Ciência e Natura, Santa Maria, v. 36 Ed. Especial, 2014, p. 58–72
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA 35 anos

Contribuições da Teoria da Resposta ao Item nas Avaliações Educacionais

Contributions of Item Response Theory in Educational Assessments

Fernando de Jesus Moreira Junior*[1]

[1] Departamento de Estatística, Universidade Federal de Santa Maria, Brasil.

Resumo

A concepção, o desenvolvimento e a aplicação de métodos avaliativos para representar os resultados da aprendizagem têm sido, há muito tempo, uma preocupação entre os pedagogos, psicólogos e educadores em geral. Tradicionalmente, as avaliações educacionais são baseadas nos escores provenientes da chamada Teoria Clássica dos Testes (TCT) ou Teoria Clássica da Medida (TCM). Nesse contexto, surge a Teoria da Resposta ao Item (TRI), uma poderosa ferramenta estatística que consegue suprir as necessidades decorrentes das limitações da TCT. O objetivo desse artigo é apresentar as principais contribuições da TRI em relação à TCT, no contexto das avaliações educacionais, e apresentar uma ilustração dessas contribuições da TRI, por meio de simulações.

***Palavras-chave**: Teoria da Resposta ao Item, Teoria Clássica dos Testes, Avaliação Educacional, Modelo Logístico de Três Parâmetros, Escala de Proficiência.*

Abstract

The design, development and implementation of evaluation methods to represent the learning outcomes have been long been a concern among pedagogues, psychologists and educators. Traditionally, educational assessments are based on scores from the so-called Classical Test Theory (CTT) and Classical Theory of Measure (CTM). In this context arises the Item Response Theory (IRT), a powerful statistical tool that can meet the needs arising from the limitations of TCT. The purpose of this paper is to present the main contributions of IRT in relation to CTT, in the context of educational evaluations, and present an illustration of these contributions IRT, through simulations.

***Keywords**: Item Response Theory, Classical Test Theory, Educational Assessment, Three-Parameter Logistic Model, Scale Proficiency.*

*fmjunior777@yahoo.com.br

Recebido: 12/03/2014 Revisado: 06/06/2014

1 Introdução

Por muito tempo, pedagogos, psicólogos e educadores em geral, se preocupam em conceber, desenvolver e aplicar métodos avaliativos que representem de forma mais fiel possível os resultados da aprendizagem (ANDRADE; LAROS; GOUVEIA, 2010). Inicialmente, as avaliações eram baseadas em escores provenientes da chamada Teoria Clássica dos Testes (TCT) ou Teoria Clássica da Medida (TCM). Esses escores, ainda muito utilizados em avaliações educacionais, geralmente são baseados nas somas ou nas médias de pontuações. No entanto, várias limitações eram observadas com o uso da TCT, entre elas, não é possível comparar grupos de alunos que respondem provas diferentes, e não é possível acompanhar o ganho, em termos de conhecimento do aluno, ao longo do tempo. Na década de 50, com os trabalhos de Lord (1952), surge outra metodologia para a avaliação de traços latentes, denominada Teoria da Resposta ao Item (TRI), a qual sugere formas de representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item (questão), os traços latentes do indivíduo e as características dos itens, por meio de modelos matemáticos (ANDRADE; TAVARES; VALLE, 2000).

A TRI é uma poderosa ferramenta estatística que surgiu para suprir as necessidades decorrentes das limitações da TCT, principalmente em relação à impossibilidade de comparar grupos de alunos que respondem provas diferentes, e à impossibilidade de acompanhar o ganho, em termos de conhecimento do aluno, ao longo do tempo. Dessa forma, a TRI começa, aos poucos, a fazer parte das avaliações educacionais no Brasil. Primeiramente, a partir de 1995, segundo Andrade, Tavares e Valle (2000), através da pesquisa AVEJU, da Secretaria de Estado da Educação de São Paulo. E, posteriormente, no Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (SARESP) e no Sistema de Avaliação da Educação Básica (SAEB) do o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) do Ministério da Educação (MEC), para montagem de instrumentos, tratamento de dados e construção de escalas a partir de resultados apresentados por alunos em provas de rendimento (SOUZA, 2005). Havia a necessidade de uma metodologia mais sofisticada e precisa que permitisse a construção de escalas de habilidade a fim de acompanhar o progresso do conhecimento adquirido ao longo do tempo (ANDRADE, D. F. e TAVARES e VALLE, 2000). Nessas aplicações, a TRI tem mostrado a sua potencialidade no que diz respeito à avaliação educacional, através da construção de uma escala comparável, permitindo o acompanhamento do progresso do conhecimento adquirido pelo aluno ao longo do tempo, como tem sido feito nos países pertencentes ao Primeiro Mundo (MOREIRA JUNIOR, 2010). No contexto internacional, a TRI vem sendo empregada amplamente por vários países: Estados Unidos, França, China, Holanda, Coreia do Sul e principalmente nos países participantes do Programa Internacional de Avaliação de Estudantes (PISA). O PISA utiliza o modelo de Rasch (RASCH, 1960) da TRI e coloca os resultados em uma mesma escala de proficiências para cada área, ao longo dos anos (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2003, 2011; KLEIN, 2011). Outro exemplo de avaliação utilizando a TRI é o exame de proficiência em língua inglesa (TOEFL). Este exame surgiu em 1964 e é largamente utilizado em todo o mundo. Desde o início de sua origem, este exame já avaliou mais de 25 milhões de alunos e tem sido administrado por mais de 4.500 centros em 165 países (MINISTÉRIO DA EDUCAÇÃO, 2012).

A partir de então, cada vez mais institutos de educação têm aderido a TRI para as suas avaliações educacionais. Por exemplo, no Sistema Mineiro de Avaliação da Educação Pública – SIMAVE (SOARES; GENOVEZ; GALVÃO, 2005) da Secretaria de Estado de Educação de Minas Gerais, no Projeto GERES – Estudo Longitudinal sobre Qualidade e Equidade no Ensino Fundamental Brasileiro – (PERRY, 2009) e, mais recentemente, no Exame Nacional do Ensino Médio – ENEM (FERREIRA, 2009) também do INEP/MEC, o qual é atualmente o grande exemplo da aplicação da TRI nas avaliações educacionais de larga escala do Brasil. Recentemente, no final de 2010, houve um caso no ENEM onde alguns alunos foram prejudicados por problemas de erro na impressão no cartão de respostas. Caso o ENEM ainda utilizasse a abordagem da TCT, o exame deveria ser anulado e todos os alunos teriam que fazer novamente a prova, pois as questões seriam diferentes. Com a utilização da TRI, que permite criar provas diferentes com o mesmo nível de dificuldade (princípio da isonomia), apenas esses alunos que foram prejudicados se submeteram a realização de outra prova novamente, sem prejuízo de nota para aqueles que não tiveram problemas na sua prova. A possibilidade de poder aplicar o exame mais de um uma vez no mesmo ano é um dos principais motivos da implementação da TRI no ENEM (MINISTÉRIO DA EDUCAÇÃO, 2012). Isso evitou o transtorno da anulação do exame e um prejuízo ainda maior para os cofres públicos.

Em relação à avaliação tradicional da TCT, a TRI apresenta algumas vantagens (EMBRETSON; REISE, 2000), dentre as quais, destacam-se: (1) a TRI fornece informações mais precisas do desempenho dos respondentes, pois o traço latente do indivíduo não depende da dificuldade das questões do teste, enquanto que na TCT o escore do indivíduo depende essencialmente dos itens que compõe o teste (ANDRADE; TAVARES; VALLE, 2000; VENDRAMINI; SILVA; CANALE, 2004); (2) a TRI permite obter melhores índices de precisão do item (função de informação do item - FII) e do teste (função de informação do teste - FIT) do que os índices utilizados pela TCT (ANDRADE; TAVARES; VALLE, 2000; BAKER, 2001); (3) a TRI permite utilizar modelos que consideram a possibilidade do acerto casual, popularmente

conhecido como "chute", algo que a TCT não conseguia contemplar (ANDRADE; TAVARES; VALLE, 2000); (4) a TRI permite, sob certas condições, a comparação através do escore entre os indivíduos que responderam questionários com itens diferentes para medir o mesmo traço latente, uma vez que os itens e os indivíduos são colocados numa mesma escala, que é o grande avanço da TRI em relação à TCT (ANDRADE; TAVARES; VALLE, 2000; EMBRETSON; REISE, 2000); (5) na TRI, uma vez estimada a proficiência do indivíduo, é possível verificar qual a probabilidade de dar certa resposta a um determinado item que ele não respondeu, probabilidade que a TCT não consegue calcular (VENDRAMINI; SILVA; CANALE, 2004); (6) na TRI, cada respondente tem seu próprio erro padrão, relacionado à sua habilidade, onde a estimação desse erro é mais precisa, enquanto que na TCT todos os respondentes têm o mesmo erro padrão estimado (EMBRETSON; REISE, 2000).

No contexto da avaliação educacional, a TRI pode ser utilizada de duas formas: (1) analisar uma única prova e um único grupo de respondentes; (2) analisar duas ou mais provas e dois ou mais grupos de respondentes. A aplicação da TRI para analisar uma única prova e um único grupo de respondentes é a mais trivial, e consiste em: (1) estimar (ou calibrar) os parâmetros dos itens; (2) avaliar estatisticamente os itens, eliminando os inadequados, se houverem; (3) re-estimar os parâmetros dos itens até que não haja mais itens inadequados; (4) construir a escala de habilidade; (5) estimar a proficiência (a nota) dos indivíduos. O segundo caso, ou seja, analisar duas ou mais provas e dois ou mais grupos de respondentes, é mais complicado e exige um planejamento prévio. Nesse caso, a situação mais simples seria dois grupos diferentes resolvendo provas diferentes. Se essas provas forem totalmente diferentes, não há como colocar as questões de ambas as provas numa mesma escala (que é o que permite a comparabilidade), e, assim, não é possível utilizar a TRI, o que configura uma das poucas desvantagens da TRI (ANDRADE; TAVARES; VALLE, 2000). Para que as questões das duas provas sejam colocadas numa mesma escala, é necessário que haja itens em comum entre as provas. Havendo itens em comum entre as provas, elas podem ser colocadas numa mesma escala por meio de um método de equalização (1) a priori, quando o processo de equalização entre as duas provas é feito simultaneamente (BOCK; ZIMOWSKI, 1997); ou (2) a posteriori, quando as provas são calibradas separadamente e, após, são colocadas na mesma escala por meio de algum método de equalização. No caso da equalização a priori, as provas são tratadas como se fossem uma única prova (nem todos os softwares permitem isso) e a aplicação da TRI consiste em nos mesmos passos vistos no primeiro caso. No caso da equalização a posteriori, a aplicação da TRI consiste em: (1) estimar (ou calibrar) os parâmetros dos itens nas duas provas separadamente; (2) avaliar estatisticamente os itens de cada prova, eliminando os inadequados, se houverem; (3) re-estimar os parâmetros dos itens, nas duas provas separadamente, até que não haja mais itens inadequados; (4) estimar a proficiência (a nota) dos indivíduos de cada grupo em escalas separadas; (5) equalizar as provas por meio de um método de equalização; (6) construir a escala de habilidade; (7) obter a proficiência (a nota) dos indivíduos de um dos grupos na escala do "grupo de referência", por meio de uma transformação linear. Depois que os itens são devidamente calibrados, eles podem ser aplicados a outros indivíduos para estimar a proficiência dos indivíduos, não sendo mais necessário estimar novamente os parâmetros dos itens.

No caso do ENEM, a TRI foi projetada durante alguns anos antes da sua aplicação definitiva no exame. Ao longo desses anos, várias provas, com itens em comum, foram aplicadas em diversas localidades do Brasil. Foi construído um grande conjunto de itens que foi calibrado numa única escala. Nessa escala foram definidos níveis denominados "níveis âncora", que são pontos selecionados pelos analistas na escala da habilidade para serem interpretados pedagogicamente. Valle (2001) ressalta que esses níveis âncoras não podem ser muito próximos nem muito distantes, podendo-se tomar como base a média e o desvio padrão. O Exame Nacional de Desempenho de Estudantes (ENADE) também têm sido objeto de estudo da TRI (FRANCISCO, 2005; OLIVEIRA, 2006; NOGUEIRA, 2008; PRIMI et al., 2009; PRIMI et al., 2010; CORREA et al., 2012), embora uma implantação da TRI no ENADE seja mais complicada do que no ENEM, já que há dezenas de cursos diferentes, o que implicaria na criação de dezenas de bancos de itens.

A TRI entrou no Brasil com o objetivo de aprimorar as avaliações educacionais e a maioria das aplicações da TRI no país têm sido na área da avaliação educacional (MOREIRA JUNIOR, 2010). Uma listagem de trabalhos desenvolvidos nessa área até o ano de 2009 pode ser encontrada em Moreira Junior (2010). O objetivo desse artigo é apresentar as principais contribuições da Teoria da Resposta ao Item em relação à Teoria Clássica dos Testes, no contexto das avaliações educacionais, e apresentar uma ilustração dessas contribuições da TRI, por meio de simulações.

2 Modelos de Resposta ao Item para Avaliação Educacional

Existem vários modelos matemáticos utilizados na TRI, diferentes quanto à sua função e à quantidade de parâmetros, e cada um deles é específico para uma (ou mais) situação. Esses modelos podem ser classificados quanto à sua dimensão (unidimensionais ou multidimensionais), quanto ao tipo de traço latente (cumulativo ou não cumulativo), quanto ao tipo de item (dicotômico ou politômico) e quanto ao número de populações envolvidas (MOREIRA JUNIOR, 2011).

No contexto da avaliação educacional, dois modelos

são adequados: (1) o Modelo Logístico de Três Parâmetros – ML3 (LORD, 1980; BIRNBAUM, 1968) e (2) o Modelo de Resposta Nominal – MRN (BOCK, 1972). Esses dois modelos são do tipo "unidimensional" (um único traço latente está relacionado com a capacidade do indivíduo) e "cumulativo" (há uma relação de dominância entre itens e indivíduos, de tal forma que se um indivíduo domina certo item, consequentemente ele domina também todos os itens que estão posicionados abaixo desse item na escala). O ML3 é dicotômico, considera apenas se o indivíduo acertou ou não a questão, mas também é o único modelo que considera a possibilidade do acerto casual. O MRN é politômico e leva em conta, não somente se o indivíduo acertou ou não a questão, mas qual foi a alternativa que ele respondeu. Dessa forma, se dois indivíduos acertam e erram as mesmas questões, porém não assinalando as mesmas alternativas quando erram, eles têm a mesma nota pelo ML3, porém diferentes notas pelo MRN. As avaliações em larga escala, que utilizam a TRI no Brasil, adotam o ML3. São raros os estudos que utilizam o MRN para a avaliação educacional. Dessa forma, o foco nesse artigo será dado ao ML3. Outros modelos mais simples, tais como, o modelos logísticos de 1 parâmetro – ML1 (WRIGHT, 1968), também conhecido como Modelo de Rasch (RASCH, 1960) e o modelo logístico de 2 parâmetros – ML2 (LORD, 1980; BIRNBAUM, 1968) também podem ser utilizados nas avaliações educacionais, como fazem alguns autores, por exemplo, Oliveira (2006) e Primi, Hutz e Silva (2011). No entanto, são mais restritos, pois não levam em conta a possibilidade do acerto casual e o ML1 ainda considera que todos os itens possuem a mesma discriminação.

O ML3, que considera a dificuldade, a discriminação e a probabilidade de acerto casual do item, é o mais indicado e aplicado nas avaliações educacionais de proficiência. O ML3 é adequado para o ajuste de itens politômicos (itens com duas ou mais categorias) com uma única opção de resposta correta, o que permite que o item seja dicotomizado em duas categorias: certa e errada. Além disso, esse modelo permite modelar a probabilidade do acerto casual, ou seja, a probabilidade de um aluno com baixa proficiência acertar um determinado item. Segundo Andrade, Tavares e Valle (2000), o ML3 é dado por:

$$P\left(U_{ij}=1 \mid \theta_j\right)=c_i+\left(1-c_i\right)\frac{1}{1+e^{-Da_i\left(\theta_j-b_i\right)}}$$

para $i=1,\ 2,\ \ldots,\ I$, e $j=1,\ 2,\ \ldots,\ n$

onde:

U_{ij} é uma variável dicotômica (assume o valor 1 quando o indivíduo *j* responde corretamente o item *i*, ou assume o valor 0, caso contrário);

θ_j é o valor do traço latente (parâmetro da habilidade) do indivíduo *j*;

$P\left(U_{ij}=1 \mid \theta_j\right)$, também chamada de Função de Resposta do Item (FRI), é a probabilidade do indivíduo *j* responder corretamente o item *i*, dado que ele tem habilidade θ_j, ou seja, é a proporção de respostas corretas do item *i* dos indivíduos da população com habilidade θ_j;

a_i é o parâmetro de discriminação (ou de inclinação) do item *i*;

b_i é o parâmetro de dificuldade (ou de posição) do item *i*, medido na mesma escala da habilidade;

c_i é o parâmetro de acerto casual, que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente o item *i*;

D é um fator de escala constante, igual a 1 se os parâmetros dos itens são estimados na métrica da Logística, ou igual a 1,7, se os parâmetros dos itens são estimados na métrica da ogiva Normal, que é a distribuição Normal acumulada, por aproximação (nesse estudo, os parâmetros serão analisados pela métrica da Logística, considerando, portanto, *D* = 1);

e é a conhecida constante matemática igual a 2,718281...;

I é o número total de itens; e

n é a quantidade total de indivíduos na amostra.

O ML3 é um modelo acumulativo, ou seja, a medida que o valor do traço latente aumenta, a probabilidade do indivíduo acertar o item também aumenta e vice-versa. A Figura 1 apresenta um exemplo de uma Curva Característica do Item (CCI) de um ML3 e a sua relação existente com os parâmetros dos itens a_i (inclinação da curva), b_i (posição do item na escala) e c_i (probabilidade de acerto casual de indivíduos com baixa habilidade). A CCI é o gráfico da função do modelo matemático, onde o eixo Y é a probabilidade de resposta correta de um indivíduo segundo o valor da sua habilidade (eixo X).

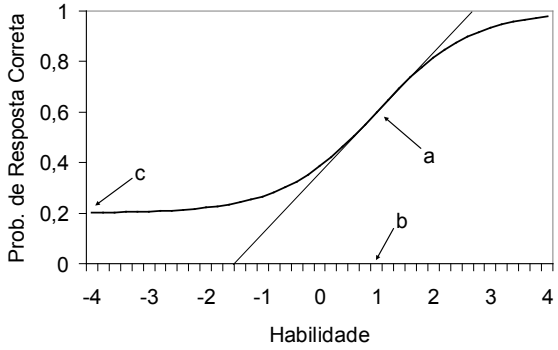

Figura 1: Relação entre os parâmetros dos itens e a CCI

O traço latente (habilidade ou proficiência) do indivíduo (θ_j) é medido em uma escala arbitrária que varia teoricamente entre $-\infty$ e $+\infty$. Porém, o importante nessa escala não é a sua magnitude, mas as relações de ordem existentes (ANDRADE; TAVARES; VALLE, 2000). O traço latente, no modelo acumulativo, é especificado como um tipo de característica que apresenta uma pro-

babilidade maior para indivíduos com θ_j maior, e uma probabilidade menor para indivíduos com θ_j menor. Ou seja, quanto maior for θ_j, maior será a probabilidade do indivíduo *j* acertar o item.

Segundo Baker (2001), o traço latente θ_j do indivíduo *j* é invariante em relação aos itens utilizados para estimá-la, desde que os itens sejam adequados, isto é, estejam calibrados (ou seja, possuam uma boa estimativa dos parâmetros), em uma métrica comum e medindo o mesmo traço latente (unidimensionalidade). Isso justifica o fato do resultado do θ_j ser o mesmo, independente dos itens que formam o questionário, o que não ocorre na TCT. Portanto, não importa se o teste é composto por itens difíceis ou fáceis, a estimativa da habilidade é a mesma. Isso é condizente com a realidade, já que a habilidade de um indivíduo, num determinado tempo *t*, é a mesma independente do grau de dificuldade do teste. Essa é a chamada propriedade de invariância do parâmetro de habilidade da TRI.

O parâmetro a_i mede a discriminação do item. Valores baixos de a_i indicam que o item tem pouco poder de discriminação, ou seja, a probabilidade de um indivíduo responder corretamente o item ou concordar com ele é aproximadamente a mesma para indivíduos com baixa ou alta proficiência. Por outro lado, valores altos de a_i indicam que o item tem grande poder de discriminação, dividindo os indivíduos praticamente em dois grupos: os que possuem habilidades abaixo do valor de b_i (com baixa probabilidade de acertar o item), e os que possuem habilidades acima do valor de b_i (com alta probabilidade de acertar o item). Não existe um valor exato de a_i para decidir se um item discrimina bem ou não. Em geral, na métrica logística, um item com a_i maior que 0,7 pode ser considerado aceitável, mas um valor maior ou igual a 1,0 indica que o item discrimina bem. Valores extremamente altos de a_i também não são adequados, pois provavelmente dividiria os indivíduos em dois grupos distintos (os que têm θ_j maior que b_i e os que têm θ_j menor que b_i), mas não faria distinção entre os indivíduos dentro dos grupos.

O parâmetro mais importante do ML3 é o b_i, parâmetro de dificuldade ou proficiência do item, que é medido na mesma unidade da escala da habilidade do indivíduo (θ_j). Ele representa o grau de dificuldade do item, ou seja, quanto maior seu valor, mais difícil o item é (somente indivíduos com habilidade alta terão uma boa probabilidade de acertá-lo), e vice-versa. Esse valor de b_i é que vai definir a posição do item na escala, por isso ele também é chamado de parâmetro de localização. Teoricamente, b_i pode assumir qualquer valor entre –∞ e +∞, entretanto, para valores muito altos ou baixos, o item pode não ser adequado, sendo usual os valores entre -3 e 3, na escala (0, 1), isto é, com média igual a zero e desvio padrão igual a um.

O parâmetro c_i é a probabilidade de um indivíduo com baixa proficiência ou com pouco (ou nenhum) conhecimento, em relação ao assunto que está sendo avaliado, responder corretamente ao item *i*. O parâmetro c_i é considerado quando existe a possibilidade de acerto casual, que é o caso do ML3, e o seu valor depende da quantidade de alternativas que o item apresenta.

Tanto para estimação dos parâmetros dos itens quanto para a estimação do traço latente, há vários métodos estatísticos sofisticados que podem ser utilizados, tais como o método de Máxima Verossimilhança Marginal (MVM), o método bayesiano da Moda a Posteriori (MAP), o método bayesiano da Média a Posteriori (EAP) e o método da Máxima Verossimilhança Conjunta (MVC) (ANDRADE; TAVARES; VALLE, 2000). Esses métodos não possuem solução explícita, o que torna necessária a utilização de algum método numérico iterativo, como o Algoritmo Newton-Raphson (ISSAC; KELLER, 1966), o Método Scoring de Fisher (RAO, 1973) e o Algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977). Essas soluções envolvem cálculos bastante complexos e, conseqüentemente, necessitam de programas computacionais específicos. Os principais softwares utilizados para análise de TRI são o BILOG-MG, o MULTILOG e o PARSCALE (TOIT, 2003), o Xcalibre (WEISS; GUYER, 2010) e o R (R DEVELOPMENT CORE TEAM, 2012). O R possui vários pacotes para análise de TRI, dentre os quais, se destacam o *ltm* (RIZOPOULOS, 2013), o *irtoys* (PARTCHEV, 2013) e o *mirt* (CHALMERS, 2012). A calibração dos itens, devido à facilidade computacional, geralmente é feita na escala (0,1), ou seja, numa escala com média igual a zero e desvio padrão igual a 1.

Todo item fornece uma informação à avaliação na TRI, através da Função de Informação do Item (FII), que permite analisar a quantidade de informação que um item fornece para a medida do traço latente analisado e reflete a qualidade do item. Segundo Andrade, Tavares e Valle (2000), a FII de um ML3 é dada por:

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

onde:

$I_i(\theta)$ é a informação fornecida pelo item *i* no nível de habilidade θ,

$P_i(\theta) = P\left(X_{ij} = 1 \mid \theta\right)$, e

$Q_i(\theta) = 1 - P_i(\theta)$

Utilizando D = 1 (métrica logística) e desenvolvendo as expressões da equação (1), chega-se a seguinte expressão, conforme Francisco (2005):

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

Quanto maior for a informação de um item, melhor será a sua qualidade. Quanto maior for o valor de a_i e menor for o valor de c_i, maior será a informação

do item (curva será mais alta e estreita) e mais acentuada será a CCI.

A soma de todas as FII da avaliação gera a Função de Informação do Teste (FIT), ou seja:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)$$

A FIT mede a qualidade do banco de itens do teste. Um teste adequado deve apresentar informação considerável em toda a extensão desejável de θ. A partir da FIT obtém-se o erro-padrão da medida da habilidade (EPM), que é o erro padrão do estimador de θ, ou seja:

$$EP(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Observa-se uma relação inversa entre a FIT e o erro padrão de estimação: quanto maior for a FIT, menor será o erro padrão de estimação e maior será a precisão da estimação da habilidade. Quanto maior o erro padrão de estimativa, menor a precisão com que é estimado o traço latente θ.

A possibilidade de equalização, ou seja, colocar diferentes populações numa mesma escala, é um dos grandes avanços da TRI em relação à TCT, pois possibilita comparar os diferentes grupos e acompanhar a evolução do traço latente ao longo do tempo. Vamos considerar a situação mais simples, ou seja, dois grupos diferentes resolvendo provas parcialmente diferentes, ou seja, com alguns itens em comum. Existem vários métodos de equalização a posteriori que podem ser utilizados para estabelecer uma relação entre os itens em comum de tal forma que permita colocar os parâmetros de um dos conjuntos de itens na escala do outro. Alguns métodos podem ser consultados em Kolen e Brennan (1995). Um método de equalização que possui um bom desempenho é o método Média-Desvio (MS), que consiste em obter a inclinação *A* por:

$$A = \frac{S_{G1}}{S_{G2}} \qquad (2)$$

onde:

S_{G1} é o desvio padrão dos parâmetros de dificuldade dos itens comuns do grupo de referência (G1), e

S_{G2} é o desvio padrão dos parâmetros de dificuldade dos itens comuns do outro grupo (G2);

e o intercepto *B* por:

$$B = M_{G1} - A.M_{G2} \qquad (3)$$

onde:

M_{G1} é a média dos parâmetros de dificuldade dos itens comuns do grupo de referência, e

M_{G2} é a média dos parâmetros de dificuldade dos itens comuns do outro grupo.

As quantidades *A* e *B* são utilizadas para colocar os parâmetros dos itens do G2 na escala do G1 a partir das relações:

$$a'_{G2} = \frac{a_{G2}}{A} \qquad (4)$$

onde:

a'_{G2} é o parâmetro de discirminação do Grupo 2 colocado na escala do grupo de referência

e

$$b'_{G2} = (A.b_{G2}) + B \qquad (5)$$

De modo semelhante, o traço latente do G2 pode ser colocado na escala do G1 por meio da transformação linear:

$$\theta'_{G2} = (A.\theta_{G2}) + B \qquad (6)$$

Outra grande vantagem da TRI em relação à TCT, é a construção e a interpretação da escala do traço latente. Conforme, Fontanive, Elliot e Klein (2007), as escalas de habilidade ordenam o desempenho dos indivíduos do menor para o maior de forma contínua e são cumulativas, isto é, os indivíduos que situam-se em um determinado nível da escala são capazes de demonstrar as habilidades descritas nesse nível e nos níveis anteriores dessa escala. A construção da escala de habilidade é efetuada após a calibração (e equalização, se necessário) dos itens, com o objetivo de encontrar uma interpretação qualitativa dos valores obtidos pela aplicação do modelo da TRI, possibilitando assim, a interpretação pedagógica dos valores das habilidades. Nesse sentido, surge a idéia dos níveis âncoras e a técnica conhecida como ancoragem (BEATON; ALLEN, 1992). Andrade, Tavares e Valle (2000) definem níveis âncora como pontos selecionados pelo analista na escala da habilidade para serem interpretados pedagogicamente. Os níveis âncoras são caracterizados pelos itens âncoras, que são itens "típicos" desse nível, ou seja, bastante respondido positivamente por indivíduos com aquele nível de habilidade e pouco respondido positivamente por indivíduos com um nível de habilidade imediatamente inferior (ANDRADE; TAVARES; VALLE, 2000). As condições para a definição de itens âncoras para os Modelos Logísticos (dicotômicos) da TRI foram propostas por Kolen e Brennan (1995), os quais definem item âncora da seguinte forma: considere dois níveis âncora consecutivos Y e Z sendo que Y < Z. Um determinado item é âncora para o nível Z se e somente se as 3 condições abaixo forem satisfeitas simultaneamente:

$$P(U = 1 \mid \theta = Z) \geq 0{,}65,$$

$$P(U = 1 \mid \theta = Y) \leq 0{,}50 \text{ e}$$

$$P(U = 1 \mid \theta = Z) - P(U = 1 \mid \theta = Y) \geq 0{,}30.$$

Na prática, às vezes um item não se caracteriza âncora por violar "levemente" uma das três condições necessárias. Nessas situações, pode-se considerar esse item como sendo âncora, se ele for importante ou se existirem poucos itens no instrumento de pesquisa. Outra alternativa é dividir os itens em grupos, segundo a quantidade de condições satisfeitas. Valle (2001) salienta que alguns níveis âncoras extremos podem ser mal caracterizados por serem definidos por itens muito fáceis ou muito difíceis, os quais geralmente são poucos.

3 Materiais e Método

Serão considerados dois grupos representando a primeira e a segunda série do Ensino Médio. Cada uma dessas séries será composta por 100 alunos simulados que responderão à uma prova de 20 itens simulados (uma prova para a primeira série e uma prova para a segunda série, totalizando 40 itens). Para fins de ilustração da equalização, dentre os 20 itens da segunda série, os cinco primeiros itens são os mesmos cinco primeiros itens da primeira série. Nessa simulação, considerou-se que o nível de dificuldade médio da prova da segunda série era um desvio padrão acima o nível de dificuldade médio da prova da primeira série. Dessa forma, os cinco itens em comum (itens de natureza da primeira série) tiveram seu parâmetro de dificuldade reduzido em um desvio padrão na prova da segunda série.

Primeiramente, foram simulados os itens de cada prova, ou seja, 20 itens do primeiro ano e 15 itens do segundo ano. Os itens foram simulados da seguinte forma: o parâmetro de discriminação a partir de uma distribuição uniforme entre 1 e 2, o parâmetro de dificuldade partir de uma distribuição normal padrão, e o parâmetro de acerto casual a partir de uma distribuição uniforme entre 0,15 e 0,25. Os itens de cada prova foram simulados separadamente em escala própria e o traço latente dos indivíduos dos dois grupos foi, primeiramente, simulado, a partir de uma distribuição normal padrão, e, posteriormente, estimado na sua própria escala.

Nessa simulação, adotou-se o método EAP para a estimação do traço latente. Na sequência, foi feita uma equalização a posteriori utilizando o método MS e os parâmetros dos itens da prova da segunda série foram colocados na escala da primeira série (grupo de referência) por meio de uma transformação linear através das equações (4) e (5). Em seguida, o traço latente dos indivíduos da segunda série também foi colocado na escala da primeira série, por meio de uma transformação linear através da equação (6).

As simulações, as estimativas do traço latente e a equalização foram realizadas com o pacote *irtoys* (PARTCHEV, 2013) do Software R (R DEVELOPMENT CORE TEAM, 2012).

4 Resultados e discussão

A Tabela 1 apresenta, para cada item da prova da primeira série, a proporção de erros e a proporção de acertos, que é chamada de Índice de Dificuldade (ID) da TCT.

Tabela 1 – Proporções de Erros e Acertos da Prova da Primeira Série

Item	Proporção de Erros	Proporção de Acertos
Item01	0,66	0,34
Item02	0,53	0,47
Item03	0,31	0,69
Item04	0,61	0,39
Item05	0,47	0,53
Item06	0,22	0,78
Item07	0,78	0,22
Item08	0,12	0,88
Item09	0,47	0,53
Item10	0,07	0,93
Item11	0,43	0,57
Item12	0,37	0,63
Item13	0,33	0,67
Item14	0,53	0,47
Item15	0,29	0,71
Item16	0,37	0,63
Item17	0,52	0,48
Item18	0,44	0,56
Item19	0,34	0,66
Item20	0,46	0,54

A Tabela 2 apresenta, para cada item da prova da segunda série, a proporção de erros e a proporção de acertos. Nota-se que os itens em comum entre as provas, que são os cinco primeiros itens de cada prova, apresentam maior proporção de acertos na prova da segunda série. Isso é coerente, uma vez que se espera que o nível de proficiência médio daqueles que estão na segunda série seja maior do que o nível de proficiência médio daqueles que estão na primeira série.

A grande limitação da TCT é colocar os alunos dessas duas provas numa mesma escala, a fim de verificar o ganho, em termos de conhecimento do aluno, ao longo do tempo, e, conseqüentemente, criar uma escala pedagógica interpretável. Com a existência de itens comuns entre as provas, é possível criar essa escala única, por meio de um processo de equalização da TRI.

Nesse estudo será utilizado um método de equalização a posteriori. Dessa forma, os itens de cada prova serão simulados independentemente (com exceção dos itens em comum) e, posteriormente, colocados na mesma escala. A Tabela 3 apresenta os parâmetros de cada item da prova da primeira série do ML3 da TRI. O nível médio de dificuldade dessa prova foi de 0,16 (média do parâmetro *b*).

Tabela 2 – Proporções de Erros e Acertos da Prova da Segunda Série

Item	Proporção de Erros	Proporção de Acertos
Item21	0,57	0,43
Item22	0,40	0,60
Item23	0,13	0,87
Item24	0,42	0,58
Item25	0,30	0,70
Item26	0,21	0,79
Item27	0,69	0,31
Item28	0,42	0,58
Item29	0,40	0,60
Item30	0,52	0,48
Item31	0,44	0,56
Item32	0,51	0,49
Item33	0,14	0,86
Item34	0,30	0,70
Item35	0,32	0,68
Item36	0,13	0,87
Item37	0,58	0,42
Item38	0,13	0,87
Item39	0,55	0,45
Item40	0,38	0,62

Tabela 3 – Parâmetros dos Itens da Prova da Primeira Série

Item	*a*	*b*	*c*
Item01	1,12	2,06	0,18
Item02	1,20	0,91	0,22
Item03	1,71	-0,38	0,21
Item04	1,03	1,24	0,18
Item05	1,48	0,20	0,16
Item06	1,29	-0,96	0,23
Item07	1,60	1,82	0,17
Item08	1,79	-1,52	0,17
Item09	1,41	0,17	0,19
Item10	1,71	-1,31	0,24
Item11	1,16	0,74	0,23
Item12	1,34	0,05	0,15
Item13	1,08	-0,50	0,21
Item14	1,15	0,82	0,15
Item15	1,41	-0,54	0,24
Item16	1,33	0,00	0,20
Item17	1,61	0,28	0,21
Item18	1,43	0,25	0,18
Item19	1,68	-0,40	0,24
Item20	1,19	0,27	0,16

A Figura 2 apresenta as CCI's dos itens da prova da primeira série do ML3 da TRI. Todos os itens apresentam comportamento satisfatório. Os parâmetros desses itens não serão modificados no processo de equalização, pois a primeira série será considerada o grupo de referência.

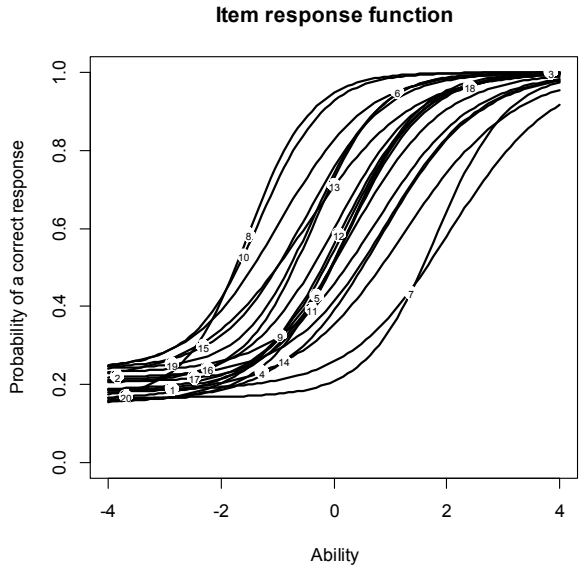

Figura 2: CCI's dos itens da prova da primeira série

A Figura 3 apresenta as FII's dos itens da prova da primeira série, enquanto que a Figura 4 apresenta a FIT. Observa-se que a informação do teste encontra-se concentrada em torno da dificuldade média. A FIT permite também identificar as regiões do traço latente onde há pouca informação contemplada pelo teste, mostrando outra vantagem da TRI em relação à TCT.

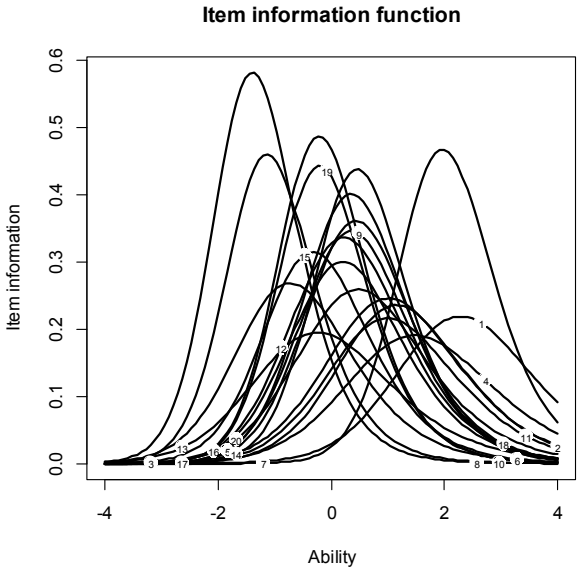

Figura 3: FII's dos itens da prova da primeira série

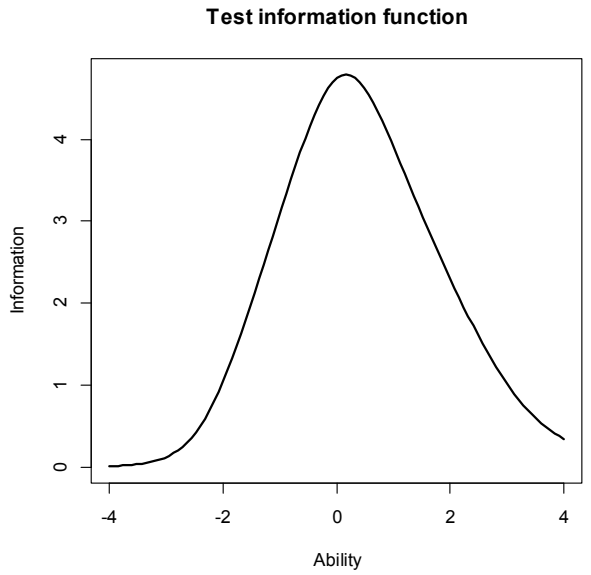

Figura 4: FIT da prova da primeira série

A Figura 5 apresenta o histograma do traço latente estimado, por meio do método EAP, dos 100 alunos da primeira série. A média do traço latente foi de 0,02 e o desvio padrão foi de 0,79. Embora o traço latente tenha sido simulado com base em uma distribuição Normal Padrão, o valor mais baixo do desvio padrão é devido ao método EAP, que tende a subestimar os traços latentes altos e superestimar os traços latentes baixos, reduzindo, assim, a variabilidade dos dados. Esses valores não serão modificados no processo de equalização.

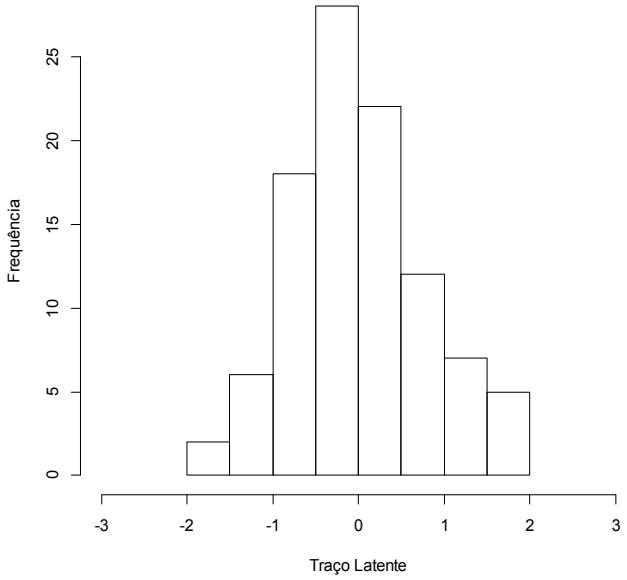

Figura 5: Histograma do traço latente estimado dos alunos da primeira série

A Tabela 4 apresenta a estimativa do traço latente de seis alunos da primeira série.

Tabela 4 - Estimativa do traço latente de seis alunos da primeira série

Aluno	Traço latente Estimado
Aluno 101	-0,5188
Aluno 102	0,3970
Aluno 103	0,0104
Aluno 104	-0,2217
Aluno 105	0,0374
Aluno 106	-0,9847

Após a calibração dos itens e a estimação do traço latente de seis alunos da primeira série, foi feito o mesmo processo com os dados da prova da segunda série. A Tabela 5 apresenta os parâmetros de cada item da prova da segunda série do ML3 da TRI. O nível médio de dificuldade dessa prova foi de -0,15 (média do parâmetro *b*). A princípio, um leigo em TRI poderia sugerir que a prova do segundo ano é mais fácil que a do primeiro ano, pois o nível médio de dificuldade da prova do segundo ano é menor do que a do primeiro ano. No entanto, essas provas foram calibradas em escalas diferentes, que não podem ser comparadas diretamente, sem um método de equalização.

Nota-se também que os parâmetros dos cinco primeiros itens (itens em comum) possuem o valor do parâmetro de dificuldade um desvio padrão menor do que na prova da primeira série (Tabela 3), como foi propositalmente simulado.

Tabela 5 – Parâmetros dos Itens da Prova da Segunda Série

Item	*a*	*b*	*c*
Item21	1,12	1,06	0,18
Item22	1,20	-0,09	0,22
Item23	1,71	-1,38	0,21
Item24	1,03	0,24	0,18
Item25	1,48	-0,80	0,16
Item26	1,68	-1,07	0,24
Item27	1,84	1,30	0,19
Item28	1,50	-0,06	0,16
Item29	1,25	0,04	0,24
Item30	1,39	0,18	0,22
Item31	1,55	0,33	0,17
Item32	1,60	0,52	0,22
Item33	1,76	-1,43	0,16
Item34	1,45	-0,34	0,17
Item35	1,05	-0,23	0,24
Item36	1,17	-1,69	0,21
Item37	1,09	1,07	0,18
Item38	1,39	-1,45	0,19
Item39	1,53	0,72	0,18
Item40	1,08	-0,01	0,21

A Figura 6 apresenta as CCI's dos itens da prova da segunda série do ML3 da TRI. Todos os itens apresentam comportamento satisfatório. No entanto, os parâmetros desses itens serão modificados no processo de equalização, pois a primeira série será considerada o grupo de referência.

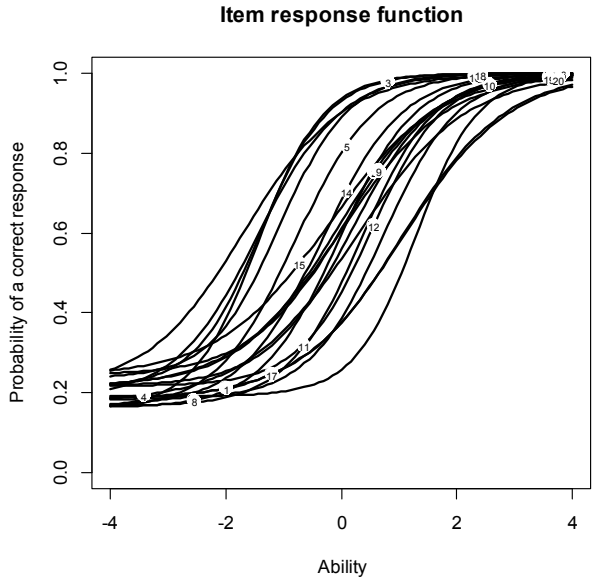

Figura 6: CCI's dos itens da prova da segunda série

A Figura 7 apresenta as FII's dos itens da prova da segunda série, enquanto que a Figura 8 apresenta a FIT. Observa-se que a informação do teste encontra-se concentrada em torno da dificuldade média.

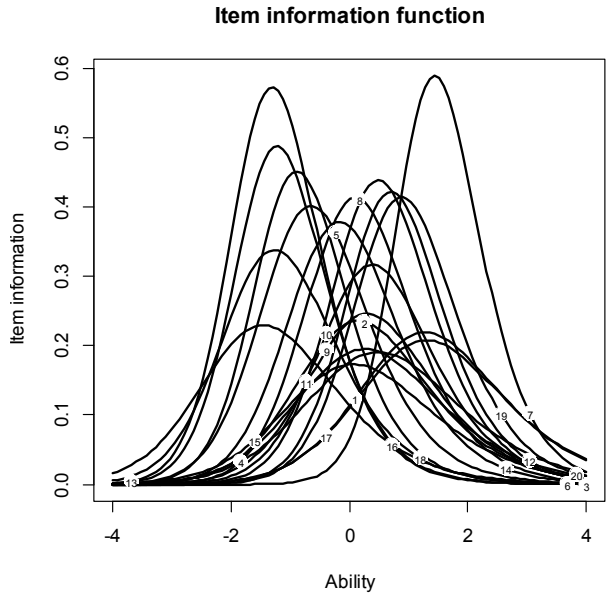

Figura 7: FII's dos itens da prova da segunda série

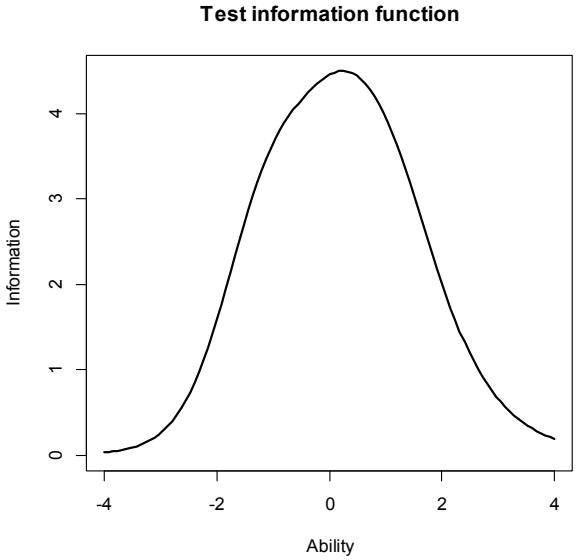

Figura 8: FIT da prova da segunda série

A Figura 9 apresenta o histograma do traço latente estimado, por meio do método EAP, dos 100 alunos da segunda série. A média do traço latente foi de -0,02 e o desvio padrão foi de 0,89. Como foi mencionado anteriormente, o método EAP tende a reduzir a variabilidade dos dados. No entanto, esses valores serão modificados no processo de equalização.

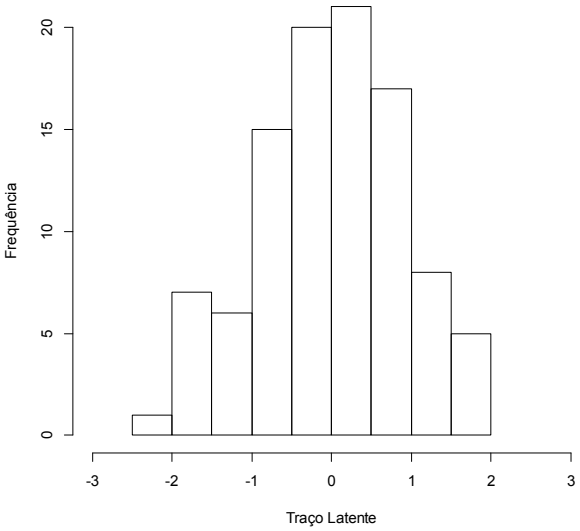

Figura 9: Histograma do traço latente estimado dos alunos da segunda série

A Tabela 6 apresenta a estimativa do traço latente de seis alunos da segunda série.

Após a calibração dos itens de ambas as séries em escalas separadas, pode-se proceder ao método de equalização a posteriori. Nesse estudo, foi utilizado o método média-desvio (*Mean-Sigma*) de equalização a posteriori.

Tabela 6 - Estimativa do traço latente de seis alunos da segunda série

Aluno	Traço latente Estimado
Aluno 201	0,6846
Aluno 202	-0,2663
Aluno 203	-0,1081
Aluno 204	0,3015
Aluno 205	0,2102
Aluno 206	-0,8788

A Tabela 7 apresenta os parâmetros de cada item da prova da segunda série na escala da primeira série. Observa-se que os itens em comum têm as mesmas estimativas do parâmetro de dificuldade nas duas provas (vide Tabela 3). No entanto, esses valores foram exatamente iguais devido à simulação, pois na prática, os valores estimados são aproximados e variam conforme o método de equalização utilizado. O nível médio de dificuldade dessa prova foi de 0,85 (média do parâmetro *b*). Ao contrário do que um leigo em TRI poderia sugerir antes da equalização, pode-se perceber que a prova do segundo ano é 0,69 desvio padrão mais difícil que a prova do primeiro ano.

Tabela 7 – Parâmetros dos Itens da Prova da Segunda Série na escala da Primeira Série

Item	*a*	*b*	*c*
Item21	1,12	2,06	0,18
Item22	1,20	0,91	0,22
Item23	1,71	-0,38	0,21
Item24	1,03	1,24	0,18
Item25	1,48	0,20	0,16
Item26	1,68	-0,07	0,24
Item27	1,84	2,30	0,19
Item28	1,50	0,94	0,16
Item29	1,25	1,04	0,24
Item30	1,39	1,18	0,22
Item31	1,55	1,33	0,17
Item32	1,60	1,52	0,22
Item33	1,76	-0,43	0,16
Item34	1,45	0,66	0,17
Item35	1,05	0,77	0,24
Item36	1,17	-0,69	0,21
Item37	1,09	2,07	0,18
Item38	1,39	-0,45	0,19
Item39	1,53	1,72	0,18
Item40	1,08	0,99	0,21

A Figura 10 apresenta as CCI's dos itens da prova da segunda série na escala da primeira série. Todos os itens apresentam comportamento satisfatório. Nota-se que, em comparação com a Figura 6, todos os itens estão deslocados para a direita. Isso se deve ao fato de terem sido colocados na escala da primeira série, ou seja, nessa escala, eles são mais difíceis, naturalmente.

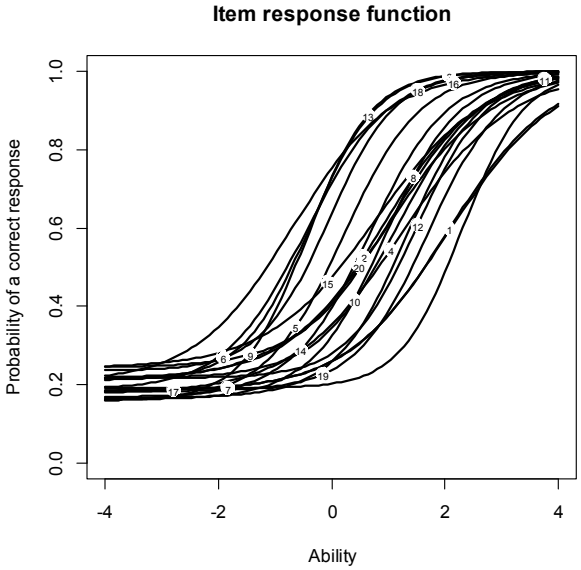

Figura 10: CCI's dos itens da prova da segunda série na escala da primeira série

A Figura 11 apresenta as FII's dos itens da prova da segunda série na escala da primeira série, enquanto que a Figura 12 apresenta a FIT. As curvas de ambas as figuras estão deslocadas para a direita, assim como as CCI's, e pelo mesmo motivo.

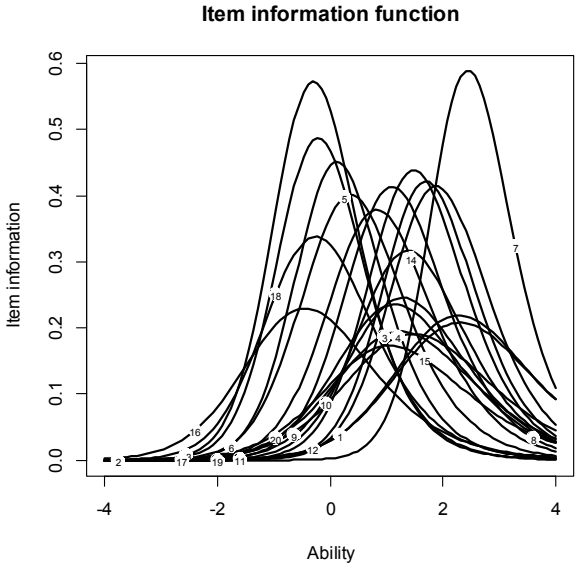

Figura 11: FII's dos itens da prova da segunda série na escala da primeira série

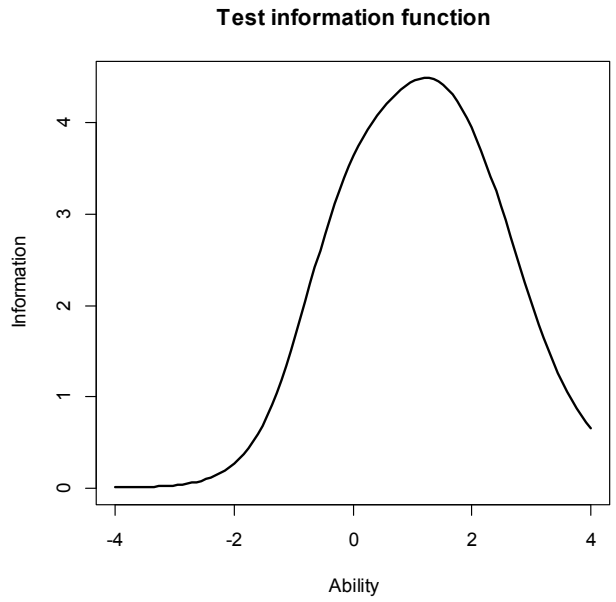

Figura 12: FIT da prova da segunda série na escala da primeira série

A Figura 13 apresenta o histograma do traço latente estimado, por meio do método EAP, dos 100 alunos da segunda série na escala da primeira série. A média do traço latente foi de 0,98 e o desvio padrão foi de 0,89. Pode-se concluir que os alunos da segunda série têm um ganho, em termos de proficiência, em média, de 1 desvio padrão, em relação à primeira série. Dessa forma, é possível comparar a evolução, em termos de ganho de proficiência, ao longo das séries.

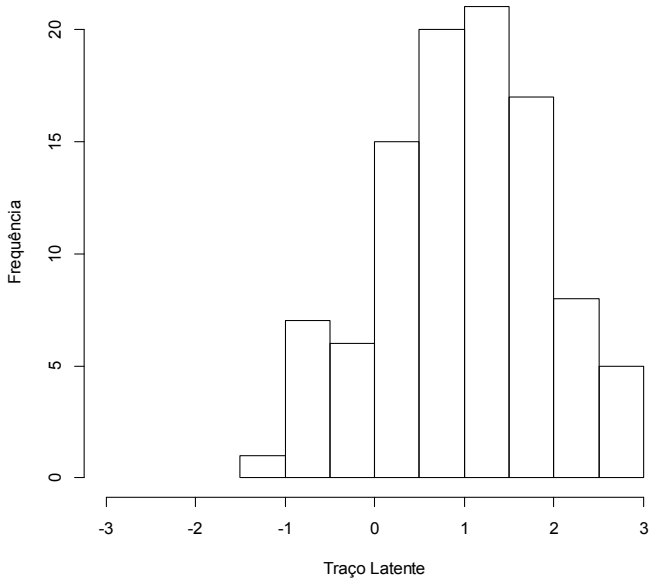

Figura 13: Histograma do traço latente estimado dos alunos da segunda série

A Tabela 8 apresenta a estimativa do traço latente de seis alunos da segunda série na escala da primeira série. Observa-se que, agora, as notas dos alunos estão maiores do que as apresentadas na tabela 6, devido ao fato de estarem na escala da primeira série. Nota-se, por exemplo, que o aluno 202, que tem nota -0,27 na escala da segunda série (Tabela 6) parecia ter uma nota inferior ao aluno 102 da primeira série, com nota 0,40 (Tabela 4), no entanto, a nota do aluno 202 colocada na escala da primeira série é de 0,73, revelando que ele têm maior nota e, consequentemente, maior proficiência que o aluno 102.

Tabela 8 - Estimativa do traço latente de seis alunos da segunda série na escala da primeira série

Aluno	Traço latente Estimado
Aluno 201	1,6846
Aluno 202	0,7337
Aluno 203	0,8919
Aluno 204	1,3015
Aluno 205	1,2102
Aluno 206	0,1212

Os itens de ambas as séries estão agora posicionados na mesma escala, nesse caso, na escala da primeira série. A Figura 14 apresenta a distribuição do traço latente de todos os alunos da primeira e da segunda série na escala da primeira série, enquanto que a Figura 15 apresenta a posição dos itens das duas provas na escala da primeira série (círculos sólidos indicam que o item é exclusivo da prova daquela série e círculos com preenchimento em branco indicam os itens em comum entre as provas).

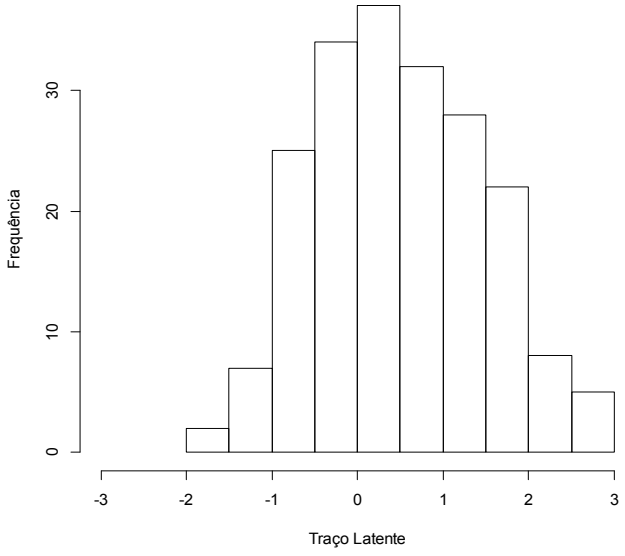

Figura 14: Distribuição do traço latente de todos os alunos na escala da primeira série

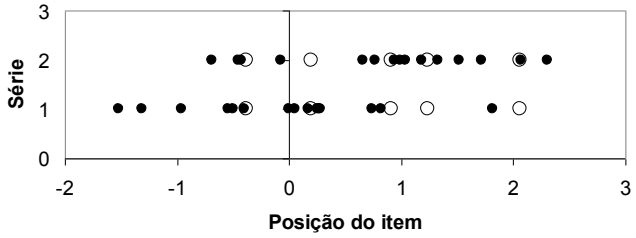

Figura 15: Posição dos itens na escala da primeira série

Dessa forma, é possível criar níveis âncoras e interpretar os níveis, que é outra grande vantagem da TRI sobre a TCT. A execução dessa tarefa precisa ser feita por um profissional que conheça o traço latente analisado e os item do construto (descrição, conteúdo), além das condições necessárias para a identificação dos itens âncoras. Como nesse estudo de simulação não há descrição dos itens (os itens foram simulados), não será possível interpretar pedagogicamente os níveis âncoras, apenas serão identificados os níveis âncoras e os itens âncoras. Observando as condições de Kolen e Brennan (1995), foram identificados os níveis âncoras e itens âncoras e quase âncoras, apresentados na Tabela 9. Em negrito, estão destacados os itens em comum das duas séries. Observa-se que os itens da primeira série tendem a se posicionar em níveis mais inferiores enquanto que os itens de segunda série tendem a se posicionar em níveis superiores.

Tabela 9 – Níveis e itens âncoras

Nível Âncora	Itens âncoras	Itens Quase âncoras
-1	8, 10	6
0	**3**, 12, 19, 26, 33	16, 38
1	28, 34	**2, 4, 5**, 14, 17, 29, 30
2	7, 39	**1**, 31, 32, 37
3		27

A interpretação é a seguinte: indivíduos que estão posicionados no nível -1 dominam ou conhecem o conteúdo dos itens 6, 8 e 10; indivíduos que estão posicionados no nível 0 dominam ou conhecem o conteúdo dos itens dos níveis anteriores e dos itens 3, 12, 16, 19, 26, 33 e 38; indivíduos que estão posicionados no nível 1 dominam ou conhecem o conteúdo dos níveis anteriores e dos itens 2, 4, 5, 14, 17, 28, 29, 30 e 34; indivíduos que estão posicionados no nível 2 dominam ou conhecem o conteúdo dos níveis anteriores e dos itens 1, 7, 31, 32, 37 e 39; e indivíduos que estão posicionados no nível 3 dominam ou conhecem o conteúdo dos níveis anteriores e do item 27. A interpretação e a criação desses níveis na escala configuram outra grande contribuição da TRI para as avaliações educacionais.

5 Conclusões

A TRI trouxe à avaliação educacional novas possibilidades, em relação à tradicional TCT. A principal vantagem foi a possibilidade de comparar indivíduos que responderam questionários com itens diferentes e colocá-los em uma escala única onde são posicionados os itens e o indivíduos, permitindo a interpretação pedagógica da mesma e possibilitando o acompanhamento do aluno ao longos dos anos, em termos de aquisição de conhecimento. Outras vantagens da utilização da TRI em relação à TCT estão relacionadas com a precisão das estimativas, informação do teste, a possibilidade de considerar o acerto casual, calcular as probabilidades de resposta dos indivíduos, etc.

Este artigo procurou apresentar os principais conceitos da TRI e as suas contribuições para as avaliações educacionais. Para ilustrar esse objetivo, foi feita uma análise, por meio de simulações, que mostrou como funciona o processo de equalização, necessário para colocar indivíduos que responderam provas parcialmente diferentes numa única escala comparável e interpretável. Também foram identificados níveis âncoras, os itens âncoras e os itens quase âncoras resultantes da simulação realizada, a fim de ilustrar a interpretação da escala.

Referências

ANDRADE, J. M.; LAROS, J. A.; GOUVEIA, V. V. O uso da teoria de resposta ao item em avaliações educacionais: diretrizes para pesquisadores. Aval. psicol., Porto Alegre, v. 9, n. 3, dez. 2010.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. Teoria da resposta ao item: conceitos e aplicações. São Paulo: ABE - Associação Brasileira de Estatística, 2000.

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

BEATON, A. E.; ALLEN, N. L. Interpreting Scales through Scale Anchoring. Journal of Educational Statistics, n. 17, p. 191-204, 1992.

BIRNBAUM, A. Some Latent Trait Models and Their Use in Infering an Examinee's Ability. In: LORD, F. M.; NOVICK, M. R. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

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

BOCK, R. D.; ZIMOWSKI, M. F. Multiple Group IRT. In Handbook of Modern Item Response Theory. W.J. van der Linder e R.K. Hambleton Eds. New York: Spring-Verlag, 1997.

CORREA, A. C.; MOREIRA JUNIOR, F. J.; ANDRADE, D. F.; BORTOLOTTI, S. L. V. Modelagem de um Instrumento de Medida de Avaliação do ENADE fundamentado na Teoria de Resposta ao Item (TRI): desenho para o MEES. In: XII Coloquio Internacional de Gestión Universitaria, 2012, Veracruz, México. Anais del XII Coloquio Internacional de Gestión Universitaria, 2012.

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

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38, 1977.

EMBRETSON, S. E.; REISE, S.P. Item Response Theory for Psychologists. New Jersey, USA: Lawrence Erlbaum Associates, 2000.

FONTANIVE, N. S.; ELLIOT, L. G.; KLEIN, R. Os desafios da apresentação dos resultados da avaliação de sistemas escolares a diferentes públicos. REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, v. 5, n. 2e, 2007.

FRANCISCO, R. Aplicação da Teoria da Resposta ao Item (TRI) no Exame Nacional de Cursos (ENC) da Unicentro. 2005. 144 f. Dissertação (Mestrado em Ciências) Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, 2005.

FERREIRA, F. F. G. Escala de Proficiência para o ENEM utilizando a Teoria da Resposta ao item. 2009. Dissertação (Mestrado em Matemática e Estatística) – Programa de Pós-Graduação em Matemática e Estatística, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2009.

KOLEN, M. J.; BRENNAN, R. L. Test Equating - Methods and Pratices. New York: Springer, 1995.

MINISTÉRIO DA EDUCAÇÃO, INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, DIRETORIA DE AVALIAÇÃO DA EDUCAÇÃO BÁSICA. Nota Técnica, 2012. Disponível em http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf. Acesso em 14/01/14.

ISSAC, E; KELLER, H. B. Analysis of Numerical Methods. New York: Wiley & Sons, 1966.

LORD, F. M. A theory of test scores (No. 7). Psychometric Monograph, 1952.

LORD, F. M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, USA: Lawrence Erlbaum Associates, Inc, 1980.

MOREIRA JUNIOR, F. J. Aplicações da Teoria da Resposta ao Item (TRI) no Brasil. Revista Brasileira de Biometria, São Paulo, v.28, n.4 , p. 137-170, out.-dez. 2010.

MOREIRA JUNIOR, F. J. Sistemática para a Implantação de Testes Adaptativos Informatizados baseados na Teoria da Resposta ao Item. 2011. 334 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2011.

OLIVEIRA, K. S. Avaliação do exame nacional de desempenho do estudante pela teoria de resposta ao item. 2006. 96 f. Dissertação (Mestrado em Psicologia) - Programa de Pós- Graduação Stricto Sensu em Psicologia, Universidade São Francisco, Itatiba, 2006.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. Pisa 2003 Data Analysis Manual. Paris, 2005. Disponível em: <http://www.oecd.org/dataoecd/53/22/35014883.pdf>. Acesso em: 30/08/2011.

_____. Pisa 2009 assessment framework: key competencies in reading, mathematics and science. Paris, 2009. Disponível em: <http://www.oecd.org/ent/44/0,3746,en_2649_35845621_44455276_1_1_1_1,00.html#TOC>. Acesso em: 30/08/2011.

PARTCHEV, I. Package irtoys: Simple interface to the estimation and plotting of IRT models, 2013. CRAN.R project, Disponível em <http://cran.r-project.org/web/packages/irtoys/irtoys.pdf> . Acesso em 16/04/2013.

PERRY, F. A. Escalas de Proficiência: Diferentes Abordagens de Interpretação na Avaliação Educacional em Larga Escala. 2009. 119 f.

Dissertação (Mestrado em Educação) - Programa de Pós-Graduação em Educação, Faculdade de Educação, Universidade Federal de Juiz de Fora, Juiz de Fora, 2009.

PRIMI, R; CARVALHO, L. F.; MIGUEL, F. K.; SILVA, M. C. R. Análise do funcionamento diferencial dos itens do Exame Nacional do Estudante (ENADE) de psicologia de 2006. Psico-USF, v. 15, n. 3, p. 379-393, set./dez. 2010.

PRIMI, R; HUTZ, C. S.; SILVA, M. C. R. A prova do ENADE de Psicologia 2006: Concepção, Construção e análise psicométrica da prova. Avaliação Psicológica, 10(3), p. 271-294, 2011.

PRIMI, R., NUNES, C. H. S. S., SILVA, M. C. R., CARVALHO, L. F., MIGUEL, F. K. ;VENDRAMINI, C. M. M. Aplicação da Teoria de Resposta ao Item na Interpretação das Notas do ENADE de Psicologia. Revista de Educação AEC, 38, 115-124, 2009.

R DEVELOPMENT CORE TEAM. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012.

RAO, C. R. Linear Statistical Inference and Its Applications. New York: Wiley & Sons, 1973.

RASCH, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.

RIZOPOULOS, D. Package ltm: Latent Trait Models under IRT. CRAN R project, 2013. Disponível em <http://cran.r-project.org/web/packages/ltm/ltm.pdf>. Acesso em 16/04/2013.

SOARES, T. M.; GENOVEZ, S. F. M.; GALVÃO, A. F. Análise do Comportamento Diferencial dos Itens de Geografia: Estudo da 4ª série avaliada no PROEB/SIMAVE. Estudos em Avaliação Educacional, v. 16, n. 32, p. 81-110, 2005.

SOUZA, S. Z. 40 Anos de Contribuição à Avaliação Educacional. Estudos em Avaliação Educacional, v. 16, n. 31, jan./jun. 2005.

TOIT, M. IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Scientific Software International, 2003.

VALLE, R. C. A Construção e a Interpretação de Escalas de Conhecimento – Considerações Gerais e uma Visão do que vem sendo feito no SARESP. Estudos em Avaliação Educacional, n. 23, p. 71-92, 2001.

VENDRAMINI, C. M. M.; SILVA, M. C.; CANALE, M. Análise de Itens de uma Prova de Raciocínio Estatístico. Psicologia em Estudo, Maringá, v. 9, n. 3, p. 487-498, set./dez, 2004.

WEISS, D. J.; GUYER, R. Manual for CATSim: Comprehensive simulation of computerized adaptive testing. St. Paul MN: Assessment Systems Corporation, 2010.

WRIGHT, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: ETS - Educational Testing Service, 1968.