Teoria da Resposta ao Item (TRI)
Modelos Multidimensionais

RONALD TARGINO NOJOSA
Professor substituto do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará; colaborador no Laboratório de Estatística e Medidas Educacionais (LEME) da Universidade Federal do Ceará.
Mestre em Estatística Aplicada, sub-área Avaliação Educacional
rtarginon@bol.com.br

Resumo
Os modelos de respostas ao item utilizados, atualmente, limitam-se a avaliações que medem um traço latente (habilidade) que pode ser representado por uma única dimensão (avaliações unidimensionais). Algumas avaliações, seja pela construção dos itens ou pela própria finalidade da aplicação, não podem, a princípio, ser consideradas unidimensionais. Este é o caso do Exame Nacional do Ensino Médio - ENEM, que foi estruturado segundo uma matriz de 5 competências. Este trabalho apresenta, primeiramente, conceitos básicos da TRI e introduz técnicas destinadas à determinação da dimensionalidade de um conjunto de dados. A Análise Fatorial de Informação Plena está entre essas técnicas e é proposta como ferramenta para a verificação da dimensionalidade das provas do ENEM. Os resultados desse estudo mostraram que a habilidade medida pelo ENEM do ano de 1999 pode ser representada por pelo menos 5 dimensões. O trabalho apresenta, ainda, modelos que tornam possível a modelagem de dados em avaliações multidimensionais.
Palavras-chave: Teoria da Resposta ao Item Multidimensional, Dimensionalidade, Análise Fatorial, Teoria da Resposta ao Item.

Resumen
Actualmente, los modelos de respuesta al ítem utilizados se limitan a evaluaciones que miden un trazo latente (habilidad) que puede ser representado por una única dimensión, evaluaciones unidimensionales. Algunas evaluaciones, sea por la construcción de cada ítem, sea por la propia finalidad de la aplicación, no pueden en primer instancia, ser consideradas unidimensionales. Es el caso del *Exame Nacional do Ensino Médio* – ENEM, estructurado según una matriz de cinco competencias. En primer plano, son presentados en este trabajo conceptos básicos de la TRI y se introducen técnicas destinadas a determinar la dimensionalidad de una base de datos. Para verificar tal dimensionalidad en las pruebas del ENEM, se propone, de entre varias técnicas, la Análisis Factorial de Información Plena. Una de sus grandes ventajas es la de permitir testar la significáncia del aumento en la dimensión de un conjunto de datos. Los resultados de este estudio ostentan que la habilidad medida por el ENEM en el año de 1999 puede ser representada al menos por cinco dimensiones. Incluso en evaluaciones de tipo unidimensionales, esta análisis es muy importante dada la dificultad en la elaboración de cada ítem que conserven la unidimensionalidad. Además, para evaluaciones multidimensionales, el trabajo también expone modelos que posibiliten amoldar sus datos.
Palabras-clave: Teoría de la Respuesta al Ítem Multidimensional, Dimensionalidad, Análisis Factorial, Teoría de la Respuesta al Ítem.

Abstract
The item response models currently applied in most educacional evaluations seek to measure a latent trait (ability) which can be represented by only one dimension (unidimensionality). Some educacional evaluations, however, do not fit into a one-dimensional framework. This is case of the 'Exame Nacional do Ensino Médio' (ENEM), which has been designed according to a matrix of five different skills. At the outset, this article presents a brief overview of the unidimensional Item Response Theory and some tecniques that are helpful in determining the dimensionality of a given set of items. One of these techniques is the Full Information Factorial Analysis, which we propose as a way to check the dimensionality of ENEM. The results suggest that this exam in 1999 is at least fivefold dimensional. We also present models designed to model multidimensional evaluations.
Keywords: Item Response Theory, Multidimensional Item Response Theory, Dimensionality, Factor Analysis.

1. Introdução

Os modelos de resposta ao item atualmente utilizados limitam-se a avaliações que medem um traço latente (habilidade) que pode ser representado por uma única dimensão (avaliações unidimensionais). Algumas avaliações, seja pela construção dos itens ou pela própria finalidade da aplicação, não podem, em princípio, ser consideradas unidimensionais. Este é o caso do Exame Nacional do Ensino Médio (ENEM), que foi estruturado segundo uma matriz de 5 competências.

Este trabalho aborda, inicialmente, conceitos importantes da Teoria da Resposta ao Item (TRI) unidimensional. Em seguida, são apresentadas técnicas destinadas à determinação da dimensionalidade de um conjunto de dados e modelos que tornam possível a modelagem para avaliações multidimensionais. A parte final do trabalho apresenta uma aplicação aos dados do ENEM do ano de 1999.

1.1 Modelos Unidimensionais

Modelos unidimensionais da Teoria da Resposta ao Item vêm sendo aplicados nos últimos anos em importantes avaliações educacionais. Esses modelos têm como suposição fundamental a unidimensionalidade do teste, ou seja, a suposição de que o teste está medindo uma única habilidade (traço latente). As principais diferenças entre os modelos está na forma matemática da Curva Característica do Item e no número de parâmetros do modelo. Esses modelos são expressões matemáticas que fornecem a probabilidade de resposta correta a um item como função da habilidade do indivíduo e dos parâmetros dos itens.

Dentre os modelos unidimensionais propostos na literatura, um dos mais utilizados é o modelo logístico de 3 parâmetros proposto por Birbaum (1957). Esse modelo é empregado em itens de múltipla escolha do tipo certo/errado e é dado por

$$P(X_{ij}=1|\theta_j)=c_i+(1-c_i)\frac{1}{1+e^{-Da_i(\theta_j-b_i)}} \quad (1.1)$$

onde

X_{ij} é uma variável dicotômica que assume os valores 1, quando o *j*-ésimo indivíduo responde corretamente ao item *i*, ou 0 quando o *j*-ésimo indivíduo não responde corretamente ao item *i*, com $i=1, 2, ..., p$ e $j=1, 2, ..., N$;

$P(X_{ij}=1\,|\,\theta_j)$ é a probabilidade de um indivíduo *j* com habilidade θ_j responder corretamente ao item *i*;

θ_j representa a habilidade (traço latente, traço ou proficiência) do *j*-ésimo indivíduo;

a_i é o parâmetro de discriminação do item *i*;

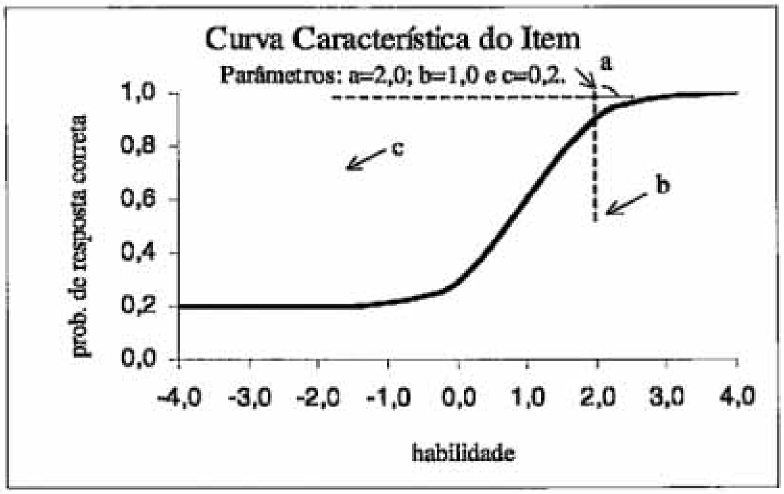
b_i é o parâmetro de dificuldade do item *i*, medido na mesma escala da habilidade;

c_i parâmetro que representa a probabilidade de um indivíduo com baixa habilidade responder corretamente ao item *i* (muitas vezes referido como a probabilidade de resposta correta dada ao acaso);

D é um fator de fator de escala, constante igual a 1.

A Figura 1.1 mostra uma curva característica de item quando a habilidade tem uma única dimensão (unidimensionalidade). Observe que indivíduos com altos valores para a habilidade têm maior probabilidade de responder corretamente ao item do que aqueles com baixa habilidade.

Figura 1.1 – Representação gráfica de uma FCI



Uma abordagem aos diversos modelos unidimensionais pode ser encontrado em Hamblenton & Swaminathan (1985) e Andrade, Tavares & Valle (2000).

1.2 Suposições

Os modelos unidimensionais têm como suposição fundamental a *unidimensionalidade* do teste, ou seja, supõem que o teste mede apenas um único traço latente (habilidade dominante no conjunto de itens). A *independência local ou condicional* fora citada em alguns trabalhos como suposição. Entretanto, com a publicação de Lord (1980), a independência não mais é vista como suposição, mas sim como conseqüência da correta determinação da dimensionalidade dos dados. A independência condicional significa que dada a habilidade do indivíduo, suas respostas aos diferentes itens do teste são independentes. Essa conseqüência é fundamental no processo de estimação dos parâmetros do modelo. A independência condicional pode ser obtida mesmo quando os dados não são unidimensionais, desde que todas as dimensões da habilidade que influenciam a performance no teste sejam consideradas. Por exemplo, considere um teste de matemática no qual os itens exigem níveis altos da habilidade verbal (interpretação de texto). Indivíduos com níveis baixos dessa habilidade podem não responder corretamente aos itens mesmo possuindo níveis altos da habilidade matemática. Dessa forma, a performance no teste é influenciada por ambas as habilidades: matemática e verbal. Se um modelo unidimensional é ajustado aos dados, considerando apenas a habilidade matemática, a independência condicional não pode ser considerada, pois as respostas aos diferentes itens dependerá da habilidade verbal. Por outro lado, se os indivíduos apresentam níveis próximos de habilidade verbal, considera-se que somente a habilidade matemática influencia na performance sobre os itens; assim, a independência condicional pode ser considerada quando um modelo unidimensional é ajustado. Em testes contendo itens que exigem o conhecimento de procedimentos específicos ou estratégicos para se obter a resposta correta ou que fornecem informações que ajudam a responder outros itens, a independência condicional também não se verifica. A habilidade para detectar procedimentos específicos pode ser vista como outra dimensão além da habilidade testada.

Outras duas suposições são: (a) o tempo para a resolução do teste é suficiente para que todos os itens possam ser respondidos por todos os indivíduos e (b) a ordem em que os itens são apresentados aos indivíduos não interfere no desempenho dos mesmos.

1.3 Estimação

Uma vez determinado o modelo da TRI a ser utilizado, é necessário determinar os valores dos parâmetros dos itens e das habilidades dos indivíduos. Nos modelos unidimensionais, cada indivíduo é caracterizado apenas por um parâmetro, θ, e para a caracterização dos itens utilizam-se de 1 a 4 parâmetros, dependendo do modelo utilizado. De modo geral, obtêm-se estimativas para os parâmetros do modelo com erros-padrão pequenos quando o número de itens é de pelo menos 30 e o número de respondentes para cada item é de pelo menos 300. No processo de estimação, defrontamos- nos com 3 casos. O primeiro é quando os parâmetros dos itens são conhecidos e deseja-se apenas estimar as habilidades dos indivíduos. No segundo caso, são conhecidas as habilidades e deseja-se apenas estimar os parâmetros dos itens. No terceiro, nem os parâmetros dos itens e nem as habilidades dos indivíduos são conhecidos; deseja-se estimar ambos. O primeiro caso começa a ser freqüente na prática e a solução é dada empregando o método da máxima verossimilhança, através da aplicação de procedimentos interativos, como, por exemplo, o método de Newton-Raphson. Outra alternativa é a utilização de métodos bayesianos. O segundo caso tem apenas caráter teórico e é solucionado usando o método da máxima verossimilhança. O terceiro caso, provavelmente o mais encontrado na prática, é abordado de duas formas: a estimação conjunta dos parâmetros de itens e das habilidades dos indivíduos; ou em duas etapas, primeiro a estimação dos itens e, em seguida, a das habilidades (Baker, 1992).

1.4 Escala de Medidas

Nos dois primeiros casos citados na subseção anterior, o conhecimento dos parâmetros (dos itens ou dos indivíduos) implica o conhecimento da escala em que eles foram medidos. No terceiro, não há nenhuma escala definida. Como a habilidade pode assumir qualquer valor real, os valores dos parâmetros que maximizam a função de verossimilhança não podem ser determinados de modo único. É necessário, então, estabelecer uma origem, que representará a média das habilidades dos indivíduos na população, e uma unidade de medida, que representará o desvio-padrão das habilidades, antes do processo de estimação. As habilidades dos indivíduos não são observadas, mas, baseado nas respostas dos indivíduos a um conjunto de itens, é possível estimá-las. A propriedade mais importante da habilidade é que ela não depende do particular conjunto de itens ao qual o indivíduo foi submetido. Essa propriedade possibilita uma comparação direta de itens, testes ou performances de diferentes grupos de indivíduos. A habilidade pode assumir teoricamente qualquer valor real, e muitas são as métricas na qual ela pode ser definida. Usualmente, utiliza-se uma escala com média 0 e desvio-padrão 1. Entretanto, poder-se-ia assumir quaisquer outros valores, pois o importante são as relações de ordem existentes entre os pontos da escala. A especificação da escala elimina a não-identificabilidade do modelo que ocorre devido à produção de um mesmo valor para a função característica do item a partir de diferentes conjuntos de parâmetros. Uma discussão e desenvolvimento completos podem ser encontrados em Baker (1992).

1.5 Comentários Adicionais

Os modelos unidimensionais da TRI têm como suposição fundamental a unidimensionalidade do teste. Esta suposição exige que os itens que irão compor o teste sejam elaborados de modo a atendê-la. Entretanto, dependendo do objetivo do teste, nem sempre isso é possível.

A violação da suposição de undimensionalidade conduz, entre outras, às seguintes conseqüências negativas: (i) a própria validade do item passa a ser questionada e (ii) assumindo um modelo unidimensional com itens multidimensionais não se pode garantir a independência condicional.

Hattie (1985) identificou mais de 30 índices sugeridos para determinar a dimensionalidade de uma medida e os categorizou nos seguintes grupos:

Grupo 1: índices baseados em padrões de respostas;
Grupo 2: índices baseados na fidedignidade;
Grupo3: índices baseados na análise de componentes principais;
Grupo 4: índices baseados na análise fatorial;
Grupo 5: índices baseados na TRI.

Pela categorização de Hattie (1985), a análise fatorial através das matrizes dos coeficientes *phi* e das correlações tetracóricas se enquadram no grupo 4 e a análise fatorial de informação plena no grupo 5. Na seção seguinte serão abordados esses dois grupos, com referência específica a essas análises.

2. Análise Fatorial

A Análise Fatorial trata do relacionamento interno de um conjunto de variáveis. Suas idéias básicas se devem, principalmente, a psicólogos como Charles Spearman, Thomson, Thurstone e Burt, que buscavam obter uma melhor compreensão para a "inteligência".

Os testes de inteligência eram - e ainda são - montados com uma grande variedade de itens que variam em graus de memorização, habilidade verbal, habilidade matemática, entre outras. A Análise Fatorial foi desenvolvida para analisar esses testes e averiguar se a "inteligência" era determinada por um único fator geral ou por vários fatores.

Na Análise Fatorial convencional, p variáveis observadas são modeladas como funções lineares de um número menor, m, de outras variáveis contínuas, denominadas variáveis latentes ou fatores. Essas variáveis são empregadas na estimação da correlação entre as variáveis observadas.

Os objetivos principais da Análise Fatorial são:

- determinar o número de fatores que fornecem um ajuste satisfatório à matriz de correlação observada;
- estimar os coeficientes de regressão das variáveis observadas nos fatores.

Com isso, almeja-se uma explicação parcimoniosa do relacionamento entre as variáveis observadas.

Basicamente, o modelo fatorial é motivado pelo seguinte argumento: suponha variáveis que possam ser agrupadas por suas correlações, isto é, suponha que todas as variáveis de um grupo particular de variáveis são fortemente correlacionadas entre si, mas têm, relativamente, baixa correlação com variáveis de outros grupos diferentes. Então, é conceitual que cada grupo de variáveis representa uma única construção básica, ou fator, que é responsável pelas correlações observadas. Por exemplo, correlações de testes de Inglês, Francês e Português sugerem um fator básico: "domínio verbal ou em línguas". Um segundo grupo de variáveis, representando escores em ciências exatas: Matemática e Física, por exemplo, se avaliado, corresponderia a outro fator. É esse tipo de estrutura que a Análise Fatorial procura determinar.

2.1 Modelo Fatorial Convencional

A Análise Fatorial é, em essência, um método para explicação da variabilidade de dados através de um modelo ajustado. A matriz de covariância ou a de correlação de um número razoável de variáveis

$Y=(Y_1,...,Y_p)$ é o objetivo da análise. Por hipótese, o interrelacionamento entre as variáveis pode ser explicado por um modelo de regressão linear múltiplo, com os Ys como variáveis dependentes. A característica que diferencia a Análise Fatorial do modelo de regressão é que nessa os preditores, $\theta = (\theta_1,..., \theta_m)$, não são observados, mas sim inferidos a partir dos dados. No modelo convencional de Análise Fatorial assume-se um vetor m-dimensional de variáveis latentes, $\theta = (\theta_1,..., \theta_m)$, na população de indivíduos. As observações de uma amostra aleatória de N indivíduos não consistem, entretanto, de valores de θ, mas de valores de p variáveis observadas $Y=(Y_1,...,Y_p)$, onde $p > m$. É assumido, também, que $\mathbf{Y}$ depende estocasticamente de θ através do seguinte sistema de equações lineares:

$$
\begin{aligned}
Y_1 &= \lambda_{11}\theta_1 + \lambda_{12}\theta_2 + ... + \lambda_{1m}\theta_m + e_1 \\
Y_2 &= \lambda_{21}\theta_1 + \lambda_{22}\theta_2 + ... + \lambda_{2m}\theta_m + e_2 \\
&\vdots \\
Y_p &= \lambda_{p1}\theta_1 + \lambda_{p2}\theta_2 + ... + \lambda_{pm}\theta_m + e_p
\end{aligned}
$$

ou, em forma matricial,

$$\boldsymbol{Y} = \boldsymbol{\Lambda\theta} + \boldsymbol{e} \qquad (2.1)$$

O coeficiente λ_{ik}, $i=1, 2, ..., p$ e $k=1, 2, ..., m$, é definido como a carga da i-ésima variável no k-ésimo fator, e a matriz Λ é denominada matriz de cargas fatoriais. θ é o vetor de fatores comuns, ou, simplesmente, vetor de fatores e $\boldsymbol{e}$ é o vetor de fatores específicos ou resíduos. Sendo possível a decomposição da matriz de covariância ou a de correlação de $\boldsymbol{Y}$, Σ, da seguinte forma:

$$\Sigma = \Lambda\Theta\Lambda' + \Psi \,, \qquad (2.2)$$

onde Θ denota a matriz de covariância de θ, e Ψ representa a matriz de covariância do resíduo $\boldsymbol{e}$, as cargas fatorias podem ser obtidas de tal modo que sejam fáceis de serem exploradas visualmente ou interpretadas, sem afetar o ajuste do modelo. Esse processo é conhecido como rotação de fatores (Johnson & Wichern, 1998; Lawley & Maxwell, 1971).

2.2 Modelo Fatorial para Dados Dicotômicos ou Dicotomizados

Nesta seção será apresentada a extensão do modelo fatorial tradicional para dados dicotômicos ou dicotomizados. A primeira idéia foi estabelecer uma matriz que substituísse a matriz de covariância ou a de correlação para variáveis contínuas e operar de acordo com o que foi apresentado na subseção anterior. As soluções iniciais foram dadas através do uso das matrizes de coeficientes *phi* e de correlações tetracóricas. Um outro método, denominado Análise Fatorial de Informação Plena, foi introduzido a partir do artigo de Bock & Aitkin (1981) e será abordado na parte final da seção.

2.2.1 O Coeficiente de Correlação Phi

Não há impedimento para o cálculo da correlação produto-momento de Pearson ($\rho(X,Y)$) entre variáveis dicotômicas - "coeficiente *phi*", como ela é chamada neste caso especial. Sua fórmula é dada por

$$\rho(X,Y) = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Dessa forma, seria natural aplicar o método da subseção anterior para ajustar um modelo fatorial a uma matriz de correlação assim obtida. Entretanto, vários pesquisadores têm demonstrado os problemas inerentes a essa prática.

O principal problema é que os valores do coeficiente *phi* dependem não só do grau de relacionamento entre as variáveis, mas, também, da média individual de cada variável.

As Tabelas 2.1, 2.2 e 2.3 ilustram os resultados obtidos numa simulação com diferentes tamanhos de amostras para o coeficiente *phi*. Na Tabela 2.1, 8 itens foram submetidos a 5000 respondentes. Cada item é uma variável que assume os valores 1 e 0: 1 para respostas corretas e 0 para respostas incorretas. O primeiro item foi respondido incorretamente por apenas um dos indivíduos; o segundo item foi respondido incorretamente por dois indivíduos: o que errou o primeiro item e mais um. Note que somente um indivíduo respondeu diferentemente os dois itens. O resultado para o coeficiente de correlação *phi* foi 0,707. O terceiro item foi respondido incorretamente pelos dois indivíduos que erraram o item anterior e mais um; o quarto item, pelos três indivíduos anteriores e mais um. Para os itens 1 e 4, apenas 3 dos 5000 indivíduos apresentam respostas diferentes para estes dois itens, isso fez com que a correlação *phi* decrescesse para 0,50. Concluindo a interpretação dos itens, o quinto foi respondido incorretamente pelos 4 indivíduos anteriores anteriores e mais 496, o sexto item, pelos 500 indivíduos anteriores e mais 500, o sétimo item, pelos 1000 indivíduos anteriores e mais 500 e o oitavo item, pelos 1500 indivíduos anteriores e mais 500.

Tabela 2.1 - Correlação *phi* para 8 itens submetidos a 5000 respondentes

Itens	1	2	3	4	5	6	7	8
2	0,707							
3	0,577	0,816						
4	0,500	0,707	0,866					
5	0,042	0,060	0,074	0,085				
6	0,028	0,040	0,049	0,057	0,667			
7	0,022	0,031	0,037	0,043	0,509	0,763		
8	0,017	0,024	0,030	0,035	0,408	0,612	0,802	1,000

Nas Tabelas 2.2 e 2.3, a mesma análise é feita para outros 8 itens submetidos a 500 e 50 indivíduos, respectivamente. Os itens 1, 2, 3 e 4 apresentam a mesma seqüência de respostas incorretas verificadas na Tabela 2.1, e os itens 5, 6, 7 e 8 apresentam a mesma seqüência e proporção de respostas incorretas. Nota-se facilmente a dependência indesejável do coeficiente de correlação *phi* da média das variáveis envolvidas.

Tabela 2.2 - Correlação *phi* para 8 itens submetidos a 500 respondentes

Itens	1	2	3	4	5	6	7	8
2	0,706							
3	0,576	0,816						
4	0,498	0,706	0,865					
5	0,134	0,190	0,233	0,269				
6	0,090	0,127	0,155	0,180	0,667			
7	0,068	0,097	0,119	0,137	0,509	0,764		
8	0,055	0,078	0,095	0,110	0,408	0,612	0,802	1,000

Tabela 2.3 - Correlação *phi* para 8 itens submetidos a 50 respondentes

Itens	1	2	3	4	5	6	7	8
2	0,700							
3	0,565	0,808						
4	0,484	0,692	0,857					
5	0,429	0,612	0,758	0,885				
6	0,286	0,408	0,505	0,590	0,667			
7	0,218	0,312	0,386	0,450	0,509	0,764		
8	0,175	0,250	0,309	0,361	0,408	0,612	0,802	1,000

Quando variáveis binárias são produzidas pela dicotomização de variáveis contínuas, a escolha do ponto de corte afeta os valores do coeficiente *phi*. A análise fatorial feita a partir da matriz dos coeficientes *phi* de variáveis binárias e produzida pela mesma estrutura de correlação, mas dicotomizadas em diferentes pontos, pode equivaler-se a modelos fatoriais com diferentes estruturas e possibilitar diferentes números de fatores.

Outro problema é que o valor de uma variável dicotômica é limitado, implicando que sua regressão com qualquer variável latente contínua com intervalo infinito não pode ser linear (McDonald & Ahlawat, 1974). Se a correlação for aplicada diretamente em variáveis dicotômicas, o modelo de análise fatorial linear dado pela Equação 2.1 fica mal especificado.

Uma alternativa ao uso da matriz de coeficientes *phi* é apresentada em seguida.

2.2.2 A Correlação Tetracórica

Diferentemente do cálculo das correlações *phi*, onde não houve nenhuma suposição sobre as variáveis observadas envolvidas, consideram-se, agora, as variáveis contínuas Y_i e Y_j que têm distribuição conjunta normal bivariada e que as variáveis observadas, X_i e X_j, que assumem os valores 0 ou 1, são resultantes dessas variáveis através de um processo de dicotomização. Nesse processo, é assumido um ponto de corte para cada uma das variáveis contínuas, denotados por γ_i e γ_j. Os valores das variáveis Y_i e Y_j não são observados diretamente, mas, somente, se assumem valores maiores ou menores do que seus respectivos pontos de corte:

$$Y_i \geq \gamma_i \text{ ou } Y_i < \gamma_i$$
$$Y_j \geq \gamma_j \text{ ou } Y_j < \gamma_j.$$

Os valores das variáveis observadas X_i e X_j são supostos resultantes das variáveis contínuas da seguinte forma:

$$X_i = \begin{cases} 1, & \text{se } Y_i \geq \gamma_i \\ 0, & \text{se } Y_i < \gamma_i \end{cases}$$

A variável X_j recebe tratamento análogo.

A idéia agora é inferir a correlação produto momento entre Y_i e Y_j, $\rho(Y_i,Y_j)$, a partir da observação de X_i e X_j. Como exemplo, poder-se-ia considerar a variável Y denotando a habilidade exigida por um item e a variável X, a respectiva resposta ao item: 1 para resposta correta e 0 para resposta incorreta.

Seja π_{ij} a proporção de indivíduos na população com $Y_i \geq \gamma_i$ e $Y_j \geq \gamma_j$. A proporção de indivíduos com $Y_i \geq \gamma_i$, independentemente de Y_j, e a proporção de indivíduos com $Y_j \geq \gamma_j$, independentemente de Y_i, são dadas por π_i e π_j, respectivamente.

Com as suposições sobre a distribuição de (Y_i, Y_j), temos

$$\pi_{ij} \equiv \int_{\gamma_i}^{\infty} \int_{\gamma_j}^{\infty} f\left(y_i, y_j, \rho\right) dy_i dy_j \quad (2.3)$$

$$\pi_i \equiv \int_{\gamma_i}^{\infty} \int_{-\infty}^{\infty} f\left(y_i, y_j, \rho\right) dy_i dy_j \quad (2.4)$$

$$\pi_j \equiv \int_{-\infty}^{\infty} \int_{\gamma_j}^{\infty} f\left(y_i, y_j, \rho\right) dy_i dy_j \quad (2.5)$$

onde $f(y_i, y_j, \rho)$ representa a função densidade da distribuição normal padrão bivariada, com ρ representando a correlação entre as variáveis Y_i e Y_j:

$$f\left(y_i, y_j, \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left[-\frac{y_i^2 + y_j^2 - 2\rho y_i y_j}{2\left(1-\rho^2\right)}\right].$$

Considerando $\pi_{i\bar{j}}$ a proporção de indivíduos que responderam corretamente o item *i*, mas responderam incorretamente o item *j*, e definindo $\pi_{\bar{i}j}$ e $\pi_{\bar{i}\bar{j}}$ analogamente, todas as proporções de respondentes numa tabela de contigência 2 x 2 estariam determinadas.

Quando γ_i, γ_j e ρ, são conhecidos, π_{ij}, π_i e π_j são determinados de forma única a partir das Equações 2.3, 2.4 e 2.5. Entretanto, o caso realista é ter conhecido π_{ij}, π_i e π_j, e daí determinar γ_i, γ_j e ρ, também de forma única. Quando ρ é calculado a partir das proporções da tabela de contigência 2 x 2, satisfazendo as expressões (2.3), (2.4) e (2.5), ele é chamado coeficiente de correlação tetracórico, que será denotado por ρ_{ij} (Pearson, 1900). Se a suposição de que (Y_i, Y_j) tem distribuição normal bivariada é verdadeira, então $\rho_{ij} = \rho(Y_i, Y_j)$.

Quando se observam proporções amostrais, ρ_{ij} é estimado por r_{ij}, o coeficiente de correlação tetracório amostral. Para o cálculo de ρ_{ij} ou r_{ij} não há fórmulas fechadas; aproximações computacionais eficientes são dadas por Divgi (1979). Calculando ρ_{ij} para todos os pares de variáveis, no nosso caso, itens, forma-se a matriz de correlações tetracóricas.

A matriz de correlações tetracóricas amostral, S^*, é uma estimativa da matriz de correlação das variáveis Y's e, assim, procedimentos padrões de análise fatorial podem ser empregados para estimar Λ e Ψ.

Da mesma forma, do modelo convencional será assumida a existência de m variáveis latentes θ. Para $p > m$ variáveis observadas (por exemplo, p itens de um teste), assume-se, também, que a estrutura linear para as p variáveis Y_i, $i = 1, 2, ..., p$, é dada por

$$Y_i = \lambda_{i1}\theta_1 + \lambda_{i2}\theta_2 + ... + \lambda_{im}\theta_m + e_i, \quad (2.6)$$

onde os e_i's são os resíduos. O contraste com a análise fatorial de variáveis observadas está no fato de que os Y's não são observados diretamente. Observa-se, sim, um vetor de variáveis dicotômicas $X=(X_1, X_2, ..., X_p)$ com valores determinados de acordo com o processo de dicotomização apresentado no início dessa subseção.

Com a análise fatorial através da matriz de correlações tetracóricas melhora-se as estimativas para as correlações, quando comparadas com a estimação através da matriz dos coeficientes *phi*. Uma desvantagem resulta do fato de que, diferentemente do caso de variáveis observadas com distribuição normal, a sumarização das variáveis dicotômicas em termos da matriz de covariância não retém, conjuntamente, toda informação sobre o relacionamento dessas variáveis; somente as informações individuais (percentual de acertos) e a informações de pares de variáveis são usadas.

2.2.3 Análise Fatorial de Informação Plena (AFIP)

Nas subseções anteriores consideramos as estimações dos parâmetros através de métodos cuja soluções são denominadas de soluções de "informações limitadas", no sentido de que somente são utilizadas informações de marginais de ordens inferiores da tabela de contingência que sumariza as respostas dos indivíduos e, portanto, toda informação avaliada para a estimação. A análise fatorial de informação plena introduzida por Bock & Aitkin (1981) não requer o cálculo de coeficientes de correlação inter-itens e supera vários problemas presentes na análise fatorial da matriz tetracórica.

Será assumido novamente o modelo fatorial para variáveis dicotômicas apresentado na subseção anterior, através da expressão (2.6). As suposições são as seguintes:

i. os resíduos e_i seguem uma distribuição normal de média 0 e variância σ_i^2;
ii. os resíduos são independentes entre si e dos θ's;
iii. $\theta = (\theta_1, \theta_2, ..., \theta_m)$, vetor dos fatores, segue uma distribuição normal multivariada com vetor de média 0 e matriz de covariâncias I_m;
iv. $Y = (Y_1, Y_2, ..., Y_p)$, vetor de variáveis pseudo-observadas, seguem uma distribuição normal multivariada com vetor de médias 0 e matriz de covariância Σ, onde $\Sigma = \Lambda\Lambda' + \psi$.

Com o intuito de aproximar a teoria da resposta ao item e a análise fatorial, o leitor pode interpretar a variável Y como habilidade geral do indivíduo e θ como o vetor de habilidades específicas.

Considera-se que a probabilidade de um indivíduo j responder corretamente ao item i, condicionado ao vetor de fatores θ_j, seja dada por

$$\Phi_i(\theta_j) = P(X_{ij}=1|\theta_j) = \frac{1}{\sqrt{2\pi}\sigma_i}\int_{-\infty}^{\gamma_i} \exp\left[-\frac{1}{2}\left(\frac{t-\sum_{k=1}^{m}\lambda_{ik}\theta_{kj}}{\sigma_i}\right)^2\right], \qquad (2.7)$$

ou seja, $\Phi_i(\theta_j)$ é a distribuição acumulada normal padrão.

No processo de estimação da matriz Σ, a matriz das cargas fatoriais, Λ, é responsável pela determinação da dimensão do vetor de fatores θ. Em outras palavras, e de modo geral, as magnitudes das cargas fatoriais indicam quantos fatores devem ser contemplados pelo modelo e quanto da variância amostral de cada variável é devida a cada fator.

Na AFIP, para a estimação das cargas fatoriais, a equação que é empregada para tal fim é função das respostas de todos os N indivíduos submetidos ao teste ou avaliação, isto é, a equação utiliza conjuntamente a informação de todos os itens para cada um dos indivíduos.

Teste de Ajuste do Modelo

Se o tamanho da amostra é suficientemente grande, uma estatística para o teste de adequação do modelo (teste da razão de verossimilhança), assumindo o modelo relacionado à alternativa multinomial, é dado por

$$G^2 = -2\sum_{l=1}^{s} r_l \log(NP_l/r_l).$$

G^2 tem, aproximadamente, distribuição qui-quadrado com $s - p(m+1) + m(m - 1)/2$ graus de liberdade, supondo um modelo com m fatores. Os graus de liberdade refletem o número de padrões de respostas observados, menos o número de parâmetros estimados, mais o número de restrições impostas para efeito de identificação. Entretanto, quando o número de parâmetros é maior que o tamanho da amostra, as freqüências esperadas podem ser próximas de zero, tornando o teste pouco confiável.

Quando na formação da matriz de correlação há dados faltantes ("missing data"), a estatística do qui-quadrado torna-se extremamente sensível, no sentido de superestimação. Esse fenômeno acontece, por exemplo, quando da utilização de delineamentos por Blocos Balanceados Incompletos. Outro fato é que em pesquisas da larga escala abrangendo diferentes localidades, os efeitos grupais podem aumentar os valores da diferença do qui-quadrado entre dois modelos. No manual do programa TESTFACT (Wilson, Wood & Gibbons, 1998) os autores recomendam, nos casos citados acima, dividir a estatística teste (diferença de qui-quadrados) por 2 ou 3 antes de avaliar a significância do teste. Neste trabalho, especificamente na seção 6, será utilizado o número 3 como divisor e a estatística teste será denominada *diferença corrigida de qui-quadrados*.

A estatística teste, G^2, é um índice de falta de ajuste entre o modelo e os dados. Entretanto, seu valor não deve ser interpretado diretamente, somente a diferença dos valores da estatística entre dois modelos é interpretável. De um modo geral, na comparação entre dois modelos: M_1 com m fatores e M_2 com $m + 1$ fatores, se $(G_1^2 - G_2^2)$ for positiva, o segundo modelo se adequa melhor aos dados, caso contrário, o primeiro modelo é preferível.

A dimensionalidade na Teoria da Resposta ao Item é vista como o espaço referente ao vetor θ. Os modelos unidimensionais assumem a unidimensionalidade desse vetor, ou seja, $m = 1$. A análise fatorial de informação plena permite testar a significância estatística dos fatores adicionados sucessivamente ao modelo (teste da razão de verossimilhança), e será utilizada nesse trabalho como método para a verificação da dimensionalidade do espaço do vetor θ.

No programa TESTFACT encontram-se implementadas a análise fatorial de informação plena e a análise fatorial através da matriz de correlações tetracóricas.

3. Modelos Multidimensionais

3.1 Introdução

Modelos unidimensionais da Teoria da Resposta ao Item vêm sendo aplicados nos últimos anos em importantes avaliações educacionais. Esses modelos têm como suposição fundamental a unidimensionalidade do teste, ou seja, assumem que somente um traço latente - chamaremos sempre habilidade - é necessário para explicar a performance de um indivíduo em um teste, ou, em outras palavras, que o teste exige de forma preponderante apenas uma habilidade. Nessa medida, habilidade, incorpora-se a personalidade, o nível de motivação, o tipo de educação, a ansiedade, a facilidade de "trabalhar sob pressão", os conhecimentos específicos e adjacentes relacionados ao teste e todos os outros fatores que possam vir a influenciar a performance do indivíduo no teste. Entre tantos fatores a existência de um dominante, referido como habilidade medida pelo teste, sustenta a suposição de unidimensionalidade.

Um ponto importante é que um teste pode ser unidimensional para uma população e não ser para outra. Supondo a aplicação de um teste, medindo as habilidades matemática e verbal, a duas populações distintas de indivíduos, se em uma delas todos os indivíduos apresentam níveis altos de habilidade verbal, somente a habilidade matemática afetará a performance no teste. Na segunda população, se a habilidade verbal ocorre de maneira heterogênea, ambas as habilidades influenciam significativamente a performance dos indivíduos no teste, e assim, o teste seria bidimensional.

3.2 Modelos Matemáticos

A Teoria da Resposta ao Item Multidimensional (TRIM) é uma metodologia relativamente nova na modelagem do relacionamento entre as habilidades de indivíduos e a respectiva matriz de respostas a um conjunto de itens.

A TRIM caminha buscando o mesmo desenvolvimento da Teoria da Resposta ao Item unidimensional, mas muitas são as lacunas a serem preenchidas. Vários modelos foram propostos e alguns deles vêm sendo testados e aplicados.

Nesse trabalho será dada maior ênfase aos modelos compensatórios. Esses modelos são generalizações diretas dos modelos logísticos unidimensionais que, atualmente, estão entre os mais aplicados.

Os dados a serem tratados pelos modelos estarão representados por uma matriz de zeros e uns, correspondendo, respectivamente, a respostas incorretas e corretas aos itens. A matriz de dados é geralmente composta por N linhas, referentes aos indivíduos, e p colunas, referentes aos itens. Assim, a interseção de uma linha com uma coluna representa a resposta do j-ésimo indivíduo, j=1, 2, ..., N, ao i-ésimo item, i=1, 2, ..., p. As suposições sobre o mecanismo que cria essa matriz de dados são as seguintes: (1) com um acréscimo em pelo menos uma das dimensões (da habilidade) medidas, a probabilidade de obter uma resposta correta a um item é crescente. Isto é freqüentemente chamado de suposição de monotonicidade; (2) a função relativa à probabilidade de resposta correta é suave, no sentido de que as derivadas da função são definidas; (3) a probabilidade da combinação de respostas pode ser determinada a partir do produto das probabilidades de resposta individual, quando as probabilidades são calculadas condicionalmente a um ponto do espaço θ. Essa é uma conseqüência da correta determinação da dimensionalidade de θ - chamada de independência condicional.

Essas suposições são consistentes para vários modelos que relacionam características de indivíduos e de itens.

3.2.1 Modelo Compensatório com Acerto Casual - MC3

A forma básica para o modelo apresentado nesta subseção é uma generalização direta do modelo logístico de 3 parâmetros (Lord, 1980), para o caso onde os indivíduos são representados por um vetor de parâmetros, ao invés de um único escalar. A forma matemática do modelo é dada por

$$P(X_{ij} = 1 \mid \theta_j) = c_i + (1 - c_i)\frac{1}{1 + \exp[-(D\boldsymbol{a}_i'\theta_j + d_i)]} \tag{3.1}$$

onde

X_{ij} é uma variável dicotômica que assume os valores 1, quando o j-ésimo indivíduo responde corretamente ao item i, ou 0 quando o j-ésimo indivíduo não responde corretamente ao item i, com i=1, 2,..., p e j=1, 2,..., N;

$\boldsymbol{\theta}_j = (\theta_{1j}, \theta_{2j},..., \theta_{mj})$ representa o vetor habilidade do j-ésimo indivíduo, sendo θ_{kj}, k=1,2,...,m, suas habilidades específicas;

$P(X_{ij}=1 \mid \theta_j)$ é a probabilidade de um indivíduo j com vetor habilidade θ_j responder corretamente o item i;

$\boldsymbol{a}'_i = (a_{i1}, a_{i2}, ..., a_{im})$ é o vetor de parâmetros relacionados à discriminação do item i;

d_i é o parâmetro relacionado à dificuldade do item i, medido na mesma escala da habilidade;

c_i é o parâmetro que representa a probabilidade de um indivíduo com baixa habilidade responder corretamente ao item i (muitas vezes referido como a probabilidade de resposta correta dada ao acaso);

D é um fator de escala, constante e igual a 1.

A denominação compensatório se deve ao fato de que um valor baixo da habilidade em uma das dimensões pode ser compensado por um valor alto em outra dimensão.

A equação para o modelo define uma superfície que fornece a probabilidade de resposta correta para o item como função da posição dos indivíduos no espaço de habilidade especificado pelo vetor θ. Quando há somente duas dimensões, a forma da superfície de probabilidade pode ser representada graficamente. As Figuras 3.1 e 3.2 mostram a superfície de probabilidade para o mesmo item i (a_{i1} = 0,9; a_{i2} = 1,5; d_i = -2,0; c_i = 0,2) usando dois diferentes métodos de representação. A Figura 3.1 usa uma superfície tridimensional que enfatiza a monotonicidade do incremento natural da superfície. A Figura 3.2 mostra a superfície como um gráfico de linhas de contorno de igual probabilidade de resposta correta. Essa representação enfatiza que as linhas equiprováveis são linhas retas, e, além disso, são paralelas entre si. Essa característica do modelo é resultado da forma linear do expoente do e na equação do modelo.

Parâmetros do modelo

Parâmetros dos indivíduos

Os parâmetros dos indivíduos, habilidades, são os elementos do vetor θ_j. O número de elementos requeridos para modelar adequadamente a matriz de dados é fornecido pela análise fatorial de informação plena introduzida na seção anterior. As dimensões podem não ter interpretação clara, mas são importantes para a especificação do modelo. O procedimento de rotação do espaço do vetor θ possibilita muitas vezes uma melhoria na interpretação das dimensões.

Parâmetro de Discriminação

As medidas clássicas para a discriminação de itens usualmente utilizadas são as correlações ponto-bisserial e bisserial. Estas medidas são usadas de um modo geral como indicadoras da qualidade do item ou para a seleção de itens na composição de um teste. Na maioria das vezes, as medidas de discriminação da TRI e os conceitos relacionados à informação do item são usados com o mesmo intuito, mas estas medidas também são usadas para especificar a precisão da medida fornecida por um item em diferentes níveis da habilidade ao longo da escala θ. Tanto as medidas de discriminação do item da Teoria Clássica quanto a da TRI unidimensional são baseadas na suposição de que o teste está medindo um único traço latente - definido pelo escore total ou pela habilidade θ.

Nos modelos da TRI unidimensional é assumido que a probabilidade de responder corretamente a um item aumenta com o aumento do nível da habilidade que está sendo medida. Da mesma forma, nos modelos multidimensionais é assumido que a probabilidade de resposta correta a um item aumenta quando há um aumento em cada uma das habilidades requeridas pelo teste.

O parâmetro de discriminação de um item i é dado por uma função dos elementos a_{ik}, $k = 1, 2,..., m$ e $i = 1, 2, ..., p$, do vetor $\boldsymbol{a}_i$. Esses elementos podem ser interpretados, muitas vezes, como k parâmetros de modelos unidimensionais (Lord, 1980). Os elementos do vetor $\boldsymbol{a}_i$ estão relacionados com a inclinação da superfície de resposta do item na direção dos correspondentes eixos no espaço $\boldsymbol{\theta}$. Os elementos, portanto, indicam a sensibilidade do item para diferenciar habilidades ao longo dos eixos de $\boldsymbol{\theta}$. No entanto, o poder de discriminação dos diferentes itens dependem da direção estabelecida no espaço de $\boldsymbol{\theta}$.

Figura 3.1 – Superfície de Resposta de um item de parâmetros

$a_1 = 0{,}9$; $a_2 = 1{,}5$; $d = -2{,}0$ e $c = 0{,}2$.

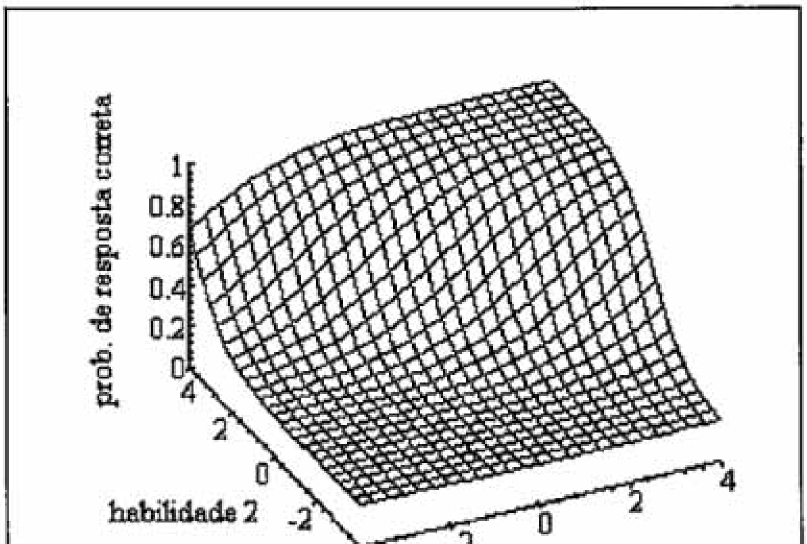
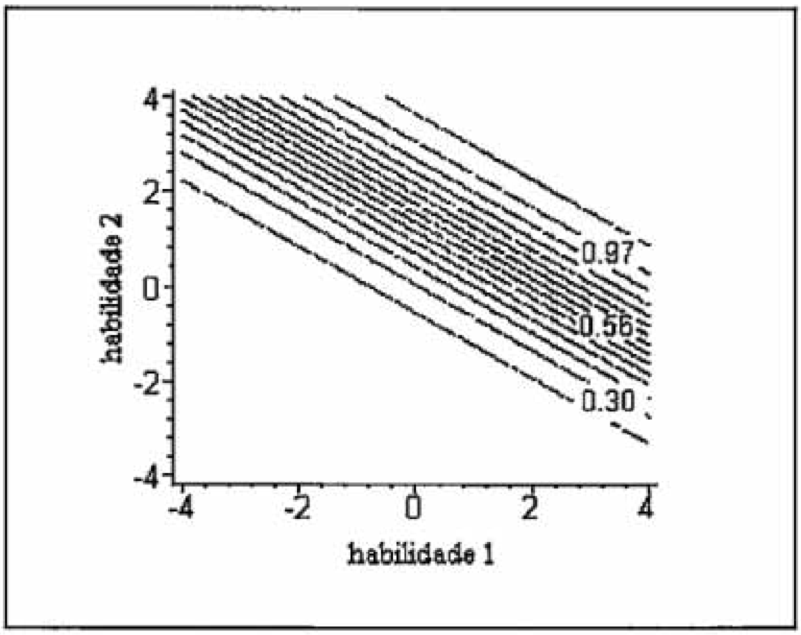


Figura 3.2 – Gráfico de Contorno da SRI dada na Figura 3.1.



O poder de discriminação de um item i multidimensional é dado por

$$DISCM_i = \sqrt{\sum_{k=1}^{m} a_{ik}^2},$$

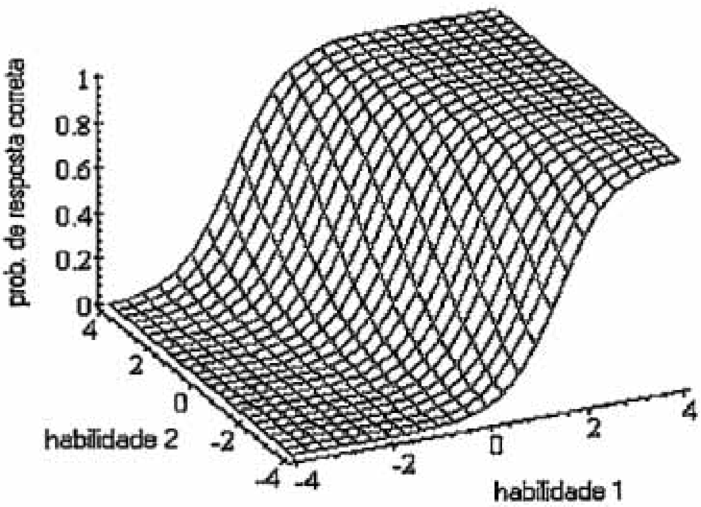
onde k é o número de dimensões do vetor θ e a_{ik} é o k-ésimo elemento do vetor a_i (Reckase, 1991).

De modo geral, o poder de discriminação de um item indica quão rápida é a transição de baixa para alta probabilidade de resposta correta a um item. Um item com alta discriminação divide claramente a região espacial em duas partes, tendo apenas uma estreita região onde as probabilidades são de magnitude intermediária.

A Figura 3.3 mostra a SRI para dois itens, o primeiro com discriminação moderada e o outro com baixa discriminação. Note que os dois itens não discriminam na mesma direção no espaço θ.

Figura 3.3 - Superfícies de Resposta de dois itens que variam na discriminação e na dimensão avaliada

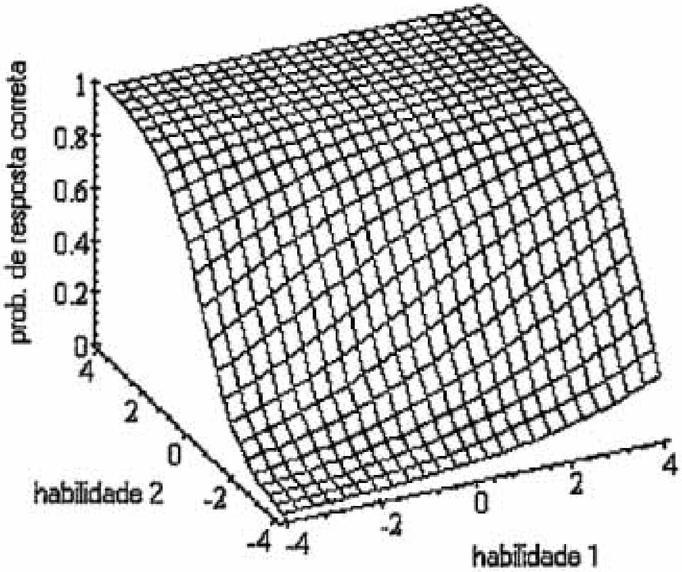
Item 1: $a_{11} = 1,9$, $a_{12} = 0,7$, $d_1 = c_1 = 0$ e $DISCM_1 = 2,02$.



A discriminação de um item está relacionada com a inclinação da superfície de resposta ao item. A inclinação difere dependendo da localização no espaço θ e da direção relativa à superfície naquela localização. Por exemplo, para o item 1 na Figura 3.3 a inclinação é quase nula no ponto (-2, -2) em qualquer direção; no ponto (2, -3) a inclinação é íngreme na direção aproximadamente paralela ao eixo θ_1 (habilidade 1), e é muito baixa, ou quase nula, na direção que passa através da superfície. O nível máximo de discriminação do item está na localização e na direção onde a inclinação é mais íngreme. O poder de discriminação de um item também pode ser descrito relativo a uma particular direção no espaço θ. Usando esta medida, dois itens podem ser comparados diretamente para determinar qual apresenta melhor medida para uma particular habilidade.

Na Figura 3.3, por exemplo, o item 2 é melhor que o item 1 para diferenciar dois indivíduos que possuem habilidades nos pontos ($\theta_{11} = 1$, $\theta_{21} = 0$) e ($\theta_{12} = 1$, $\theta_{22} = 1$) no espaço θ. Note que estes dois indivíduos somente diferem na habilidade 2 (θ_2: $\theta_{21} \neq \theta_{22}$). Assim, o item 2 apresenta mais informação para medir θ_2 na região especificada no espaço θ. De fato, o item 2 apresenta uma mudança maior nas probabilidades para os dois indivíduos do que o item 1, considerando a direção que é paralela a θ_2.

Item 2: $a_{21}=0,4$, $a_{22}=1,2$, $d_2=2$, $c_2=0$ e $DISCM_2=1,26$.



Parâmetro de dificuldade

Estatísticas que descrevem características de itens são comumente empregadas no processo de construção de testes. Estas estatísticas são freqüentementes usadas para produzir formas equivalentes de testes ou para produzir testes com características específicas. De modo geral, essas estatísticas assumem que o item está medindo uma única habilidade. Entretanto, os itens são, geralmente, multidimensionais em algum sentido e, dependendo da intensidade das dimensões, as estatísticas unidimensionais não são apropriadas. Alguns itens medem ou exigem de forma mais dominante uma só habilidade. Para estes itens as estatísticas unidimensionais são razoáveis. Por outro lado, itens que requerem claramente mais de uma habilidade necessitam de um tratamento diferenciado, ou seja, necessitam de medidas que levem em consideração as diferentes dimensões da habilidade.

Os problemas de matemática que envolvem um texto a ser interpretado são exemplos comuns desses tipos de itens. Ambas habilidade verbal (interpretação de texto) e habilidade matemática são necessárias ou exigidas para se obter a resposta correta. Medidas unidimensionais são inapropriadas para esse tipo de item. Como exemplo, poder-se-ia considerar o interesse em ordenar um conjunto de itens desse tipo segundo o parâmetro de dificuldade unidimensional. A ordenação alternar-se-ia dependendo de quão intensas fossem as habilidades verbal ou matemática da amostra de indivíduos.

Na Teoria da Resposta ao Item unidimensional o parâmetro de dificuldade b representa a habilidade necessária para uma probabilidade de acerto ao item igual a $(1 + c)/2$, no modelo de 3 parâmetros e, obviamente, $1/2$ no modelo de 2 parâmetros. O parâmetro b é indicado pelo ponto na escala θ onde a inclinação da curva característica do item é máxima. Esse ponto coincide com o ponto de inflexão da curva.

No caso multidimensional, o parâmetro d_i está relacionado à dificuldade do item. No entanto, o valor desse parâmetro não pode ser interpretado da mesma forma como no modelo unidimensional correspondente (modelo logístico de 3 parâmetros). O parâmetro d_i corresponde ao termo $-a_i b_i$ no modelo unidimensional (veja Equação 1.1). Um valor que é equivalente em interpretação ao parâmetro de dificuldade unidimensional, b_i, é dado por

$$DIFICM_i = \frac{-d_i}{DISCM_i}$$

$DIFICM_i$ indica a distância da origem do espaço θ ao ponto de maior inclinação na direção de máxima inclinação (Reckase, 1985). Esse é um significado análogo ao parâmetro b_i do modelo unidimensional.

Parâmetro de Acerto Casual

O parâmetro de acerto casual, c_i, tem o mesmo significado dado no modelo logístico unidimensional de 3 parâmetros. O valor desse parâmetro indica a probabilidade de resposta correta ao item por indivíduos com baixa habilidade em todas as dimensões.

Escala de Habilidades

De forma análoga ao caso unidimensional, faz-se necessário estabelecer uma métrica na qual serão interpretados os parâmetros do modelo. Neste trabalho, considera-se que as dimensões do vetor habilidade são ortogonais e que para cada uma delas o valor médio das respectivas habilidades é igual a 0 e o desvio-padrão é igual a 1. No caso multidimensional também não faz diferença alguma quais valores são estabelecidos para as medidas de posição e dispersão da habilidade; o importante são as relações de ordem existentes entre seus pontos. Os parâmetros multidimensionais de dificuldade e de (*DIFICM*) discriminação (*DISCM*) também estão na métrica (0;1), ou seja, possuem média 0 e desvio-padrão 1.

Os valores de *DIFICM* e *DISCM* são interpretados da mesma forma que os respectivos parâmetros no caso unidimensional. Dessa forma, esperam-se valores entre 0 e 3,0 para *DISCM* e valores entre -3,0 e 3,0 para *DIFICM*. O parâmetro de acerto casual depende sempre do número de alternativas de resposta para o item. Por exemplo, para um item com 5 alternativas, valores plausíveis estariam no intervalo (0,1;0,3), independente da escala utilizada: logística ou normal. Os modelos multidimensionais Compensatório e Ogiva Normal (Subseção 3.2.4) apresentam a mesma equivalência encontrada nos respectivos modelos unidimensionais, ou seja, para D=1,702 e o mesmo vetor de parâmetros os dois modelos fornecem resultados bem próximos.

Suposições do Modelo

O modelo descrito acima pressupõe que a dimensionalidade do espaço θ está adequadamente especificada. Isso nos garante a independência condicional. Como introduzido na Seção 1, a independência condicional é fundamental no processo de estimação dos parâmetros do modelo. Outras suposições são: (a) o tempo para a resolução do teste é suficiente para que todos os itens possam ser respondidos por todos os indivíduos e (b) a ordem em que os itens são apresentados aos indivíduos não interfere no desempenho dos mesmos.

3.2.2 Modelo Compensatório sem Acerto Casual - MC2

Quando não é possível, ou coerente, considerar a possibilidade de acerto casual ao item, o parâmetro c torna-se igual a zero e tem-se o chamado Modelo Compensatório sem Acerto Casual - MC2, dado por

$$P\left(X_{ij} = 1 \mid \theta_j\right) = \frac{1}{1 + exp[-(D\mathbf{a}_i'\theta_j + d_i)]}.$$

Um exemplo seria o caso de testes com itens abertos corrigidos como certo ou errado. Os parâmetros deste modelo foram definidos na subseção anterior.

3.2.3 Modelos Não-Compensatórios

O modelo não-compensatório (com acerto casual) foi desenvolvido por Sympson (1978) e é citado em Ackerman (1996). Neste modelo há um parâmetro relacionado à discriminação e outro à dificuldade para cada uma das dimensões. Como os termos no modelo são multiplicativos, a probabilidade condicional de resposta correta é limitada pela menor das probabilidades obtidas nas m dimensões do espaço θ. O modelo é dado por

$$P(X_{ij}=1|\theta_j)=c_i+(1-c_i)\prod_{k=1}^{m}\frac{\exp[D\,a_{ik}(\theta_{kj}+b_{ik})]}{1+\exp[D\,a_{ik}(\theta_{kj}+b_{ik})]},$$

onde b_{ik}, $i = 1, 2, ..., p$ e $k=1, 2, ..., m$, é o parâmetro relacionado à dificuldade do item i na dimensão k.

A denominação não-compensatório se deve ao fato de que um nível de habilidade alto em uma dimensão não compensa um nível baixo em outra dimensão. Ackerman (1996) propõe comparações entre este modelo e o modelo compensatório (Equação 3.1) e cita Ackerman (1989), Lim (1993) e Hsu (1995) como referências para a estimação dos parâmetros dos modelos não-compensatórios e compensatórios. Entretanto, programas computacionais ainda não estão disponíveis para este modelo, o que torna sua utilização restrita. Neste trabalho, esses modelos não serão vistos com maiores detalhes. O modelo não-compensatório sem acerto casual é obtido fazendo $c_i = 0$.

3.2.4 Modelos Ogiva-Normal Multidimensionais

O modelo ogiva-normal multidimensional com acerto casual - MN3 é dado por

$$P(X_{ij}=\theta_j)=c_i+(1-c_i)\int_{-\infty}^{Z_i(\theta_j)}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt, \qquad (3.2)$$

onde $Z_i(\theta_j) = a_{i1}\theta_{1i} + a_{i2}\theta_{2i} + ... + a_{im}\theta_{mi} + d_i$. As definições dos parâmetros desse modelo foram dadas na Subseção 3.2.1. A relação entre este modelo e o modelo compensatório com acerto casual é a mesma dos respectivos modelos unidimensionais, ou seja, fazendo D=1,702, com o mesmo conjunto de parâmetros, os dois modelos fornecem resultados bem próximos.

Note que a Equação 3.2 possui a mesma forma da Equação 2.7. A relação entre análise fatorial e teoria da resposta ao item é tema já discutido e apresentado por alguns pesquisadores. O modelo ogiva-normal está implementado nos dois programas mais utilizados na modelagem de itens multidimensionais: NOHARM (Fraser & McDonald, 1988) e TESTFACT (Wilson, Wood & Gibbons, 1998). Estes programas são comumente empregados para estimação de parâmetros de modelos multidimensionais.

Sem a possibilidade de acerto casual ao item, faz-se o parâmetro c igual a zero e tem-se o chamado modelo ogiva-normal multidimensional sem acerto casual - MN2, dado por

$$P(X_{ij} = \theta_j) = \int_{-\infty}^{Z_i(\theta_j)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

4. Aplicação a Dados Reais

4.1. Introdução

O Exame Nacional do Ensino Médio (ENEM) introduziu um novo conceito de avaliação educacional no Brasil. Diferentemente de avaliações disciplinares como, por exemplo, o Sistema Nacional de Avaliação do Ensino Básico (SAEB) e o VESTIBULAR, onde cada conhecimento é medido em testes individuais, o ENEM é um exame interdisciplinar, onde os vários conhecimentos associados aos conteúdos do ensino fundamental e médio são avaliados de uma só vez por um único teste. O grande diferencial do exame pode ser atribuído aos itens que o compõem. Cada um deles é elaborado de modo a avaliar até 5 competências, mesclando os conhecimentos de diferentes disciplinas.

As análises tradicionais de itens tornaram-se, em alguns casos, obsoletas com o surgimento e desenvolvimento da Teoria da Resposta ao Item (TRI). Entretanto, a aplicação dessa teoria estava restrita aos modelos unidimensionais. A grande importância do ENEM no contexto nacional motivou o estudo dos modelos multidimensionais da TRI.

O ENEM caracteriza-se como o exame do perfil de saída da escolaridade básica e tem como um dos principais objetivos fornecer ao participante subsídios para a sua auto-avaliação (INEP, 1999; INEP 2000). Entretanto, com o aumento do número de instituições de ensino superior que vêm aderindo à utilização dos resultados do exame como parte de processos seletivos, o número de participantes tem aumentado significativamente.

Este seção apresenta a análise da dimensionalidade e a modelagem dos dados do Exame Nacional de Ensino Médio do ano de 1999 (ENEM-99). A verificação da dimensionalidade foi feita através da Análise Fatorial de Informação Plena introduzida no Seção 2. Na modelagem dos dados foi adotado o modelo compensatório com acerto casual (MC3) apresentado na Subseção 3.2.1.

4.2 Características do ENEM

O Exame Nacional do Ensino Médio é aplicado desde 1998 em todo território nacional. É constituído de um teste único contendo 63 itens de múltipla escolha e uma proposta para redação. Os itens objetivos e a redação destinam-se a avaliar as competências desenvolvidas pelos participantes ao longo da escolaridade básica.

O exame tem caráter voluntário e dele podem participar, mediante inscrição, os concluintes do ensino médio, no ano de realização do exame, e também os que já o concluíram em anos anteriores, em qualquer de suas modalidades.

O ENEM é estruturado por uma matriz de competências que define claramente os pressupostos do exame e delineia suas características operacionais. O modelo da matriz contempla a indicação das competências gerais próprias do aluno, na fase de desenvolvimento cognitivo correspondente ao término da escolaridade básica, associadas aos conteúdos do ensino fundamental e médio.

De forma resumida, as cinco competências tratadas pelo ENEM são:

I. Dominar linguagens;
II. Compreender fenômenos;
III. Enfrentar situações-problema;
IV. Construir argumentação;
V. Elaborar propostas.

4.3 Dados do ENEM

O ENEM é aplicado anualmente em todo território nacional a voluntários que concluíram ou estão em fase de conclusão do ensino médio. A parte objetiva do exame é composta de 63 itens. A ordem dos itens e/ou as opções de resposta correta são alteradas com o intuito de formar 4 testes "diferentes" cada um contendo os mesmos 63 itens. Esses testes são identificados pelas cores amarela, branca, rosa e verde, e a cada um deles corresponde aproximadamente 25,0% do total de inscritos. A distribuição dos testes segue uma seqüência alternada de cores de modo que todos eles estão presentes em todas as localidades de realização do exame.

O número de inscritos no ano de 1999 foi de 315.960 alunos em todo o Brasil. A distribuição para cada um dos 4 testes foi a seguinte: 80.251 para a cor amarela, 79.194 para a cor branca, 78.803 para a cor rosa e 77.712 para a cor verde.

Nesse trabalho serão utilizados os dados referentes à parte objetiva do teste de cor amarela (teste amarelo). Como a distribuição dos testes segundo as cores é bem heterogênea em cada uma das localidades de aplicação do exame, tem-se uma amostra representativa da população e de tamanho significativo para a estimação dos parâmetros necessários à determinação da dimensionalidade dos dados e também para a estimação dos parâmetros dos itens. Para a estimação das habilidades dos participantes dispõe-se inicialmente de 63 itens, que é um número adequado - a literatura sugere um valor maior que 30.

4.4 Dimensionalidade do ENEM

O estudo da dimensionalidade do ENEM-99 foi precedido por uma análise preliminar dos 63 itens que o compõem. Estatísticas descritivas e algumas medidas da teoria clássica foram obtidas. Os resultados envolvendo o item 45 nos levaram a questionar sua utilização no estudo. Este item apresentou correlação bisserial negativa - a saber $corr_b$= - 0,117. Isto significa que dos indivíduos que o responderam corretamente a maior parte foi do grupo de indivíduos de pior desempenho no exame. é provável - isto merece um estudo por especialistas em itens - que a formulação do item tenha induzido os candidatos a não optar pela alternativa correta. Além da correlação bisserial, todas as correlações tetracóricas envolvendo este item apresentaram valores negativos. Isto poderia comprometer as estimativas dos parâmetros necessários para a determinação da dimensionalidade. Estes resultados levaram à exclusão do item 45 do estudo. Foram excluídos também 41 indivíduos que entregaram o exame sem responder nenhum item. Dessa forma, a base de dados do ENEM utilizada neste trabalho será composta por uma matriz de respostas referente a 62 itens e 80.210 indivíduos.

Continuando a análise descritiva, foram obtidos os percentuais de resposta para cada item. O objetivo foi verificar a não-resposta aos últimos itens do exame. Um percentual grande - considera-se grande qualquer valor acima de 20,0% - poderia indicar que o tempo especificado para a realização do teste foi insuficiente. Nos dados analisados o menor percentual de respostas a um item foi de 99,4%, o que sugere não ter ocorrido problema quanto ao controle de tempo no exame. Os passos seguintes referem-se propriamente à verificação da dimensionalidade.

Procedimentos

Na análise da dimensionalidade dos itens o primeiro passo foi a organização do banco de dados. O programa *Statistical Package for the Social Sciences - SPSS*, versão 10.0, foi utilizado para esse fim. Com ele procedeu-se à formatação da base de dados de forma conveniente à utilização dos programas BILOG 3 (Mislevy & Bock, 1990) e TESTFACT, versão 2.0.

O segundo passo foi a utilização do programa TESTFACT para determinar a dimensionalidade dos dados. Este programa tem como base teórica os artigos de Bock & Aitkin (1981) e Dempster, Laird & Rubin (1977). O TESTFACT vem sendo utilizado, principalmente, na verificação da dimensionalidade de testes. Nele se encontram implementadas a análise fatorial através da matriz de correlações tetracóricas e a análise fatorial de informação plena. O programa fornece ainda estatísticas descritivas dos itens e algumas medidas utilizadas na teoria clássica. O TESTFACT permite o uso de modelos com parâmetro de acerto casual, mas não possibilita a estimação destes parâmetros; exige a entrada destes já estimados. Utilizou-se, então, o programa BILOG 3 para este fim. O BILOG 3 é um programa computacional específico para análise de itens dicotômicos ou dicotomizados via TRI. Neste programa estão implementados os modelos unidimensionais logístico e ogiva-normal de 1, 2 e 3 parâmetros.

Análise Fatorial de Informação Plena - AFIP

A Análise Fatorial de Informação Plena será utilizada para determinar a dimensionalidade dos dados do ENEM-99 (teste amarelo), ou seja, para a determinação do número de fatores necessários ou adequados para a explicação desses dados. Uma grande vantagem da AFIP é a possibilidade de testar a significância estatística dos fatores adicionados sucessivamente ao modelo.

Os critérios utilizados para determinar a dimensionalidade são:

i. a diferença nos valores corrigidos de qui-quadrado (veja Subseção 2.2.3);
ii. a magnitude das cargas fatoriais após a rotação VARIMAX.

Laros et al. (2000) utilizaram um índice de unidimensionalidade definido como a razão entre a mudança corrigida no qui-quadrado e os seus graus de liberdade. Um índice positivo maior que 2,0 implica que o modelo de $k + 1$ fatores se adequa melhor aos dados do que o modelo com k fatores. Índices positivos menores do que 2,0 indicam que o modelo de $k + 1$ fatores se ajusta melhor aos dados, mas que esta melhoria não é significativa. Índices negativos indicam que o modelo de k fatores se ajusta melhor aos dados do que o modelo de $k + 1$ fatores. Neste trabalho também será considerado tal índice - chamaremos índice de dimensionalidade (*ID*).

O programa TESTFACT utiliza as cargas fatoriais obtidas a partir da análise fatorial principal sobre a matriz de correlações tetracóricas (Divgi, 1979) como valores iniciais para o algoritmo EM na análise fatorial de informação plena. Este programa permite soluções para os principais problemas encontrados na construção da matriz de correlações tetracóricas. Essas soluções dizem respeito a possíveis correções antes da aplicação da análise fatorial. As correções são as seguintes: (i) correção para o acerto casual; (ii) correção para respostas omitidas; (iii) correção para obter uma matriz positiva definida e (iv) correção para evitar casos Heywood[2].

Resultados

Os resultados da análise para a verificação da dimensionalidade do Exame Nacional do Ensino Médio do ano de 1999 são apresentados na Tabela 4.1.

A Tabela 4.1 mostra os valores da estatística qui-quadrado, X^2, para os modelos de 1 a 5 fatores com os respectivos graus de liberdade, *gl*, as diferenças dos qui-quadrados, X^2_{dif}, entre um modelo de k fatores e outro de $k + 1$ fatores, $k = 1,2,3,4$, os graus de liberdade para essas diferenças, os valores corrigidos das diferenças dos qui-quadrados, $X^2_{difcorr}$, e o índice de dimensionalidade, *ID*.

Foram ajustados sucessivamente os modelos de 1 a 5 fatores. Para cada fator adicionado ao modelo a estatística $X^2_{difcorr}$ foi calculada. Até o último modelo ajustado, modelo de 5 fatores, esta estatística apresentou significância, ou seja, houve melhoria no ajuste do modelo aos dados.

---

[2] Na análise fatorial de informação plena, esses casos caracterizam-se quando um ou mais parâmetros a aumentam continuamente com o aumento do número de ciclos do algoritmo EM.

Tabela 4.1 – Estatísticas para o Número de Fatores do Modelo para o ENEM - 99

	Número de Fatores do Modelo				
	1	2	3	4	5
X^2	4.006.317,82	3.992.022,79	3.981.784,75	3.976.992,41	3.974.241,79
gl	80.081	80.020	79.960	79.901	79.843
X^2_{dif}		14.295,03	10.238,04	4.792,34	2.750,63
gl_{dif}		61	60	59	58
$X^2_{difcorr}$		4.765,01	3.412,68	1.597,45	916,88
ID		78,11	56,88	27,08	15,81

Observe na Tabela 4.1 que os valores de X^2 diminuem à medida em que são acrescentados fatores ao modelo. Conseqüentemente, as diferenças corrigidas também diminuem, variando de forma decrescente de 4.765,01 a 916,88. A menor destas diferenças corresponde à diferença entre os valores de X^2 dos modelos de 4 e 5 fatores dividida por 3 (veja Subseção 2.2.3: Teste de Ajuste do Modelo).

Segundo o índice de dimensionalidade, *ID*, também calculado para cada fator adicionado ao modelo, a conclusão é a mesma. Os índices decrescem de 78,11 (valor referente aos modelos de 1 e 2 fatores) até 15,81 (valor referente aos modelos de 4 e 5 fatores). Este último valor ainda está bem acima do valor crítico 2,0 citado na seção anterior, indicando a significância do modelo de 5 fatores.

O segundo critério utilizado foi a análise das cargas fatoriais. Estas cargas são apresentadas no Apêndice. As cargas fatoriais possibilitam informar o quanto da variância de cada variável (item) é explicada por cada fator e também quais itens se relacionam a cada fator - quanto maior a carga fatorial melhor é a relação entre item e fator. Neste trabalho considera-se que cargas fatoriais inferiores a 0,20 não contribuem para a mensuração dos respectivos fatores. Analisando as cargas fatoriais (após rotação VARIMAX), para cada um dos 3 primeiros fatores, nota-se que a grande maioria delas apresenta valores expressivos. Isto indica que quase todos os itens estão relacionadas a pelo menos um dos três fatores. Por exemplo, os itens *I3*, *I4* e *I5* estão bem relacionados aos fatores 1, 2 e 3. Já os itens *I21* e *I27* se relacionam melhor aos fatores 1 e 3. Para a mensuração do quarto fator contribuem quase 50% dos itens. Ao quinto fator estão relacionados apenas os itens *I9*, *I28*, *I29*, *I36* e *I43*.

Um critério complementar para a decisão do número de fatores no modelo será o percentual de explicação da variância das variáveis (itens) devida a cada fator. A Tabela 4.2 mostra esses percentuais. O valor acumulado até o quinto fator foi de 45,33%. No caso dos fatores não rotacionados, o primeiro fator é responsável por 39,85% e os demais fatores contribuem com percentuais inferiores a 2,06%. Sob essa análise poder-se-ia sugerir a utilização de um modelo de 1 fator (modelo unidimensional). Entretanto, a própria estrutura do exame sugere um modelo multidimensional. Objetivando uma melhor interpretação, os fatores foram submetidos à rotação VARIMAX. Com a rotação, os três primeiros fatores passaram a apresentar percentuais muito próximos. O percentual acumulado dos três primeiros foi de 38,54%. O quarto fator contribuiu com 5,55% e o quinto fator com 1,24%. Estes resultados retratam coerentemente os resultados obtidos anteriormente pela análise das cargas fatoriais.

Tabela 4.2 – Percentual da Variância Explicada por cada Fator do Modelo

	Fatores				
	1°	2°	3°	4°	5°
Fator(es) não rotacionado(s)	39,85	2,05	1,54	1,06	0,82
Fator(es) após rotação varimax	13,49	12,31	12,74	5,55	1,24

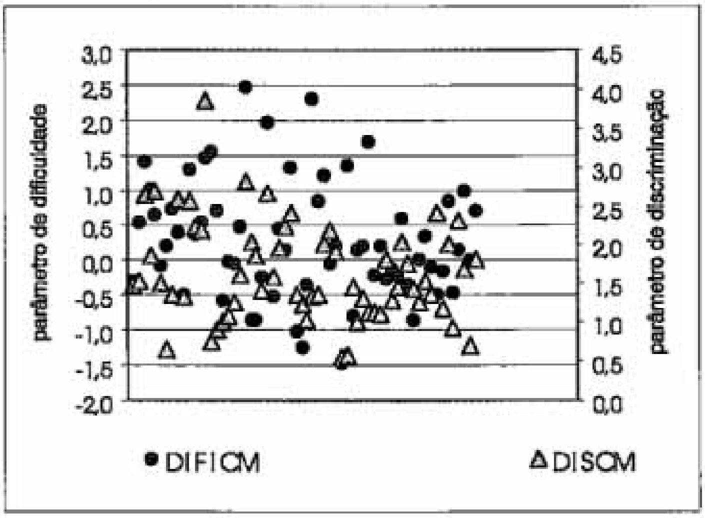
A decisão final sobre o número de fatores do modelo, isto é, a dimensionalidade do referido instrumento de avaliação (ENEM), deve, entretanto, ser tomada em conjunto com os especialistas que o construíram.

4.5 Análise dos Itens do ENEM

Nesta seção é apresentado o resultado da estimação dos parâmetros dos itens do ENEM-99. O modelo adotado foi o MC3 (veja Subseção 3.2.1). Os itens são caracterizados pelos parâmetros de acerto casual (*c*), discriminação (*DISCM*) e dificuldade (*DIFICM*). No apêndice estes 3 parâmetros são apresentados para cada um dos 62 itens do exame. Os resultados estão representados graficamente na Figura 4.1.

Com exceção do item 14, com parâmetro de discriminação igual a 3,84, todos os demais itens apresentaram discriminação no intervalo esperado de 0 a 3,00. Para o parâmetro de dificuldade *DIFICM*, todos os itens apresentaram valores no intervalo -3,00 a 3,00, mais precisamente entre -1,48 e 2,48. Na Figura 4.1 vê-se claramente que os itens estão bem distribuídos segundo esses dois parâmetros.

Figura 4.1 – Representação gráfica dos parâmetros DIFICM e DISCM para os 62 itens do ENEM - 99



Os parâmetros *c*, *d* e *a*'s do modelo MC3 também estão presentes no Apêndice. Estes resultados permitem identificar, por exemplo, quais itens são comparáveis segundo seus parâmetros de dificuldade. Diferentemente do caso unidimensional, nem sempre a comparação é possível ou adequada (veja Seção 3). Para exemplificar, considere dois grupos de itens. O primeiro deles formado pelos itens *I1* e *I8* e o segundo composto pelos itens *I28* e *I29*. Analisando os valores dos parâmetros do modelo (especificamente os parâmetros *a*'s) é razoável comparar as dificuldades para os itens *I1* e *I8*. As dificuldades para os itens *I28* e *I29* também são comparáveis. Entretanto, as dificuldades entre itens de grupos diferentes não são comparáveis, pois os grupos não estão medindo as mesmas habilidades (fatores). Os itens *I28* e *I29* estão relacionados aos fatores 1, 2, 3 e 5, enquanto os itens *I1* e *I8* estão relacionados apenas aos fatores 1 e 2. Com esta análise é possível estruturar testes que possam medir combinações de habilidades em vários níveis de dificuldade.

5. Comentários e Sugestões Finais

A aplicação e o desenvolvimento da Teoria da Resposta ao Item (TRI) em muito dependem da disponibilização de programas computacionais que possam facilitar ou viabilizar sua utilização. Na Europa e nos Estados Unidos, a TRI é uma ferramenta comumente utilizada em diversas áreas profissionais, entre elas a Educacional. No

Brasil, a TRI é bem recente. Sua primeira aplicação foi na análise do Sistema Nacional de Avaliação Básica (SAEB), em 1995. Desde então, os órgãos governamentais, através do Ministério da Educação (MEC), vêm valorizando e incentivando o uso dessa teoria nas avaliações educacionais brasileiras. Entre outras avaliações podemos citar o Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP) e a Avaliação das Escolas Públicas do Estado do Rio Grande do Norte.

Este trabalho é pioneiro na utilização da TRI multidimensional no Brasil. Este estudo inicial possibilitou avaliar sob este novo enfoque o Exame Nacional do Ensino Médio caracterizado como um exame interdisciplinar e que por isto sugeria a utilização de modelos onde a habilidade fosse expressa por mais de uma dimensão. Até então, as aplicações se restringiam às avaliações unidimensionais.

Relativamente a outras áreas da psicometria, a TRI multidimensional ainda está na sua "infância". Vários problemas precisam ser solucionados e metodologias ainda precisam ser desenvolvidas para ajudar na solução desses problemas. Contudo, os estudos têm ratificado que o relacionamento entre itens e indivíduos é bem mais complexo que o suposto nos procedimentos psicométricos usuais.

Entre as lacunas a serem preenchidas podemos citar, por exemplo, a questão do efeito sobre o escore total quando da inclusão de itens medindo múltiplas dimensões no teste (Reckase, 1989). Alguns trabalhos foram feitos no sentido de compreender melhor esta questão, mas esta pesquisa ainda está no seu início. Outro ponto pouco explorado diz respeito à equalização utilizando modelos multidimensionais. Esta é uma área de grande necessidade de pesquisa porque muitos dos testes educacionais utilizados, provavelmente, avaliam a habilidade em mais de uma dimensão e o estabelecimento de uma métrica comum no sentido de comparar os resultados desses testes ainda não está definida.

Neste trabalho apresentamos alguns dos modelos multidimensionais citados na literatura. Outros modelos estão presentes em algumas referências citadas, por exemplo, Linden et al. (1997). Há vários outros modelos ainda pouco explorados, como, por exemplo, a versão multidimensional para modelos de respostas a itens politômicos, os modelos multidimensionais longitudinais e os modelos multidimensionais multivariados. Tavares (2001) e Matos (2001) desenvolveram, respectivamente, modelos longitudinais e modelos multivariados unidimensionais.

Outra área que necessita de pesquisas adicionais é a de estimação dos parâmetros dos modelos multidimensionais. Embora bons programas existam (por exemplo, TESTFACT e NOHAM), pouco é conhecido sobre o número de itens e/ou de respondentes necessários para a especificação adequada da dimensão do espaço latente. Perguntas como: (i) Qual é o relacionamento entre tamanho da amostra, a heterogeneidade da população de indivíduos e o número de dimensões que pode ser identificada? e (ii) O que significa dizer que duas dimensões são altamente correlacionadas mas distintas? ainda necessitam de estudos para que sejam respondidas. Sem dúvida, essa é uma área rica para futuras pesquisas.

A aplicação feita neste trabalho considerou apenas os resultados de um dos 4 testes do Exame Nacional do Ensino Médio. Como a "diferença" nos testes está na ordem de apresentação dos itens e/ou das alternativas de respostas, uma questão importante seria avaliar se haveria mudanças significativas nos resultados obtidos para a análise da dimensionalidade dependendo do teste utilizado.

Outro ponto que necessita de pesquisas adicionais refere-se ao estabelecimento de critérios para a qualidade do item multidimensional. No caso unidimensional, um exemplo de item de má qualidade é aquele com parâmetro de discriminação negativo ou com valor positivo baixo. Para o caso multidimensional, as aplicações estão no início e um estudo detalhado faz-se necessário.

Por fim, gostaríamos de ressaltar que a aplicação apropriada dessa teoria exige fundamentalmente a integração de especialistas das áreas de estatística e educação.

Apêndice

Este Apêndice apresenta as cargas fatoriais, antes e após a rotação VARIMAX, fornecidas pelo programa TESTFACT. Os valores para os parâmetros c, d, a's, DIFICM e DISCM para cada um dos itens, segundo o modelo MC3, são apresentados em seguida.

Resultados do programa TESTFACT: cargas fatoriais não rotacionadas

		FATORES				
	ITEM	1	2	3	4	5
1	I1	0.597	-0.111	-0.230	-0.110	0.015
2	I2	0.803	0.127	-0.242	-0.053	-0.091
3	I3	0.784	0.022	0.180	-0.165	-0.181
4	I4	0.876	0.111	0.158	-0.114	-0.191
5	I5	0.819	0.113	-0.018	-0.118	-0.124
6	I6	0.828	0.027	-0.021	-0.049	-0.204
7	I7	0.332	-0.101	0.016	-0.042	-0.055
8	I8	0.577	-0.127	-0.155	-0.102	-0.068
9	I9	0.794	0.011	-0.142	-0.055	0.208
10	I10	0.508	-0.320	-0.017	-0.124	-0.084
11	I11	0.794	0.243	0.023	-0.063	0.017
12	I12	0.779	0.122	-0.004	-0.077	-0.081
13	I13	0.756	-0.164	0.085	-0.121	0.039
14	I14	0.893	0.131	-0.026	-0.069	-0.128
15	I15	0.345	0.082	0.184	-0.056	-0.007
16	I16	0.456	-0.013	0.010	-0.019	0.112
17	I17	0.462	-0.139	0.056	-0.051	-0.070
18	I18	0.514	-0.101	-0.078	-0.071	0.050
19	I19	0.642	0.071	-0.236	0.004	-0.035
20	I20	0.646	0.216	-0.055	-0.046	0.006
21	I21	0.787	0.325	0.021	-0.075	0.005
22	I22	0.689	-0.268	-0.137	-0.140	0.043
23	I23	0.647	-0.308	-0.059	-0.148	-0.069
24	I24	0.615	0.117	-0.122	0.001	0.055
25	I25	0.820	0.140	-0.125	-0.039	-0.032
26	I26	0.676	0.014	-0.077	-0.019	-0.006

		FATORES				
	ITEM	1	2	3	4	5
27	I27	0.705	0.240	-0.074	-0.010	-0.101
28	I28	0.764	-0.005	0.027	-0.075	0.206
29	I29	0.771	0.044	-0.007	-0.140	0.223
30	I30	0.578	0.091	-0.189	0.053	0.072
31	I31	0.523	-0.265	-0.008	-0.014	-0.024
32	I32	0.462	0.148	0.100	-0.006	0.063
33	I33	0.593	0.081	0.098	-0.048	0.114
34	I34	0.553	0.139	0.256	-0.015	0.014
35	I35	0.743	0.096	-0.126	-0.017	0.057
36	I36	0.765	0.033	0.179	0.000	0.131
37	I37	0.664	0.019	0.324	-0.022	0.109
38	I38	0.278	0.030	0.077	0.064	-0.085
39	I39	0.297	-0.000	-0.130	0.023	-0.020
40	I40	0.622	-0.183	0.021	-0.006	-0.037
41	I41	0.502	-0.037	0.024	-0.013	0.064
42	I42	0.549	-0.249	-0.054	-0.025	0.080
43	I43	0.517	0.028	-0.006	0.066	0.183
44	I44	0.513	0.084	0.190	0.057	0.015
45	I46	0.523	-0.098	0.068	0.066	0.004
46	I47	0.685	0.132	-0.131	0.157	-0.029
47	I48	0.574	-0.181	0.025	0.036	-0.028
48	I49	0.594	-0.190	0.187	0.022	-0.047
49	I50	0.759	0.065	0.014	0.052	0.086
50	I51	0.670	-0.038	-0.060	0.251	-0.028
51	I52	0.579	0.055	-0.151	0.228	-0.011
52	I53	0.551	-0.190	0.056	0.098	0.027
53	I54	0.619	-0.043	0.215	0.127	0.029
54	I55	0.567	-0.045	0.054	0.219	-0.124
55	I56	0.762	-0.068	0.011	0.287	-0.036
56	I57	0.520	0.023	0.217	0.083	-0.043
57	I58	0.76	0.005	0.017	0.056	0.048
58	I59	0.414	-0.123	-0.087	0.189	0.042
59	I60	0.771	-0.165	0.142	0.083	-0.046
60	I61	0.658	0.119	-0.090	0.179	-0.076
61	I62	0.345	-0.019	-0.040	0.163	0.002
62	I63	0.670	-0.264	-0.080	0.067	0.002

Resultados do programa TESTFACT: cargas fatoriais após rotação VARIMAX

		FATORES				
	ITEM	1	2	3	4	5
1	I1	0.438	0.452	0.122	0.125	0.089
2	I2	0.567	0.260	0.188	0.153	-0.022
3	I3	0.372	0.445	0.589	0.111	-0.107
4	I4	0.363	0.305	0.538	0.111	-0.130
5	I5	0.552	0.389	0.480	0.160	-0.042
6	I6	0.397	0.342	0.340	0.175	-0.138
7	I7	0.133	0.269	0.164	0.089	-0.017
8	I8	0.375	0.452	0.163	0.128	0.003
9	I9	0.532	0.419	0.327	0.214	0.294
10	I10	0.156	0.563	0.172	0.104	-0.020
11	I11	0.560	0.241	0.532	0.177	0.090
12	I12	0.515	0.343	0.467	0.181	-0.004
13	I13	0.300	0.551	0.433	0.160	0.118
14	I14	0.599	0.399	0.513	0.229	-0.039
15	I15	0.127	0.119	0.360	0.046	0.019
16	I16	0.236	0.245	0.250	0.133	0.159
17	I17	0.175	0.379	0.264	0.136	-0.017
18	I18	0.292	0.373	0.192	0.123	0.109
19	I19	0.489	0.250	0.134	0.190	0.028
20	I20	0.511	0.188	0.382	0.150	0.067
21	I21	0.602	0.178	0.556	0.153	0.074
22	I22	0.353	0.622	0.194	0.144	0.128
23	I23	0.288	0.637	0.221	0.129	0.011
24	I24	0.475	0.235	0.276	0.197	0.119
25	I25	0.615	0.345	0.397	0.231	0.053
26	I26	0.435	0.355	0.312	0.214	0.067
27	I27	0.569	0.198	0.404	0.207	-0.032
28	I28	0.404	0.408	0.437	0.180	0.283
29	I29	0.463	0.394	0.439	0.115	0.305
30	I30	0.474	0.225	0.189	0.238	0.137
31	I31	0.164	0.488	0.191	0.201	0.040
32	I32	0.276	0.125	0.385	0.132	0.104

		FATORES				
	ITEM	1	2	3	4	5
33	I33	0.309	0.245	0.426	0.136	0.168
34	I34	0.218	0.163	0.542	0.144	0.056
35	I35	0.541	0.325	0.337	0.227	0.135
36	I36	0.314	0.344	0.554	0.238	0.200
37	I37	0.170	0.303	0.614	0.183	0.163
38	I38	0.116	0.103	0.213	0.151	-0.059
39	I39	0.244	0.161	0.058	0.129	0.017
40	I40	0.243	0.472	0.293	0.230	0.033
41	I41	0.242	0.288	0.278	0.159	0.116
42	I42	0.211	0.488	0.175	0.195	0.147
43	I43	0.283	0.213	0.273	0.225	0.235
44	I44	0.197	0.168	0.443	0.207	0.058
45	I46	0.174	0.323	0.310	0.247	0.059
46	I47	0.507	0.220	0.295	0.368	0.043
47	I48	0.207	0.432	0.266	0.252	0.037
48	I49	0.114	0.442	0.400	0.241	0.014
49	I50	0.432	0.322	0.435	0.293	0.161
50	I51	0.351	0.311	0.274	0.470	0.047
51	I52	0.409	0.203	0.188	0.409	0.055
52	I53	0.156	0.400	0.266	0.299	0.089
53	I54	0.164	0.298	0.470	0.323	0.086
54	I55	0.230	0.270	0.309	0.406	-0.063
55	I56	0.337	0.365	0.365	0.536	0.045
56	I57	0.152	0.214	0.445	0.243	0.002
57	I58	0.402	0.371	0.417	0.306	0.126
58	I59	0.189	0.258	0.094	0.334	0.093
59	I60	0.244	0.500	0.462	0.356	0.035
60	I61	0.459	0.209	0.305	0.381	-0.006
61	I62	0.161	0.148	0.130	0.273	0.041
62	I63	0.277	0.543	0.206	0.327	0.085

Os percentuais presentes na Tabela 4.2 são calculados pela fórm $PVE_m = \left(\sum_{i=1}^{p} \lambda_{im}^2 \right) / n$, onde PVE_m representa o percentual de variância explicada pelo m-ésimo fator, m=1, 2, 3, 4 e 5, e λ_{im} denota a carga fatorial para o i-ésimo item, i=1, 2, ..., n (n=62), no fator m.

Parâmetros dos itens segundo o modelo MC3

	ITEM	c	d	a1	a2	a3	a4	a5	DISCM	DIFICM
1	I1	0.139	0.437	1.137	0.701	-0.235	-0.563	0.261	1.491	-0.293
2	I2	0.207	-0.821	1.278	0.276	-0.576	-0.388	0.411	1.536	0.534
3	I3	0.142	-3.745	2.448	0.216	0.519	-0.194	0.803	2.644	1.416
4	I4	0.262	-1.885	1.732	-0.029	0.225	-0.034	0.630	1.857	1.015
5	I5	0.073	-1.760	2.446	0.359	-0.161	-0.388	0.952	2.682	0.656
6	I6	0.352	0.092	1.484	0.191	-0.087	-0.110	0.270	1.508	-0.061
7	I7	0.130	-0.131	0.572	0.249	0.115	-0.082	0.071	0.643	0.203
8	I8	0.222	-1.007	1.152	0.555	-0.095	-0.419	0.140	1.356	0.743
9	I9	0.251	-1.055	1.763	1.227	-0.342	-0.590	1.263	2.583	0.408
10	I10	0.389	0.647	1.048	0.669	0.357	-0.333	-0.096	1.339	-0.483
11	I11	0.126	-3.308	2.072	0.309	-0.287	-0.203	1.434	2.563	1.291
12	I12	0.237	-0.867	2.002	0.354	-0.160	-0.236	0.897	2.241	0.387
13	I13	0.149	-1.161	1.756	0.922	0.469	-0.335	0.692	2.178	0.533
14	I14	0.396	-5.671	3.497	0.601	-0.368	-0.359	1.384	3.843	1.476
15	I15	0.109	-1.159	0.571	0.002	0.220	0.059	0.430	0.750	1.545
16	I16	0.150	-0.648	0.642	0.423	0.024	-0.105	0.470	0.908	0.714
17	I17	0.041	0.583	0.896	0.386	0.219	-0.083	0.150	1.014	-0.575
18	I18	0.008	0.021	0.836	0.642	0.000	-0.287	0.307	1.083	-0.019
19	I19	0.025	0.053	1.005	0.387	-0.508	-0.274	0.334	1.266	-0.042
20	I20	0.225	-0.797	1.285	0.199	-0.337	-0.192	0.849	1.601	0.498
21	I21	0.234	-6.958	2.214	0.110	-0.420	-0.221	1.655	2.807	2.479
22	I22	0.064	1.736	1.510	1.133	0.137	-0.666	0.277	2.026	-0.857
23	I23	0.163	1.568	1.499	0.907	0.333	-0.512	0.039	1.856	-0.845
24	I24	0.047	0.356	1.093	0.435	-0.372	-0.199	0.673	1.420	-0.251
25	I25	0.129	-5.224	2.263	0.618	-0.553	-0.380	1.074	2.665	1.960
26	I26	0.130	0.817	1.345	0.550	-0.189	-0.201	0.560	1.581	-0.517
27	I27	0.172	-0.899	1.693	0.130	-0.499	-0.114	0.836	1.961	0.459
28	I28	0.062	-0.338	1.550	0.960	0.109	-0.349	1.226	2.236	0.151
29	I29	0.287	-3.211	1.622	0.936	0.049	-0.579	1.410	2.415	1.329
30	I30	0.127	1.376	0.971	0.528	-0.494	-0.178	0.571	1.351	-1.019
31	I31	0.136	1.544	0.964	0.722	0.219	-0.130	0.057	1.232	-1.253
32	I32	0.157	0.359	0.761	0.157	0.002	0.043	0.676	1.031	-0.348
33	I33	0.231	-3.106	0.984	0.380	0.099	-0.078	0.822	1.343	2.313
34	I34	0.164	-1.169	1.034	0.095	0.277	0.203	0.817	1.365	0.857
35	I35	0.235	-2.420	1.584	0.662	-0.416	-0.304	0.899	2.005	1.207
36	I36	0.139	0.121	1.609	0.744	0.300	0.082	1.225	2.177	-0.056
37	I37	0.161	-0.412	1.336	0.494	0.614	0.229	1.097	1.914	0.215
38	I38	0.035	0.819	0.499	0.064	0.016	0.169	0.152	0.553	-1.482
39	I39	0.026	-0.789	0.465	0.238	-0.218	-0.101	0.111	0.586	1.347
40	I40	0.112	1.178	1.234	0.699	0.185	-0.083	0.251	1.455	-0.810

	ITEM	c	d	a1	a2	a3	a4	a5	DISCM	DIFICM
41	I41	0.046	-0.142	0.784	0.442	0.057	-0.077	0.432	1.003	0.142
42	I42	0.158	-0.264	0.919	0.868	0.149	-0.237	0.220	1.313	0.201
43	I43	0.314	-1.822	0.899	0.577	-0.107	-0.003	0.668	1.131	1.890
44	I44	0.128	0.241	0.865	0.216	0.150	0.245	0.621	1.140	-0.211
45	I46	0.219	-0.226	0.897	0.509	0.146	0.126	0.334	1.101	0.205
46	I47	0.105	0.479	1.454	0.646	-0.620	0.125	0.670	1.806	-0.205
47	I48	0.074	0.176	1.074	0.667	0.150	0.004	0.211	1.290	-0.137
48	I49	0.117	0.347	1.205	0.612	0.464	0.160	0.305	1.469	-0.236
49	I50	0.160	-1.219	1.593	0.749	-0.159	0.008	1.020	2.045	0.596
50	I51	0.148	0.607	1.392	0.839	-0.371	0.359	0.438	1.763	-0.344
51	I52	0.036	1.210	1.068	0.604	-0.566	0.216	0.414	1.429	-0.846
52	I53	0.047	0.009	0.965	0.743	0.163	0.129	0.264	1.256	-0.002
53	I54	0.134	-0.534	1.173	0.574	0.275	0.305	0.644	1.533	0.348
54	I55	0.230	0.150	1.166	0.485	-0.103	0.486	0.235	1.361	-0.110
55	I56	0.080	1.218	1.928	1.135	-0.295	0.622	0.618	2.421	-0.503
56	I57	0.001	0.195	0.968	0.249	0.226	0.329	0.494	1.165	-0.156
57	I58	0.136	-1.711	1.636	0.797	-0.088	0.023	0.855	2.013	0.850
58	I59	0.041	0.445	0.616	0.643	-0.101	0.153	0.145	0.934	-0.477
59	I60	0.261	-0.373	1.962	1.006	0.372	0.267	0.580	2.325	0.161
60	I61	0.267	-1.649	1.402	0.459	-0.530	0.231	0.554	1.679	0.993
61	I62	0.015	-0.003	0.519	0.365	-0.183	0.163	0.162	0.709	0.004
62	I63	0.165	-1.272	1.398	1.120	0.026	-0.101	0.166	1.804	0.705

Referências

ACKERMAN, T. A. Unidimensional IRT calibration of compensatory and noncompensatory items. **Applied Psychological Measurement, 13,** 1989, 113-127.

______. Graphical representation of multidimensional item response theory analysis. **Applied Psychological Measurement, 20,** 1996, 311-329.

ANDRADE, D. F., TAVARES, H. R., VALLE, R. C. **Teoria da Resposta ao Item: Conceitos e Aplicações.** 14 SINAPE. 2000.

BAKER, F. B. **Item Response Theory – Parameter Estimation Techniques.** New York: Marcel Dekker, Inc., 1992.

BIRNBAUM, A. **Efficient design and use of a mental ability for various decision-making problems**, (Series Report No. 58-16. Project No. 7755-23). USAF School of Aviation Medicine, Texas: Randolph Air Force Base), 1957.

BOCK, R.D., AITKIN, M. Marginal maximum likelihood estimation of item parameters: application of an em algorithm. **Psychometrika, 46**, 1981, 443-459.

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

DIVGI, D. R. Calculation of the tetrachoric correlation coefficient. **Psychometrika, 44**, 1979, 169-172.

FRASER, C., McDONALD, R. P. **NOHARM II: A FORTRAN program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory** [Computer software]. Armidale, Australia: University of News Wngland, Centre for Behavioral Studies, 1988.

HAMBLETON, R. K., SWAMINATHAN, H. **Item Response Theory: Principles and Applications**, Boston: Kluwer Academic Publishers, 1985.

HATTIE, J. A. Methodology Review: assessing undimensionality of tests and items: **Applied psychological Measurement, 9**, 1985, 139-164.

HSU, Y. Item parameter estimation of a two-dimensional generalized MIRT model (Doctoral dissertation, University of Illinois, 1995). **Dissertation Abstracts International**, 57, 1995, 1584.

Instituto Nacional de Estudos e Pesquisas Educacionais - INEP/MEC **Exame Nacional do Ensino Médio - ENEM: Relatório Final**. Brasília: O Instituto, 1999.

Instituto Nacional de Estudos e Pesquisas Educacionais - INEP/MEC **Exame Nacional do Ensino Médio - ENEM: Documento Básico}**. Brasília: O Instituto, 2000.

JOHNSON, R. A., Wichern, D. W. **Applied Multivariate Statistical Analysis**, 4nd Ed. New Jersey: Prentice Hall, 1998.

LAROS, J. A., PASQUALI, L., Rodrigues, M. M. M. **Análise da Unidimensionalidade das Provas do SAEB**. Universidade de Brasília, 2000.

LAWLEY, D. N., Maxwell, A. E. Factor **Analysis as a Statistical Method**, 2nd Ed. London: Butterworth, 1971.

LINDEN, W. J. van der, HAMBLETON, R. K. **Handbook of Modern Item Response Theory**. New York: Springer-Verlag, 1997.

LIM, C. An application of the joint maximum likelihood estimation procedure to a two-dimensional case of Sympson's noncompensatory IRT model (Doctotal dissertation, University of Iowa, 1993. **Dissertation Abstracts Interntaional, 54,** 1993, 2549.

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

MATOS, G. S. **Teoria da Resposta ao Item: Uma Proposta de Modelo Multivariado.** Dissertação de Mestrado apresentada à Universidade Federal de Pernambuco. Pernambuco, 2001.

MCDONALD, R. P., AHLAWAT, W. S. Difficulty factors in binary data. **British Journal of Mathematical and Statistical Psychology, 27,** 1974, 82-99.

MISLEVY, R. J., Bock, R. D. **BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models.** Chicago: Scientific Software, Inc., 1990.

PEARSON, K. On the correlation of characters not quantitatively measurable. **Royal Society Philosophical Transactions,** Series A, **195,** 1900, 1-47.

RECKASE, M. D. **The difficult of test items that measure more than one ability.** Applied Psychological Measurement, 9, 1985, 401-412.

RECKASE, M. D. **Controlling the psychometric snake: Or, how I learned to love multidimensionality.** Invited address at the meeting of the American Psychological Association, New Orleans, 1989.

RECKASE, M. D., McKinkey, R. L. **The discriminating power of items that measure more than one dimension.** Applied Psychological Measurement, 15, 1991, 361-373.

SYMPSON, J. B. A model for testing multidimensional items. In D. J. Weiss (Ed.) **Proceedings of the 1977 computerized adaptive testing conference** (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods program, 1978.

TAVARES, H. R. **Teoria da Resposta ao Item para Dados Longitudinais.** Tese de Doutorado apresentada à Universidade de São Paulo. São Paulo, 2001.

WILSON, D.T., Wood, R., Gibbons, R. **TESTFACT: Test Scoring, Item Statistics and Item Factor Analysis.** Chicago: Scientific Software, Inc, 1998.