



## XIV COLÓQUIO INTERNACIONAL DE GESTÃO UNIVERSITÁRIA – CIGU

### A Gestão do Conhecimento e os Novos Modelos de Universidade

Florianópolis – Santa Catarina – Brasil  
3, 4 e 5 de dezembro de 2014.

ISBN: 978-85-68618-00-4

## UMA APLICAÇÃO DA TEORIA DA RESPOSTA AO ITEM NA AVALIAÇÃO DO ENADE DO CURSO DE ADMINISTRAÇÃO

**Vinícius Teodoro Scher**

UFSM

[profvinimat@hotmail.com](mailto:profvinimat@hotmail.com)

**Vitor Francisco Schuch Junior**

UFSM

[vfschuch@gmail.com](mailto:vfschuch@gmail.com)

**Fernando de Jesus Moreira Junior**

UFSM

[fmjunior777@yahoo.com.br](mailto:fmjunior777@yahoo.com.br)

**Dalton Francisco de Andrade**

UFSC

[dandrade@inf.ufsc.br](mailto:dandrade@inf.ufsc.br)

**Angela Cristina Correa**

UFSC

[angelacorrea@gmail.com](mailto:angelacorrea@gmail.com)

**Silvana Ligia Vicenzi Bortolotti**

UTFPR

[sligie@globo.com](mailto:sligie@globo.com)

### Resumo

Tradicionalmente considerar os escores brutos ou padronizados como o resultado de uma avaliação ou de uma seleção de certo indivíduo é um fato comum. No entanto, os resultados obtidos dependem dos itens ou questões que compõem essas avaliações. O Exame Nacional de Desempenho de Estudantes (ENADE), tem o objetivo de aferir o rendimento dos alunos dos cursos de graduação em relação aos conteúdos programáticos, suas habilidades e competências. O exame é obrigatório para os alunos selecionados e condição indispensável para a emissão do histórico escolar. Seus resultados produzem dados por instituição de educação superior, categoria administrativa, organização acadêmica, município, estado, região geográfica e Brasil, construindo referenciais que permitam a definição de ações voltadas à melhoria da qualidade dos cursos de graduação por parte dos professores, técnicos, dirigentes e autoridades educacionais. O presente trabalho apresenta uma análise da prova do ENADE de 2009 que foi respondida por 230.486 alunos ingressantes e concluintes do Curso de Administração de diversas instituições do país por meio da Teoria da Resposta ao Item (TRI). Foi possível verificar a viabilidade do uso da TRI como instrumento de medida de avaliação dos itens do ENADE, assim como verificar o ganho, em termos de proficiência, entre ingressantes e concluintes.

Palavras-chave: ENADE, curso de administração, avaliação educacional, Teoria da Resposta ao Item.

## 1. Introdução

No âmbito da avaliação educacional é mais do que comum a utilização dos escores brutos ou padronizados das avaliações tradicionais como parâmetro para seleção ou classificação de indivíduos em determinadas áreas. Porém, os resultados encontrados dependem particularmente dos itens ou questões que compõem o instrumento avaliativo. Assim, torna-se inviável a comparação entre indivíduos que não foram submetidos às mesmas provas ou avaliações.

Atualmente, existe um grande interesse, por parte das áreas do conhecimento onde exista avaliação educacional, na aplicação de técnicas derivadas da Teoria de Resposta ao Item (TRI), que propõe modelos para os traços latentes, ou seja, características dos indivíduos que não podem ser observadas diretamente. Esse tipo de variável deve ser inferida a partir da observação de variáveis secundárias que estejam relacionadas a ela.

O Exame Nacional de Desempenho de Estudantes (ENADE) é um subsistema de avaliação do Sistema Nacional de Avaliação da Educação Superior (SINAES), ambos ligados ao Ministério da Educação, tem o propósito de aferir o desempenho dos estudantes dos cursos de graduação em relação aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso, suas habilidades e competências. O ENADE foi instituído como um componente curricular obrigatório dos cursos de graduação, onde participam da avaliação os estudantes em final de primeiro ano (ingressantes) e de último ano (concluintes) das áreas e cursos a serem avaliados.

Com o advento do Conceito Preliminar de Curso - CPC, instituído pela Portaria N° 4/2008 do MEC, tornou-se possível a substituição das tradicionais avaliações in loco para renovação de reconhecimento dos cursos de graduação por um cálculo aonde os resultados do ENADE são preponderantes. Com base nos resultados dos CPC's de todos os cursos de uma instituição, passou a ser calculado o Índice Geral de Cursos (IGC). Trata-se de um indicador geral de qualidade da instituição. Em função destes novos conceitos e critérios, independentemente dos questionamentos levantados segundo Schwartzman (2008), o ENADE passou a ter muito maior repercussão.

Tradicionalmente, o instrumento de medida de avaliação das provas utilizado se fundamenta na Teoria Clássica dos Testes (TCT). A Teoria da Resposta ao Item (TRI) é uma ferramenta estatística que surge como alternativa, para suprir as necessidades decorrentes das limitações da TCT. Baseado nos estudos de Andrade et al. (2000), Embretson e Reise (2000), Hambleton et al. (1991), Sao Paulo et al. (2007), Olea et al. (1996) e Vendramini et al. (set./dez 2004) podemos citar as principais desvantagens da TCT:

- O escore do indivíduo depende essencialmente dos itens que compõem o teste; Não possui a possibilidade de considerar o acerto casual;
- Os indivíduos que acertam a mesma quantidade de itens possuem o mesmo escore, tendo acertado itens diferentes (fáceis ou difíceis) na mesma prova;
- Não permite a comparação através do escore entre os indivíduos que responderam questionários com itens diferentes para medir o mesmo traço latente;
- Não consegue determinar qual seria a resposta de um indivíduo a um determinado item que ele não respondeu em um questionário;
- Os índices de dificuldade e de discriminação, por serem uma proporção de acertos, dependem da quantidade e da qualidade dos respondentes, ou seja, depende da amostra;
- O erro padrão de medida (EPM) é o mesmo para todos os escores dos respondentes;
- Formatos mesclados de itens (por exemplo, dicotômicos e poliatômicos nominais e graduais) na TCT conduzem a impacto desequilibrado nos escores total do teste;
- Analisa a confiabilidade em função de uma medida como um todo, geralmente por meio do Alfa de Cronbach.

As avaliações educacionais tradicionalmente têm sido analisadas segundo os critérios da Teoria Clássica dos Testes (TCT). O surgimento da Teoria da Resposta ao Item (TRI) possibilitou uma nova forma de elaborar avaliações educacionais. O uso da TRI apresenta potencialidades efetivas que tem incentivado o incremento de seu uso como instrumento de mensuração, principalmente em sistemas de avaliação educacional. De acordo com ANDRADE et al. (2000), TRI é uma metodologia que sugere formas de representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item e seus traços latentes. Sob este prisma a avaliação é realizada a partir de observações de variáveis secundárias relacionadas a determinada característica do indivíduo que não pode ser observada diretamente. Essa relação é possível por meio da utilização de modelos matemáticos e métodos complexos de estimação.

Levando em conta a importância que a TRI vem ganhando no meio acadêmico como forma de mensuração de conhecimento, o presente trabalho apresenta uma análise da prova do ENADE de 2009 que foi respondida por 230.486 alunos ingressantes e concluintes do Curso de Administração de diversas instituições do país. Verificou-se a viabilidade do uso da TRI como instrumento de medida de avaliação do ENADE e foi avaliado o quanto a prova do ENADE do Curso de Administração esta mensurando o verdadeiro traço latente dos seus acadêmicos, fazendo um comparativo entre o desempenho entre ingressantes e concluintes.

## **2. Estado da Arte: Teoria da Resposta ao Item (TRI)**

A TRI é um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade do respondente. Essa relação é expressa de tal forma que quanto maior a habilidade, maior a probabilidade de acerto no item (ANDRADE et al., 2000).

A Teoria da Resposta ao Item começou a ser estudada na década de 50, e eram modelos em que se considerava que uma única habilidade, de um único grupo, estava sendo medida por um teste onde os itens eram corrigidos de maneira dicotômica. Os primeiros modelos desenvolvidos, propostos por Lord (1952), foram o modelo unidimensional de dois parâmetros e o modelo unidimensional de três parâmetros, baseados na distribuição normal acumulada (ogiva normal). Birnbaum (1968) aprimorou os modelos de Lord (1952), substituindo, em ambos os modelos, a função ogiva normal pela função logística, que é mais conveniente matematicamente.

Paralelamente, porém independentemente do trabalho de Lord (1952), Rasch (1960) propôs o modelo unidimensional de um parâmetro, expresso também como modelo de ogiva normal e, posteriormente aprimorado por Wright (1968), substituindo também, a função ogiva normal pela função logística. Samejima (1969) propôs o modelo de resposta gradual com o objetivo de obter mais informação das respostas dos indivíduos do que simplesmente avaliar se eles deram respostas corretas ou incorretas aos itens. Bock (1972), Andrich (1978), Masters (1982) e Muraki (1992), também propuseram modelos para mais de duas categorias de resposta, assumindo diferentes estruturas entre essas categorias. Bock e Zimowski (1997) introduziram os modelos logísticos de 1, 2 e 3 parâmetros para duas ou mais populações de respondentes, o que trouxe novas possibilidades para as comparações de rendimentos de duas ou mais populações submetidas a diferentes testes com itens comuns.

A TRI começou a ser utilizada no Brasil em 1996, primeiramente na 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 INEP/MEC (Andrade et al., 2000).

### **2.1. Estudos de TRI em avaliações Educacionais**

A partir de 1995 a TRI passou a ser utilizada na avaliação das provas do Exame Nacional

do Ensino Médio (ENEM), sob a responsabilidade do Instituto Nacional de Estudos e Pesquisas Educacionais do Ministério da Educação (INEP/MEC). Algumas das provas do ENEM já tinham sido analisadas previamente em estudos que utilizaram a TRI. Francisco (2005) realizou um estudo de caso por meio da aplicação da TRI, com a finalidade de verificar o desempenho dos alunos formandos do curso de Matemática da Universidade Estadual do Centro-Oeste (UNICENTRO), em Guarapuava-PR, no período de 2000 até 2003, no Exame Nacional de Cursos (ENC). Foi estudado a estimação dos parâmetros dos modelos logísticos unidimensionais de um, dois e três parâmetros da TRI. Os resultados mostraram que, para a população em estudo (alunos formandos do curso de Matemática da UNICENTRO), em todos os anos analisados, a grande maioria dos itens era difícil. No entanto, observaram que a quantidade de itens difíceis diminuiu ao longo do período analisado. Uma limitação desse estudo foi a amostra utilizada para a calibração dos itens. Foram analisadas quatro provas (uma para cada ano), com número de itens igual a 25, 40, 30 e 40, e com número de respondentes igual a 46, 59, 41 e 41, respectivamente. Essa quantidade pequena de respondente pode afetar seriamente a estimativas dos parâmetros dos itens, tornando os resultados duvidosos. Outra restrição desse estudo é que os parâmetros estimados não podem ser utilizados para avaliar a proficiência dos alunos de outra IES, que poderiam apresentar DIF (*Differential item functioning*), já que a população não é a mesma.

Oliveira (2006) utilizou a TRI para proceder a uma análise das propriedades psicométricas da prova do ENADE de 2004, aplicada aos alunos dos cursos de medicina. A amostra utilizada foi de 8.124 estudantes e as análises foram realizadas apenas com as questões objetivas tanto do componente de Formação Geral (FG) quanto do componente de Formação Específica (FE), totalizando 28 itens. Os resultados, baseados no coeficiente de fidedignidade que utiliza a fórmula KR-21 Kuder e Richardson (1937) mostraram que os dados se ajustam com maior confiabilidade ao Modelo Logístico Unidimensional de um parâmetro (Modelo de Rasch) do que nos Modelos Logísticos Unidimensionais de dois e três Parâmetros.

Nogueira (2008) aplicou a TRI para avaliar as questões da prova de formação geral (FG) do ENADE, em especial aquelas que envolvem conceitos estatísticos, visando estimar a proficiência dos estudantes nos conteúdos avaliados e o ajuste dos itens ao modelo de Rasch. A amostra utilizada foi de 403.512 estudantes de vários cursos que zeraram a prova do ENADE nos anos de 2004 e de 2005. Os resultados mostraram que as questões objetivas da prova de 2004 apresentam parâmetros de dificuldade mais altos em relação a prova de 2005, exigindo maior habilidade do estudante. Por outro lado, as questões discursivas de 2004 e 2005 apresentam parâmetros de dificuldade mais baixos, embora não sejam equiparáveis. O Modelo da TRI utilizado neste último estudo é mais restrito, pois considera que todos os itens possuem o mesmo nível de discriminação e apenas a dificuldade dos itens é avaliada, não levando em conta a possibilidade do acerto casual.

Primi et al. (2010), apresentam outro enfoque da aplicação da TRI no ENADE para a determinação de pontos de corte, formando grupos de competências requeridas para a resolução de itens. A amostra utilizada foi de 26.613 estudantes que realizaram a prova do ENADE de Psicologia em 2006 e foram analisadas somente as questões referentes ao conteúdo específico. Os resultados mostraram que, de maneira geral, os estudantes concluintes concentram-se na competência mínima, enquanto que os ingressantes distribuem-se, em sua maioria, em fraca e competência mínima. Primi et al. (2010) expõem um estudo utilizando no ENADE. O DIF pode ser definido com sendo a observação, em pessoas com a mesma habilidade, de uma chance diferenciada de acerto de um item. Os parâmetros estimados pela TRI devem ser invariantes em relação a habilidade da amostra, mas quando isso não ocorre, tem-se essa situação anômala chamada de DIF.

O uso da TRI no ENEM proporcionou o seu destaque em âmbito nacional. No contexto internacional, a TRI vem sendo empregada amplamente por vários países: Estados Unidos,

Franca, China, Holanda, Coreia do Sul e principalmente nos países participantes do Programa Internacional de Avaliação de Estudantes (PISA) que conta com a participação de 32 países, tais como Estados Unidos, Coreia do Sul, Canadá e Brasil. O PISA utiliza o modelo de Rasch da Teoria de Resposta ao Item (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 (KARINO; ANDRADE, 2010). As avaliações educacionais tradicionalmente têm sido analisadas segundo os critérios da TCT. A TRI surge como uma metodologia alternativa utilizada em diversas áreas do conhecimento, e em larga escala em avaliações educacionais.

## 2.2. O modelo matemático

Existem diversos modelos matemáticos que são utilizados nas aplicações da TRI, propõe-se neste estudo a utilização do Modelo Logístico Unidimensional de Três Parâmetros (MLU3), por ser mais completo e o mais indicado na aplicação das avaliações educacionais de proficiência. O MLU3 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 avaliar a dificuldade do item, a discriminação e a probabilidade de resposta correta dada por indivíduos de baixa habilidade.

O modelo logístico unidimensional de três parâmetros (ML3) é dado por:

$$P(U_{ij} = 1 | \theta_j) = c_i + (1 - c_i) + \frac{1}{1 + e^{-D a_i (\theta_j - b_i)}}, \text{ para } i = 1, 2, \dots, I \quad e \quad j = 1, 2, \dots, n$$

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);

$\theta_j$  é o valor do traço latente (parâmetro da habilidade) do indivíduo  $j$ ;

$P(U_{ij} = 1 | \theta_j)$ , 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  $\theta_j$ , ou seja, é a proporção de respostas corretas do item  $i$  dos indivíduos da população com habilidade  $\theta_j$ ;

$a_i$  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... (número de Euler);

$i$  é o número total de itens; e

$n$  é a quantidade total de indivíduos na amostra.

O MLU3 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 MLU3 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 (Figura 1) é o gráfico da função do modelo matemático, onde o eixo Y das ordenadas é a probabilidade de resposta correta de um indivíduo segundo o valor da sua habilidade (eixo X das abcissas).

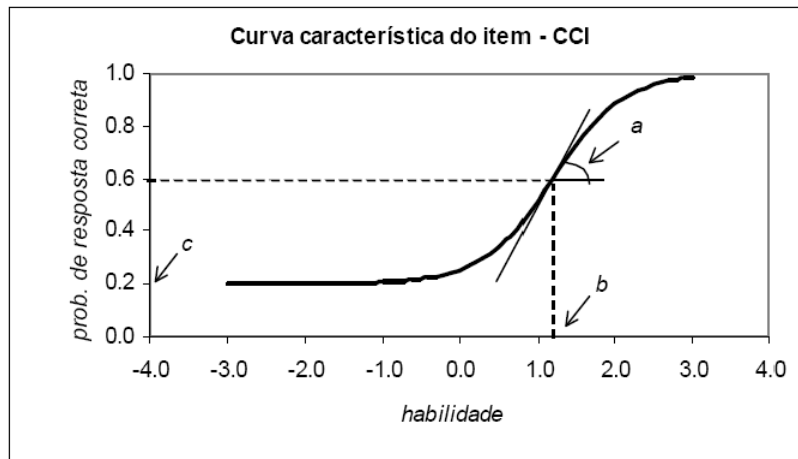


Figura 1: Relação entre os parâmetros dos itens e a CCI  
Fonte: Andrade et al. (2000)

O traço latente (habilidade ou pro ciência) do indivíduo ( $\theta_j$ ) é medido em uma escala arbitrária que varia teoricamente em toda a reta real. Porém, o importante nessa escala não é a sua magnitude, mas as relações de ordem existentes (ANDRADE et al., 2000). O traço latente, no modelo acumulativo, é especificado como um tipo de característica que apresenta uma probabilidade maior para indivíduos com  $\theta_j$  maior, e uma probabilidade menor para indivíduos com  $\theta_j$  menor. Ou seja, quanto maior for  $\theta_j$ , maior será a probabilidade do indivíduo  $j$  acertar o item.

Segundo Baker (2001), o traço latente  $\theta_j$  do indivíduo  $j$  é invariante em relação aos itens utilizados para estima-lo, 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  $\theta_j$  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 pro ciência. 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  $\theta_j$  maior que  $b_i$  e os que têm  $\theta_j$  menor que  $b_i$ ), mas não faria distinção entre os indivíduos dentro dos grupos.

O parâmetro mais importante do MLU3 é o  $b_i$ , parâmetro de dificuldade ou pro ciência do item, que é medido na mesma unidade da escala da habilidade do indivíduo  $\theta_j$ . 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 -1 e +1, 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 media igual a zero e desvio padrão igual a um.

O parâmetro  $c_i$  é a probabilidade de um indivíduo com baixa pro ciência ou com pouco (ou nenhum) conhecimento, em relação ao assunto que esta 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 MLU3, e o seu valor depende da quantidade de alternativas que o item apresenta. Por exemplo, para um item com 5 alternativas, espera-se valores entre 0,1 e 0,3. O índice de discriminação  $a$ , refere-se a inclinação da CCI no ponto de inflexão, isto é, quando a curva corta a linha que corresponde a probabilidade de  $(1 + c)/2$  de resposta correta. Quanto maior for a inclinação da curva, maior será o seu valor. Ele é proporcional ao coeficiente angular da reta tangente ao ponto de inclinação máxima, ou seja, onde a probabilidade de acerto for igual a  $(1 + c)/2$ . Dessa forma, itens com  $a$  negativo não são esperados para esse modelo, uma vez que indicariam que a probabilidade de responder corretamente o item diminui com o aumento da habilidade.

Vale ressaltar que a capacidade de discriminação dos itens varia de acordo com o nível de habilidade avaliado Primi et al. (2010). Todo item fornece uma informação a 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. Maiores detalhes podem ser encontrados em Andrade et al. (2000).

### 2.3. Métodos de estimação

Os modelos da TRI carão completamente especificados a partir da estimação dos parâmetros dos itens e das habilidades dos respondentes. O processo de estimação dos parâmetros dos itens é chamado de calibração (BAKER, 2001). No processo de estimação podem acontecer os seguintes casos:

1. os parâmetros dos itens são conhecidos e precisa-se estimar as habilidades;
2. os parâmetros das habilidades dos respondentes são conhecidos e precisa-se estimar os parâmetros dos itens;
3. os parâmetros das habilidades e dos itens são estimados simultaneamente

No primeiro caso a solução é dada empregando o método da máxima verossimilhança ou métodos bayesianos, ambos através da aplicação de procedimentos iterativos, como, por exemplo, o método de Newton-Raphson ou "Scoring" de Fisher. O segundo caso tem apenas caráter teórico e é solucionado usando o método da máxima verossimilhança. O terceiro caso, o mais encontrado na pratica e presente neste trabalho, pode ser resolvido de duas formas também usando o método da máxima verossimilhança: a estimação conjunta dos parâmetros dos itens e das habilidades dos indivíduos; ou em duas etapas, primeiro a estimação dos parâmetros dos itens e, em seguida, a estimação das habilidades. A TRI nos possibilita analisar individualmente os itens de um teste fornecendo parâmetros referentes aos itens e parâmetros referentes aos indivíduos. Bem interpretados esses parâmetros fornecem uma

grande quantidade de informação necessária para a análise dos itens, identificando itens por sua dificuldade e por seu poder de discriminação entre os indivíduos de maior ou menor nível de habilidade. Informações mais detalhadas sobre a estimação dos parâmetros podem ser encontradas em Baker (2001), Lord (1952), Birnbaum (1968).

#### **2.4. Construção e interpretação de escalas de habilidade**

Uma escala é constituída por uma sequência numérica. O número em uma escala esta sempre associado a uma interpretação. Na escala utilizada pelo SAEB, por exemplo, cada intervalo numérico representa um nível de desempenho de grupos de estudantes e vem acompanhado de uma interpretação das competências e habilidades que eles já construíram no seu processo de desenvolvimento. A escala é definida por níveis âncora, que por sua vez são caracterizados por um conjunto de itens denominados itens âncora. Níveis âncora são pontos selecionados pelo analista na escala da habilidade para serem interpretados pedagogicamente. Já os itens âncora são itens selecionados para cada um dos níveis âncoras segundo um critério de definição, ou seja, sejam dois níveis âncora consecutivos  $Y$  e  $Z$  com  $Y < Z$ . Um determinado item é âncora para o nível  $Z$  somente se as 3 condições abaixo forem satisfeitas simultaneamente (Andrade et al., 2000):

- $P(U = 1 | \theta = Z) \geq 0,65$
- $P(U = 1 | \theta = Y) < 0,50$
- $P(U = 1 | \theta = Z) - P(U = 1 | \theta = Y) \geq 0,30$

Dessa forma, para um item ser âncora em um determinado nível âncora da escala, ele precisa ser respondido corretamente por uma grande proporção de indivíduos (pelo menos 65%) com este nível de habilidade e por uma proporção menor de indivíduos (no máximo 50%) com o nível de habilidade imediatamente anterior. Além disso, a diferença entre a proporção de indivíduos com esses níveis de habilidade que acertam a esse item deve ser de pelo menos 30%. Assim, itens-âncora são itens que caracterizam um ponto ou nível das escalas para o qual a grande maioria dos alunos situados naquele nível acerta o item, ao passo que um percentual considerável de alunos situados ao nível abaixo da escala erra o item. Esses itens têm o objetivo de discriminar pontos na escala que separam alunos que construíram daqueles que não construíram determinadas competências ou habilidades. Para auxiliar na interpretação das escalas, é preciso identificar um número razoável de itens âncora para cada um dos níveis da escala. É importante que o conjunto de itens-âncora cubra a maior extensão possível da matriz de referência para que a análise das competências e habilidades seja rica.

### **3. Metodologia**

#### **3.1. Participantes**

Alunos que realizaram a prova do ENADE de 2009 que foi respondida por 230.486 participantes, entre ingressantes e concluintes do Curso de Administração de diversas instituições de ensino superior do país. A idade dos universitários variou entre 16 e 79 anos tendo uma predominância do sexo feminino de (55%).

#### **3.2. Instrumento Avaliativo**

A prova do ENADE do curso de Administração é composta por 35 questões de múltipla escolha e 5 questões discursivas, onde 10 questões são de formação geral e 30 são de formação específica da área. A tabela 1 descreve de forma exata a divisão das questões bem como os pesos atribuídos a cada área:



Tabela 1: Estrutura da prova do ENADE

Partes	No das questões	Peso/questões	Peso/componentes
Formação Geral/Múltipla. Escolha	01 a 08	60%	
Formação. Geral/Discursiva.	09 e 10	40%	25%
Comp. Específica./ Múltipla Escolha	11 a 37	85%	
Comp. Específica./ Discursiva.	38 a 40	15%	75%

### 3.3. Compilação dos dados

Como visto anteriormente a prova do ENADE é composta por 40 questões ao todo. O presente trabalho contempla a análise somente das 35 questões objetivas. Foi realizado uma estimativa previa na qual foram excluídas 8 questões, pois os valores dos seus parâmetros eram inadequados para o estudo, podendo causar possíveis distorções nos resultados. O tempo de trabalho computacional acabou sendo demorado devido ao tamanho do banco de dados.

### 3.4. Recursos Computacionais

Tanto para a estimação dos parâmetros dos itens quanto para a estimação dos traços latentes dos indivíduos utilizou-se o software estatístico livre R (R DEVELOPMENT CORE TEAM, 2012) juntamente com o pacote *irt* (PARTCHEV, 2014). A construção dos gráficos e tabelas também contaram com o Software R como ferramenta.

## 4. Resultados

As figuras 2(a) e 2(b) nos mostram a distribuição da quantidade de acertos que os candidatos conseguiram dentre as 27 questões selecionadas e uma simples caracterização da amostra respectivamente.

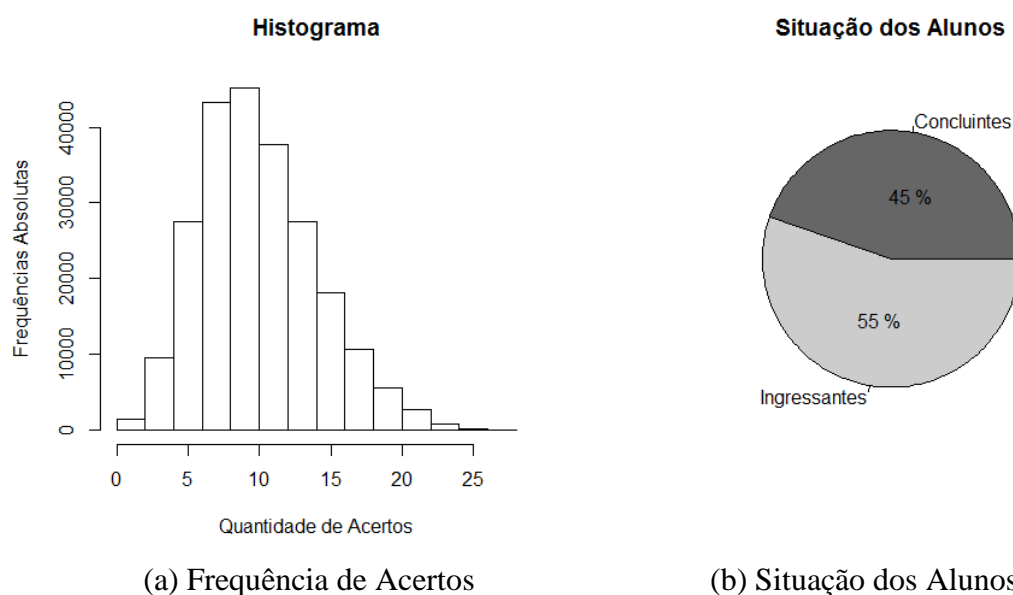


Figura 2: Caracterização da Amostra

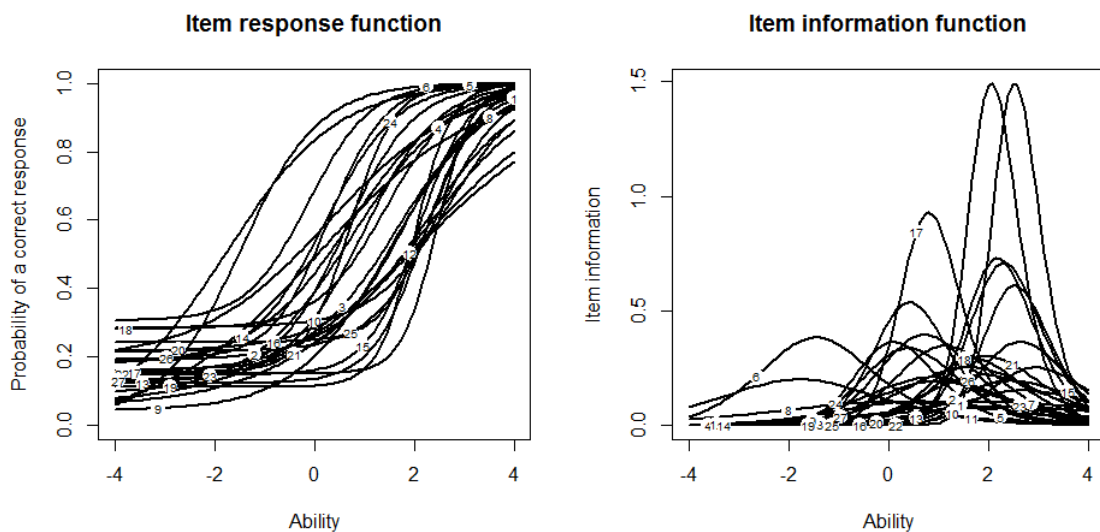
O processo de estimação dos parâmetros é denominado calibração. Na tabela 2 podemos visualizar as estimativas dos parâmetros dos itens ( $a$ ,  $b$  e  $c$ ) para cada item que compõe a prova. Esses parâmetros foram obtidos através das respostas dos participantes a uma das cinco alternativas de cada questão. Indicadas as respostas corretas, os dados foram transformados em itens do tipo certo/errado (itens dicotômicos) e estimados com o auxílio do software estatístico livre R (R DEVELOPMENT CORE TEAM, 2012) por meio do pacote IRTOYS. O

parâmetro  $a$  indica o quanto determinado item consegue discriminar quem construiu de quem não construiu determinadas competências, ou seja quanto maior o valor de  $a$  melhor o item irá discriminar. Pela tabela 2 podemos verificar que os 3 itens com maior valor de  $a$  em ordem decrescente foram os itens Q32, Q28 e Q25. Já os itens com menor poder de discriminação foram Q19, Q13 e Q11 em ordem decrescente, ou seja, esses itens não conseguem discriminar quem construiu de quem não construiu determinada competência. O parâmetro  $b$  determina o grau de dificuldade do item e tem seus valores entre -3 e 3, na escala (0, 1), isto é, com media igual a zero e desvio padrão igual a um. Analisando a tabela 2 percebemos que os 3 itens com maior grau de dificuldade em ordem crescente foram Q32, Q35 e Q15, por possuírem valores maiores do que 2 para o parâmetro  $b$ .

Tabela 2: Estimativas dos parâmetros dos itens

	a	b	c		a	b	c		a	b	c
Q3	0,78	0,28	0,19	Q15	1,32	2,69	0,29	Q28	2,72	2,00	0,11
Q4	1,05	2,22	0,19	Q18	0,90	-1,78	0,01	Q30	0,96	2,40	0,22
Q5	1,09	1,67	0,16	Q19	0,60	2,10	0,05	Q31	1,28	1,80	0,16
Q6	1,00	0,71	0,09	Q21	1,89	2,21	0,12	Q32	2,83	2,45	0,15
Q7	1,62	-0,12	0,31	Q22	1,04	0,70	0,18	Q33	1,42	0,59	0,13
Q8	1,27	-1,48	0,02	Q23	1,97	2,08	0,15	Q34	1,39	0,21	0,18
Q11	0,67	2,05	0,05	Q24	1,97	2,37	0,24	Q35	1,49	2,45	0,22
Q13	0,64	0,07	0,00	Q25	2,24	0,71	0,16	Q36	1,77	0,27	0,19
Q14	1,05	1,53	0,04	Q27	1,43	1,45	0,28	Q37	1,34	0,98	0,13

A figura 3(a) trás as curvas características dos itens da prova do ENADE, percebe-se que existe uma grande variação no que diz respeito a posição do ponto de inflexão de cada curva, ponto este que esta diretamente relacionado ao parâmetro  $b$ , de dificuldade de cada item. Observa-se também a existência de variações na inclinação que cada curva apresenta, inclinação essa que corresponde ao valor que o parâmetro  $a$  assume, ou seja, parâmetro de discriminação. Já na figura 3(b) temos as funções de informação de todos os itens que compõem a prova, visualmente e perceptível que itens que possuem maior valor de informação estão localizados no intervalo onde a habilidade assume valores entre 0 e 3.



(a) Curva Característica dos Itens

(b) Função de informação dos itens

Figura 3: Curva e Função dos Itens

As figuras 4(a) e 4(b) apresentam a curva característica dos itens Q32 e Q13 respectivamente. Estes itens proporcionam uma boa visualização gráfica do quanto a prova apresentou questões heterogêneas com relação a dificuldade e discriminação. Através da análise visual da figura 4(a) percebemos que a questão 32 possui um valor elevado para a caracterizando um item de boa discriminação e por isso sua inclinação vertical no gráfico é mais forte que na figura 4(b). A questão 32 também possui um valor alto para o parâmetro  $b$ , e por isso seu deslocamento do ponto de inflexão da curva característica para a direita ao contrario da questão 13 que tem o ponto de inflexão próximo a zero.

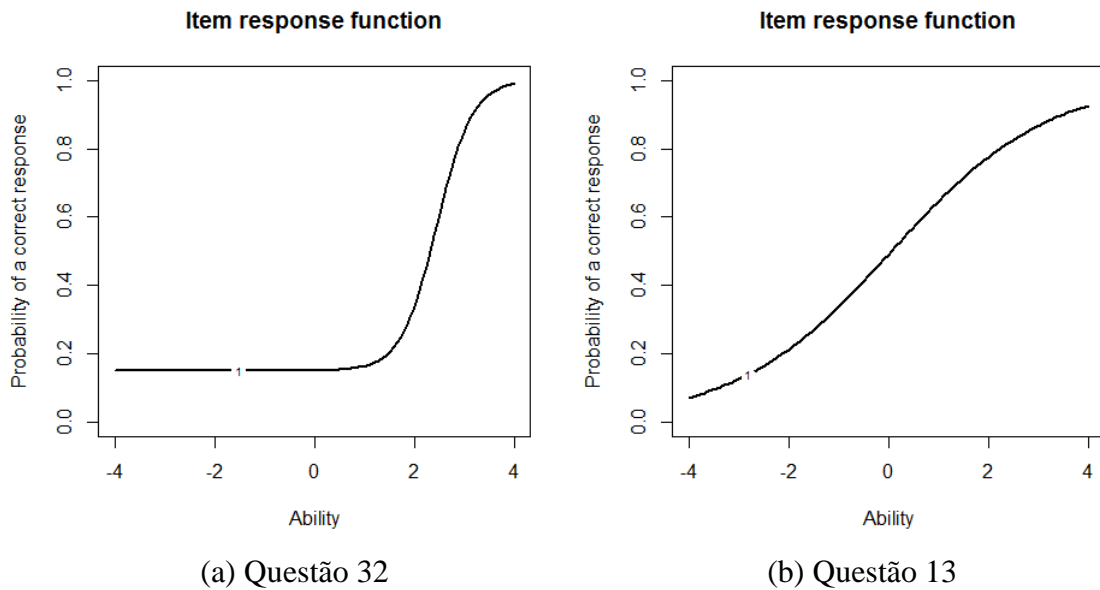


Figura 4: Curva característica dos itens

A figura 5 exibe a Função de Informação do Teste. Este gráfico mostra que a prova do ENADE de 2009 apresentou uma dificuldade alta, pois somente indivíduos que possuem habilidade igual ou superior a 2 tem uma grande probabilidade de ter um bom desempenho ou seja o instrumento é mais adequado para medir o traço latente de indivíduos que tem habilidade próxima a 2.

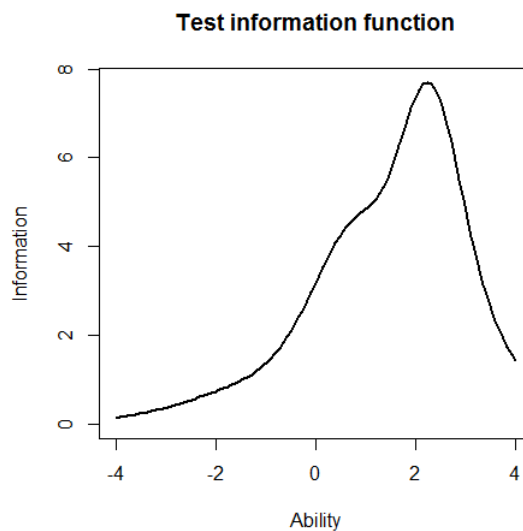


Figura 5: Função de informação do teste

Depois de terminada a fase de calibração dos parâmetros dos itens, é feita a estimação das habilidades dos respondentes. O pacote IRTOYS oferece três métodos de estimação: por máxima verossimilhança, por esperança a posteriori (EAP) e por máximo a posteriori (MAP), o qual foi escolhido para essa estimação.

Na figura 6 temos o histograma das habilidades de todos os participantes da prova do ENADE de 2009, percebe-se através da análise visual que uma grande maioria dos participantes tem habilidades em torno de zero, diferentemente da função informação do teste que consegue mensurar com maior precisão indivíduos com habilidades em torno de 2.

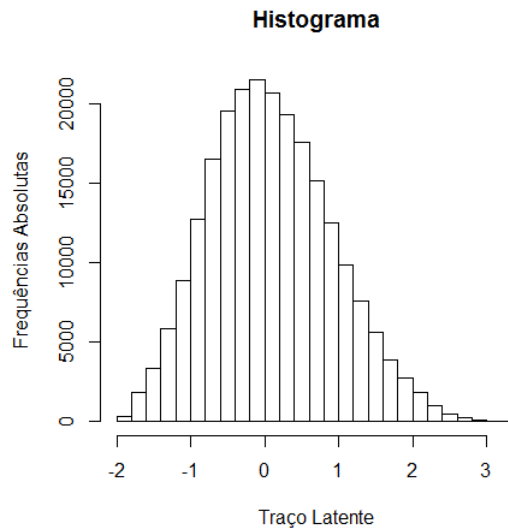
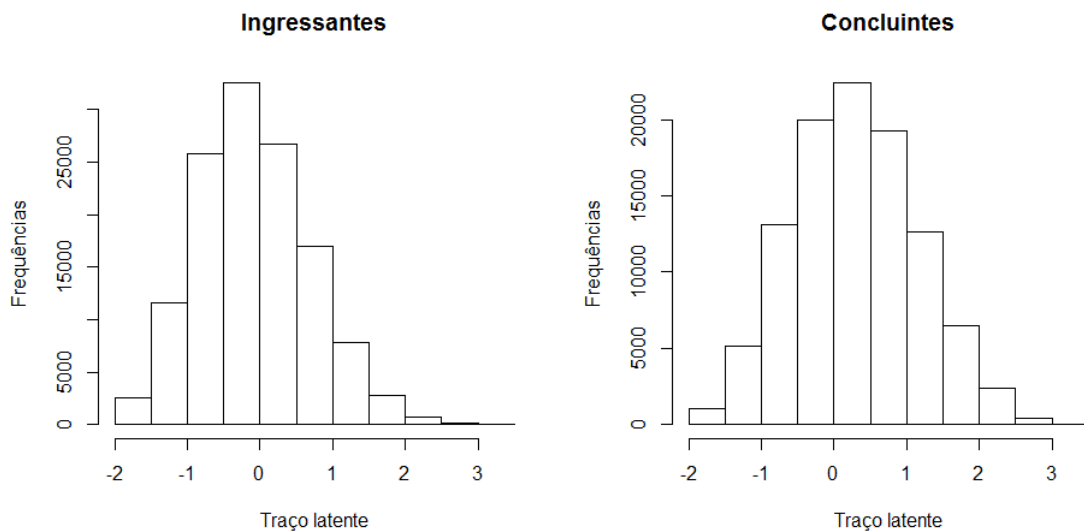


Figura 6: Traço Latente dos Participantes

A figura 7(a) temos o histograma das habilidades dos indivíduos ingressantes no curso de administração. Visualmente percebe-se que pelo menos metade dos alunos tem habilidades menores do que zero. Já na figura 7(b), onde temos o histograma de habilidades dos indivíduos concluintes do curso, percebe-se através da análise visual que pelo menos metade dos alunos tem habilidade acima de zero.



(a) Habilidades estimadas ingressantes      (b) Habilidades estimadas concluintes

Figura 7: Habilidades Estimadas dos Respondentes

Através do gráfico de Box-Plot ( figura 8) essa análise visual fica mais evidente e clara entre os indivíduos ingressantes e concluintes. Esse resultado mostra, que apesar da prova não ser tão precisa para mensurar o traço latente de indivíduos que possuem habilidades estimadas menores do que 0, ela consegue mostrar que há um ganho em termos de traço latente entre ingressantes e concluintes, ou seja, o aluno ao sair da universidade construiu habilidades e conhecimento.

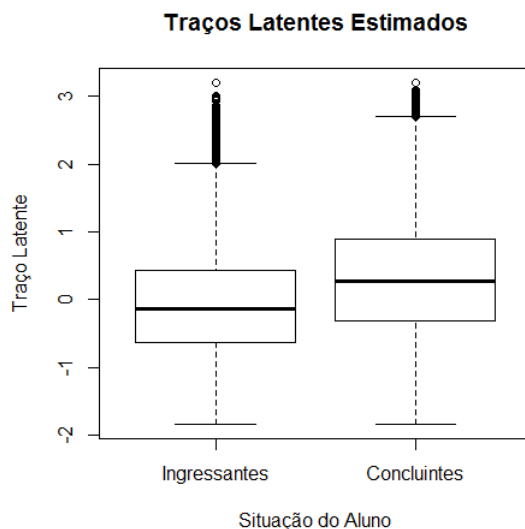


Figura 8: Box-Plot das Habilidades

Também é possível perceber essa diferença quando comparamos os valores médios do traço latente entre ingressantes e concluintes, que são de -0,081 e 0,30 respectivamente.

A obtenção dos itens âncoras, ou seja itens que conseguem avaliar de forma mais precisa indivíduos que construíram determinada competência dos que não construíram, três níveis foram encontrados 1, 2 e 3, onde as questões ficaram distribuídas nos níveis conforme abaixo:

- nível 1: Q7, Q34 e Q36
- nível 2: Q27
- nível 3: Q21, Q23, Q24, Q28, Q31, Q32 e Q35

Percebemos que nenhum item âncora foi colocado no nível 0, -1, -2 e -3, o que é aceitável e torna ainda mais verdadeiro o fato da prova conseguir mensurar de forma mais precisa, indivíduos que tenham habilidades nos níveis citados anteriormente, maiores do que zero.

## 5. Conclusão

De maneira geral, este estudo teve como objetivo mostrar os benefícios do estudo da TRI em uma avaliação educacional, através da verificação do quanto a prova do ENADE vem mensurando o verdadeiro traço latente, se a prova é adequada para o publico alvo, identificar a existência de itens problemáticos e comparar o desempenho entre ingressantes e concluintes.

O resultado das análises da Prova do ENADE do curso de Administração do ano de 2009 de varias instituições de ensino superior do país, mostra que é viável a utilização da TRI como instrumento avaliativo. Apesar da prova apresentar itens problemáticos, ou seja, itens que não tem uma boa calibração, apos a eliminação desses itens do banco de dados, não ocorreram problemas que impedissem a estimação de nenhum parâmetro ou informação de algum item. A prova consegue mensurar de maneira mais precisa, somente indivíduos que tem habilidades superior a zero, isso não quer dizer que ela não consiga mensurar as habilidades de indivíduos que possuem valores abaixo de 0, apenas o traço latente estimado terá um erro padrão maior para esses indivíduos. O resultado mais interessante fica por conta do comparativo entre

ingressantes e concluintes. Houve um ganho em termos de traço latente entre esses dois grupos de alunos, ficou claro que os concluintes ao término do período acadêmico, tiveram em média um traço latente superior aos ingressantes e construíram de certa forma habilidades acadêmicas durante o período letivo. Esse tipo de comparação é uma das grandes vantagens da TRI com relação a TCT, a qual não permite verificar o ganho de traço latente. A análise da prova através da TRI possibilitou também mostrar que os itens que compõem o teste possuem níveis de dificuldade elevado, não sendo muito adequada para avaliar indivíduos que possuem conhecimento abaixo de 0.

## REFERÊNCIAS

- ANDRADE, D. F., TAVARES, H. R., VALLE, R. C. **Teoria da resposta ao item: conceitos e aplicações**. São Paulo, 2000.
- ANDRICH, D. A rating formulation for ordered response categories. **Psychometrika**, 43., 561-573, 1978.
- BAKER, F. B. **The Basics of Item Response Theory**. 2 ed. USA, 2001.
- BIRNBAUM, A. **Some Latent Trait Models and Their Use in Inferring an Examinee's Ability**, 1968.
- BOCK, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. **Psychometrika**, 37., 29-51, 1972.
- BOCK, R. D., ZIMOWSKI, M. F. **Multiple Group IRT**. In: Handbook of Modern Item Response Theory. New York: Springer-Verlag, 1997.
- EMBRETSON, S. E., REISE, S. P. **Item Response Theory for Psychologists**. New Jersey, USA: Lawrence Erlbaum Associates, 2000.
- FRANCISCO, R. **Aplicação da teoria da resposta ao item (TRI) no exame nacional de cursos (ENC) da Unicentro**. Dissertação de Mestrado, Universidade Federal do Paraná, Curitiba, 2005.
- HAMBLETON, R. K., SWAMINATHAN, H., ROGERS, H. J. **Fundamentals of item response theory**. Newbury Park, CA: Sage, 1991.
- KARINO, C. A., ANDRADE, D. F. (2010). **Entenda a teoria de respostas ao item (tri), utilizada no enem**. Disponível em: [http://www.senado.gov.br/comissoes/CE/AP/AP20101116\\_NotaTecnica\\_INEPMEC\\_TeoriaRespostasAoItem.pdf](http://www.senado.gov.br/comissoes/CE/AP/AP20101116_NotaTecnica_INEPMEC_TeoriaRespostasAoItem.pdf). Acesso em: 01/06/2014.
- KLEIN, R. Uma re-análise dos resultados do PISA: problemas de comparabilidade. **Ensaio: aval.pol.públ.Educ**. vol.19 no.73 Rio de Janeiro Oct./Dec. 2011.
- KUDER, G. F., RICHARDSON, M. W. **The theory of estimation of test reliability**, 1973.
- LORD, F. M. **A theory of test scores** (Nº. 7). Psychometric Monograph, 1952.
- MASTERS, G. N. A rasch model for partial credit scoring. **Psychometrika**, 47., 149-174, 1982.
- MURAKI, E. A generalized partial credit model: Application of an em algorithm. **Applied Psychological Measurement**, 16, 159-176, 1992.
- NOGUEIRA, S. O. **ENADE: Análise de itens de formação geral e de estatística pela TRI**. Dissertação de Mestrado, Universidade São Francisco, Itatiba, 2008.
- OLEA, J., PONSODA, V., REVUELTA, J., BELCHI, J. Propiedades psicométricas de un test adaptativo informatizado de vocabulario inglés. **Estudios de Psicología**, n. 55, p. 61-73, 1996.
- OLIVEIRA, K. S. **Avaliação do exame nacional de desempenho do estudante pela teoria de resposta ao item**. Dissertação de Mestrado, Universidade São Francisco, Itatiba, 2006.
- PARTCHEV, I. **Simple interface to the estimation and plotting of irt models**, 2014.
- 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, 15, n. 3, p. 379-393, 2010.

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

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

SAMEJIMA, F. A. **Estimation of latent ability using a response pattern of graded scores.** Psychometric Monograph, N. 17. 1969.

SCHWARTZMAN, S. O "conceito preliminar" e as boas praticas de avaliação do ensino superior. **Revista da Associação Brasileira de Mantenedoras de Ensino Superior**, n. 38, pp. 9-32, 2008.

SAO PAULO, E., MIRANDA, B. S., MOREIRA NETO, J. G., PAIXAO, L. A. R. Aplicação do modelo de credito parcial generalizado na avaliação do projeto sesi por um brasil alfabetizado. **Revista Eletronica Iberoamericana sobre Calidad, Eficacia y Cambio en Educacion**, v. 5, n. 2e, p. 24-38, 2007.

VENDRAMINI, C. M. M., SILVA, M. C., CANEL, M. Analise de itens de uma prova de raciocínio estatístico. **Psicologia em Estudo**, Maringá , v. 9, n. 3, p. 487-498. (set./dez 2004).

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