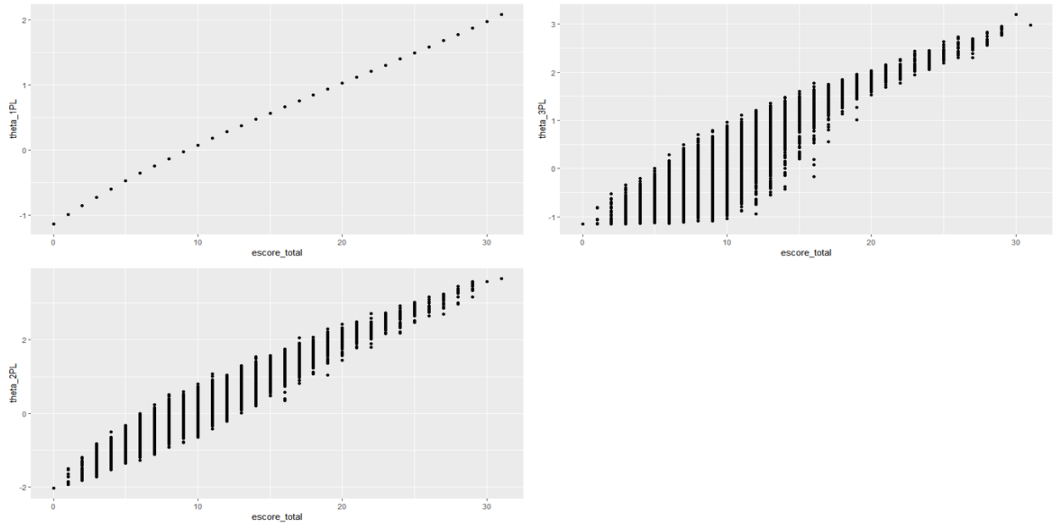
Fonte: Da pesquisa (2019).

Figura 15 – Dispersão de pontos entre o escore total dos participantes pela TCT e a habilidade pela TRI da prova de Matemática, Enem 2017.



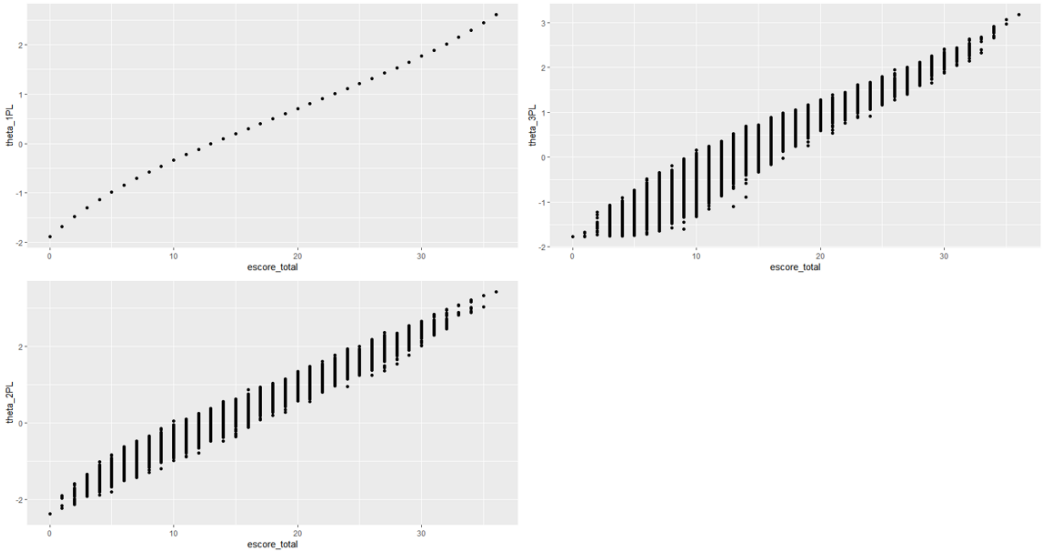
Fonte: Da pesquisa (2019).

Figura 16 – Dispersão de pontos entre o escore total dos participantes pela TCT e a habilidade pela TRI da prova de Ciências da Natureza, Enem 2017.



Fonte: Da pesquisa (2019).

Figura 17 – Dispersão de pontos entre o escore total dos participantes pela TCT e a habilidade pela TRI da prova de Ciências Humanas, Enem 2017.



Fonte: Da pesquisa (2019).

A comparabilidade entre os dois modelos pode ser identificada em estudos teóricos e empírico. Discussões teóricas dos modelos podem ser encontrados em Hambleton e Jones (1993), Navas (1994), Muñiz (2010) e Sartes e Sousa-Formigoni (2013).

Nesses textos há um certo consenso em relação as limitações da TCT e os avanços oferecidos pela TRI. Entre os principais problemas da TCT é que as pontuações são dependentes do teste e da amostra (HAMBLETON; JONES, 1993; MUÑIZ, 2010; NAVAS, 1994). Isso quer dizer que os itens terão parâmetros de dificuldade e discriminação diferentes quando aplicados a amostras diferentes, assim como quando um mesmo grupo é submetido a dois testes terão escores totais distintos. Soluções para esses problemas são propostos pelos modelos de TRI, uma vez que os parâmetros dos itens são invariantes (KLEIN, 2013). Com isso, é possível equalizar testes com itens diferentes, mas que tenham o mesmo nível de dificuldade.

No entanto, embora a TRI apresente soluções para os problemas da TCT, seus pressupostos são muito restritivos e necessita de amostras relativamente maiores, o que dificulta o ajuste do modelo aos dados. Com isso, o modelo clássico parece mais viável de aplicação prática, considerando que é possível ajustar um modelo para amostras menores e o modelo matemático é mais simples de compreensão (HAMBLETON; JONES, 1993).

As propriedades psicométricas oferecidas pela TRI são particularmente úteis quando se deseja a comparabilidade dos resultados entre os anos. No caso do Enem, que no modelo atual, não é parâmetro para monitoramento e avaliação do sistema educacional

brasileira, considerando que já existe o Saeb para esse objetivo, as propriedades da TRI parecem não ser úteis aos objetivos do exame. Isso porquê os candidatos que realizam a prova podem utilizar sua pontuação apenas no ano seguinte, quando na concorrência aos cursos de graduação nas IES brasileiras.

Outras possibilidades permitidas pela TRI são a criação de um grande banco de itens comparáveis e a possibilidade de realização de testes adaptativos computadorizados (PASQUALI, 2009), de modo que o exame pudesse ser aplicado a qualquer momento ou várias vezes ao ano.

Por outro lado, alguns autores argumentam fortemente pela relação complementar entre a TCT e TRI (BECHGER et al., 2003; COSTA; FERRÃO, 2015). Bechger et al. (2003) alega pela complementaridade entre os modelos e indica que parâmetros de TCT podem ser utilizados quando um modelo de TRI é bem ajustado aos dados. Costa e Ferrão (2015) defendem a utilização da TCT na análise de itens na fase de pré-teste e pré-calibração como forma de minimização de custos no desenvolvimento de banco de itens. Esse argumento é colocado ao indicar que há alta correlação entre os parâmetros dos pela TCT e TRI.

Em outros estudos são realizadas comparações entre os modelos a partir de estudos empíricos ou por simulação Monte Carlo (ADEDOYIN; ADEDOYIN, 2013; ADEGOKE, 2013; AWOPEJU; AFOLABI, 2016; COSTA; FERRÃO, 2015; COURVILLE, 2004; ERGÜVEN, 2014; FAN, 1998; KOHLI; KORAN; HENN, 2014; MACDONALD; PAUNONEN, 2002; PROGAR; SOČAN, 2008). Nesses estudos são identificados comparabilidade entre as pontuações dos itens e dos sujeitos entre TCT e TRI.

Essas pesquisas também utilizam o coeficiente de correlação para realizar as comparações entre os parâmetros de dificuldade e discriminação dos itens e das pontuações dos sujeitos entre os dois modelos. Em uma das pesquisa (ADEDOYIN; ADEDOYIN, 2013) foi utilizado também o teste “t” dependente para as comparações. Adicionalmente foi realizado um teste de regressão linear simples considerando o escore os sujeitos pela TCT como variável independente e o escore pela TRI como variável dependente.

Em acordo com os resultados desta pesquisa, em todos os trabalhos, fortes correlações foram identificadas entre os parâmetros de dificuldade e discriminação dos itens, assim como entre as pontuações dos sujeitos pela TCT e TRI. Ao comparar os escores dos sujeitos pela TCT e TRI correlações fortes também foram encontradas (COURVILLE, 2004; FAN, 1998; MACDONALD; PAUNONEN, 2002; PROGAR; SOČAN, 2008). Corroborando com os resultados desta pesquisa, em dois trabalhos (AWOPEJU; AFOLABI, 2016;

COURVILLE, 2004) essas correlações diminuem quando se compara TCT com o modelo TRI de 3 parâmetros.

Diante disso, parece haver maior comparabilidade entre os parâmetros dos itens e as pontuações dos sujeitos pela TCT e TRI com apenas um parâmetro, ou seja, considerando apenas a dificuldade dos itens. Isso pode ocorrer devido a interferência do terceiro parâmetro, o de acerto casual, utilizado no modelo de TRI adotado no exame. Esse parâmetro faz com que candidatos com baixa habilidade em uma prova que não cumpre o pressuposto de unidimensionalidade não pontue em determinados itens, mesmo que os candidatos os acertem sistematicamente (PRIMI; CICCHETTO, 2018). Esse estudo levanta uma questão importante, a de que candidatos estão sendo prejudicados por falta de ajuste de um modelo matemático a realidade empírica do exame.

Na presente pesquisa a aplicação da regressão linear indicou que o escore de TCT é um ótimo previsor dos escores calculados pela TRI, mesmo no modelo com três parâmetros. Isso indica que a nota da TRI é bem explicada pela nota a partir da TCT.

Em relação a estabilidade dos parâmetros dos itens há resultados contraditórios entre as pesquisas. No estudo de Adegoke (2013) os parâmetros de dificuldade e discriminação pela TRI se apresentaram mais estáveis. No entanto, nos trabalhos de Fan (1998) os parâmetros dos itens foram invariantes na TCT e TRI. MacDonald e Paunonen (2002) identificaram invariância e mais consistência nos parâmetros dos itens pela TCT.

Dessa forma, no que se refere aos parâmetros dos itens e os escores dos sujeitos entre a TCT e TRI, esses apresentaram-se altamente comparáveis nos estudos analisados. Em dois estudos a invariâncias dos parâmetros foi identificada nos dois modelos, que é um dos principais argumentos pela superioridade teórica dos modelos de resposta ao item.

Com base nos resultados encontrados e nas discussões realizadas, na próxima seção é apresentada as principais conclusões extraídas dessa pesquisa. Além disso, é realizada algumas reflexões e recomendações com base nessas conclusões.

7 CONCLUSÕES

Esta tese teve como objetivo principal avaliar os resultados do Enem 2017 a partir dos modelos de TCT e TRI. As hipóteses iniciais eram que os itens das provas do exame não apresentam o pressuposto de unidimensionalidade. No entanto, quando possível ajustar um modelo unidimensional, os parâmetros dos itens e os escores dos participantes são comparáveis.

Os resultados indicaram que os itens das provas do exame não atenderam o pressuposto de unidimensionalidade, pois só foi possível ajustar um modelo unidimensional após a exclusão de alguns itens. Na prova de matemática, por exemplo, um modelo unidimensional foi ajustado após a exclusão de 12 dos 45 itens da prova.

No que se refere as comparações dos modelos, foi identificado fortes correlações entre os parâmetros de dificuldade dos itens da TCT e nos três modelos de TRI. O mesmo ocorreu na comparação entre os escores dos participantes, além do teste de regressão linear simples ter se ajustado muito bem aos dados, indicando forte comparabilidade entre os resultados gerados pelos dois modelos. No entanto, essa comparabilidade diminui com o modelo de TRI com três parâmetros, provavelmente em decorrência ao parâmetro de acerto casual. Um quadro síntese dos resultados de forma mais detalhada é apresentado a seguir.

Quadro 5 – Síntese dos resultados.

Objetivo: Analisar a dimensionalidade das provas das quatro áreas do Exame Nacional do Ensino Médio, Linguagens e Códigos, Ciências Humanas, Ciências da Natureza e Matemática, a partir da análise paralela e análise fatorial de informação completa.	
Método	**Resultado**
Análise Paralela e Análise fatorial de Informação Plena	Os itens das provas do exame não apresentaram o pressuposto de unidimensionalidade. Um modelo com apenas um fator foi possível de ser ajustado após a exclusão de 10 itens nas provras de LC e MT, 12 itens na prova de CN e 8 itens na prova de CH.
Objetivo: Estimar os parâmetros de dificuldade e discriminação dos itens e o escore total dos participantes do Enem 2017 via Teoria Clássica dos Testes.	
Método	**Resultado**
Proporção de acertos para a dificuldade, correlação ponto bisserial para a discriminação e somatório de itens acertados para a escore do participante.	- A maioria dos itens apresentaram proporção de acerto abaixo de 0,50, indicando elevada dificuldade. - Para todas a provas a maioria dos itens apresentou discriminação adequada, mas alguns itens seriam excluídos por apresentarem valores abaixo de 0,20. - O escore total variou entre 362,00 e 406,41, numa escala de 0 a 1000.

Quadro 7 (Continuação)

Objetivo: Estimar os parâmetros de dificuldade, discriminação e acerto casual dos itens das provas do Enem 2017 e a habilidade dos sujeitos via Teoria de Resposta ao Item.	
Método	**Resultado**
Modelos de TRI de 1, 2 e três parâmetros	**Linguagens e Códigos:** - Para o modelo 1PL *b* variou entre -2,04 e 1,33. - Para o modelo 2PL *b* variou entre -1,97 e 1,53 e 14 itens apresentam valores *a* baixos (<0,65). - Para o modelo 3PL *b* variou entre -4,39 e 1,30 e apenas um item apresentou baixo valor de *a*. O valor *c* variou entre 0,01 e 0,34. **Matemática:** - Para o modelo 1PL *b* variou entre -2,81 e 0,60. - Para o modelo 2PL *b* variou entre -2,72 e 0,67 e 24 itens tem valor *a* baixo (<0,65). - Para o modelo 3PL *b* variou entre -7,93 e 0,44, apenas um item com valor *a* baixo (<0,65), e *c* entre 0,01 e 0,34. **Ciências da Natureza:** - Para o modelo 1PL *b* variou entre -2,19 a 0,22. - Para o modelo 2PL *b* variou entre -2,15 e 0,22 e 23 itens tem valor *a* baixo (<0,65). - Para o modelo 3PL *b* variou entre -12,24 e 0,15, todos os itens com valor *a* acima de 0,65, e *c* entre 0,01 e 0,39. **Ciências Humanas:** - Para o modelo 1PL *b* variou entre -2,26 e 0,42. - Para o modelo 2PL b variou entre -2,09 e 0,53 e 13 itens tem valor *a* baixo (<0,65). - Para o modelo 3PL *b* variou entre -6,09 e 0,42, apenas dois item com valor *a* baixo (<0,65), e *c* entre 0,01 e 0,26.
Objetivo: Comparar e correlacionar os parâmetros dos itens da Teoria Clássica dos Testes com os parâmetros dos modelos logísticos de 1, 2 e 3 parâmetros da Teoria de Resposta ao Item	
Método	**Resultado**
Coeficiente de correlação de Pearson (*r*)	- Para todas as provas fortes correlações (>0,9) foram identificadas entre os parâmetros de dificuldade dos itens pela TCT e modelos 1PL e 2PL. A correlação decresce levemente com o modelo 3PL. - Para todas as provas fortes correlações (>0,9) foram identificadas entre os parâmetros de discriminação dos itens pela TCT e modelo 2PL. A correlação decresce consideravelmente com o modelo 3PL.
Objetivo: Comparar e correlacionar o desempenho dos estudantes estimados a partir da Teoria Clássica dos Testes com o desempenho estimado a partir dos modelos logísticos de 1, 2 e 3 parâmetros da Teoria de Resposta ao Item	
Método	**Resultado**
Coeficiente de correlação de Pearson (*r*) e regressão linear simples	- Para todas as provas fortes correlações (>0,9) foram identificadas entre o escore total pela TCT e a habilidade dos participantes pelos modelos 1PL e 2PL. A correlação decresce levemente com o modelo 3PL. - O gráfico de dispersão de pontos indica maior dispersão nas correlações entre o escore total pela TCT e TRI a medida que se acrescenta mais parâmetros ao modelo. - Todos os modelos apresentaram bom ajuste e significância estatística, indicando que os escores dos participantes calculados pela TCT são bons previsores dos escores calculados para os três modelos de TRI (1, 2 e 3 parâmetros).

Fonte: Da pesquisa (2019).

A partir disso, considerando os resultados desta pesquisa, indica-se que as provas do Enem podem ser analisadas a partir do paradigma da psicometria clássica, pois esse modelo é suficiente para as finalidades desse exame que atualmente não é utilizado para o

acompanhamento das redes de ensino, bem como não tem sido útil como ferramenta pedagógica de análise do desempenho dos participantes. Na hipótese de os itens medirem apenas uma dimensão, o modelo de TRI com um parâmetro seria suficiente e adequado para a análise dos resultados.

Acredita-se que os resultados desta tese podem contribuir para a discussão da utilização do modelo de TRI unidimensional no Enem, considerando a falta de evidência sobre a sua estrutura latente. No entanto, ressalta-se a necessidade de mais estudos empíricos que possibilitem melhores informações sobre essa questão, sobretudo, debates e reflexões em relação a real necessidade desse modelo para os atuais objetivos do exame. Como discutido, não é apenas uma questão de modelo matemático, mas de adequação a pressupostos que, se violados, podem prejudicar milhões de candidatos que almejam uma vaga nas IES de todo o país.

8 REFERÊNCIAS

ADEDOYIN, O. O.; ADEDOYIN, J. A. Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. **Herald Journal of Education and General Studies**, v. 2, n. 3, p. 107–114, 2013.

ADEGOKE, B. A. Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. **Journal of Education and Practice**, v. 4, n. 22, p. 87–96, 2013.

AERA; APA; NCME. **Standards for educational and psychological testing**. New York: [s.n.].

AMARO, I. Avaliação externa da escola: repercussões, tensões e possibilidades. **Estudos em Avaliação Educacional**, v. 24, n. 54, p. 32–55, 2013.

ANDRADE, D. F. DE; TAVARES, H. R.; VALLE, R. DA C. **Teoria de Resposta ao Item: Conceitos e Aplicações**. [s.l.] ABE – Associação Brasileira de Estatística, 2000.

ANDRADE, J. M. DE; LAROS, J. A.; GOUVEIA, V. V. O uso da teoria de resposta ao item em avaliações educacionais: diretrizes para pesquisadores. **Avaliação Psicológica**, v. 9, n. 3, p. 421–435, 2010.

ANDRADE, G. G. A metodologia do Enem: uma reflexão. **Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB**, n. 33, p. 67–76, 2012.

ANDRIOLA, W. B. Avaliação da aprendizagem: uma análise descritiva segundo a teoria de resposta ao item (TRI). **Educação em Debate**, v. 20, n. 36, p. 93–102, 1998.

ANDRIOLA, W. B. Psicometria Moderna:características e tendências. **Estudos em Avaliação Educacional**, v. 20, n. 43, p. 319, 30 ago. 2009.

ARIAS, M. R. M.; LLOREDA, M. J. H.; LLOREDA, M. V. H. **Psicometría**. Alianza Ed ed. Madrid: [s.n.].

AWOPEJU, O. A.; AFOLABI, E. R. I. Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. **European Scientific Journal**, v. 12, n. 28, p. 263–284, 2016.

AYALA, R. J. **The theory and practice of item response theory**. New York: THE GUILFORD PRESS, 2009.

BAKER, F. B. **Item response theory: parameter estimation techniques**. New York: Marcel Dekker, 1992.

BAKER, F. B. **The basics of Item Response Theory**. Second Edi ed. United States of America: [s.n.].

BARTHOLOMEW, D. J. Factor Analysis for Categorical Data. **Journal of the Royal Statistical Society**, v. 42, n. 3, p. 293–321, 1980.

BAUER, A.; ALAVARSE, O. M.; OLIVEIRA, R. P. DE. Avaliações em larga escala : uma sistematização do debate. **Educ. Pesqui.**, v. 41, n. especial, p. 1367–1382, 2015.

BECHGER, T. M. et al. Using Classical Test Theory in Combination with Item Response Theory. **Applied Psychological Measurement**, v. 27, n. 5, p. 319–334, 27 set. 2003.

BEJAR, I. I. A procedure for investigating the unidimensionality of achieviment tests based on item parameter estimates. **Journal of educational measurement**, v. 17, n. 4, p. 283–296, 1980.

BÉLAND, S. et al. Impact of simple substitution methods for missing data on classical test theory difficulty and discrimination. **The Quantitative Methods for Psychology**, v. 14, n. 3, p. 180–192, 2018.

BOCK, R. D.; AITKIN, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. **Psychometrika**, v. 46, n. 4, p. 443–459, 1981.

BOCK, R. D.; GIBBONS, R.; MURAKI, E. Full-Information Item Factor Analysis. **Applied Psychological Measurement**, v. 12, n. 3, p. 261–280, 1988.

BONAMINO, A.; SOUSA, S. Z. L. Três gerações de avaliação da educação básica no Brasil: interfaces com o currículo da/na escola. **Educação e Pesquisa**, v. 38, n. 2, p. 373–388, 2012.

BRASIL. **Portaria nº 438, de 28 de maio de 1998**, 1998.

BRASIL. **Portaria nº 109, de 27 de maio de 2009**, 2009.

BRASIL. **Teoria de resposta ao item avalia habilidade e minimiza o "chute" de candidatos**. Disponível em: <http://portal.mec.gov.br/ultimas-noticias/389-ensino-medio-2092297298/17319-teoria-de-resposta-ao-item-avalia-habilidade-e-minimiza-o-chute>.

CANÇADO, R.; CASTRO, M. J. P.; OLIVEIRA, I. F. DE. **Análise pedagógica de itens de teste por meio da teoria de resposta ao item**, 2013. (Nota técnica).

CASASSUS, J. Política y metáforas: un análisis de la evaluación estandarizada em el contexto de la política educativa. In: BAUER, A.; GATTI, B. A.; TAVARES, M. R. (Eds.). . **Vinte e cinco anos de avaliação de sistemas educacionais no Brasil – Origens e pressupostos**. 1. ed. Florianópolis: Insular, 2013. p. 192.

CHALMERS, R. P. mirt : A Multidimensional Item Response Theory Package for the R Environment. **Journal of Statistical Software**, v. 48, n. 6, p. 1–29, 2012.

CONDE, F. N.; LAROS, J. A. Unidimensionalidade e a propriedade de invariância das estimativas da habilidade pela TRI. **Avaliação Psicológica**, v. 6, n. 2, p. 205–215, 2007.

CORTI, A. P. As diversas faces do ENEM: análise do perfil dos participantes (1999-2007). **Estudos em Avaliação Educacional**, v. 24, n. 55, p. 198–221, 2013.

COSTA, C. E. S. **Análise da dimensionalidade e modelagem multidimensional pela TRI no Enem (1998-2008)**. [s.l.] Universidade Federal de Santa Catarina, 2015.

COSTA, P.; FERRÃO, M. E. On the complementarity of classical test theory and item response models: item difficulty estimates and computerized adaptive testing. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 23, n. 88, p. 593–610, 2015.

COURVILLE, T. G. **An empirical comparison of item response theory and classical test theory item/person statistics**. [s.l.] Texas A&M University, 2004.

DALBEN, A.; ALMEIDA, L. C. Para uma avaliação de larga escala multidimensional. **Est. Avali. Educ.**, v. 26, n. 61, p. 12–28, 2015.

DESJARDINS, C. **Psychometric validity and reliability statistics in R**, [s.d.].

DRASGOW, F.; LISSAK, R. I. Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. **Journal of Applied**

Psychology, v. 68, n. 3, p. 363–373, 1983.

EMBRETSON, S. E.; REISE, S. P. **Item response theory for psychologists**. Mahwah, N.J: Lawrence Erlbaum Associates, Inc., 2000.

ERGÜVEN, M. An empirical evaluation and comparison of classical test theory and Rasch model. **Journal of Education**, v. 3, n. 1, p. 33–38, 2014.

FAN, X. Item response theory and classical test theory: an empirical comparison of their item/person statistics. **Educational and Psychological Measurement**, v. 58, n. 3, 1998.

FERREIRA, F. F. G. **Escala de proficiência para o Enem utilizando teoria de resposta ao item**. [s.l.] Universidade Federal do Pará, 2009.

FONTANIVE, N. S. A divulgação dos resultados das avaliações dos sistemas escolares: limitações e perspectivas. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 21, n. 78, p. 83–100, 2013.

FREITAS, D. N. T. DE. Avaliação da educação básica e ação normativa federal. **Cadernos de Pesquisa**, v. 34, n. 123, p. 663–689, dez. 2004.

GÜLER, N.; UYANIK, G. K.; TEKER, G. T. Comparison of classical test theory and item response theory in terms of item parameters. **European Journal of Research on Education**, v. 2, n. 1, p. 1–6, 2014.

GULLIKSEN, H. A course in the theory of mental tests. **Psychometrika**, v. 8, n. 4, p. 223–245, 1943.

GULLIKSEN, H. **Theory of mental tests**. New York: John Wiley & Sons, 1950.

HAMBLETON, R. K.; JONES, R. W. Comparison of classical test theory and item response theory and their applications to test development. **Educational Measurement Issues and Practice**, p. 38–47, 1993.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, D. J. **Fundamentals of item response Theory**. London: Sage Publications, 1991.

HATTIE, J. Methodology Review: Assessing Unidimensionality of Tests and Items. **Applied**

Psychological Measurement, v. 9, n. 2, p. 139–164, 1985.

HAYTON, J. C.; ALLEN, D. G.; SCARPELLO, V. Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. **Organizational Research Methods**, v. 7, n. 2, p. 191–205, 2004.

HORTA NETO, J. L. Avaliação externa de escolas e sistemas: questões presentes no debate sobre o tema. **R. bras. Est. pedag.**, v. 91, n. 227, p. 84–104, 2010.

KLEIN, R. Utilização da teoria de resposta ao item no Sistema Nacional de Avaliação da Educação Básica (SAEB). **Meta: Avaliação**, v. 1, n. 2, p. 125–140, 2009.

KLEIN, R. Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências. **Ensaio: Aval. Pol. Públ. Educ.**, v. 21, n. 78, p. 35–56, 2013.

KOHLI, N.; KORAN, J.; HENN, L. Relationships among classical test theory and item response theory frameworks via factor analytic models. **Educational and Psychological Measurement**, v. 75, n. 3, p. 389–405, 2014.

LAROS, J. A. Análise gráfica de itens. In: PASQUALI, L. (Ed.). . **Psicometria: teoria dos testes na psicologia e na educação**. 3. ed. Petrópolis, Rio de Janeiro: Vozes, 2009.

LEDESMA, R. D.; VALERO-MORA, P. Determining the number of factors to retain in EFA: an easy-to- use computer program for carrying out parallel analysis. **Practical assessment, research and evaluation**, v. 12, n. 2, p. 1–3, 2007.

LINDEN, W. J. VAN DER. Introduction. In: LINDEN, W. J. VAN DER (Ed.). . **Handbook of item response theory - volume one**. Boca Raton, FL: CRC Press, 2015.

LOPES, A. CASIMIRO; LÓPEZ, S. B. A performatividade nas políticas de currículo: o caso do Enem. **Educar em Revista**, v. 26, n. 01, p. 89–110, 2010.

LORD, F. M. An approach to mental test theory. **Psychometrika**, v. 24, n. 4, p. 283–302, 1959.

MACDONALD, P.; PAUNONEN, S. V. A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. **Educational and Psychological Measurement**, v. 62, n. 6, p. 921–943, 2002.

MACHADO, C. Impactos da avaliação externa nas políticas de gestão educativa. **REICE - Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**, v. 11, n. 1, p. 40–55, 2013.

MACHADO, C.; ALAVARSE, O. M. Qualidade das Escolas: tensões e potencialidades das avaliações externas. **Educação & Realidade**, v. 39, n. 2, p. 413–436, 2014.

MARCONI, M. DE A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. 5ª Edição ed. São Paulo: Editora Atlas, 2003.

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

MCDONALD, R. P. The dimensionality of tests and items. **British Journal of Mathematical and Statistical Psychology**, v. 34, n. 1, p. 100–117, 1981.

MESQUITA, S. Os resultados do Ideb no cotidiano escolar. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 20, n. 76, p. 587–606, 2012.

MINHOTO, M. A. P. Política de Avaliação da Educação Brasileira : limites e perspectivas. **Jornal de Políticas Educacionais**, v. 10, n. 19, p. 77–90, 2016.

MUNER, L. C. **Análise fatorial exploratória e comfirmatória do Enem 2010 com estudantes paulistas**. [s.l.] Universidade São Francisco, 2013.

MUÑIZ, J. **Teoría de repuesta a los ítems**. Madrid: Pirámede, 1990.

MUÑIZ, J. **Teoría clássica de los testes**. Madrid: Pirámede, 1994.

MUÑIZ, J. Las teorias de los tests: Teoria clasica y teoria de respuesta a los tems. **Papeles del Psicologo**, v. 31, n. 1, p. 57–66, 2010.

NAVAS, M. J. Teoría clásica de los tests versus teoria de respuesta al item. **Psicológica**, v. 15, p. 175–208, 1994.

NEWMAN, D. A. Missing data: five practical guidelines. **Organizational Research Methods**, v. 17, n. 4, p. 372–411, 2014.

NOJOSA, R. T. Teoria da resposta ao item (TRI): modelos multidimensionais. **Estudos em Avaliação Educacional**, n. 25, p. 123–166, 2002.

NOVICK, M. R. The axioms and principal results of classical test theory. **Journal of Mathematical Psychology**, v. 3, n. 1, p. 1–18, 1966.

NYLUND, K. L.; ASPAROUHOV, T.; MUTHÉN, B. O. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. **Structural Equation Modeling: A Multidisciplinary Journal**, v. 14, n. 4, p. 535–569, 23 out. 2007.

OLIVEIRA, B. A. **Interdisciplinaridade e dimensionalidade das provas do EnemReuniões da ABAVE**São Paulo, 2015. Disponível em: <http://www.abave.com.br/ojs/index.php/Reunioes_da_Abave/article/view/336/109>

PASQUALI, L. **Psicometria: teoria dos testes na psicologia e na educação**. 3. ed. Petrópolis, Rio de Janeiro: Vozes, 2009.

PICCIRILLI, G. P.; SOUZA, A. D. P. DE. **Teoria da resposta ao item multidimensional: análise da dimensionalidade da prova do Enem 201630º Congresso de Iniciação Científica da Unesp**Presidente Prudente, 2018. Disponível em: <http://prope.unesp.br/cic/admin/ver_resumo.php?area=100093&subarea=29291&congresso=40&CPF=46993869896>

PRIMI, R.; CICCHETTO, A. A. **Como os escores do ENEM são atribuídos pela TRI?Anais do VI CONBRATRI: Métodos para detecção de fraudes em testes**Juiz de Fora, 2018. Disponível em: <https://even3.blob.core.windows.net/anais/92067.pdf>

PROGAR, Š.; SOČAN, G. An empirical comparison of item response theory and classical test theory item/person statistics. **Horizons of Psychology**, v. 17, n. 3, p. 5–24, 2008.

RECKASE, M. D. **Multidimensional Item Response Theory**. New York, NY: Springer, 2009.

REISE, S. P.; COOK, K. F.; MOORE, T. M. Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In: **Handbook of item response theory modeling : applications to typical performance assessment**. New York, NY: Routledge, 2015.

REQUENA, C. S. **Psicometria: teoria y practica em la construcción de tests**. Madrid: Ediciones Norma, 1990.

REVELLE, W. **psych: Procedures for Personality and Psychological Research**, 2017. Disponível em: <https://cran.r-project.org/package=psych>

RICHARDSON, R. J. **Pesquisa social: métodos e técnicas**. 3ª edição ed. São Paulo: Editora Atlas, 2012.

RIZOPOULOS, D. ltm: An R package for latent variable modeling and item response theory analyses. **Journal of Statistical Software**, v. 17, n. 5, p. 1–25, 2006.

SANTOS, J. M. C. T. Exame nacional do ensino médio : entre a regulação da qualidade do ensino médio e o vestibular. **Educar em Revista**, n. 40, p. 195–205, 2011.

SARTES, L. M. A.; SOUSA-FORMIGONI, M. L. O. DE. Avanços na Psicometria: aa teoria clássica dos testes à teoria de resposta ao item. **Psicologia: Reflexão e Crítica**, v. 26, n. 2, p. 241–250, 2013.

SCHNEIDER, M. P.; ROSTIROLA, C. R. Estado-Avaliador: reflexões sobre sua evolução no Brasil. **Revista Brasileira de Política e Administração da Educação - Periódico científico editado pela ANPAE**, v. 31, n. 3, p. 493, 1 jun. 2015.

SILVEIRA, F. L. DA. Considerações sobre o índice de discriminação de itens em testes educacionais. **Educação & Seleção**, v. 7, 1983.

SOUSA, S. Z. L. Concepções de qualidade da educação básica forjadas por meio de avaliações em larga escala. **Avaliação**, v. 14, n. 2, p. 407–420, 2014.

STEVENS, S. S. On the theory of scales of measurement. **Science**, v. 103, n. 2684, p. 677–680, 1946.

STOUT, W. F. A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. **Psychometrika**, v. 55, n. 2, p. 293–325, 1990.

TAVARES, C. Z. Teoria da resposta ao item: uma análise crítica dos pressupostos epistemológicos. **Estudos em Avaliação Educacional**, v. 24, n. 54, p. 56–76, 2013.

TOFFOLI, S. F. L. et al. Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. **Educação e Pesquisa**, v. 42, n. 2, p. 343–358, jun. 2016.

TRAN, U. S.; FORMANN, A. K. Performance of parallel Analysis in retrieving unidimensionality in the presence of binary data. **Educational and Psychological Measurement**, v. 69, n. 1, p. 50–61, 2009.

TRAUB, R. E. Classical test theory in historical perspective. **Educational Measurement: Issues and Practice**, p. 8–14, 1997.

URBINA, S. **Essentials of psychological testing**. New Jersey: John Wiley & Sons, Inc., 2004.

VALLE, R. DA C. Teoria de resposta ao item. **Estudos em Avaliação educacional**, v. 21, p. 07–92, 2000.

VIANNA, H. M. **Testes em educação**. São Paulo: IBRASA, 1976.

VIANNA, H. M. Avaliações nacionais em larga escala: análises e propostas. **Estudos em Avaliação Educacional**, n. 27, p. 41–76, 2003.

VIEIRA, N. N. **As provas das quatro áreas do ENEM vista como prova única na ótica de modelos da teoria da resposta ao item uni e multidimensional**. [s.l.] Universidade Federal de Santa Catarina, 2016.

WENG, L. J.; CHENG, C. P. Parallel analysis with unidimensional binary data. **Educational and Psychological Measurement**, v. 65, n. 5, p. 791–810, 2005.

WICKHAM, H. et al. **Create elegant data visualisations using the grammar of graphics**, 2019. Disponível em: <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>

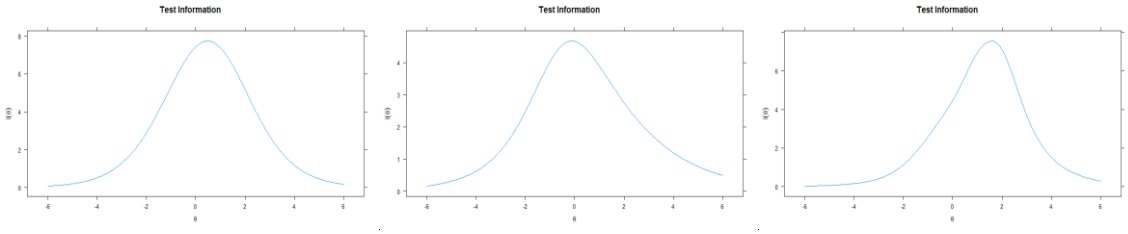
ZIEGLER, M.; HAGEMANN, D. Testing the unidimensionality of items: Pitfalls and loopholes. **European Journal of Psychological Assessment**, v. 31, n. 4, p. 231–237, 2015.

ZWICK, W. R.; VELICER, W. F. Comparison of five rules for the number of factors to retain. **Psychological Bulletin**, v. 99, n. 3, p. 432–442, 1986.

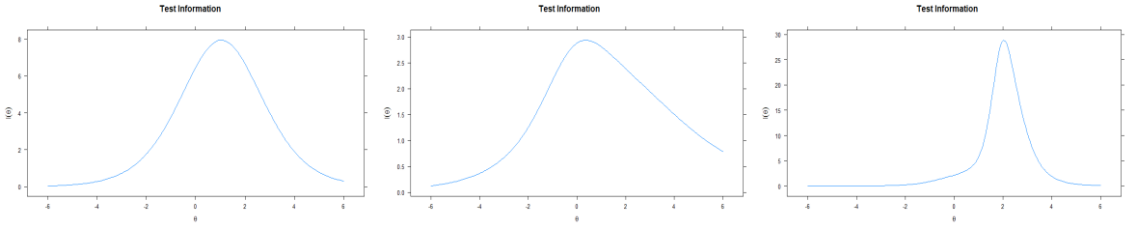
9 APÊNDICE A – CURVAS DE INFORMAÇÃO DO TESTE

Curvas de Informação das provas do Enem nos modelos de TRI de 1, 2 e 3 parâmetros, respectivamente.

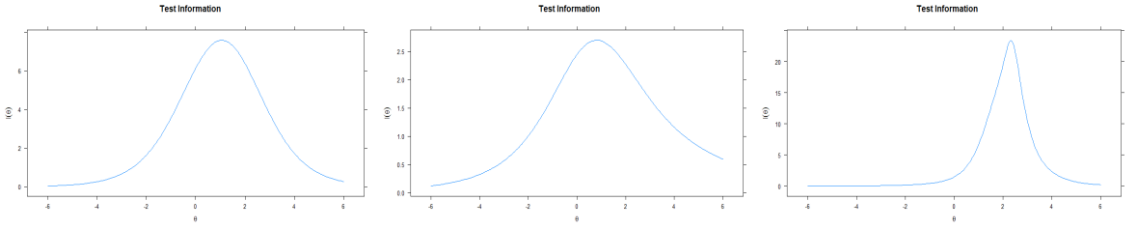
Linguagens e Códigos



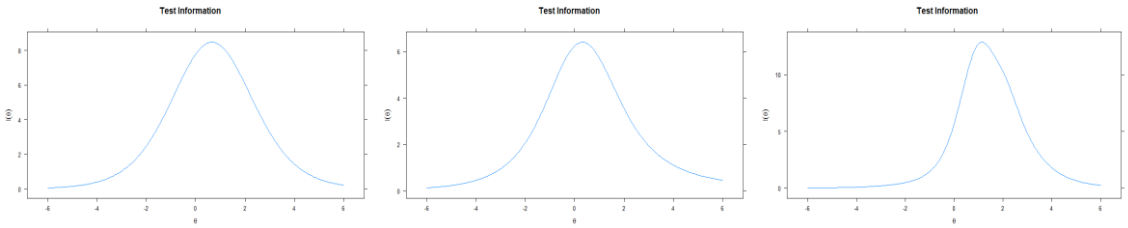
Matemática



Ciências da Natureza



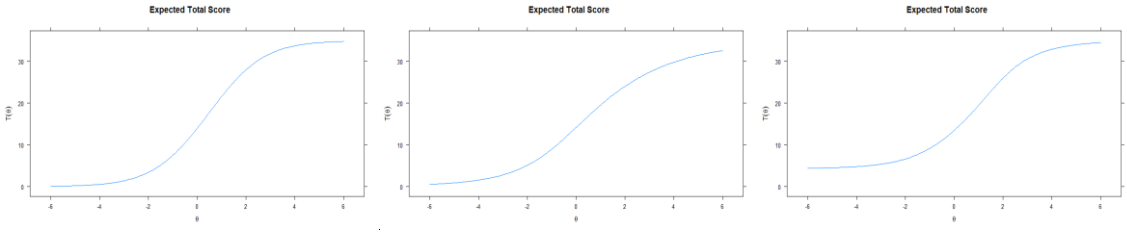
Ciências Humanas



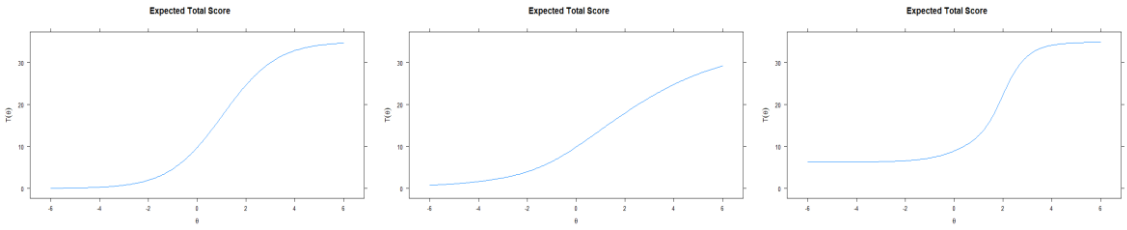
10 APÊNDICE B – CURVA CARACTERÍSTICA DO TESTE

Curvas Características das provas do Enem nos modelos de TRI de 1, 2 e 3 parâmetros, respectivamente.

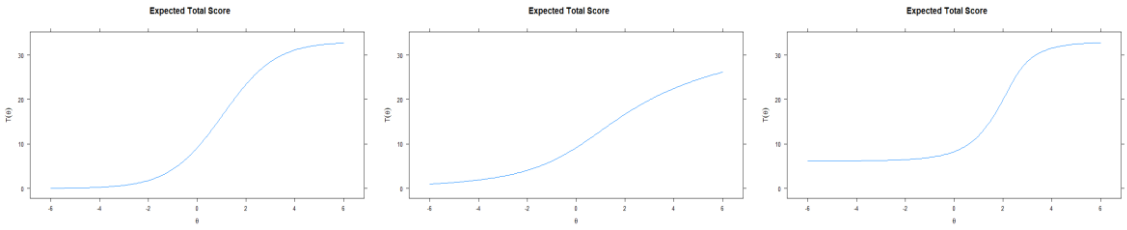
Linguagens e Códigos



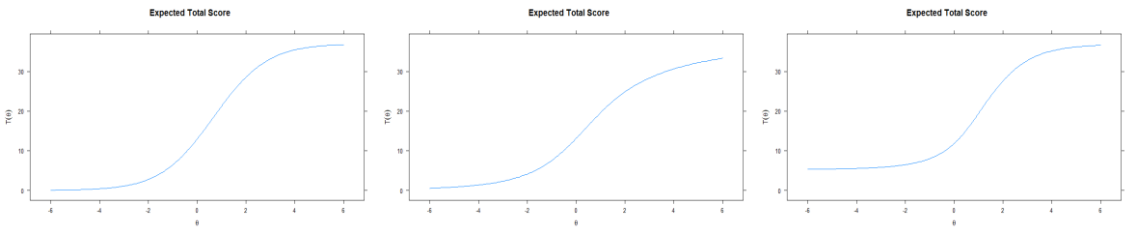
Matemática



Ciências da Natureza



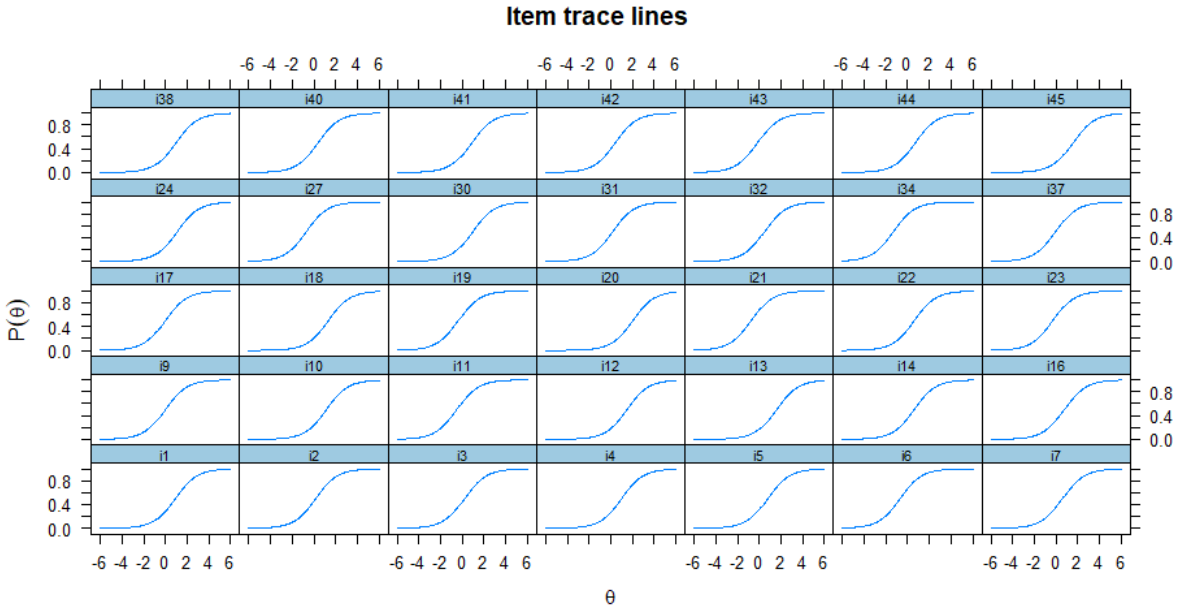
Ciências Humanas

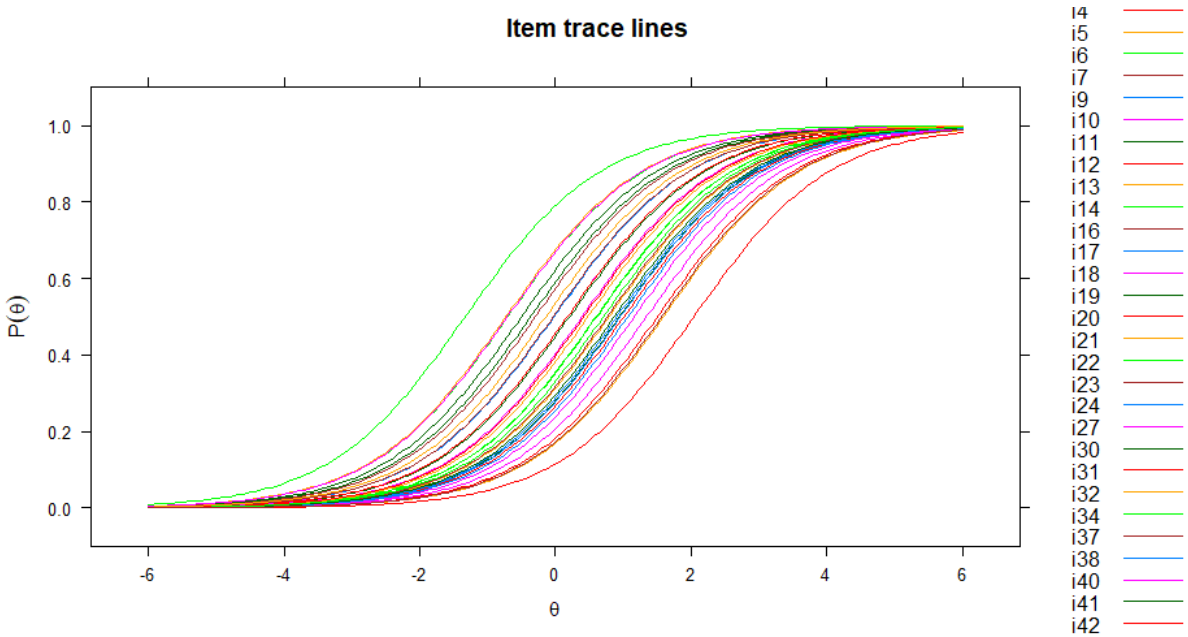


11 APÊNDICE C – CURVA CARACTERÍSTICA DOS ITENS

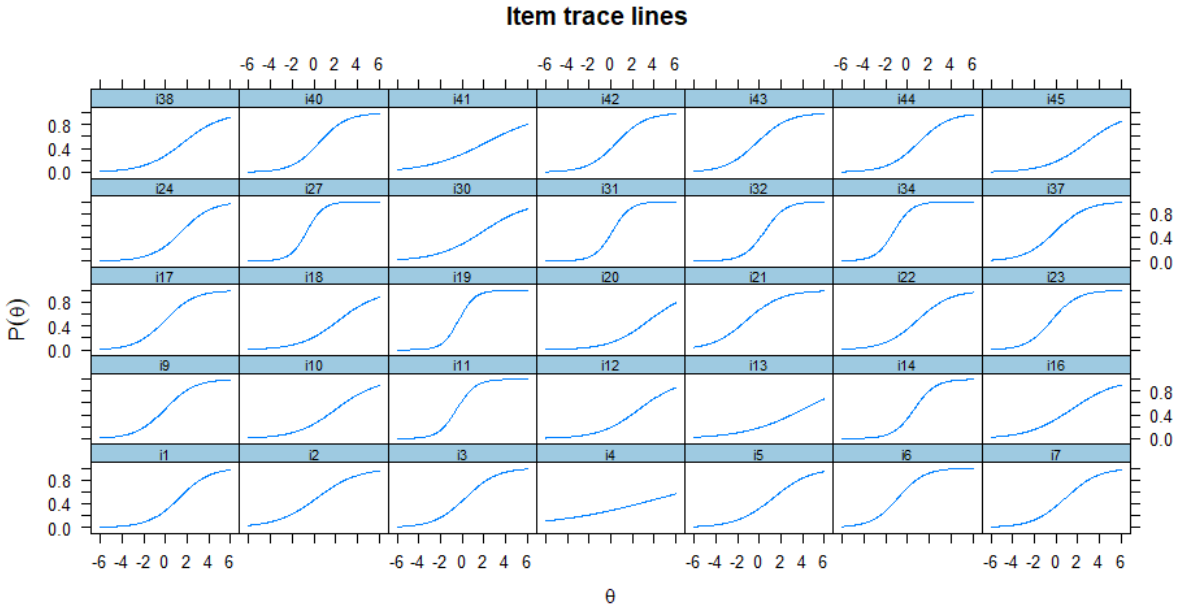
LINGUAGENS E CÓDIGOS

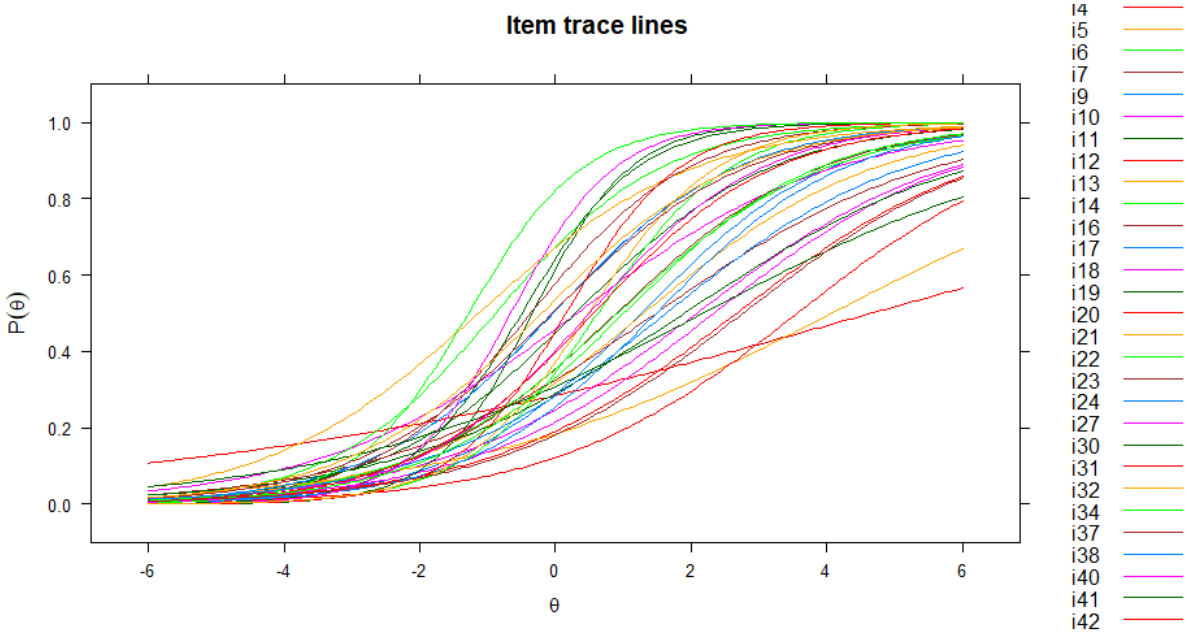
1 parâmetro



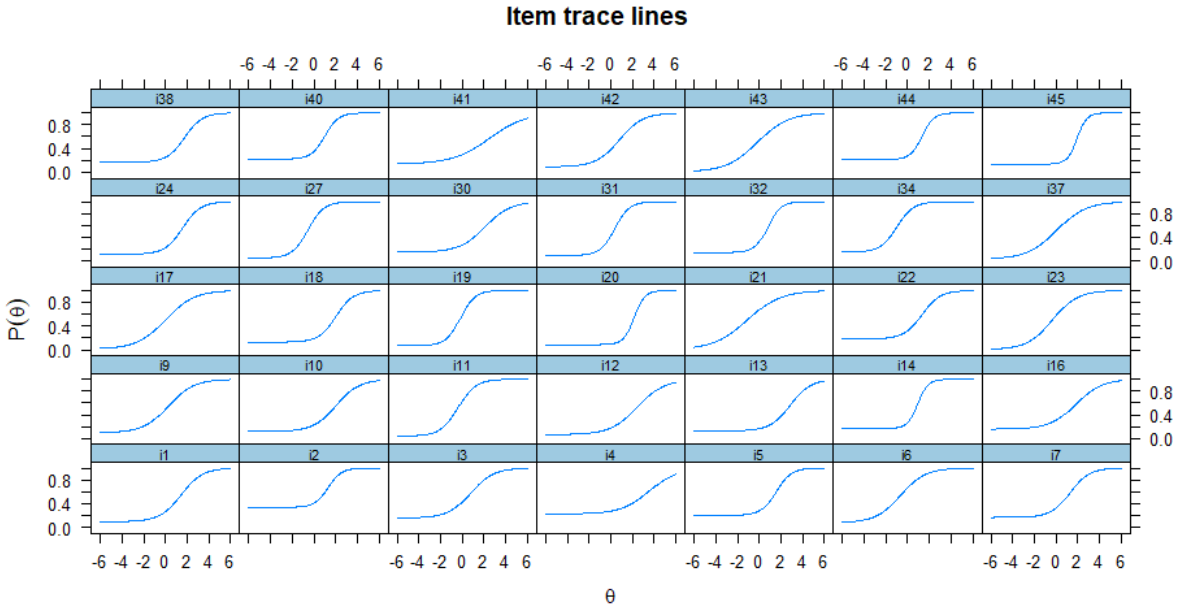


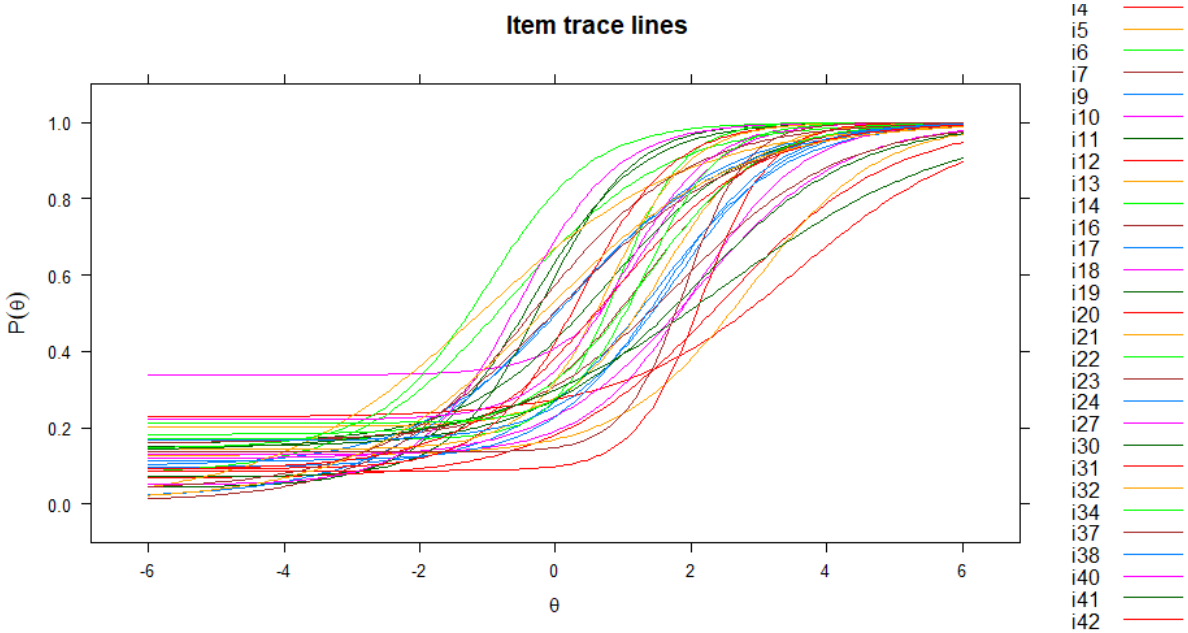
2 parâmetros





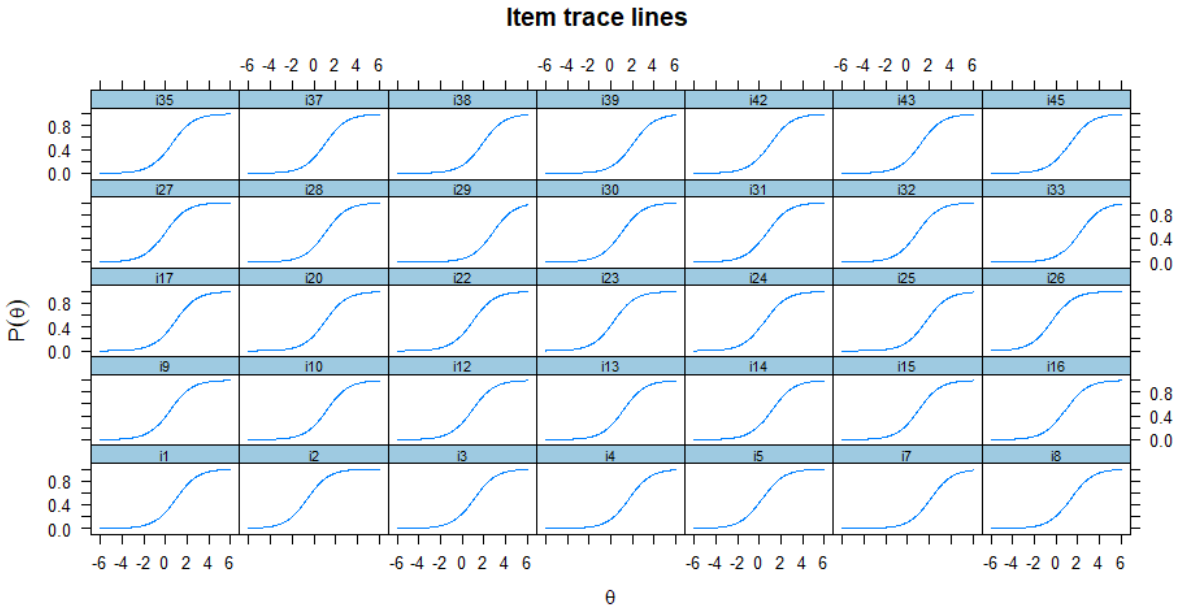
3 parâmetros

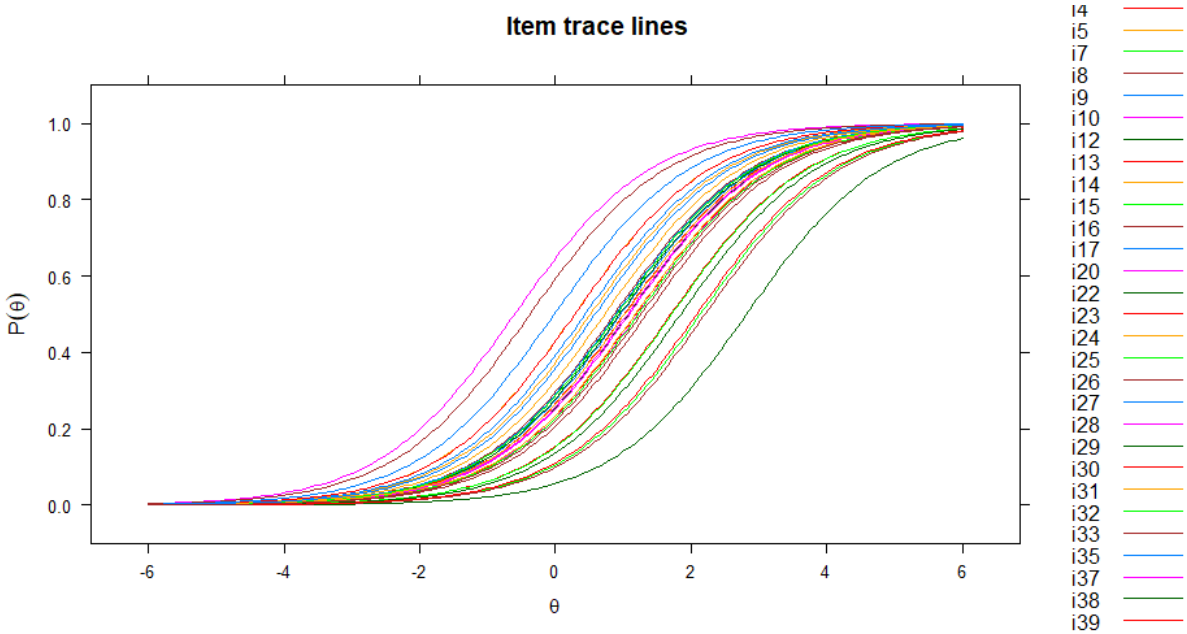




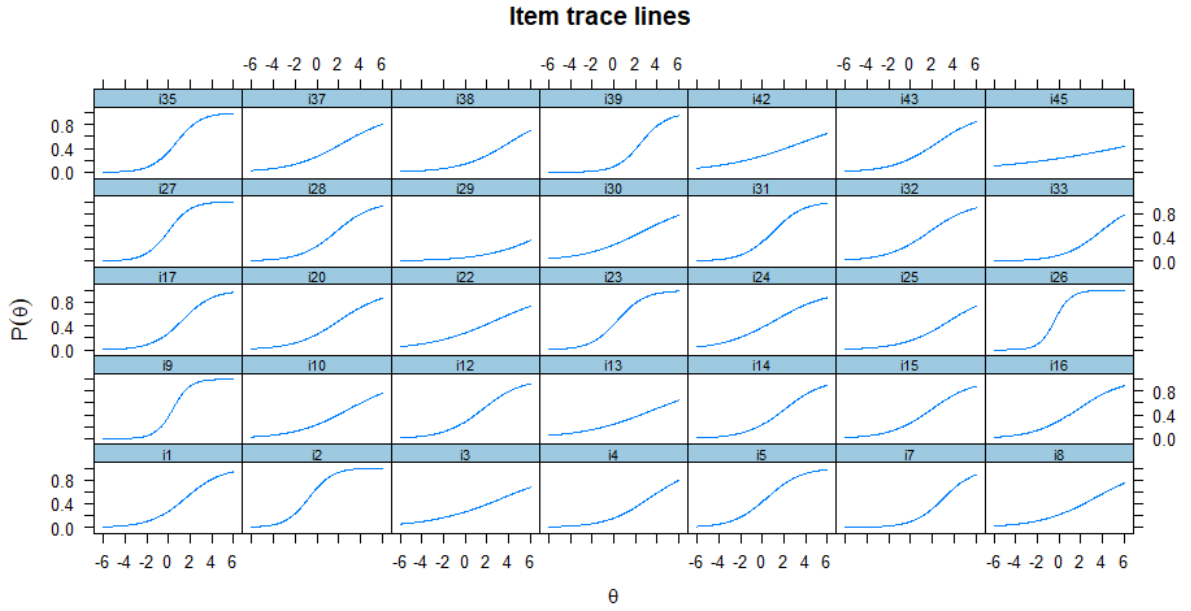
MATEMÁTICA

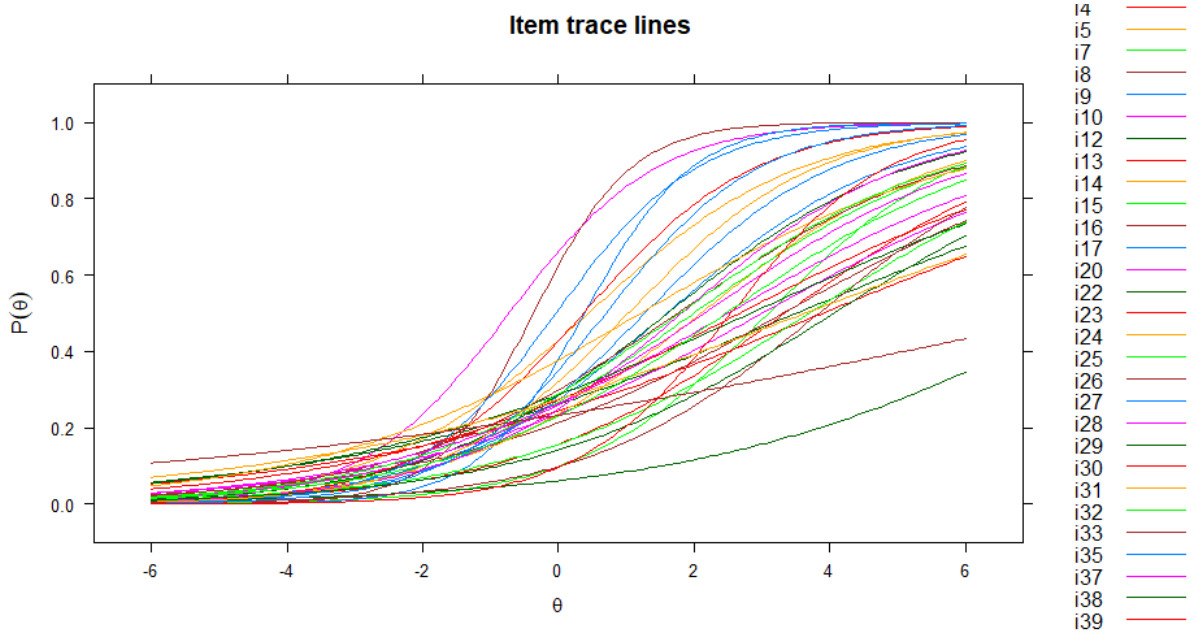
1 parâmetro



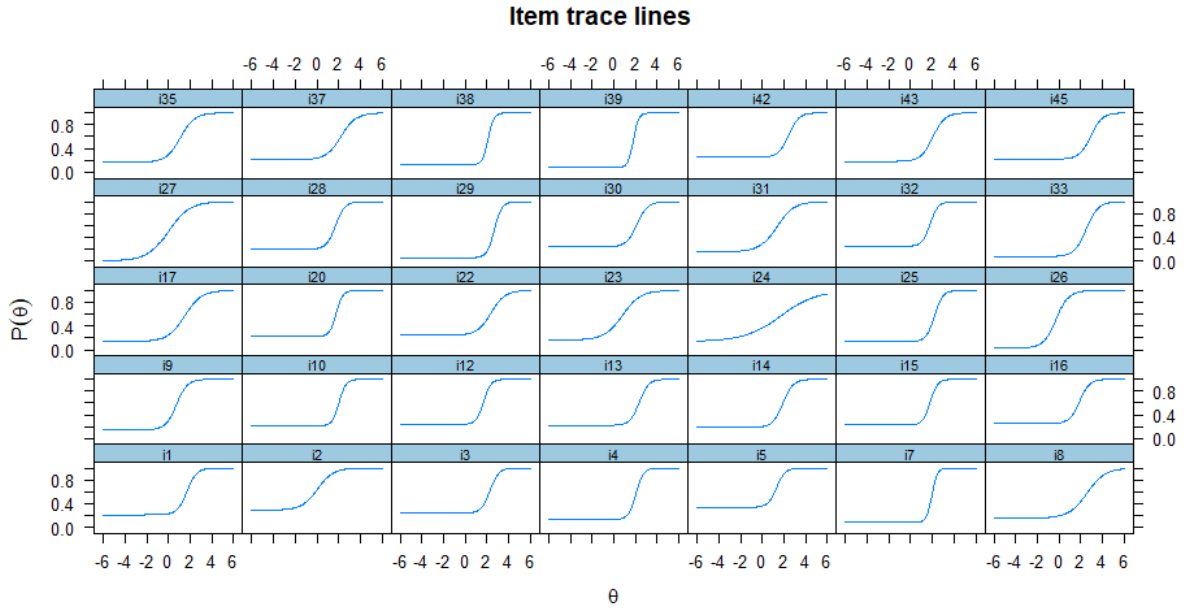


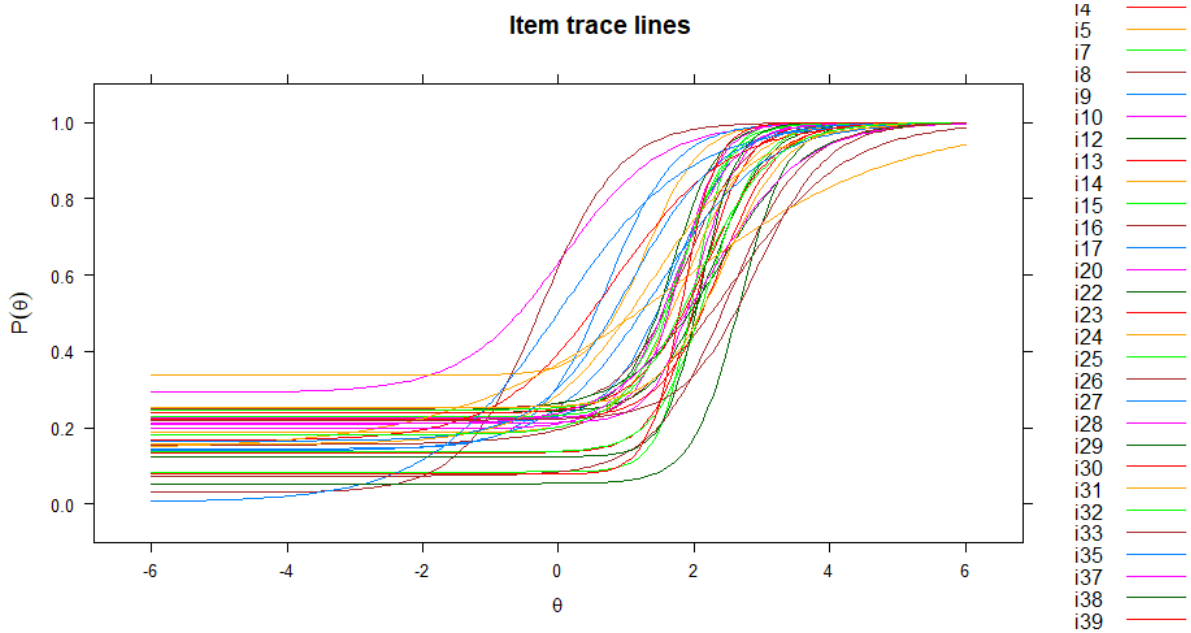
2 parâmetros





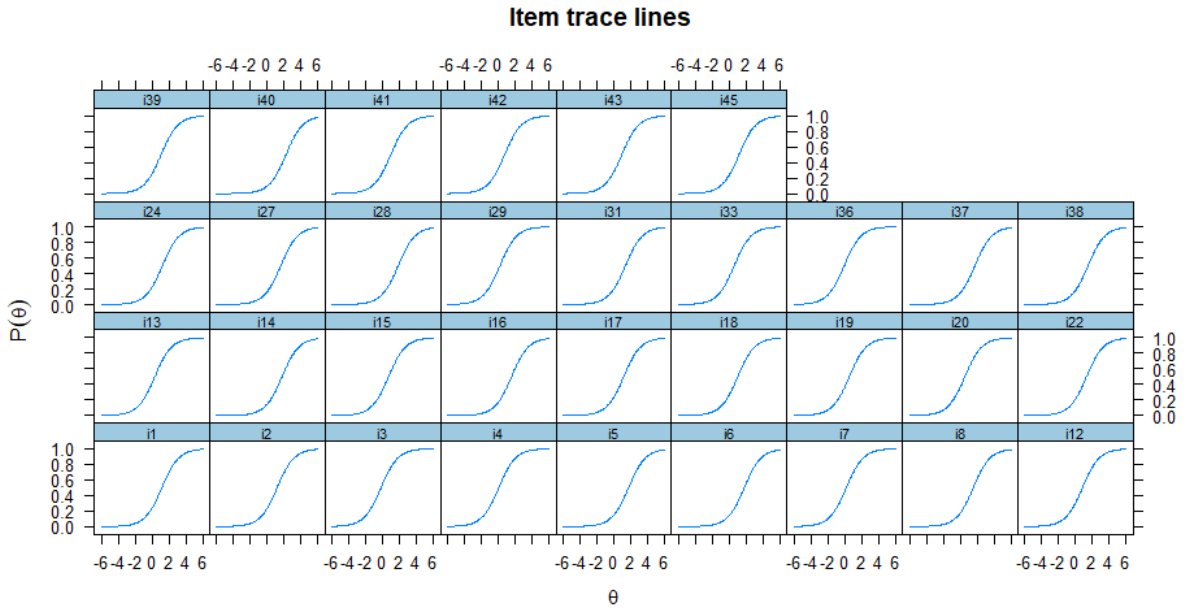
3 parâmetros

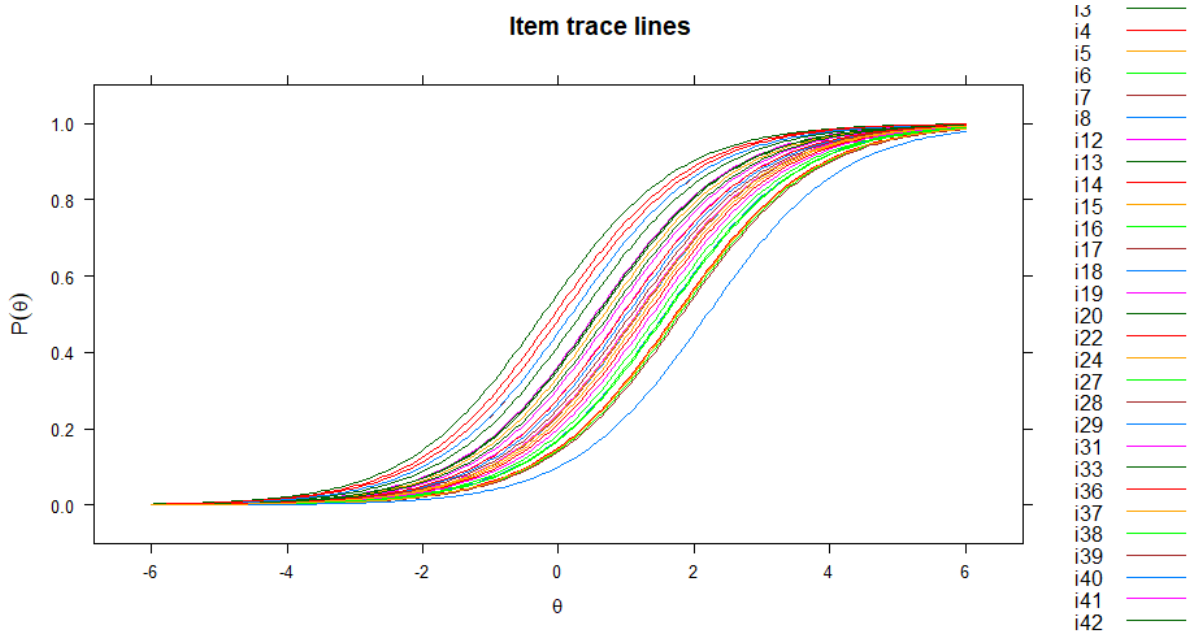




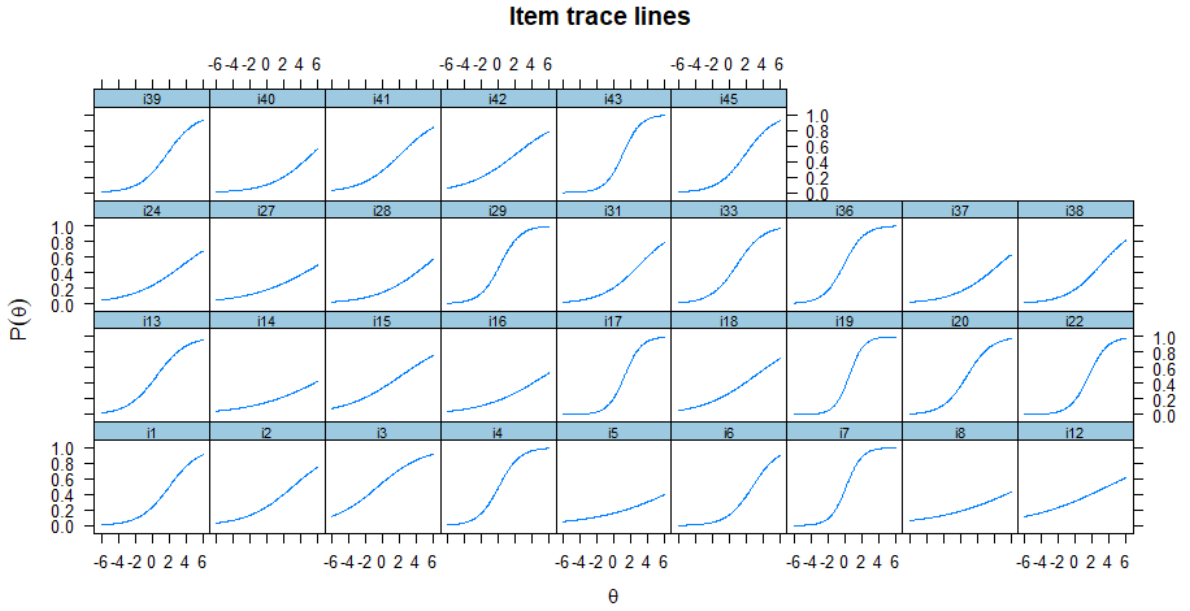
CIÊNCIAS DA NATUREZA

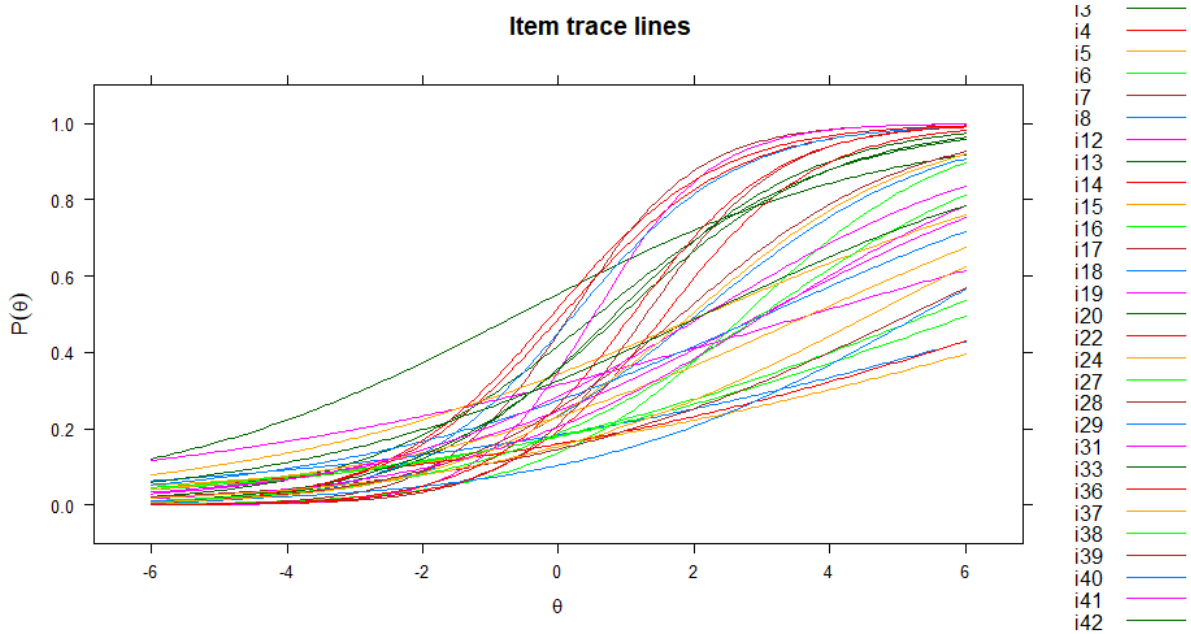
1 parâmetro



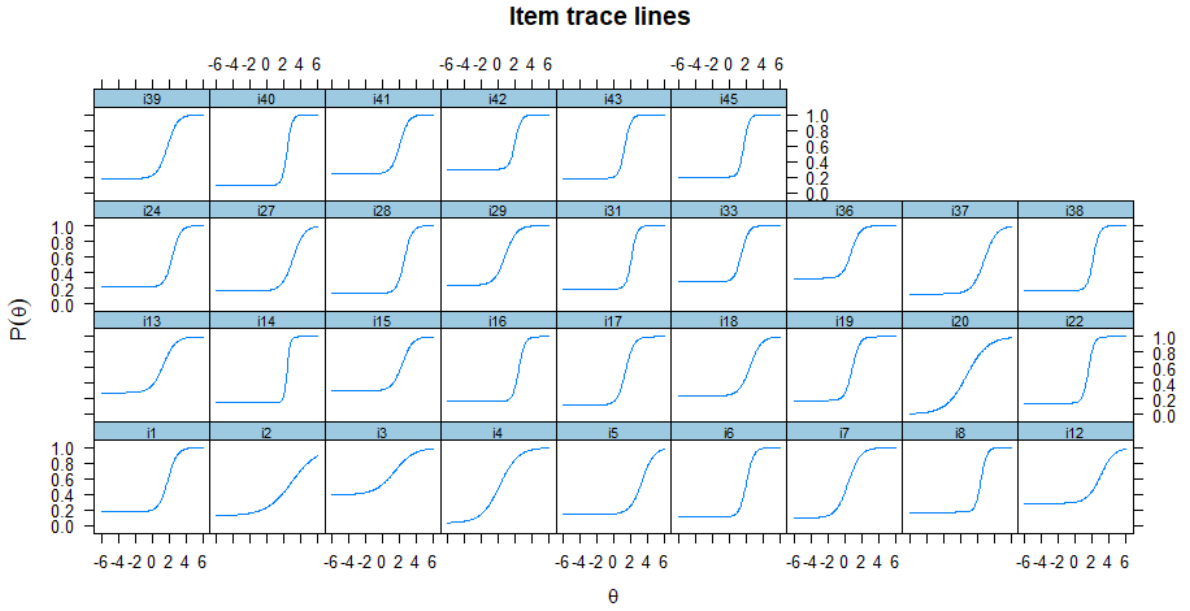


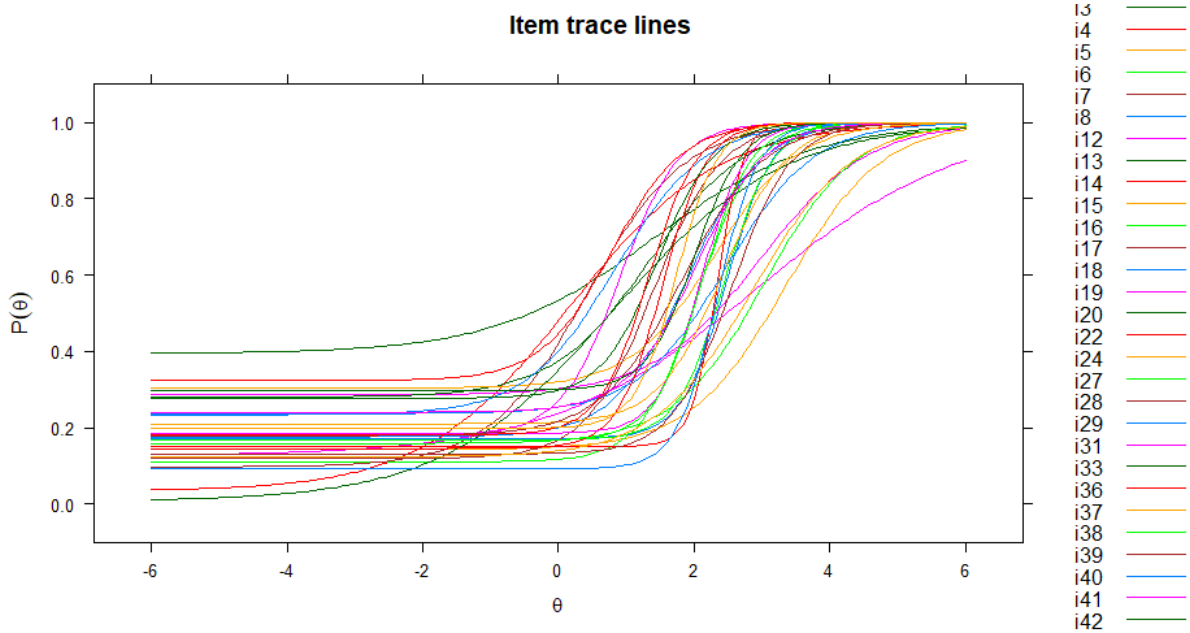
2 parâmetros





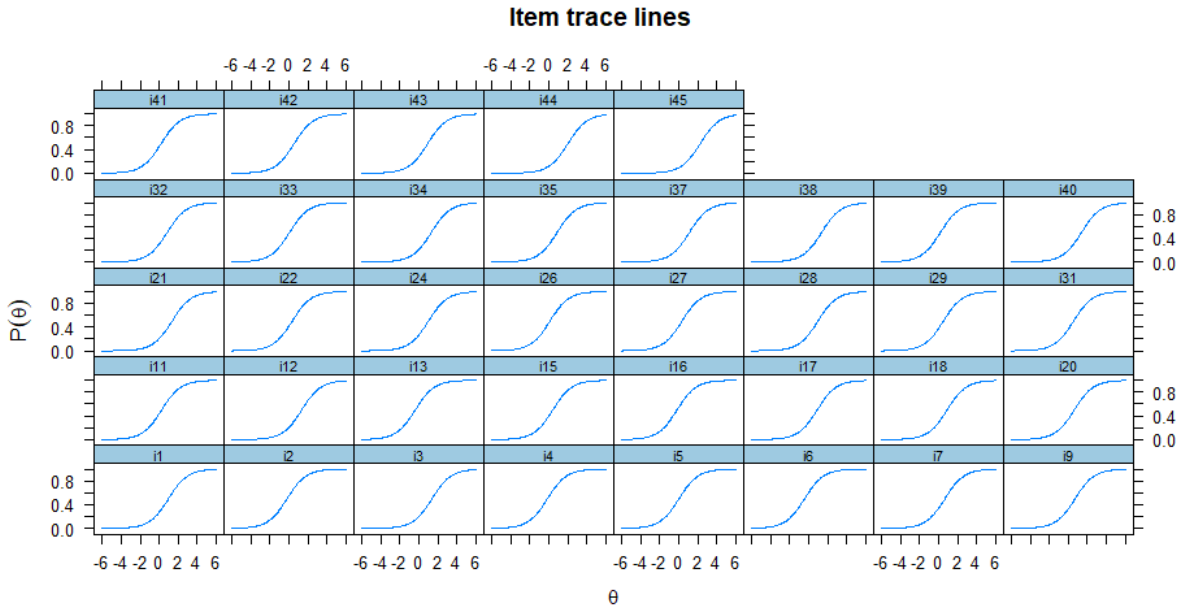
3 parâmetros

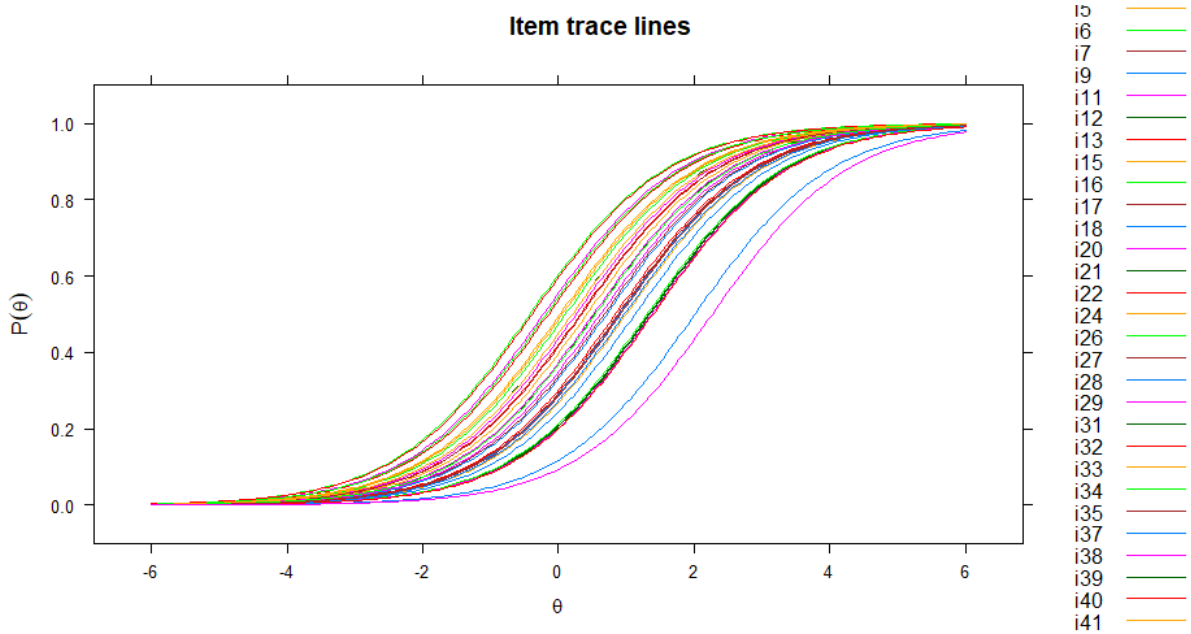




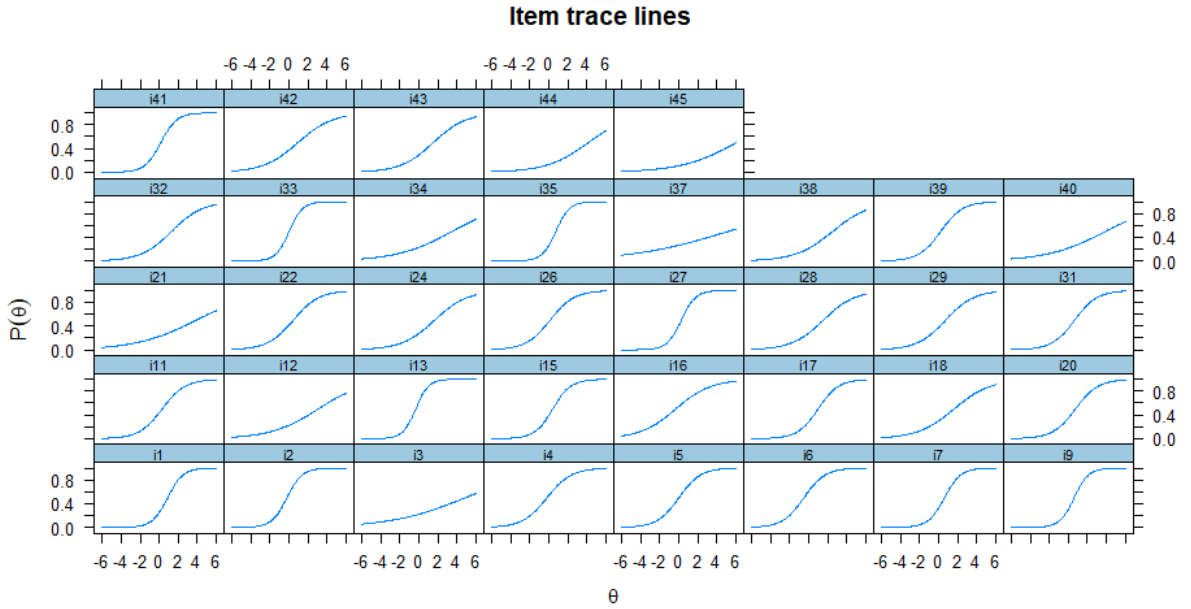
CIÊNCIAS HUMANAS

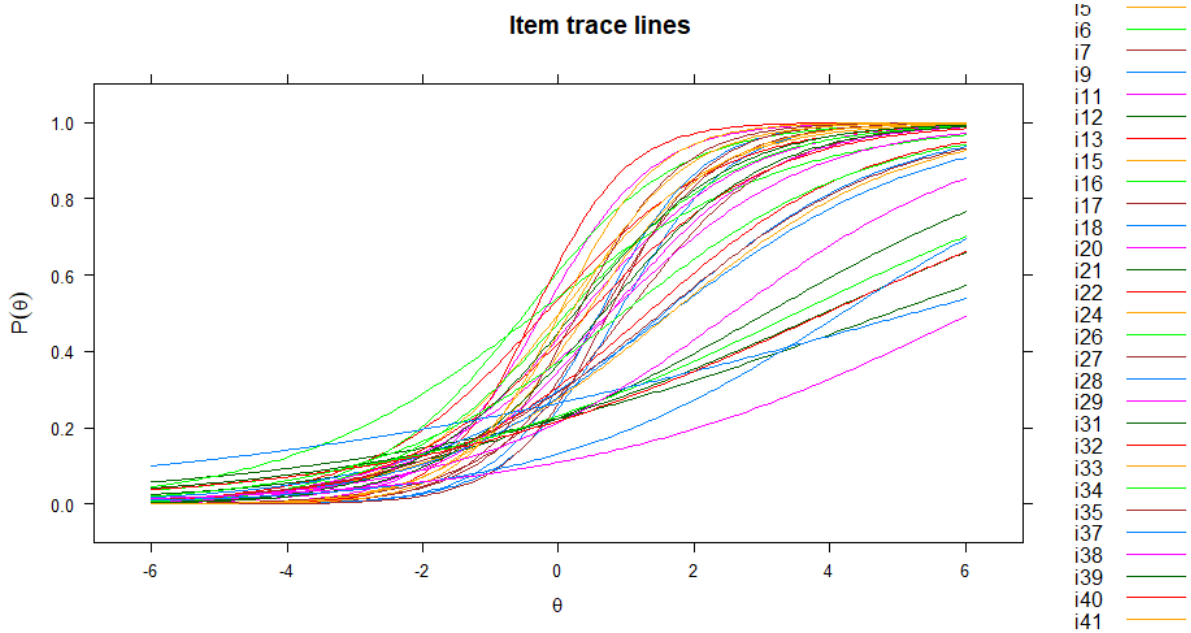
1 parâmetro





2 parâmetros





3 parâmetros