COMPARANDO DESEMPENHOS DE GRUPOS DE ALUNOS POR INTERMÉDIO DA TEORIA DE RESPOSTA AO ITEM

Dalton Francisco de Andrade
Professor Titular de Estatística do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará – UFC.
Ex-professor da Universidade de São Paulo – USP

Resumo

O objetivo principal do presente trabalho é de apresentar e comparar três formas de eqüalização (ou ligação) de diferentes populações (obtenção de valores dos parâmetros dos itens e das habilidades na mesma escala), a partir de um estudo de simulação, levando em consideração também o problema do número de itens comuns às diferentes provas. Uma apresentação de modelos matemáticos com suas suposições básicas e dos principais métodos de estimação dos parâmetros dos itens (calibração) e de habilidades dos respondentes é realizada. A seguir, são apresentados estudos de simulação e seus principais resultados. Uma aplicação destes resultados na análise do Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo – SARESP – dos anos 1996 e 1997 é mostrado e, finalmente, são apresentadas conclusões e sugestões.

1. Introdução

Um dos grandes problemas em avaliações educacionais é a comparação do desempenho de diferentes grupos (populações) de respondentes. Por exemplo, como verificar se houve ganho de conhecimento dos alunos da escola pública estadual de São Paulo do ano de 1996 para o ano de 1997? Ou, ainda, como comparar o desempenho dos alunos da 4ª série do ensino fundamental de um certo estado da federação com o desempenho médio nacional obtido pelo Sistema Nacional de Avaliação do Ensino Básico (SAEB) implementado pelo MEC (ver Ministério da Educação e do Desporto (1995)).

A Teoria Clássica de Medida, a qual se baseia nos resultados obtidos em provas expressos apenas por seus escores brutos ou padronizados, permite estes tipos de comparações desde que os indivíduos sejam submetidos todos à mesma prova ou, pelo menos, ao que se denomina de provas paralelas. Os resultados encontrados dependem do particular conjunto de itens que compõem o instrumento de medida, ou seja, as análises e interpretações estão sempre associadas à prova como um todo. O leitor encontrará maiores detalhes sobre esta metodologia, incluindo a sua fundamentação matemática em Gulliksen (1950), Lord and Novick (1968) e Vianna (1987), por exemplo.

Atualmente, na área educacional, vem crescendo o interesse pela aplicação de técnicas derivadas da Teoria de Resposta ao Item (TRI) que propõem modelos de variáveis latentes para representar a relação entre a probabilidade de um aluno responder corretamente a um item e seus traços latentes ou habilidades na área do conhecimento avaliada, os quais não são observados diretamente. Tendo como elemento central os itens e não a prova como um todo, a TRI permite, por exemplo, a comparação entre populações distintas submetidas a provas diferentes, mas com alguns itens comuns ou, ainda, a comparação entre indivíduos da mesma população que tenham sido submetidos a diferentes provas, com ou sem itens comuns. Em Lord (1980) e Hambleton, Swaminathan e Rogers (1991) e, mais recentemente, em Andrade e Valle (1998) e Andrade et al. (2000), por exemplo, o leitor encontrará maiores detalhes sobre os fundamentos e aplicações desta teoria.

Os modelos propostos dependem de parâmetros associados aos itens e das habilidades dos respondentes e as comparações desejadas

somente são possíveis de serem realizadas com todos os valores destes parâmetros e das habilidades na mesma escala de medida. Em geral, tanto os parâmetros dos itens quanto as habilidades precisam ser estimados a partir de provas aplicadas a amostras aleatórias de elementos das populações em estudo. As provas aplicadas às diferentes populações deverão ter itens comuns. Os desempenhos das diferentes populações são comparados a partir das distribuições, ou de parâmetros dessas distribuições, das habilidades de seus elementos. Em Mislevy (1992), o leitor encontrará uma completa discussão sobre as mais diferentes formas de ligação de avaliações educacionais. Uma forma bastante utilizada, ver por exemplo Kolen e Brennan (1995) e Ministério da Educação e do Desporto (1995), é a estimação dos parâmetros dos itens em separado para cada uma das populações em estudo e, através do princípio da invariância dos itens, determinar a partir dos itens comuns os coeficientes apropriados para a conversão das diferentes escalas das diferentes populações em uma única escala. Uma segunda forma, recentemente introduzida por Bock e Zimowski (1997), seria a estimação dos parâmetros de todos os itens aplicados nas diferentes provas, com alguns itens comuns, às várias populações simultaneamente, a partir de modelos para grupos múltiplos. Uma outra forma, mais simples do que as duas anteriores, e sugerida por Hambleton, Swaminathan e Rogers (1991), seria considerar todos os respondentes como pertencentes a uma única população e todas as provas formando uma única prova. No processo de estimação dos parâmetros dos itens, os itens não comuns às populações seriam tratados como não apresentados aos respondentes das outras populações.

2. Modelos para um ou mais grupos

A TRI baseia-se em modelos que representam a probabilidade de um indivíduo responder corretamente a um item como função dos parâmetros do item e da(s) habilidade(s) do respondente. Os vários modelos propostos na literatura dependem, fundamentalmente, do tipo do item. Um dos mais utilizados, e que será utilizado neste trabalho, é o **modelo logístico unidimensional de 3 parâmetros** para itens de múltipla escolha dicotômicos ou dicotomizados (do tipo certo/errado).

2.1 Modelos para um único grupo

O modelo logístico unidimensional de 3 parâmetros para um

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

determinado item i é dado por

i = 1,2,...,I, j = 1,2,...,n, onde

X_{ij} é uma variável dicotômica que assume os valores 1 (quando o indivíduo j responde corretamente ao item i) ou 0 (quando o indivíduo j não responde corretamente ao item i),

θ_j representa a habilidade ou proficiência (traço latente) do indivíduo j,

$P(X_{ij}=1|\theta)$ é a probabilidade de um indivíduo j com habilidade igual a θ responder corretamente ao item i,

D é um fator de escala constante conhecido, igual a 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal,

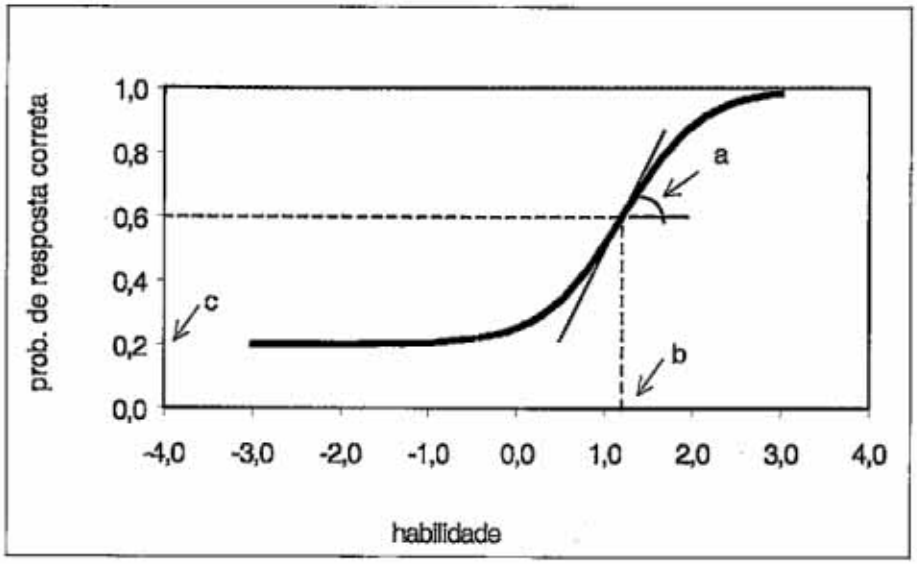
b_i é o parâmetro de dificuldade (ou de posição) do item i, medido na mesma escala da habilidade,

a_i é o parâmetro de discriminação (ou de inclinação) do item i, com valor proporcional à inclinação da curva característica do item - CCI no ponto b_i, e

c_i é o parâmetro de acerto ao acaso do item i.

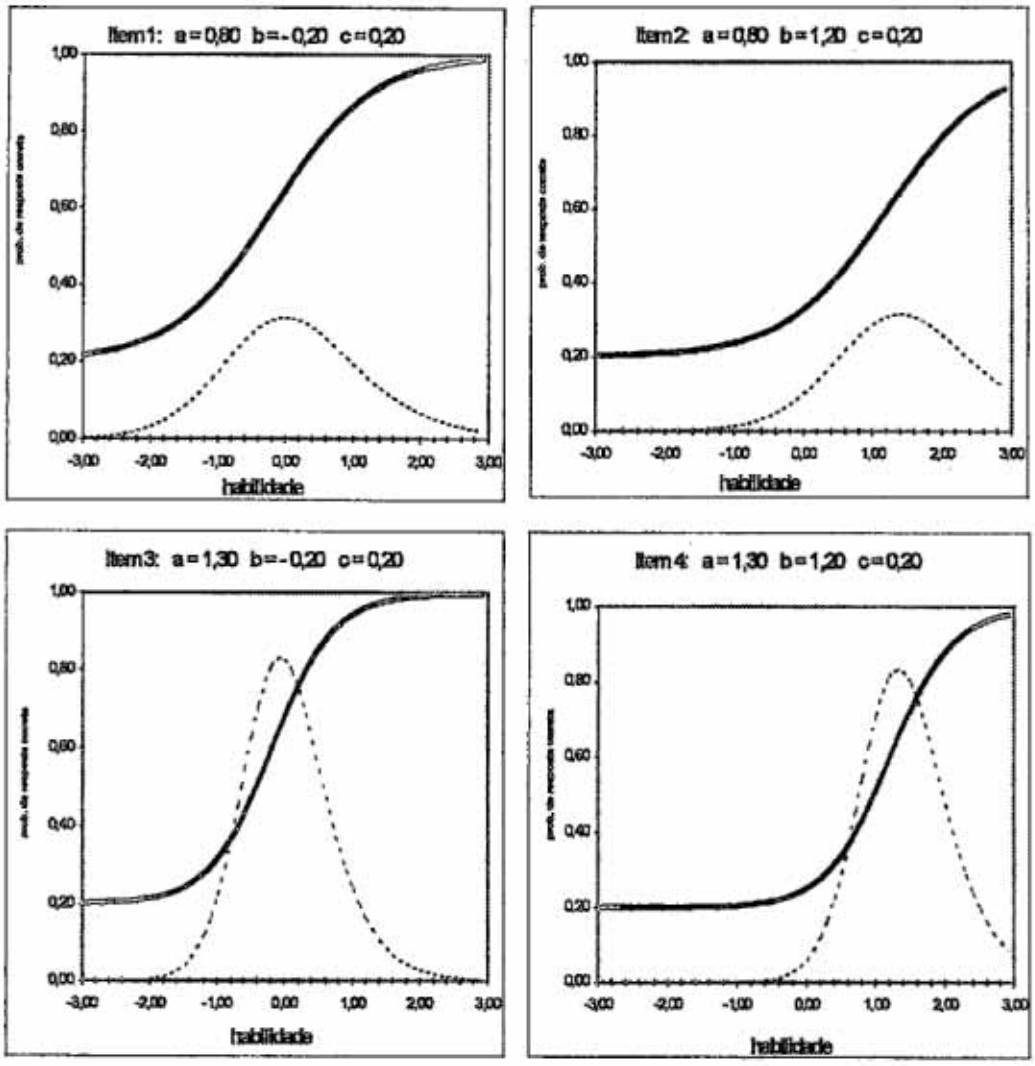
Note que $P(X_i=1|\theta)$ pode ser vista também como a proporção de resposta correta ao item dentre todos os indivíduos da população com a mesma habilidade θ. A figura abaixo exemplifica a relação existente entre $P(X_i=1|\theta)$ e os parâmetros do modelo.

Figura 2.1
Curva característica do item - CCI



O modelo proposto baseia-se no fato de que indivíduos com maior habilidade possuem maior probabilidade de acertar ao item e que esta relação não é linear. De fato, pode-se perceber a partir do gráfico acima que a CCI tem forma de "S" com inclinação e deslocamento na escala de habilidade definidos pelos parâmetros do item. Na Figura 2.2 apresentam-se curvas características e também curvas de informação (traçado pontilhado) de quatro itens com diferentes combinações de valores dos parâmetros **a** e **b**.

Figura 2.2: Curvas carcterísticas e de informação de vários itens



Comparando-se os itens 1 e 3 e também os itens 2 e 4 pode-se perceber que o item com maior valor do parâmetro **a** tem a curva característica com inclinação mais acentuada. A conseqüência disto é que a diferença entre as probabilidades de resposta correta de dois indivíduos com habilidades 2,00 e 1,00, por exemplo, é maior no item 4 (0,37=0,88-0,51) do que no item 2 (0,25=0,80-0,55). Em outras palavras, o item 4 é mais apropriado para discriminar estes dois indivíduos do que o item 2. Por este motivo é que o parâmetro **a** é

denominado de **parâmetro de discriminação (ou de inclinação)** do item. Por outro lado, comparando-se os itens 1 e 2 e também os itens 3 e 4, pode-se perceber que o item com maior valor do parâmetro **b** exige uma habilidade maior para uma mesma probabilidade de resposta correta. Por exemplo, a habilidade requerida para uma probabilidade de resposta correta de 0,60 é igual a –0,20 no item 1 e igual a 1,20 no item 2. Isto é, o item 2 é mais difícil do que o item 1. Assim, o parâmetro **b** é denominado de **parâmetro de dificuldade (ou de posição)** do item e seu valor está na mesma escala da habilidade. Na realidade, o parâmetro **b** representa a habilidade necessária para uma probabilidade de acerto igual a $(1+c)/2$. O parâmetro **c** representa a probabilidade de um aluno com baixa habilidade responder corretamente ao item (muitas vezes referido como a probabilidade de acerto ao acaso). Note que a cada item está associado um intervalo na escala de habilidade no qual o item tem maior poder de discriminação. Este intervalo é definido em torno do valor do parâmetro **b** e está mostrado nos gráficos pelas curvas de informação (traçados pontilhados). Deste modo, a discriminação entre bons alunos é feita a partir de itens considerados difíceis e não de itens considerados fáceis. Apesar de receberem a mesma denominação da Teoria Clássica, o parâmetro de dificuldade do item não é medido por uma proporção (valor entre 0 e 1) e o parâmetro de discriminação não é uma correlação (valor entre -1 e 1). Na TRI, estes dois parâmetros podem, teoricamente, assumir qualquer valor real entre $-\infty$ e $+\infty$. É claro que não se espera um valor negativo para o parâmetro a.

Na prática, as habilidades e os parâmetros dos itens são estimados a partir das respostas de um grupo de respondentes submetidos a esses itens mas, uma vez estabelecida a escala de medida da habilidade, os valores dos parâmetros dos itens não mudam, isto é, seus valores são invariantes a diferentes grupos de respondentes, desde que os indivíduos destes grupos tenham suas habilidades medidas na mesma escala.

2.1.1 Escala de habilidade/Indeterminação

Diferentemente da medida escore em um teste com n questões do tipo certo/errado, que assume valores inteiros entre 0 e n, na TRI a habilidade pode teoricamente assumir qualquer valor real entre $-\infty$ e $+\infty$. Assim, precisa-se estabelecer uma origem e uma unidade de medida para a definição da escala. Esses valores são escolhidos de

modo a representar, respectivamente, o valor médio e o desvio padrão das habilidades dos indivíduos da população em estudo. Para os gráficos mostrados anteriormente, utilizou-se a escala com média igual a 0 e desvio padrão igual a 1, que será representada ao longo deste trabalho por escala (0,1). Em termos práticos, não faz a menor diferença estabelecer-se estes valores ou outros quaisquer. O importante são as relações de ordem existentes entre os pontos da escala. Por exemplo, na escala utilizada acima um indivíduo com habilidade 1,20 está 1,20 desvios padrões acima da habilidade média. Este mesmo indivíduo teria a habilidade 92,00 e, conseqüentemente, estaria também 1,20 desvios padrões acima da habilidade média, se a escala utilizada para esta população fosse a escala(80,10). Isto pode ser visto a partir da transformação de escala

$$a(\theta-b) = (a/10)[(10\theta+80) - (10b+80)] = a^*(\theta^*-b^*)$$

onde $a(\theta-b)$ é a parte do modelo probabilístico proposto envolvida na transformação.
Assim, tem-se que:

1. $\theta^* = 10\theta+80$
2. $b^* = 10b+80$
3. $a^* = a/10$
4. $P(X_i=1\,|\,\theta) = P(X_i=1\,|\,\theta^*)$

Por exemplo, os valores dos parâmetros **a** e **b** do item 4, mostrado anteriormente, na escala(0,1) são, respectivamente, 1,30 e 1,20 e seus correspondentes na escala (80,10) são, respectivamente, 0,13=1,30/10 e 92,00=10x1,20+80. Além disso, um indivíduo com habilidade θ = 1,00 medida na escala(0,1) tem sua habilidade representada por θ^* = 10X1,00+80 = 90,00 na escala(80,10) e

$$P(X_4 = 1|\theta = 1{,}00) = 0{,}20 + (1-0{,}20)\frac{1}{1+e^{-1{,}7x1{,}30(1{,}00-1{,}20)}}$$

$$= 0{,}20 + (1-0{,}20)\frac{1}{1+e^{-1{,}7x0{,}13(90{,}00-92{,}00)}}$$

$$= P(X_4 = 1|\theta^* = 90{,}00) = 0{,}51$$

ou seja, a probabilidade de um indivíduo responder corretamente a um certo item é sempre a mesma, independentemente da escala utilizada

para medir a sua habilidade ou, ainda, a habilidade de um indivíduo é invariante a escala de medida.

Assim, não faz qualquer sentido querermos analisar itens a partir dos valores de seus parâmetros **a** e **b** sem conhecer a escala na qual eles foram determinados. Na escala(0,1), valores mais apropriados para o parâmetro **a** estão no intervalo [1,00;2,50] e para o parâmetro **b** no intervalo [-2,00;2,00]. É claro que estes valores dependem muito do objetivo da avaliação. Por exemplo, um item com **a** igual a 2,00 serve, basicamente, para discriminar os indivíduos em dois grupos de habilidade, os que têm habilidade menor do que o valor de **b** dos que têm habilidade maior do que o valor de **b**. Note que, o valor do parâmetro **c** não se altera com a mudança de escala porque ele mede a probabilidade de acerto para indivíduos com baixa habilidade, qualquer que seja a escala de medida.

2.1.2 Unidimensionalidade/independência local

O modelo proposto pressupõe que o número de traços latentes medidos pela prova é igual a 1, isto é, o modelo supõe que a prova mede uma única habilidade. Tradicionalmente, tem-se utilizado a técnica de análise fatorial a partir da matriz de correlações tetracóricas para a verificação da dimensionalidade de provas. Mislevy(1986) discute as deficiências da aplicação deste procedimento e sugere um outro procedimento baseado no método de máxima verossimilhança.

Uma outra suposição do modelo é a chamada independência local ou independência condicional, a qual assume que para uma dada habilidade as respostas aos diferentes itens da prova são independentes. Esta suposição é fundamental para o processo de estimação dos parâmetros do modelo. Na realidade, como a unidimensionalidade implica independência local, tem-se somente uma e não duas suposições a serem verificadas. Assim, itens devem ser elaborados de modo a satisfazer a suposição de unidimensionalidade.

2.1.3 Outros modelos

Dois outros modelos podem ser facilmente obtidos do modelo apresentado acima. Por exemplo, quando não existe possibilidade de resposta correta ao acaso pode-se considerar $\mathbf{c} = 0$ no modelo acima e tem-se o chamado **modelo logístico unidimensional de 2 parâmetros**. Se além de não existir resposta ao acaso ainda tivermos todos os itens com o mesmo poder de discriminação, tem-se o chamado modelo de 1

parâmetro, o qual possui somente o parâmetro de dificuldade do item. Este modelo é conhecido como modelo de Rasch.

Existem também modelos para itens com mais de 2 categorias de resposta. Por exemplo, itens abertos podem ser corrigidos de modo a ter-se uma ou mais categorias intermediárias ordenadas entre as categorias certo e errado. Estes modelos são chamados de modelos de resposta gradual e foram introduzidos por Samejima(1969).

2.2 Modelos para dois ou mais grupos

Uma generalização desses modelos logísticos para 1, 2 e 3 parâmetros, para o caso de duas ou mais populações, foi recentemente proposta por Bock e Zimowski (1997) e é dada por

$$P(X_{ijk}=1|\theta_{jk})=c_i+(1-c_i)\frac{1}{1+e^{-Da_i(\theta_{jk}-b_i)}}$$

$i=1,2,...,I$, $j=1,2,...,n_k$, $k=1,2,...,K$, onde

X_{ijk} é uma variável dicotômica que assume os valores 1 (quando o indivíduo j da população k responde corretamente ao item i a ele apresentado) ou 0 (quando indivíduo não responde corretamente ao item),

θ_{jk} representa a habilidade ou proficiência (traço latente) do indivíduo j da população k,

$P(X_{ijk}=1|\theta)$ é a probabilidade do indivíduo jk com habilidade igual a θ responder corretamente ao item i, e

D, b_i, a_i e c_i estão descritos em **2.1**.

Em geral, os indivíduos pertencentes às diferentes populações não são submetidos todos aos mesmos itens mas, para permitir a comparação entre populações, alguns deles deverão ser comuns a mais de uma das populações. Assim, I representa o número total de itens apresentados. Nos tópicos 4 e 5 apresentaremos exemplos de diferentes provas aplicadas a diferentes populações.

3. Estimação dos parâmetros dos itens (calibração) e das habilidades

Uma das etapas mais importante da TRI é a estimação dos parâmetros dos itens e/ou das habilidades dos respondentes. Em algumas situações, os parâmetros dos itens já são conhecidos e o que se deseja é estimar as habilidades; já em outras situações menos freqüentes, conhecem-se as habilidades dos respondentes e o que se deseja é a estimação dos parâmetros dos itens. Porém, as situações mais comuns são aquelas em que se deseja estimar tanto os parâmetros dos itens quanto as habilidades dos respondentes simultaneamente. O processo de estimação dos parâmetros dos itens é conhecido por **calibração**. Em todas estas situações, assume-se como verdadeiro o modelo proposto e, a partir do conjunto de respostas dadas por um certo número de respondentes de uma ou mais populações, estima-se os parâmetros e/ou habilidades a partir do método de máxima verossimilhança ou de métodos bayesianos. Ambos os métodos exigem procedimentos iterativos que envolvem cálculos bastante complexos e, conseqüentemente, programas de computador específicos. É importante ressaltar que, em qualquer uma destas situações, os valores das habilidades e dos parâmetros dos itens estarão todos na mesma escala de medida. Vários autores têm sugerido que cada respondente seja submetido a pelo menos 30 itens e que cada item seja submetido a pelo menos 300 respondentes, para que se obtenham estimativas com erros padrões pequenos. Note que, apesar de estarmos sempre nos referindo à habilidade de um indivíduo, na prática, em geral, o que se deseja é estimar a habilidade média de uma população de indivíduos, por exemplo, a população dos alunos da 3ª série do ensino fundamental da escola pública estadual de São Paulo. A seguir, apresenta-se uma discussão dos principais resultados desses métodos de estimação necessários para a equalização de diferentes populações. O leitor interessado no desenvolvimento completo destes métodos de estimação poderá consultar, entre outros, Baker(1992).

3.1 Um único grupo

Sejam X_{ij} a variável aleatória dicotômica (1=acerto e 0=erro) que representa a resposta do j-ésimo indivíduo, j=1,2,...,n, ao i-ésimo item, i=1, 2,..., I, $\mathbf{X}_j=(X_{1j}, X_{2j}, ..., X_{Ij})^t$ o vetor aleatório (Ix1) que representa as respostas do j-ésimo indivíduo a todos os itens, $\mathbf{a} = (a_1,$

$a_2, \ldots, a_I)^t$ o vetor (Ix1) dos parâmetros de discriminação, $\mathbf{b} = (b_1, b_2, \ldots, b_I)^t$ o vetor (Ix1) dos parâmetros de dificuldade, $\mathbf{c} = (c_1, c_2, \ldots, c_I)^t$ o vetor (Ix1) dos parâmetros de acerto ao acaso, $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^t$ o vetor (nx1) das habilidades (parâmetros) de n respondentes de uma determinada população e P_{ij} o modelo logístico de 3 parâmetros apresentado em 2.1.

3.1.1 Estimação de máxima verossimilhança conjunta

Esta forma de estimação tem sido a base para a estimação dos n parâmetros de habilidade e dos 3I parâmetros de itens para os mais diferentes programas computacionais desenvolvidos para a TRI. Sob a suposição de independência local, ver **2.1.2**, a probabilidade do vetor de resposta $\mathbf{x}_j$ do respondente j, condicionado na sua habilidade θ_j e nos parâmetros dos itens, é dada por

$$P_j(\boldsymbol{x}_j|\theta_j,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \prod_{I=1}^{I} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}}$$

e a função de verossimilhança, baseada nas respostas de uma amostra aleatória de n indivíduos de uma população, pode ser escrita como

$$L(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n|\theta,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \prod_{j=1}^{n} P_j(\boldsymbol{x}_j|\theta_j,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$$

Devido à indeterminação associada ao modelo, apresentada em **2.1.1.**, os valores dos parâmetros que maximizam a função de verossimilhança acima não podem ser determinados de modo único. Este problema não ocorre quando se conhecem as habilidades e deseja-se estimar os parâmetros dos itens e também quando se conhecem os parâmetros dos itens e deseja-se estimar as habilidades. Nesses casos, o conhecimento dos parâmetros implica no conhecimento da escala em que eles foram medidos. O problema da indeterminação é solucionado definindo-se uma escala arbitrária para os valores das habilidades ou para os valores dos parâmetros de dificuldade b, haja vista que tanto a habilidade quanto o parâmetro de dificuldade são medidos na mesma escala. Usualmente, estabelecem-se os valores 0 para a média e 1 para o desvio padrão dos valores das

habilidades. Métodos iterativos do tipo Newton-Raphson e Scoring de Fisher são utilizados para obtenção dos estimadores de máxima verossimilhaça. As habilidades de indivíduos que acertaram ou erraram todos os itens ou os parâmetros dos itens respondidos corretamente ou erroneamente por todos os indivíduos não podem ser estimados por este procedimento. Além do mais, como o número de parâmetros a ser estimado aumenta com o aumento do tamanho da amostra, as propriedades assintóticas dos estimadores de máxima verossimilhança não se aplicam. Atualmente, este procedimento de estimação tem sido utilizado somente como base para outros procedimentos de estimação.

3.1.2 Estimação de máxima verossimilhança marginal

Devido às dificuldades de estimarem-se conjuntamente os parâmetros dos itens e as habilidades, este procedimento sugere que os parâmetros dos itens sejam estimados em uma primeira fase e, supondo-se que esses valores obtidos sejam os verdadeiros valores dos parâmetros dos itens, estimam-se as habilidades em uma segunda fase. Ao contrário do procedimento descrito acima que não faz qualquer suposição sobre a distribuição da habilidade, este procedimento de estimação assume que os respondentes representam uma amostra aleatória de uma população na qual a habilidade é distribuída segundo uma determinada função densidade $g(\theta|\tau)$, onde τ é o vetor dos parâmetros desta distribuição. Estimativas de máxima verossimilhança para os parâmetros dos itens são obtidas a partir da maximização da função de verossimilhança marginal que depende das habilidades somente através da distribuição a priori $g(\theta)$. Note que a mesma distribuição a priori é assumida para todos os θ's. Neste caso, o problema da indeterminação do modelo é resolvido ao estabelecer-se a distribuição a priori, isto é, no final do processo de estimação tem-se as estimativas dos parâmetros dos itens em uma métrica definida pelos parâmetros de locação e escala da priori. Em geral, utiliza-se como priori a distribuição normal com média 0 e desvio padrão 1.

$$L(x_1, x_2, \ldots, x_n | a, b, c) = \prod_{j=1}^{n} \int P_j(x_j|\theta_j, a, b, c) g(\theta|\tau) d\theta$$

Apesar de ter-se que estimar somente os parâmetros dos itens, a obtenção dos valores de **a**, **b** e **c** que maximizam a função acima é computacionalmente bastante trabalhosa e inapropriada quando o número de itens é grande. Uma reformulação deste enfoque de

máxima verossimilhança marginal dentro da estrutura do algorítmo EM, produz estimativas consistentes para os parâmetros dos itens e é computacionalmente muito mais simples. Neste caso, a distribuição a posteriori da habilidade tem um papel fundamental, apesar deste procedimento não ser um procedimento bayesiano, tendo em vista que os parâmetros dos itens são considerados fixos. Em cada ciclo do algorítmo, as estimativas dos parâmetros dos itens são calculadas em uma métrica definida a partir da normalização e reescalonamento da distribuição a posteriori da habilidade, de modo a fazer com que os parâmetros de locação e escala da distribuição a posteriori tenham os mesmos valores dos correspondentes parâmetros da distribuição a priori. Ao final, o procedimento fornece as estimativas de máxima verossimilhança dos parâmetros dos itens e também uma estimativa da distribuição a posteriori das habilidades, todos na mesma escala.

Com as estimativas dos parâmetros dos itens consideradas como sendo os verdadeiros valores dos parâmetros, estimam-se as habilidades dos respondentes através do método da máxima verossimilhança ou bayesiano, na mesma métrica dos parâmetros dos itens. Pode-se também obter uma nova estimativa da distribuição a posteriori das habilidades. Um resultado importante é que a distribuição de habilidade estimada e a distribuição das estimativas das habilidades dos respondentes não são as mesmas.

Em algumas situações, estes procedimentos podem não fornecer resultados satisfatórios. Isto ocorre principalmente na estimação do parâmetro **c** do modelo de 3 parâmetros, devido à própria natureza do parâmetro que está associado à probabilidade de acerto de indivíduos com habilidade muito pequena, que em geral não são muitos. Um problema similar ocorre com a estimação do parâmetro **b** de itens muito fáceis ou muito difíceis para a população em estudo. Para o processo de estimação ser bem sucedido, é importante terem-se respondentes com habilidades cobrindo todo o espectro do conhecimento a ser avaliado. Nestas situações problemáticas, sugere-se que procedimentos bayesianos sejam utilizados a partir da incorporação de distribuições a priori também para os parâmetros dos itens. Os procedimentos bayesianos fornecem estimativas para todos os itens e habilidades, mesmo para os indivíduos que acertaram ou erraram todos os itens ou para os itens respondidos corretamente ou erroneamente por todos os indivíduos.

A aplicação destes procedimentos depende fundamentalmente da disponibilidade de programas computacionais específicos. Um programa computacional disponível comercialmente que tem implementado todas estas técnicas, inclusive com a opção de estimativas bayesianas para as habilidades dos respondentes, é o programa BILOG 3 (ver Mislevy e Bock (1990)).

3.2 Dois ou mais grupos

A estimação na mesma métrica dos parâmetros dos itens e das habilidades dos respondentes das populações envolvidas no estudo, tem sido realizada a partir da estimação dos parâmetros dos itens em separado para cada população envolvida, usando os procedimentos descritos anteriormente em **3.1**, assumindo-se que pelo menos alguns itens são comuns às populações. Apesar de tecnicamente obterem-se estimativas em métricas com parâmetro de locação 0 e parâmetro de escala 1 para cada uma das populações envolvidas no estudo, estas estimativas estão em métricas diferentes porque estes parâmetros de locação e escala são relativos às habilidades das diferentes populações e, conseqüentemente, representam diferentes escalas. Este fato pode ser facilmente verificado observando-se os valores das estimativas do parâmetro **b** de um mesmo item aplicado a mais de uma população. Para cada população, conforme descrito anteriormente em **2.1**, o valor do parâmetro **b** está associado ao grau de dificuldade do item e poderá, por exemplo, assumir valores positivos para as populações onde ele é considerado mais difícil e valores negativos naquelas onde ele é considerado mais fácil. Como a habilidade necessária para responder corretamente a um determinado item com uma certa probabilidade independe da métrica utilizada, valores diferentes para o parâmetro **b** de um mesmo item indicam que as estimativas obtidas para as populações estão em métricas diferentes. Existem várias técnicas disponíveis para a equalização, isto é, obtenção de todas as estimativas na mesma métrica, das diferentes populações. Neste trabalho usaremos o método denominado **Mean-Sigma**, o qual se baseia nas relações lineares existentes entre os parâmetros de um mesmo item medidos em escalas diferentes. Conforme discutido em **2.1.1**, dado que o modelo é adequado aos dados, os parâmetros **a** e **b** de um certo item apresentado a dois grupos de respondentes devem satisfazer, a menos de flutuações amostrais, as seguintes relações lineares:

$$b_{G1} = \alpha\, b_{G2} + \beta \qquad e \qquad a_{G1} = (1/\alpha)\, a_{G2}$$

onde b_{G1} e b_{G2} são os valores do parâmetro de dificuldade e a_{G1} e a_{G2} são os valores do parâmetro de discriminação nos grupos 1 e 2, respectivamente. Uma vez determinados os coeficientes α e β, as estimativas dos parâmetros dos itens do grupo 2 podem facilmente ser colocadas na mesma escala das estimativas do grupo 1. O método Mean-Sigma utiliza

$$\alpha = S_{G1} / S_{G2} \qquad e \qquad \beta = M_{G1} - \alpha\, M_{G2}$$

onde S_{G1} e S_{G2} são os desvios padrão e M_{G1} e M_{G2} as médias amostrais das estimativas dos parâmetros de dificuldade dos itens comuns nos grupos 1 e 2, respectivamente. Da mesma forma, as habilidades dos respondentes do grupo 2 podem ser colocadas na mesma escala das habilidades dos respondentes do grupo 2 a partir da relação

$$\theta^{1}_{G2} = \alpha\, \theta_{G2} + \beta$$

onde θ^{1}_{G2} é o valor da habilidade θ_{G2} na escala do grupo 1. O leitor poderá encontrar maiores detalhes deste e dos outros métodos em Kolen e Brennan (1995), por exemplo.

A utilização do modelo de duas ou mais populações descrito em **2.2.** permite que todos os parâmetros dos itens e, conseqüentemente, as habilidades dos respondentes das populações envolvidas no estudo sejam estimados diretamente na mesma métrica, desde que existam itens comuns às populações. O método de estimação dos parâmetros dos itens é o da máxima verossimilhança marginal descrito em **3.1.**, estabelecendo-se uma das populações como referência. A média e o desvio padrão da distribuição a posteriori da habilidade da população de referência são fixados como sendo iguais a 0 e 1, respectivamente, e os parâmetros das outras populações são obtidos dentro desta métrica. Como mencionado em Hedges e Vevea (1997), este procedimento deve, pelo menos teoricamente, fornecer estimativas menos viesadas do que aquelas produzidas pelos procedimentos referidos acima. Esta alternativa de equalização será denominada de **Grupos múltiplos**. Atualmente, pelo menos comercialmente, o único programa de computador disponível para a aplicação desta metodologia é o programa BILOG-MG, ver Zimowisk, Muraki, Mislevy, e Bock (1996),

que também pode ser utilizado para as situações de uma única população.

Uma terceira alternativa poderia ser a estimação simultânea de todos os parâmetros dos itens, através dos métodos descritos em **3.1.**, considerando-se todos os respondentes como sendo provenientes de uma única população. Neste caso, todos os parâmetros estariam em uma mesma métrica definida por uma população diferente daquelas consideradas no estudo. Esta alternativa de equalização será denominada de **Uma única população**.

Qualquer que seja o procedimento utilizado para estimação dos parâmetros dos itens na mesma escala, as habilidades dos respondentes podem ser estimadas separadamente para cada população através dos métodos descritos em **3.1**. Com as duas primeiras alternativas de equalização pode-se obter também uma estimativa da distribuição a posteriori das habilidades. Por outro lado, com a terceira alternativa, somente se pode obter, a posteriori, a distribuição das estimativas das habilidades.

No próximo tópico apresentaremos um estudo de simulação para comparar estas três alternativas de equalização.

4. Estudo de simulação

Neste estudo de simulação consideraram-se duas populações de respondentes P_1 e P_2 com habilidades distribuídas segundo duas distribuições normais com médias 0 e 1 e desvios padrão iguais a 1, respectivamente. Assim, a distância entre os centros das duas populações na métrica da população 1 é de 1 desvio padrão. De cada uma das populações selecionou-se uma amostra aleatória simples de 1000 respondentes. Com relação aos itens, foram considerados itens com o parâmetro **a** variando de 0,6 (baixa discriminação) a 1,5 (alta discriminação), o parâmetro **b** variando de –2.0 (item fácil) a 3.0 (item difícil) e o parâmetro **c** assumindo os valores 0,15 e 0,25. Os valores dos parâmetros dos itens estão apresentados nas Tabelas A.1-A.3 no Apêndice. Com o objetivo de também avaliar a influência do número de itens comuns às populações na equalização, simularam-se dados considerando-se provas compostas de 48 itens do tipo certo/errado com 5 alternativas de resposta, em 3 diferentes situações: a primeira com 24 itens comuns e 24 não comuns, perfazendo um total de 72 itens aplicados (item1-item48 para P_1 e item25-item72 para P_2 da Tabela A.1), a segunda com 12 itens comuns e 36 não comuns,

perfazendo um total de 84 itens aplicados (item1-item48 para P_1 e item37–item84 para P_2 da Tabela A.2) e a terceira com 6 itens comuns e 42 não comuns, perfazendo um total de 90 itens aplicados (item1-item48 para P_1 e item43-item90 para P_2 da Tabela A.3). A alocação dos itens às populações procurou evitar a aplicação de itens muito fáceis ou muito difíceis em cada uma das populações. Por exemplo, a aplicação de um item com parâmetro **b** igual a –2,0 na população 2 poderia trazer problemas de estimação, pois espera-se que poucos, ou quase nenhum dos 1000 elementos selecionados, tenham habilidade em torno deste valor. O mesmo poder-se-ia dizer sobre a aplicação de um item com grau de dificuldade 3,0 na população 1.

A partir das habilidades das amostras de respondentes, do modelo logístico de três parâmetros e dos itens descritos acima, geraram-se, para cada população, respostas para os 48 itens apresentados a cada um dos 1000 respondentes. Este procedimento foi repetido 1000 vezes, para cada uma das 3 situações de itens comuns descritas acima. Para cada uma das 1000 iterações, os parâmetros dos itens e as habilidades foram estimados através da aplicação de cada uma das três alternativas de equalização de populações descritas em **3.2.**, utilizando-se o programa de computador BILOG-MG.

4.1 Método Um único grupo

Os parâmetros dos itens e as habilidades foram todos estimados supondo-se que os 2000 respondentes pertenciam a um único grupo, diferente dos dois grupos gerados. Os dados obtidos de habilidade dos respondentes do grupo 1 foram padronizados (média 0 e desvio padrão 1) através de uma transformação apropriada e a mesma transformação foi realizada nas estimativas dos parâmetros **a** e **b** dos itens, comuns ou não, e das habilidades dos respondentes do grupo 2.

4.2 Método Mean-Sigma

Cada um dos grupos foi analisado em separado, supondo-se uma distribuição a priori normal com média 0 e desvio padrão 1, fazendo com que as estimativas dos parâmetros dos itens e das habilidades associados à população 1 já fossem obtidas na escala apropriada. A partir da relação linear existente entre os parâmetros dos itens comuns, colocou-se na escala do grupo 1 as estimativas dos parâmetros dos itens e das habilidades associadas ao grupo 2. Mesmo

estando na mesma escala, é possível que os valores dos parâmetros dos itens comuns sejam diferentes. Nestes casos, a estimativa final é obtida da média aritmética destes valores.

4.3 Método Grupos múltiplos

Os parâmetros dos itens e as habilidades foram todos estimados supondo-se que os respondentes pertenciam a dois grupos diferentes, com o grupo 1 considerado como o grupo de referência. Desta forma, as estimativas dos parâmetros de todos os itens e habilidades foram automaticamente obtidos na escala da distribuição a priori; neste caso, a normal com média 0 e desvio padrão 1, estabelecida para o grupo de referência.

4.4 Resultados

O desempenho das três alternativas de equalização foi estudado comparando-se as estimativas dos parâmetros dos itens e das médias e variâncias das habilidades com os verdadeiros valores definidos no processo de geração dos dados. Para as habilidades, a comparação é feita a partir das estimativas da média e variância obtidas para os parâmetros da distribuição da habilidade do grupo 2. Os valores são 1 para a média e 1 para a variância. Nas figuras 4.1 a 4.9 apresentamos os histogramas dos valores das estimativas das médias obtidos nas 1000 iterações, para cada método e situação de número de itens comuns.

Figura 4.1
Histograma da estimativa da média da habilidade do grupo 2.
Método Um único grupo com 24 itens comuns

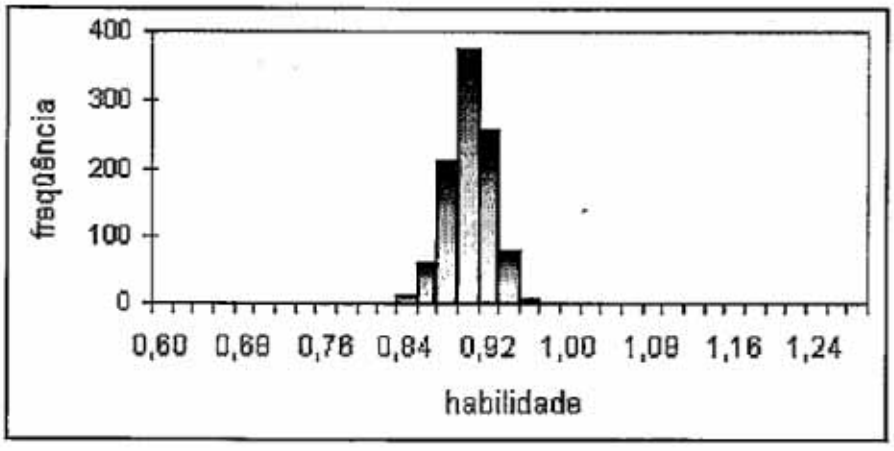


Figura 4.2
Histograma da estimativa da média da habilidade do grupo 2.
Método Mean-Sigma com 24 itens comuns.

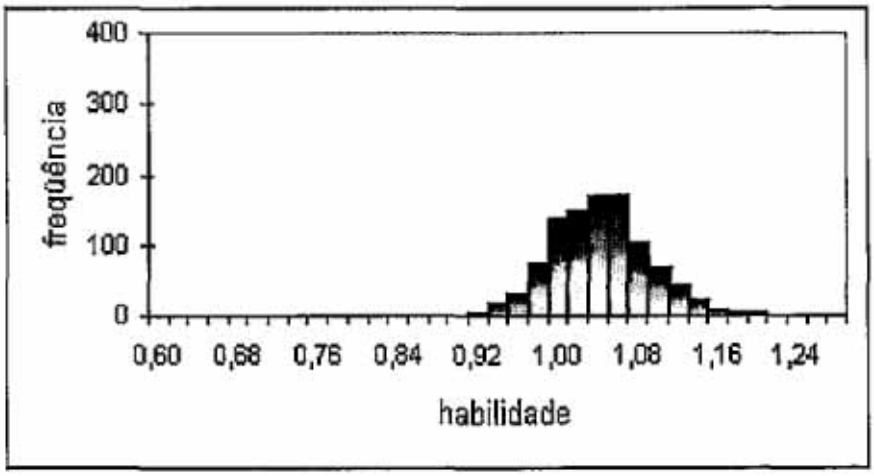


Figura 4.3
Histograma da estimativa da média da habilidade do grupo 2.
Método Grupos múltiplos com 24 itens comuns

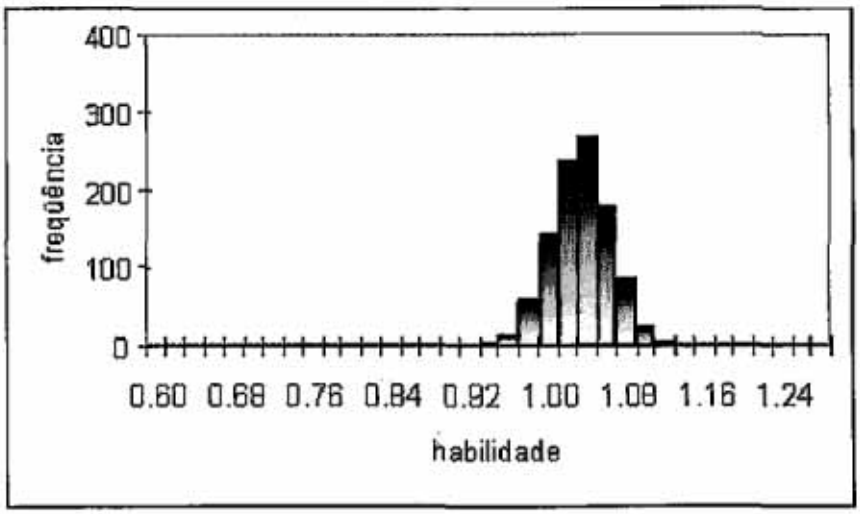


Figura 4.4
Histograma da estimativa da média da habilidade do grupo 2.
Método Um único grupo com 12 itens comuns.

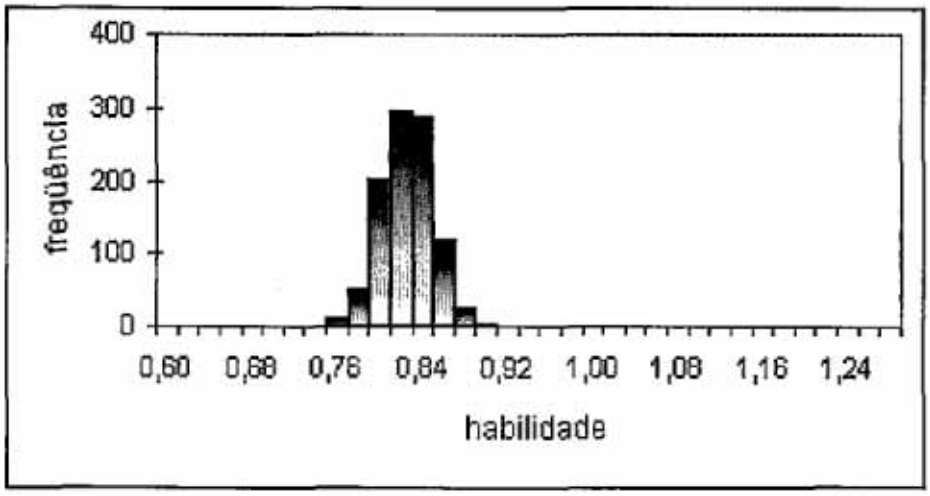


Figura 4.5
Histograma da estimativa da média da habilidade do grupo 2.
Método Mean-Sigma com 12 itens comuns

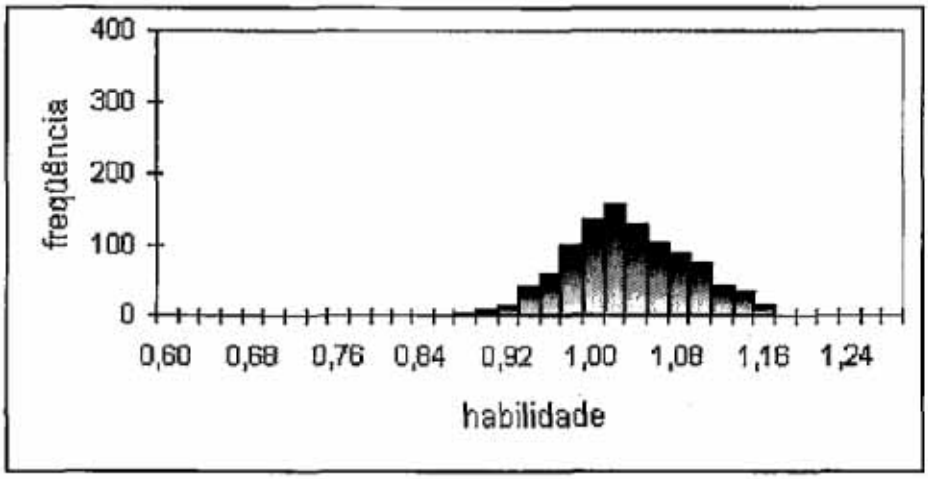


Figura 4.6
Histograma da estimativa da média da habilidade do grupo 2.
Método Grupos múltiplos com 12 itens comuns

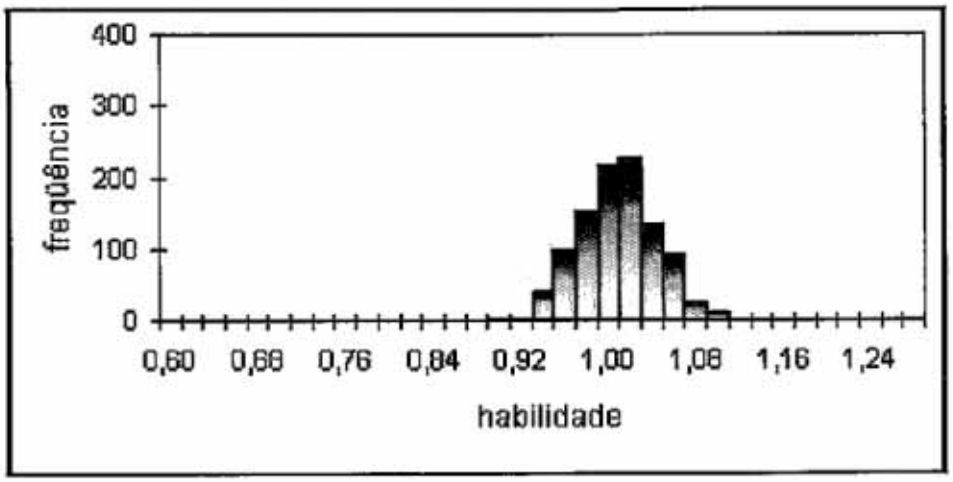


Figura 4.7
Histograma da estimativa da média da habilidade do grupo 2.
Método Um único grupo com 6 itens comuns

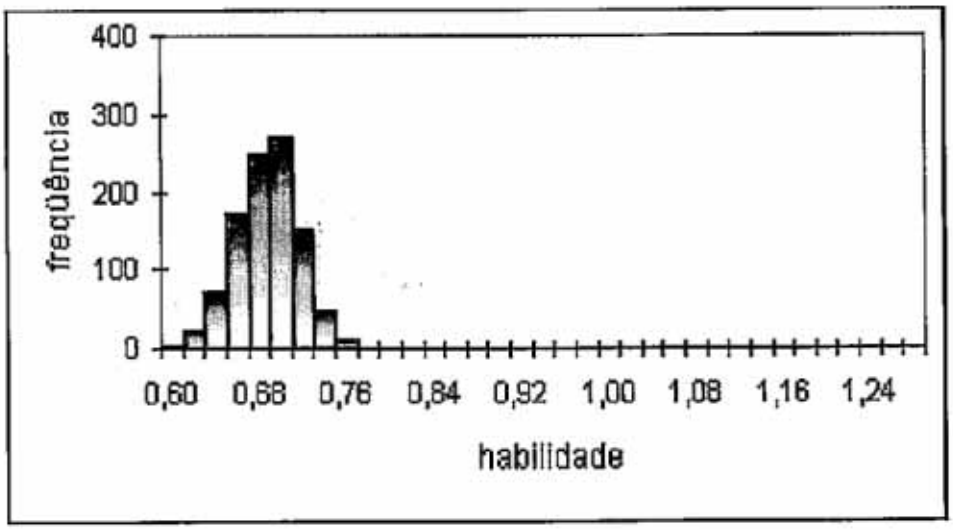


Figura 4.8
Histograma da estimativa da média da habilidade do grupo 2.
Método Mean-Sigma com 6 itens comuns

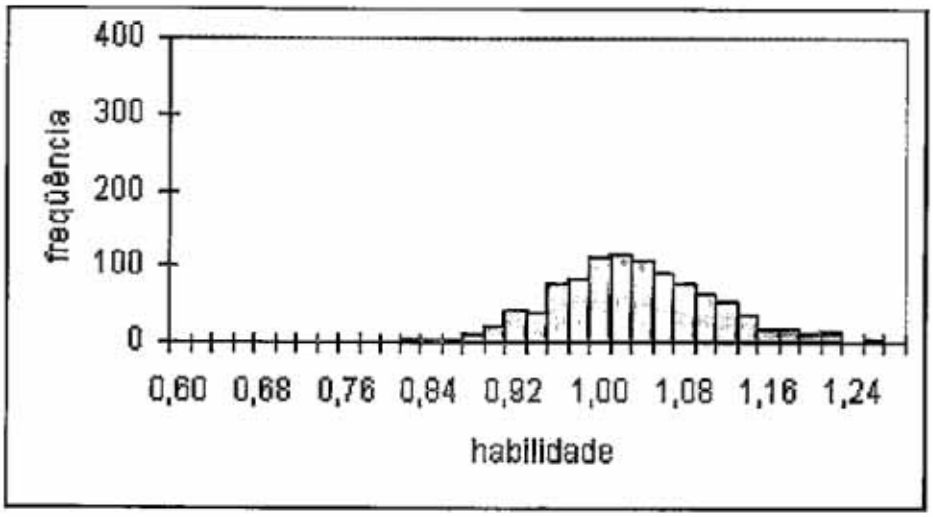
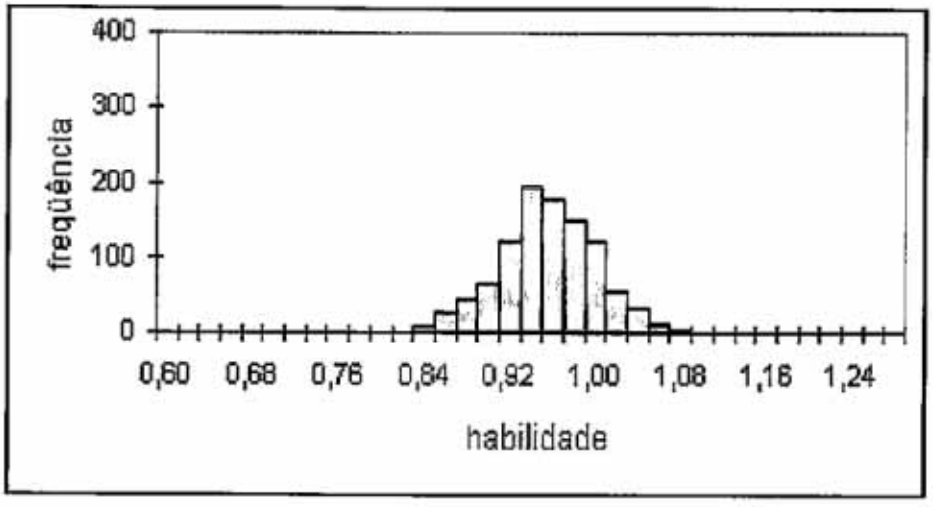


Figura 4.9
Histograma da estimativa da média da habilidade do grupo 2.
Método Grupos múltiplos com 6 itens comuns



Os histogramas indicam que o método Um único grupo teve o pior desempenho com relação à estimação da média da habilidade, fornecendo valores consistentemente menores, viés negativo, do que o verdadeiro valor 1. Quanto menor o número de itens comuns, maior o

viés. No caso do método Mean-Sigma, os resultados mostraram um viés positivo e uma variabilidade maior do que aquelas apresentadas pelos outros dois métodos. O viés apresentado por este método é bem menor do que o viés apresentado pelo método Um único grupo. O viés, ao contrário da variabilidade, não se alterou com a diminuição do número de itens comuns. Por último, o método Grupos múltiplos forneceu os melhores resultados, apresentando simultaneamente pequena variabilidade e pequeno viés. Note que o viés foi positivo para 24 itens comuns e negativo para 12 itens comuns. Nas Tabelas 4.1 e 4.2 apresentamos as estimativas das médias e os erros quadráticos médios associados a estas estimativas, para os diferentes métodos.

Tabela 4.1
Média das estimativas dos parâmetros da distribuição da habilidade do grupo 2

		Número de itens comuns		
Parâmetro	Método	24	12	6
média	Um único grupo	0,892	0,815	0,677
	Mean-Sigma	1,031	1,023	1,023
	Grupos múltiplos	1,023	0,999	0,945
variância	Um único grupo	1,025	0,994	0,987
	Mean-Sigma	1,001	1,047	1,108
	Grupos múltiplos	1,037	0,995	0,973

Tabela 4.2
Erro quadrático médio das estimativas dos parâmetros da distribuição da habilidade do grupo 2

		Número de itens comuns		
Parâmetro	Método	24	12	6
média	Um único grupo	12,201	34,754	104,959
	Mean-Sigma	3,121	3,658	6,197
	Grupos múltiplos	1,411	1,206	4,905
variância	Um único grupo	1,903	1,273	0,487
	Mean-Sigma	19,101	36,058	15,006
	Grupos múltiplos	6,941	6,094	2,508

Os resultados apresentados nas tabelas acima mostram que os métodos Mean-Sigma e Grupos múltiplos forneceram estimativas com pequenos vieses, mas com vantagens para o método Grupos múltiplos que apresentou um erro quadrático menor. Das tabelas, pode-se observar também que os três métodos forneceram boas estimativas para a variância do grupo 2, sendo que o método Um único grupo foi o que

apresentou o menor erro quadrático médio, seguido pelo método Grupos múltiplos.

Para o estudo do desempenho dos três métodos com relação a estimação dos parâmetros dos itens, criaram-se duas medidas. Uma medida da distância entre o valor médio das estimativas obtidas nas 1000 iterações e o verdadeiro valor do parâmetro, cujos resultados estão na Tabela 4.3, e a média dos erros quadráticos médios obtidos para cada um dos parâmetros, cujos resultados estão apresentados na Tabela 4.4. Nos cálculos dessas duas medidas foram considerados todos os itens aplicados às duas populações, ou seja, 72 itens para o caso de 24 itens comuns, 84 para o caso de 12 itens comuns e 90 para o caso de 6 itens comuns.

Tabela 4.3

Média dos quadrados dos desvios da média das 1000 estimativas de cada um dos parâmetros com relação ao seu verdadeiro valor

		Número de itens comuns		
Parâmetro	Método	24	12	6
A	Um único grupo	3,486	3,906	4,470
	Mean-Sigma	2,130	3,872	3,302
	Grupos múltiplos	6,733	8,477	9,432
B	Um único grupo	3,218	14,345	41,468
	Mean-Sigma	2,010	5,897	8,294
	Grupos múltiplos	3,881	5,003	9,967
C	Um único grupo	0,054	0,683	0,861
	Mean-Sigma	0,143	0,733	0,896
	Grupos múltiplos	0,061	0,689	0,983

(*) Valores multiplicados por 1000.

Tabela 4.4

Média dos erros quadráticos médios

		Número de itens comuns		
Parâmetro	Método	24	12	6
a	Um único grupo	20,628	22,773	21,069
	Mean-Sigma	23,072	27,215	26,764
	Grupos múltiplos	23,930	26,899	25,173
b	Um único grupo	11,033	29,308	62,155
	Mean-Sigma	15,080	30,142	48,023
	Grupos múltiplos	13,696	22,792	34,201
c	Um único grupo	0,556	1,802	2,083
	Mean-Sigma	0,562	1,854	2,126
	Grupos múltiplos	0,576	1,891	2,273

(*) Valores multiplicados por 1000.

Como já era esperado, quanto menor o número de itens comuns, maior é o erro cometido na estimação dos parâmetros, sendo que os comportamentos não são os mesmos para todos os três parâmetros. Para o parâmetro de discriminação, as medidas variaram um pouco na passagem de 24 para 12 itens comuns e apresentaram praticamente os mesmos valores para 12 e 6 itens comuns. Com relação ao parâmetro **c**, as medidas também apresentaram praticamente os mesmos valores para 12 e 6 itens comuns, mas apresentaram uma grande variação na passagem de 24 para 12 itens comuns. De um modo geral, pode-se dizer que os três métodos de equalização apresentaram os mesmos resultados para esses dois parâmetros. Na estimação do parâmetro de dificuldade os resultados foram bem diferentes. Tanto na passagem de 24 para 12 itens comuns, quanto na passagem de 12 para 6 itens comuns há uma perda significativa na precisão das estimativas e o método **Um único grupo** mostrou um desempenho inferior aos outros dois, que por sua vez apresentaram um desempenho bastante próximo, com uma certa superioridade para o método **Grupos múltiplos.**

5. Aplicação aos dados do SARESP 1996/97

A TRI vem sendo aplicada em diversas avaliações educacionais no Brasil e no exterior. Várias dessas aplicações estão citadas em Andrade e Valle (1998) e Andrade et al. (2000). Nesta seção, pretende-se discutir com mais detalhes uma aplicação da TRI que achamos de relevância para exemplificar uma das muitas contribuições que a aplicação desta teoria pode dar para as nossas avaliações educacionais. Ressalta-se que outras áreas do conhecimento, como, por exemplo, a pesquisa médica e o marketing, também já estão fazendo uso desta teoria.

O Sistema de Avaliação do Rendimento das Escolas do Estado de São Paulo -SARESP (ver Secretaria da Educação (1996 e 1997)) tem utilizado a TRI na análise dos seus dados dos anos de 1996 (3ª série e 7ª série do ensino fundamental), 1997 (4ª série e 8ª série do ensino fundamental) e 1998 (5ª série do ensino fundamental e 1ª série do ensino médio). Em cada um desses anos foram aplicadas provas de matemática, língua portuguesa, ciências, história e geografia no ensino fundamental e matemática, língua portuguesa, física, química e biologia no ensino médio. Iremos utilizar somente os dados de língua

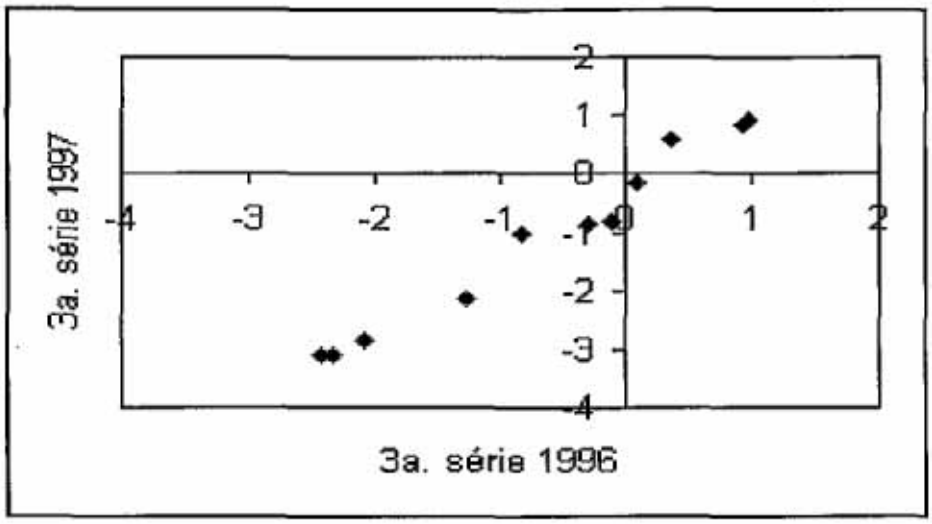
portuguesa da 3ª e 4ª séries. No momento da conclusão deste trabalho a análise dos dados de 1998 estava em andamento.

Em abril de 1996, 1/3 dos alunos da 3ª série do ensino fundamental, selecionados aleatoriamente em cada classe de todas as escolas do estado de São Paulo, responderam a uma prova de língua portuguesa com 28 questões de múltipla escolha, baseada no conteúdo da 2ª série. Em abril de 1997, utilizando o mesmo procedimento de amostragem, 1/3 dos alunos da 4ª série responderam a uma prova de língua portuguesa com 30 questões de múltipla escolha, baseada no conteúdo da 3ª série. As duas provas não tinham qualquer item em comum e , conseqüentemente, qualquer análise somente poderia ser feita para cada população considerada isoladamente, isto é, uma escala de habilidade diferente para cada uma das duas populações. Assim, nenhuma comparação entre as duas populações poderia ser feita.

De modo a viabilizar comparações entre as duas populações de interesse, elaborou-se uma terceira prova, chamada de prova de ligação, com 32 itens (11 itens da prova de 1996 e 21 itens da prova de 1997) que foi aplicada em outubro de 1997 a uma amostra de alunos de uma terceira população (3ª série de 1997). A amostra utilizada para a terceira população, chamada de população de ligação, foi definida de modo a ter-se uma boa calibração de todos os 58 itens, em uma mesma escala de habilidade. Não havia qualquer interesse em estimar a distribuição das habilidades dos indivíduos dessa terceira população. O uso de um número diferente de itens de 1996 e de 1997 na terceira prova deveu-se ao melhor desempenho dos itens da prova de 1997; desempenho este avaliado pela aplicação da TRI nos dados isolados de cada um dos dois anos.

As três formas de equalização descritas no tópico 4 foram aplicadas aos dados de uma amostra aleatória dos alunos que realizaram as provas, com o grupo da 3ª série de 1997 considerado como referência. Para efeito de ilustração do método **Mean-Sigma** apresentamos as estimativas dos parâmetros dos itens comuns nas Tabelas A.4 e A.5 no Apêndice. Por exemplo, a partir das estimativas do parâmetro de dificuldade apresentadas na Tabela A.4, pode-se concluir que o desempenho dos respondentes da 3ª série de 1996 foi inferior ao desempenho dos respondentes da 3ª série de 1997, tendo em vista que estes valores foram maiores para o grupo de 1996 do que para o grupo de 1997. Em outras palavras, o mesmo item foi mais difícil para o grupo de 1996. A figura a seguir mostra a relação linear existente entre os valores do parâmetro **b** nos dois grupos.

Figura 5.1
Gráfico de dispersão das estimativas do parâmetro de dificuldade dos itens comuns aos grupos 3ª série 1996 e 3ª série 1997



A equalização dos valores obtidos para a 3ª série 1996 na escala da 3ª série de 1997, o grupo de referência com média 0 e desvio padrão 1, pode ser realizada a partir dos resultados apresentados na seção **3.2.**, notando-se que as médias e desvios padrão amostrais das estimativas do parâmetro de dificuldade são $M_{3s96} = -0,638$, $M_{3s97} = -1,101$, $S_{3s96} = 1,246$ e $S_{3s97} = 1,544$, respectivamente. Os coeficientes da transformação são dados por

$$\alpha = S_{3s97} / S_{3s96} = 1,544/1,246 = 1,239$$

e

$$\beta = M_{3s97} - \alpha M_{3s96} = -1,101 - 1,239x(-0,638) = -0,311$$

fazendo com que as estimativas da média e do desvio padrão da distribuição das habilidades da 3ª série de 1996, na escala da 3ª série de 1997, sejam –0,311 (=1,239x0-0,311) e 1,239 (=1,239x1), respectivamente, notando-se que, no processo de calibração, estes valores haviam sido fixados como iguais a 0 e 1. A Tabela 5.1 apresenta os resultados obtidos com os três métodos de equalização.

Tabela 5.1
Estimativa dos parâmetros da distribuição da habilidade

	Média		Desvio padrão	
Método	3ª série 1996	4ª série 1997	3ª série 1996	4ª série 1997
Um único grupo	-0,128	0,137	1,136	1,072
Mean-Sigma	-0,311	0,279	1,239	0,858
Grupos múltiplos	-0,375	0,299	1,277	0,940

Os resultados obtidos nesta equalização correspondem aos resultados obtidos com a simulação, onde tivemos estimativas das habilidades médias bem diferentes com o método **Um único grupo** e valores próximos para a medida de variabilidade. A diferença entre as estimativas das habilidades médias obtidas com os métodos **Mean-Sigma** e **Grupos múltiplos** é uma indicação de que estes dois procedimentos de equalização podem fornecer estimativas diferentes, dependendo do número de grupos envolvidos no estudo. Vale a pena ressaltar que, enquanto o método **Mean-Sigma** exige o uso de várias transformações, dependendo do número de grupos, para terem-se todos os parâmetros dos itens e habilidades na mesma escala, o método **Grupos múltiplos** resolve este problema em uma única calibração fornecendo, pelo menos teoricamente, estimativas com menores erros padrão.

Para finalizar este exemplo de aplicação da TRI na análise dos dados do SARESP, apresentamos, nas figuras a seguir, os histogramas gerados a partir das estimativas das distribuições a posteriori das habilidades dos respondentes de 1996 e 1997, fornecidas pelo programa Bilog-MG.

Figura 5.2
Histograma das habilidades em língua portuguesa dos alunos
da 3ª série de 1996

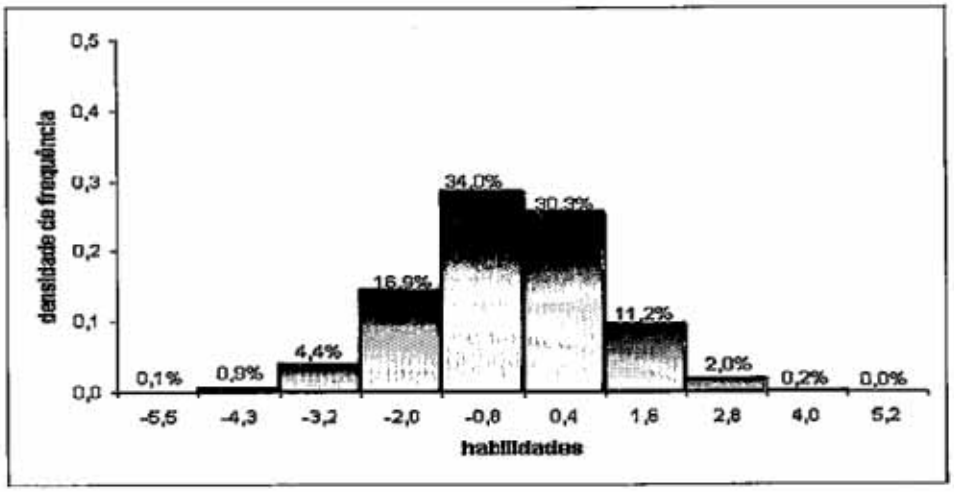
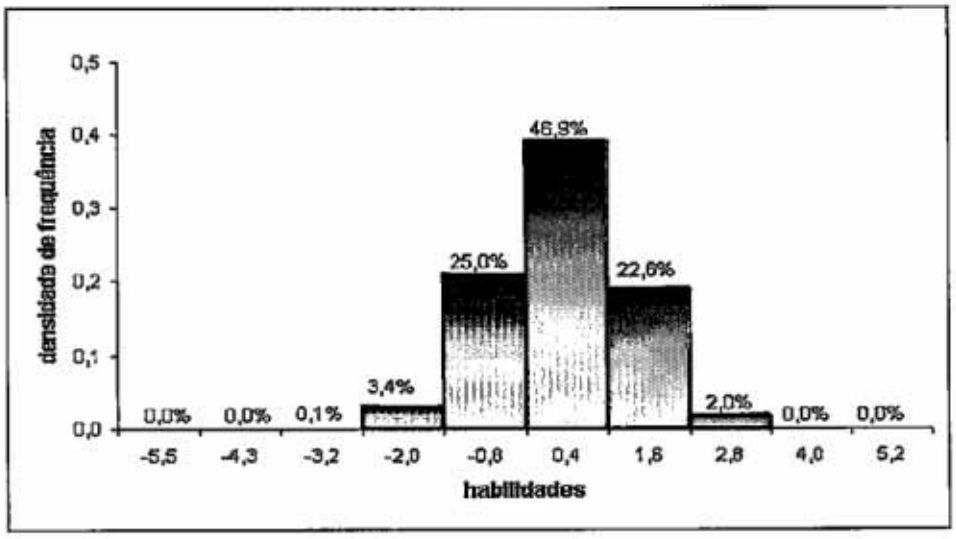


Figura 5.3
Histograma das habilidades em língua portuguesa dos alunos
da 4ª série de 1997



A partir dos histogramas, pode-se concluir que a habilidade média dos alunos da 4ª série de 1997 é maior do que a habilidade

média dos alunos da 3ª série de 1996, e também que a variabilidade das habilidades dos alunos da 4ª série de 1997 é menor do que a variabilidade das habilidades dos alunos da 3ª série de 1996. Dos histogramas pode-se também obter, a partir do cálculo de áreas de retângulos, a porcentagem de alunos em determinado intervalo de habilidade. Por exemplo, enquanto que, em 1996, a porcentagem de alunos da 3ª série do ensino fundamental que tinham habilidade maior ou igual a 1 foi de 13,5% (=11,2% + 2,0% + 0,2%), em 1997 a porcentagem de alunos da 4ª série do ensino fundamental que tinha habilidade maior ou igual a 1 foi de 24,6% (=22,6% + 2,0%). Estes valores mostram nos que houve um aumento de 11,1 pontos percentuais na porcentagem de alunos com habilidade igual ou maior do que 1 de 1996 para 1997.

Com todos os 58 itens calibrados na mesma escala, pode-se também construir uma interpretação pedagógica para esta escala nos níveis âncora –2, -1, 0, 1, 2. Sugere-se que o leitor consulte Beaton e Allen (1992) para a definição e obtenção de níveis âncora e Secretaria da Educação (1996,1997), para maiores resultados sobre as escalas construídas e suas interpretações pedagógicas.

6. Conclusões e sugestões

Neste trabalho procurou-se apresentar as idéias e modelos básicos da Teoria da Resposta ao Item, com o objetivo de mostrar o grande potencial da aplicação desta teoria em avaliação educacional, em particular na solução do problema da comparação do desempenho de dois ou mais grupos de alunos. Apesar desta teoria ter mais de 50 anos, somente nos últimos 15 anos é que ela vem sendo aplicada em larga escala nas principais avaliações educacionais em diferentes países. Atribui-se este fato à complexidade matemática dos métodos envolvidos e também à ausência de programas computacionais eficientes. A aplicação apropriada desta teoria exige necessariamente o envolvimento de especialistas em educação e em estatística. Sua primeira aplicação no Brasil foi na análise do SAEB95.

Alguns pontos têm sido levantados na literatura sobre a adequação desta teoria. Dois deles que consideramos importantes são a dimensionalidade do espaço de traços latentes envolvidos na avaliação e o assunto tratado neste trabalho, ou seja a equalização de diferentes avaliações. Como exemplos do segundo ponto, destacamos as equalizações dos resultados do SAEB95 e do SAEB97, do SARESP96

e do SARESP97 e de resultados de avaliações estaduais com o SAEB. Com relação ao primeiro ponto, alguns autores têm defendido a tese de que os modelos unidimensionais têm fornecido bons resultados mesmo em situações multidimensionais, mas com uma das dimensões predominante. Mais recentemente, modelos para mais de uma dimensão têm sido propostos. Com relação ao problema da equalização, conforme demonstrado neste trabalho, a proposta recente de modelos para Grupos múltiplos de Bock e Zimowski (1997) deu um novo rumo à solução deste problema, tendo em vista que os modelos anteriores envolvem outros erros de modelagem além daqueles da própria teoria. Sugerimos a leitura de Goldstein e Wood (1989), Mislevy (1992), Goldstein (1994) e Hedges e Vevea (1997), entre outros, para um melhor entendimento destes problemas e suas soluções.

Para finalizar, gostaríamos de ressaltar que, apesar de não termos dúvidas de que a aplicação desta teoria tem muito a contribuir para a melhora de nossas avaliações educacionais, sua disseminação dependerá muito da integração de especialistas das áreas de estatística e educação. A criação de programas de pós-graduação, envolvendo departamentos de estatística e de medidas em educação em algumas de nossas universidades, seria de fundamental importância.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Programa Nacional de Excelência (PRONEX), contrato nº 23800.96/021-05-01, e pelo Projeto Temático da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), contrato nº 96/01741-7.

Gostaria de agradecer ao Prof. Helinton R. Tavares e a Raquel C. Valle pelos cálculos que tornaram possíveis os itens 4 e 5 e, também, à Secretaria de Estado da Educação de São Paulo pela utilização de parte dos dados SARESP 96/97.

APÊNDICE

Tabela A.1: Parâmetros dos itens das provas com 24 itens comuns

	Parâmetro				Parâmetro		
Item	a	b	c	Item	a	b	c
1	1,0	-2,0	0,15	37	1,2	0,6	0,15
2	1,0	-2,0	0,25	38	1,2	0,6	0,25
3	1,0	-1,5	0,15	39	1,2	1,0	0,15
4	1,0	-1,5	0,25	40	1,2	1,0	0,25
5	1,0	-1,0	0,15	41	1,4	0,0	0,15
6	1,0	-1,0	0,25	42	1,4	0,0	0,25
7	1,0	-0,5	0,15	43	1,4	0,3	0,15
8	1,0	-0,5	0,25	44	1,4	0,3	0,25
9	1,2	-2,0	0,15	45	1,4	0,6	0,15
10	1,2	-2,0	0,25	46	1,4	0,6	0,25
11	1,2	-1,5	0,15	47	1,4	1,0	0,15
12	1,2	-1,5	0,25	48	1,4	1,0	0,25
13	1,2	-1,0	0,15	49	1,0	1,5	0,15
14	1,2	-1,0	0,25	50	1,0	1,5	0,25
15	1,2	-0,5	0,15	51	1,0	2,0	0,15
16	1,2	-0,5	0,25	52	1,0	2,0	0,25
17	1,4	-2,0	0,15	53	1,0	2,5	0,15
18	1,4	-2,0	0,25	54	1,0	2,5	0,25
19	1,4	-1,5	0,15	55	1,0	3,0	0,15
20	1,4	-1,5	0,25	56	1,0	3,0	0,25
21	1,4	-1,0	0,15	57	1,2	1,5	0,15
22	1,4	-1,0	0,25	58	1,2	1,5	0,25
23	1,4	-0,5	0,15	59	1,2	2,0	0,15
24	1,4	-0,5	0,25	60	1,2	2,0	0,25
25	1,0	0,0	0,15	61	1,2	2,5	0,15
26	1,0	0,0	0,25	62	1,2	2,5	0,25
27	1,0	0,3	0,15	63	1,2	3,0	0,15
28	1,0	0,3	0,25	64	1,2	3,0	0,25
29	1,0	0,6	0,15	65	1,4	1,5	0,15
30	1,0	0,6	0,25	66	1,4	1,5	0,25
31	1,0	1,0	0,15	67	1,4	2,0	0,15
32	1,0	1,0	0,25	68	1,4	2,0	0,25
33	1,2	0,0	0,15	69	1,4	2,5	0,15
34	1,2	0,0	0,25	70	1,4	2,5	0,25
35	1,2	0,3	0,15	71	1,4	3,0	0,15
36	1,2	0,3	0,25	72	1,4	3,0	0,25

Tabela A.2: Parâmetros dos itens das provas com 12 itens comuns

Item	Parâmetro			Item	Parâmetro		
	a	b	c		a	b	c
1	0,6	-2,0	0,15	43	1,2	0,6	0,15
2	0,6	-1,5	0,15	44	1,2	1,0	0,15
3	0,6	-1,5	0,25	45	1,4	0,0	0,15
4	0,6	-1,0	0,15	46	1,4	0,3	0,15
5	0,6	-1,0	0,25	47	1,4	0,6	0,15
6	0,6	-0,6	0,15	48	1,4	1,0	0,15
7	0,6	-0,6	0,25	49	0,6	1,3	0,15
8	0,6	-0,3	0,15	50	0,6	1,3	0,25
9	0,6	-0,3	0,25	51	0,6	1,6	0,15
10	1,0	-2,0	0,15	52	0,6	1,6	0,25
11	1,0	-1,5	0,15	53	0,6	2,0	0,15
12	1,0	-1,5	0,25	54	0,6	2,0	0,25
13	1,0	-1,0	0,15	55	0,6	2,5	0,15
14	1,0	-1,0	0,25	56	0,6	2,5	0,25
15	1,0	-0,6	0,15	57	0,6	3,0	0,15
16	1,0	-0,6	0,25	58	1,0	1,3	0,15
17	1,0	-0,3	0,15	59	1,0	1,3	0,25
18	1,0	-0,3	0,25	60	1,0	1,6	0,15
19	1,2	-2,0	0,15	61	1,0	1,6	0,25
20	1,2	-1,5	0,15	62	1,0	2,0	0,15
21	1,2	-1,5	0,25	63	1,0	2,0	0,25
22	1,2	-1,0	0,15	64	1,0	2,5	0,15
23	1,2	-1,0	0,25	65	1,0	2,5	0,25
24	1,2	-0,6	0,15	66	1,0	3,0	0,15
25	1,2	-0,6	0,25	67	1,2	1,3	0,15
26	1,2	-0,3	0,15	68	1,2	1,3	0,25
27	1,2	-0,3	0,25	69	1,2	1,6	0,15
28	1,4	-2,0	0,15	70	1,2	1,6	0,25
29	1,4	-1,5	0,15	71	1,2	2,0	0,15
30	1,4	-1,5	0,25	72	1,2	2,0	0,25
31	1,4	-1,0	0,15	73	1,2	2,5	0,15
32	1,4	-1,0	0,25	74	1,2	2,5	0,25
33	1,4	-0,6	0,15	75	1,2	3,0	0,15
34	1,4	-0,6	0,25	76	1,4	1,3	0,15
35	1,4	-0,3	0,15	77	1,4	1,3	0,25
36	1,4	-0,3	0,25	78	1,4	1,6	0,15
37	1,0	0,0	0,15	79	1,4	1,6	0,25
38	1,0	0,3	0,15	80	1,4	2,0	0,15
39	1,0	0,6	0,15	81	1,4	2,0	0,25
40	1,0	1,0	0,15	82	1,4	2,5	0,15
41	1,2	0,0	0,15	83	1,4	2,5	0,25
42	1,2	0,3	0,15	84	1,4	3,0	0,15

Tabela A.3: Parâmetros dos itens das provas com 6 itens comuns

	Parâmetro				Parâmetro		
Item	a	b	c	Item	a	b	c
1	0,6	-2,0	0,15	46	1,2	0,0	0,15
2	0,6	-2,0	0,25	47	1,2	0,5	0,15
3	0,6	-1,5	0,15	48	1,2	1,0	0,15
4	0,6	-1,5	0,25	49	0,6	1,3	0,15
5	0,6	-1,3	0,15	50	0,6	1,3	0,25
6	0,6	-1,3	0,25	51	0,6	1,5	0,15
7	0,6	-1,0	0,15	52	0,6	1,5	0,25
8	0,6	-1,0	0,25	53	0,6	1,7	0,15
9	0,6	-0,7	0,15	54	0,6	1,7	0,25
10	0,6	-0,7	0,25	55	0,6	2,0	0,15
11	0,6	-0,5	0,15	56	0,6	2,0	0,25
12	0,6	-0,5	0,25	57	0,6	2,3	0,15
13	0,6	-0,3	0,15	58	0,6	2,3	0,25
14	0,6	-0,3	0,25	59	0,6	2,5	0,15
15	1,0	-2,0	0,15	60	0,6	2,5	0,25
16	1,0	-2,0	0,25	61	0,6	3,0	0,15
17	1,0	-1,5	0,15	62	0,6	3,0	0,25
18	1,0	-1,5	0,25	63	1,0	1,3	0,15
19	1,0	-1,3	0,15	64	1,0	1,3	0,25
20	1,0	-1,3	0,25	65	1,0	1,5	0,15
21	1,0	-1,0	0,15	66	1,0	1,5	0,25
22	1,0	-1,0	0,25	67	1,0	1,7	0,15
23	1,0	-0,7	0,15	68	1,0	1,7	0,25
24	1,0	-0,7	0,25	69	1,0	2,0	0,15
25	1,0	-0,5	0,15	70	1,0	2,0	0,25
26	1,0	-0,5	0,25	71	1,0	2,3	0,15
27	1,0	-0,3	0,15	72	1,0	2,3	0,25
28	1,0	-0,3	0,25	73	1,0	2,5	0,15
29	1,2	-2,0	0,15	74	1,0	2,5	0,25
30	1,2	-2,0	0,25	75	1,0	3,0	0,15
31	1,2	-1,5	0,15	76	1,0	3,0	0,25
32	1,2	-1,5	0,25	77	1,2	1,3	0,15
33	1,2	-1,3	0,15	78	1,2	1,3	0,25
34	1,2	-1,3	0,25	79	1,2	1,5	0,15
35	1,2	-1,0	0,15	80	1,2	1,5	0,25
36	1,2	-1,0	0,25	81	1,2	1,7	0,15
37	1,2	-0,7	0,15	82	1,2	1,7	0,25
38	1,2	-0,7	0,25	83	1,2	2,0	0,15
39	1,2	-0,5	0,15	84	1,2	2,0	0,25
40	1,2	-0,5	0,25	85	1,2	2,3	0,15
41	1,2	-0,3	0,15	86	1,2	2,3	0,25
42	1,2	-0,3	0,25	87	1,2	2,5	0,15
43	1,0	0,0	0,15	88	1,2	2,5	0,25
44	1,0	0,5	0,15	89	1,2	3,0	0,15
45	1,0	1,0	0,15	90	1,2	3,0	0,25

Tabela A.4: Estimativas dos parâmetros dos itens comuns aos grupos 3ª série 96 e 3ª série 97

	Parâmetro a		Parâmero b		Parâmetro c	
Item	3ª. 96	3ª 97	3ª. 96	3ª 97	3ª. 96	3ª 97
C3S01	1,37	1,04	-1,27	-2,18	0,01	0,01
C3S02	2,29	1,33	-0,30	-0,90	0,01	0,01
C3S03	2,24	1,18	0,09	-0,18	0,01	0,01
C3S04	1,25	1,08	-2,33	-3,12	0,20	0,25
C3S05	1,63	1,54	-2,09	-2,90	0,21	0,24
C3S06	1,32	1,57	-2,43	-3,14	0,19	0,24
C3S07	1,03	0,79	0,35	0,54	0,22	0,19
C3S08	1,04	0,80	0,96	0,88	0,29	0,25
C3S09	1,37	1,70	0,94	0,82	0,29	0,27
C3S10	0,85	1,17	-0,83	-1,05	0,19	0,23
C3S11	0,99	1,56	-0,12	-0,88	0,22	0,17

Tabela A.5: Estimativas dos parâmetros dos itens comuns aos grupos 4ª série 97 e 3ª série 97.

	Parâmetro a		Parâmero b		Parâmetro c	
Item	4ª 97	3ª 97	4ª 97	3ª 97	4ª 97	3ª 97
C34S01	1,38	1,55	1,93	1,75	0,16	0,18
C34S02	1,69	0,86	-0,03	-0,54	0,25	0,26
C34S03	0,92	0,74	0,61	1,26	0,21	0,25
C34S04	1,34	1,39	-0,53	-0,12	0,18	0,23
C34S05	1,54	1,05	-0,12	0,26	0,22	0,19
C34S06	1,43	1,59	1,73	1,77	0,12	0,12
C34S07	1,19	0,99	0,11	0,18	0,12	0,13
C34S08	2,22	1,65	0,17	-0,03	0,23	0,13
C34S09	1,83	1,58	1,72	1,04	0,25	0,18
C34S10	0,94	1,97	1,89	1,82	0,11	0,12
C34S11	0,97	1,25	-1,00	-0,86	0,27	0,26
C34S12	1,50	1,37	-0,12	-0,06	0,34	0,23
C34S13	0,48	0,52	-0,50	0,42	0,21	0,27
C34S14	1,64	2,29	1,05	1,37	0,25	0,30
C34S15	1,43	1,42	-0,93	-0,07	0,13	0,18
C34S16	1,25	1,70	-0,64	-0,09	0,20	0,18
C34S17	1,12	1,26	-0,60	0,23	0,14	0,13
C34S18	1,65	1,75	0,31	0,51	0,14	0,10
C34S19	1,81	1,81	1,26	1,78	0,13	0,12
C34S20	1,99	2,39	1,13	1,26	0,10	0,13
C34S21	1,79	1,60	0,11	0,47	0,21	0,16

Referências bibliográficas

ANDRADE, Dalton F.; VALLE, Raquel da C. Introdução à teoria da resposta ao item: conceitos e aplicações. *Estudos em Avaliação Educacional*. São Paulo: Fundação Carlos Chagas. (18), p. 13-32, 1998.

ANDRADE, Dalton F.; KLEIN, Rubem. (1999). Métodos estatísticos para avaliação educacional: teoria da resposta ao item. *Boletim da ABE*, (43), p. 21-28, 1999.

ANDRADE, Dalton F.; TAVARES, H. R.; VALLE, Raquel da C. Teoria da Respostas ao Item – Conceitos e Aplicações. ABE/Associação Brasileira de Estatística. *14ª SINAPE*, Caxambu, 154 p, 2000.

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

BEATON, A. E.; ALLEN, N. L. Interpreting scales through scale anchoring. *Journal of Educational Statistics*. (17), p. 191-204, 1992.

BOCK, R. D.; ZIMOWSKI, M. F. Multiple Group IRT. In: W. J. van der LINDEN; R. K. HAMBLETON (eds). *Handbook of Modern Item Response Theory*. Eds. New York: Spring-Verlag. 1997.

GOLDSTEIN, H. Recontextualizing mental measurement. *Educational Measurement: Issues and Practice*. (13), p. 16-43. 1994.

GOLDSTEIN, H.; WOOD, R. Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*. (42), p. 139-167. 1989.

GULLIKSEN, H. *Theory of Mental Tests*. New York: John Wiley and Sons. 1950.

HAMBLETON R. K.; SWAMINATHAN, H.; Rogers, H. J. *Fundamentals of Item Response Theory*. Newburry Park: Sage Publications. 1991.

HEDGES, L. V.; VEVEA, J. L. A study of equating in NAEP. Paper presented at *The NAEP* Validity Studies Panel. Palo Alto: American Institutes for Research. 1997.

KOLEN, Michael J.; BRENNAN, Robert L. *Test Equating – Methods and Practices*. New York : Springer, 1995.

LORD, F. M. *Applications of Item Response Theory to Practical Testing Problems*. 1980. Hillsdade: Lawrence Erlbaum Associates. 1980.

LORD, F. M.; NORVICK, M. R. *Statistical Theories of Mental Test Score*. Reading: Addison-Wesley. 1968.

Ministério da Educação e do Desporto. *Sistema Nacional de Avaliação da Educação Básica: SAEB 95 – relatório técnico*. São Paulo/Rio de Janeiro: Fundação Carlos Chagas/Fundação Cesgranrio. 1995.

MISLEVY, R. J. Recent Developments in the Factor Analysis of Categorical Variables. *Journal of Educational Statistics*. (11), p. 3-31. 1986.

MISLEVY, R. J. *Linking Educational Assessments: concepts, issues, methods, and prospects*. Princeton: Educational Testing Service. 1992.

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

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

SECRETARIA DA EDUCAÇÃO DO ESTADO DE SÃO PAULO. *Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo – SARESP: relatório final dos resultados*, 3v. São Paulo: SEE. 1996.

SECRETARIA DA EDUCAÇÃO DO ESTADO DE SÃO PAULO. *Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo – SARESP: relatório final dos resultados*, 4v. São Paulo: SEE. 1997.

VALLE, Raquel da C. Teoria da Resposta ao Item. *Estudos em Avaliação Educacional*. São Paulo : Fundação Carlos Chagas, (21), p. 7 - 91. 2000.

VIANNA, Heraldo M. *Testes em Educação*. São Paulo : IBRASA. 1987.

ZIMOWISK, M. F.; MURAKI, E.; MISLEVY, R. J.; BOCK, R. D. *Bilog-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items*. Chicago: Scientific Software Inc., 1996.