TEORIA DE RESPOSTA AO ITEM

RAQUEL DA CUNHA VALLE[1]

1.0 INTRODUÇÃO

Resultados obtidos em provas, expressos apenas por seus escores brutos ou padronizados, têm sido tradicionalmente utilizados nos processos de avaliação e seleção de indivíduos. No entanto, 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 que é a característica principal da Teoria Clássica das Medidas. Assim, torna-se inviável a comparação entre indivíduos que não foram submetidos às mesmas provas, ou pelo menos, ao que se denomina de formas paralelas de testes. Maiores detalhes sobre essa metodologia, incluindo sua fundamentação matemática, podem ser encontrados em Gulliksen (1950), Lord e Novick (1968) e Vianna (1987), entre outros.

Atualmente, na área educacional, vem crescendo o interesse na aplicação de técnicas derivadas da Teoria da Resposta ao Item – TRI, que propõe modelos para os traços latentes, ou seja, características do indivíduo que não podem ser observadas diretamente. Esse tipo de variável deve ser inferida a partir da observação de variáveis secundárias que estejam relacionadas a ela. O que esta metodologia sugere são algumas formas de representar a relação entre a probabilidade de um aluno responder corretamente a um item e seus traços latentes ou habilidades na área de conhecimento avaliada.

Uma das grandes vantagens da TRI sobre a Teoria Clássica é que ela permite a comparação entre populações, desde que submetidas a provas que tenham alguns itens comuns, ou ainda, a comparação entre indivíduos da mesma população que tenham sido submetidos a provas

1 Do Departamento de Pesquisas Educacionais da Fundação Carlos Chagas, São Paulo – SP.

totalmente diferentes. Isto porque uma das principais características da TRI é que ela tem como elementos centrais os itens, e não a prova como um todo.

Assim, várias questões de interesse prático na área da Educação podem ser respondidas. É possível por exemplo, avaliar o desenvolvimento de uma determinada série de um ano para outro ou comparar o desempenho entre escolas públicas e privadas.

O objetivo principal deste trabalho é apresentar os conceitos básicos envolvidos na TRI e algumas de suas aplicações em avaliações educacionais brasileiras. Em Lord (1980) e Hambleton, Swaminathan e Rogers (1991), por exemplo, pode-se encontrar maiores detalhes sobre os fundamentos e aplicações desta teoria. No Item 2 são apresentados os modelos matemáticos, com suas interpretações e suposições básicas. No Item 3, discute-se o processo de estimação dos parâmetros dos itens e das habilidades dos respondentes. No Item 4, é introduzido o conceito de equalização ("equating"), a partir do qual torna-se possível a comparação entre populações. Aqui também discute-se a construção e interpretação de escalas de habilidades por meio desta teoria. No Item 5, discutem-se os recursos computacionais disponíveis. No Item 6, apresenta-se uma aplicação da TRI na análise de dados obtidos pelo Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP), nos anos de 1996 e 1997. Finalmente, as conclusões e sugestões estão no Item 7.

2.0 Modelos Matemáticos

A TRI propõe a utilização de modelos que representam a probabilidade de um indivíduo responder corretamente a um item como função dos parâmetros do item e da habilidade (ou habilidades) do respondente. Essa relação é sempre expressa de tal forma que quanto maior a habilidade, maior a probabilidade de acerto no item. Os vários modelos propostos na literatura dependem fundamentalmente de três fatores:

(i) da natureza do item – dicotômicos ou não dicotômicos;
(ii) do número de populações envolvidas – apenas uma ou mais de uma;
(iii) e da quantidade de traços latentes que está sendo medida – apenas um ou mais de um.

Com relação ao ponto (iii), cabe ressaltar que neste trabalho estaremos sempre considerando modelos que avaliam apenas um traço latente ou habilidade. Alguns modelos que consideram que mais de uma habilidade está sendo medida, chamados de modelos multidimensionais, podem ser encontrados, por exemplo em Linden e Hambleton (1997).

A seguir, apresentaremos alguns dos modelos mais utilizados.

2.1 Modelos envolvendo um único grupo

Em primeiro lugar, é importante definir os conceitos de grupo e população, que serão largamente utilizados neste e nos demais tópicos. Quando usarmos o termo grupo, estaremos nos referindo a uma amostra de indivíduos de uma população. Neste trabalho, o conceito de grupo está diretamente ligado ao processo de amostragem – e estaremos sempre considerando o processo de amostragem aleatória simples. Portanto, quando falarmos em um único grupo de respondentes, nos referimos a uma amostra de indivíduos retirada de uma mesma população. Consequentemente, dois grupos – ou mais – de respondentes são dois conjuntos distintos de indivíduos, que foram amostrados de duas – ou mais – populações.

Na área de Avaliações Educacionais é comum que uma população seja definida por determinadas características que podem variar, dependendo dos objetivos do estudo, e portanto, podem ou não ser relevantes para a diferenciação de uma população de outra, dependendo do caso.

Por exemplo, pode-se considerar que a 5ª série do Ensino Fundamental de São Paulo é a população alvo. Daí, toma-se uma única amostra dos alunos dessa população, composta de alunos do período diurno e do noturno. Nesse caso, temos então um único grupo de respondentes. Já em outro estudo, poderíamos considerar a 5ª série diurna e a 5ª série noturna do Ensino Fundamental de São Paulo como duas populações de interesse. Então, seriam tomadas duas amostras: uma dos alunos do período diurno e outra dos alunos do noturno. Nessa situação, teríamos dois grupos de alunos. Portanto, é pelo próprio processo de amostragem do estudo que pode-se identificar quantas (e quais) populações estão envolvidas.

Exemplos do que usualmente são consideradas como populações distintas são: séries distintas (3ª série e 4ª série); períodos distintos (diurno

e noturno); uma mesma série, mas em anos distintos (3ª série de 1996 e 3ª série de 1997), etc.

A seguir, apresentaremos então os modelos mais utilizados quando um teste é aplicado a um único grupo de respondentes.

2.1.1 Modelos para itens dicotômicos

Aqui são incluídos tanto os modelos para a análise de itens de múltipla escolha dicotomizados (corrigidos como certo ou errado) quanto para a análise de itens abertos (de resposta livre) que são avaliados de forma dicotomizada.

Os primeiros modelos de resposta ao item surgiram na década de 50, e eram modelos em que se considerava que uma única habilidade, de um único grupo, estava sendo medida por um teste onde os itens eram corrigidos de maneira dicotômica. Estes modelos foram primeiramente desenvolvidos na forma de uma função ogiva normal e depois, foram descritos para uma forma matemática mais conveniente, e que vem sendo usada até então: a logística.

Na prática, esses modelos logísticos para itens binários são os modelos de resposta ao item mais utilizados, sendo que há basicamente três tipos, que se diferenciam pelo número de parâmetros que utilizam para descrever o item – os modelos logísticos de 1, 2 e 3 parâmetros, que consideram, respectivamente:

(i) somente a dificuldade do item;

(ii) a dificuldade e a discriminação;

(iii) a dificuldade, a discriminação e a probabilidade de resposta correta dada por indivíduos de baixa habilidade.

Lord (1952) foi o primeiro a desenvolver o modelo unidimensional de 2 parâmetros, baseado na distribuição normal acumulada (ogiva normal). Após algumas aplicações desse modelo, o próprio Lord sentiu a necessidade da incorporação de um parâmetro que tratasse do problema do acerto casual. Assim, surgiu o modelo de 3 parâmetros. Anos mais tarde, Birnbaum (1968) substituiu, em ambos os modelos, a função ogiva normal pela função logística, matematicamente mais conveniente, pois é uma função explícita dos parâmetros do item e de habilidade e não envolve integração.

Já o modelo unidimensional de 1 parâmetro veio depois. Foi inicialmente proposto por Rasch (1960), expresso também como modelo de ogiva normal e, mais tarde foi descrito como um modelo logístico por Wright (1968).

Neste trabalho, daremos maior ênfase à explicação do modelo logístico de 3 parâmetros, uma vez que é o mais completo e portanto os outros dois podem ser facilmente obtidos a partir dele.

2.1.1.1 Modelo logístico de 3 parâmetros

Definição:

Dentre os modelos propostos pela TRI, o **modelo logístico unidimensional de 3 parâmetros** é atualmente o mais utilizado e é dado por:

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

com $i = 1, 2, ..., I$, $j = 1, 2, ..., n$, 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 (traço latente) do j-ésimo indivíduo.

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

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.

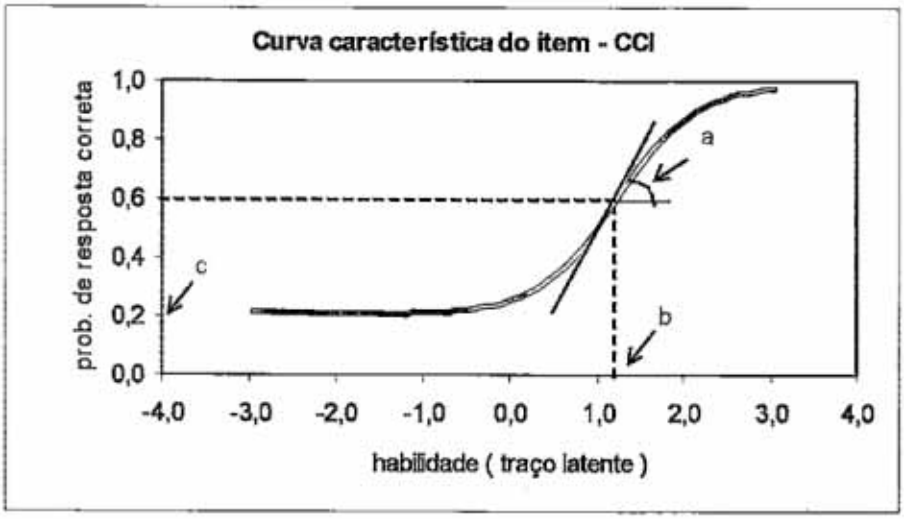
c_i é o parâmetro do item que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente ao item i (muitas vezes referido como a probabilidade de acerto casual).

D é um fator de escala, constante e igual a 1. Utiliza-se o valor 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal.

Interpretação e representação gráfica

Note que $P(X_{ij} = 1 \mid \theta_j)$ pode ser vista como a proporção de respostas corretas ao item i dentre todos os indivíduos da população com habilidade θ_j. A relação existente entre $P(X_{ij} = 1 \mid \theta_j)$ e os parâmetros do modelo é mostrada na figura a seguir, que é chamada de Curva Característica do Item – CCI.

Figura 2.1 Exemplo de uma 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.

A escala da habilidade é uma escala **arbitrária** onde o importante são as relações de ordem existentes entre seus pontos e não necessariamente sua magnitude. O parâmetro **b** é medido na mesma unidade da habilidade e o parâmetro **c** não depende da escala, pois trata-se de uma probabilidade, e como tal, assume sempre valores entre 0 e 1.

Na realidade, o parâmetro **b** representa a habilidade necessária para uma probabilidade de acerto igual a (1+c)/2. Assim, quanto maior o valor de **b**, mais difícil é o item, e vice-versa.

O parâmetro **c** representa a probabilidade de um aluno com baixa habilidade responder corretamente ao item e é muitas vezes referido como a probabilidade de acerto ao acaso. Então, quando não é permitido "chutar", **c** é igual a 0 e **b** representa o ponto na escala da habilidade onde a probabilidade de acertar ao item é 50%.

O parâmetro **a** é proporcional à derivada da tangente da curva no ponto de inflexão. Assim, itens com **a** negativo não são esperados sob esse modelo, uma vez que indicariam que a probabilidade de responder corretamente ao item diminui com o aumento da habilidade. Baixos valores de **a** indicam que o item tem pouco poder de discriminação (alunos com habilidades bastante diferentes têm aproximadamente a mesma probabilidade de responder corretamente ao item) e valores muito altos indicam itens com curvas características muito "íngremes", que discriminam os alunos basicamente em dois grupos: os que possuem habilidade abaixo do valor do parâmetro **b** e os que possuem habilidades acima do valor do parâmetro **b**.

Função de Informação do Item

Uma medida bastante utilizada em conjunto com a CCI é a **função de informação do item**. Ela permite analisar quanto um item (ou teste) traz de informação para a medida de habilidade. A função de informação de um item é dada por:

$$I_i(\theta_j) = \frac{[\frac{d}{d\theta_j} P_i(\theta_j)]^2}{P_i(\theta_j) Q_i(\theta_j)},$$

onde:

$I_i(\theta_j)$ é a "informação"[4] fornecida pelo item **i** no nível de habilidade θ_j;

$P_i(\theta_j) = P(X_{ij} = 1 \mid \theta_j)$ e

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

No caso do modelo logístico de 3 parâmetros, a equação pode ser escrita como:

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

Esta equação mostra a importância que têm os três parâmetros sobre o montante de informação do item. Isto é, a informação é maior:

(i) quando $\mathbf{b_i}$ se aproxima de θ_j;

(ii) quanto maior for o $\mathbf{a_i}$;

(iii) e quanto mais $\mathbf{c_i}$ se aproximar de 0.

Função de Informação do Teste

A informação fornecida pelo teste é simplesmente a soma das informações fornecidas por cada item que compõe o mesmo:

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

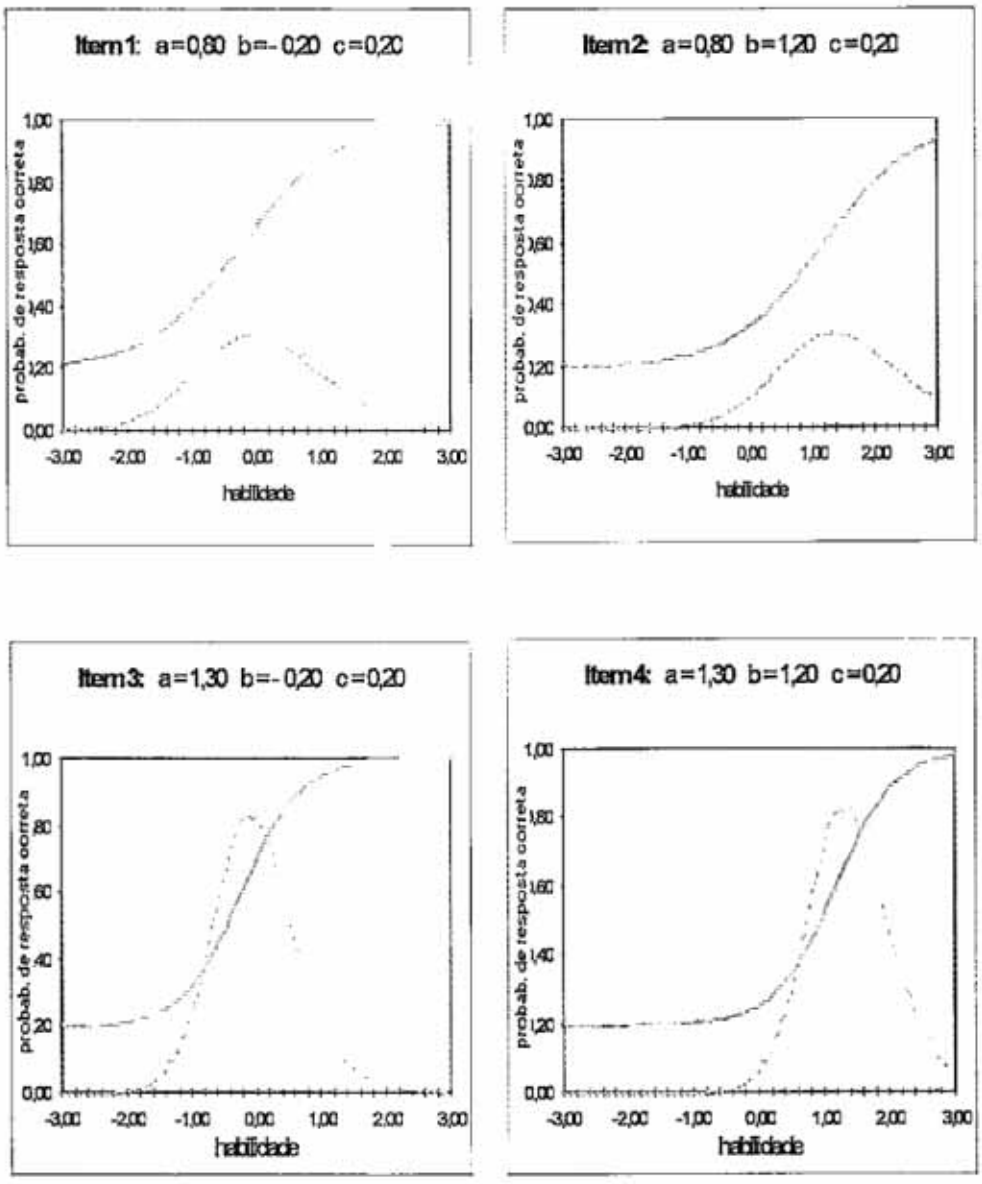
Outra maneira de representar esta função de informação do teste é através do erro padrão de medida, chamado na TRI de erro padrão de estimação. A $I(\theta_j)$, na verdade, é o inverso deste erro:

$$EP(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}} = \text{ erro padrão de estimação}$$

Similarmente ao erro padrão de medida da Teoria Clássica, o EP permite estabelecer intervalos de confiança em torno das habilidades θ_j dos sujeitos.

A seguir, apresentamos as curvas características e também as curvas de informação (traçado pontilhado) de quatro itens com diferentes combinações de valores dos parâmetros **a** e **b**.

Figura 2.2 Curvas características e de informação de vários itens



Comparando-se os itens 2 e 4 (e também os itens 1 e 3) pode-se perceber que os itens com maior valor do parâmetro **a** têm 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 os itens com maior valor do parâmetro **b** exigem 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.

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 $+\infty$. Mas como já foi dito, 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 indivíduos 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.

A Escala de Habilidade

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 por escala(0;1) . Essa escala é bastante utilizada pela TRI, e nesse caso, os valores do parâmetro **b** variam (tipicamente) entre -2 e +2. Com relação ao parâmetro **a**, esperam-se valores entre 0 e +2, sendo que os valores mais apropriados de **a** seriam aqueles maiores do que 0,6 e menores do que 1,7, quando utiliza-se D=1,7.

Apesar da frequente utilização da 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 seus pontos. Por exemplo, na escala(0;1) um indivíduo com habilidade 1,20 está 1,20 desvios padrão acima da habilidade média. Este mesmo indivíduo teria a habilidade 248, e consequentemente estaria também 1,20 desvios padrão acima da habilidade média, se a escala utilizada para esta população fosse a escala(200;40). Isto pode ser visto a partir da transformação de escala:

$$a(\theta - b) = (a / 40)\,[\,(40 \times \theta + 200) - (40 \times b + 200)\,] = a^{*}(\theta^{*} - b^{*})$$

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

1. $\theta^{*} = 40 \times \theta + 200$
2. $b^{*} = 40 \times b + 200$
3. $a^{*} = a / 40$
4. $P(X_i=1|\theta) = P(X_i=1|\theta^{*})$

Por exemplo, os valores dos parâmetros **a** e **b** do item 1 mostrado anteriormente, na escala(0;1) são, respectivamente, 0,80 e -0,20 e seus correspondentes na escala(200;40) são, respectivamente, 0,02 = 0,80 / 40 e 192 = 40 x -0,20 + 200. Além disso, um indivíduo com habilidade θ = 1,00 medida na escala(0;1) tem sua habilidade representada por θ^{*} = 40 X 1,00 + 200 = 240 na escala(200;40) e

$$P(X_1 = 1 \mid \theta = 1) = 0,20 + (1 - 0,20)\frac{1}{1 + e^{-1,7 \times 0,80 \times (1 - (-0,20))}} =$$

$$= 0,20 + (1 - 0,20)\frac{1}{1 + e^{-1,7 \times 0,02 \times (240 - 192)}} =$$

$$= P(X_1 = 1 \mid \theta^{*} = 240) = 0,87$$

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** à 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.

Suposições do Modelo: Unidimensionalidade e Independência Local

Unidimensionalidade

O modelo proposto pressupõe a unidimensionalidade do teste, isto é, a homogeneidade do conjunto de itens que supostamente devem estar medindo um único traço latente. Em outras palavras, deve haver apenas uma habilidade responsável pela realização de todos os itens da prova. Parece claro que qualquer desempenho humano é sempre multideterminado ou multimotivado, dado que mais de um traço latente entra na execução de qualquer tarefa. Contudo, para satisfazer o postulado da unidimensionalidade, é suficiente admitir que haja uma habilidade **dominante** (um fator dominante) responsável pelo conjunto de itens. Este fator é o que se supõe estar sendo medido pelo teste.

Tipicamente, a dimensionalidade do teste é verificada através da análise fatorial, feita a partir da matriz de correlações tetracóricas. Mislevy(1986) discute as deficiências da aplicação deste procedimento e sugere um outro procedimento baseado no método de máxima verossimilhança.

Independência local

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.

2.1.1.2 Outros modelos para itens dicotômicos

Dois outros modelos podem ser facilmente obtidos a partir do modelo logístico de 3 parâmetros. Por exemplo, quando não existe possibilidade de acerto ao acaso, pode-se considerar c = 0 no modelo anterior e tem-se o chamado **modelo logístico unidimensional de 2 parâmetros**, dado por:

$$P(X_{ij} = 1 \mid \theta_j) = \frac{1}{1+e^{-Da_i(\theta_j - b_i)}}$$

com $i = 1, 2, ..., I, \quad j = 1, 2, ..., n$.

Mas, se além de não existir resposta ao acaso ainda tivermos todos os itens com o mesmo poder de discriminação (pode-se considerar a = 1), tem-se o chamado **modelo logístico unidimensional de 1 parâmetro**, também conhecido como modelo de Rasch. Este modelo é dado por:

$$P(X_{ij} = 1 \mid \theta_j) = \frac{1}{1+e^{-D(\theta_j - b_i)}}$$

com $i = 1, 2, ..., I, \quad j = 1, 2, ..., n$.

2.1.2 Modelos para itens não dicotômicos

Aqui são incluídos os modelos tanto para a análise de itens abertos (de resposta livre) quanto para a análise de itens de múltipla escolha que são avaliados de forma graduada, ou seja, itens que são elaborados ou corrigidos de modo a ter-se uma ou mais categorias intermediárias ordenadas entre as categorias certo e errado. Nesse tipo de item não se considera somente se o indivíduo respondeu à alternativa correta ou não, mas também leva-se em conta qual foi a resposta dada por ele.

2.1.2.1 Modelo de Resposta Nominal (Nominal Categories Model)

Bock (1972) desenvolveu um modelo logístico de dois parâmetros que pode ser aplicado a todas as categorias de resposta escolhidas em um teste com itens de múltipla escolha. O propósito deste **modelo de resposta nominal** foi maximizar a precisão da habilidade estimada usando toda a informação contida nas respostas dos indivíduos, e não apenas se o item foi respondido corretamente ou não. Bock assumiu que a probabilidade com que um indivíduo j selecionaria uma particular opção **k** (de **m** opções avaliáveis) do item i seria representada por:

$$P_{i,k}(\theta_j) = \frac{e^{a^+_{i,k}(\theta_j - b^+_{i,k})}}{\sum_{h=1}^{m_i} e^{a^+_{i,h}(\theta_j - b^+_{i,h})}}$$

com $i=1,2,...,I$; $j=1,2,...,n$; $k=1,2,...,m_i$

Em cada θ_j, a soma das probabilidades sobre as m_i opções, $\sum_{k=1}^{m_i} P_{i,k}(\theta_j)$, é 1. As quantidades $(b^+_{i,k}; a^+_{i,k})$ são parâmetros do item i relacionados a k-ésima opção. O modelo assume que não há nenhuma ordenação a priori das opções de resposta.

2.1.2.2 Modelo de Resposta Gradual (Graded Response Model)

O **modelo de resposta gradual** de Samejima (1969) assume que as categorias de resposta de um item podem ser ordenadas entre si. Este modelo, como o modelo de Bock, tenta obter mais informação das respostas dos indivíduos do que simplesmente se eles deram respostas corretas ou incorretas.

Suponha que os escores das categorias de um item i são arranjados em ordem do menor para o maior e denotados por $k = 0, 1, ..., m_i$ onde (m_i+1) é o número de categorias do i-ésimo item. A probabilidade de um indivíduo j escolher uma particular categoria ou outra mais alta do item i pode ser dada por uma extensão do modelo logístico de 2 parâmetros:

$$P_{i,k}^{+}(\theta_j) = \frac{1}{1+e^{-Da_i(\theta_j - b_{i,k})}}$$

com $i=1,2,...,I$; $j=1,2,...,n$; $k=0,1,...,m_i$, onde:

a_i é o parâmetro de inclinação comum a todas as categorias do item i.

$b_{i,k}$ é o parâmetro de dificuldade da k-ésima categoria do item i.

Os demais parâmetros no modelo são análogos aos já definidos anteriormente.

No caso dos modelos para itens dicotômicos, o parâmetro de inclinação do item pode ser chamado de discriminação do item. Entretanto, no caso de modelos para itens não dicotômicos, a discriminação de uma categoria específica de resposta depende tanto do parâmetro de inclinação quanto da distância das categorias de dificuldade adjacentes.

Cabe ressaltar, que da definição, devemos ter:

$$b_{i,1} \leq b_{i,2} \leq ... \leq b_{i,mi}$$

ou seja, devemos ter necessariamente uma ordenação entre o nível de dificuldade das categorias de um dado item, de acordo com a classificação de seus escores.

A probabilidade de um indivíduo j receber um escore k no item i é dada então pela expressão:

$$P_{i,k}(\theta_j) = P_{i,k}^{+}(\theta_j) - P_{i,k+1}^{+}(\theta_j)$$

Samejima também define $P_{i,0}^{+}(\theta_j)$ e $P_{i,m_i+1}^{+}(\theta_j)$ de modo que:

$$P_{i,0}^{+}(\theta_j) = 1$$

e

$$P_{i,m_i+1}^{+}(\theta_j) = 0$$

Portanto,

$$P_{i,0}(\theta_j) = P_{i,0}^{+}(\theta_j) - P_{i,1}^{+}(\theta_j) = 1 - P_{i,1}^{+}(\theta_j)$$

e

$$P_{i,m}(\theta_j) = P_{i,m}^{+}(\theta_j) - P_{i,m_i+1}^{+}(\theta_j) = P_{i,m}^{+}(\theta_j)$$

Então, temos que:

$$P_{i,k}(\theta_j) = \frac{1}{1+e^{-Da_i(\theta_j - b_{i,k})}} - \frac{1}{1+e^{-Da_i(\theta_j - b_{i,k+1})}}$$

Note que num item com (m_i+1) categorias, m_i valores de dificuldade necessitam ser estimados, além do parâmetro de inclinação do item. Assim, para cada item, o número de parâmetros a ser estimado será dado pelo seu número de categorias de resposta. Se, por exemplo, tivermos um teste com I itens, cada um com (m_i+1) categorias de resposta, teremos então $[\ \sum_{i=1}^{I} m_i + I\ ]$ parâmetros de item a serem estimados.

2.1.2.3 Modelo de Escala Gradual (Rating Scale Model)

Um caso particular do modelo de resposta gradual de Samejima é o **modelo de escala gradual.** Analogamente ao modelo de resposta gradual, este modelo também é adequado para itens com categorias de resposta ordenadas. No entanto, aqui é feita uma suposição a mais: a de que os escores das categorias são igualmente espaçados.

Este modelo, proposto por Andrich (1978), é dado por:

$$P_{i,k}(\theta_j) = \frac{1}{1+e^{-Da_i(\theta_j - b_i + d_k)}} - \frac{1}{1+e^{-Da_i(\theta_j - b_i + d_{k+1})}}$$

com $i=1,2,...,I$; $j=1,2,...,n$; $k=0,1,...,m$, onde:

a_i é o parâmetro de inclinação comum a todas as categorias do item i.

b_i é agora o parâmetro de locação do item i e

d_k o parâmetro de categoria.

Como $P^+_{i,k}(\theta_j) - P^+_{i,k+1}(\theta_j) \geq 0$ então $d_k - d_{k+1} \geq 0$.

Ou seja, devemos ter:

$$d_1 \geq d_2 \geq ... \geq d_m$$

Note que a maior distinção entre o modelo de resposta gradual e o modelo de escala gradual está na hipótese de neste último os escores das categorias de resposta devem ser eqüidistantes. Assim, no modelo de

escala gradual o parâmetro $b_{i,k}$ é decomposto em um parâmetro b_i de locação do item e num parâmetro de categoria d_k, isto é:

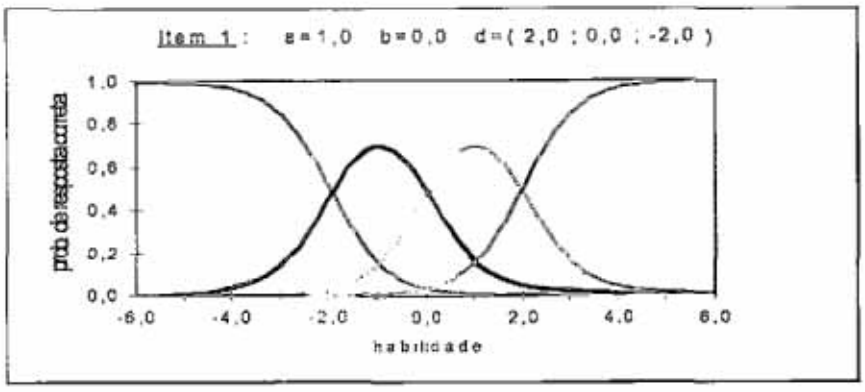
$$b_{i,k} = b_i - d_k$$

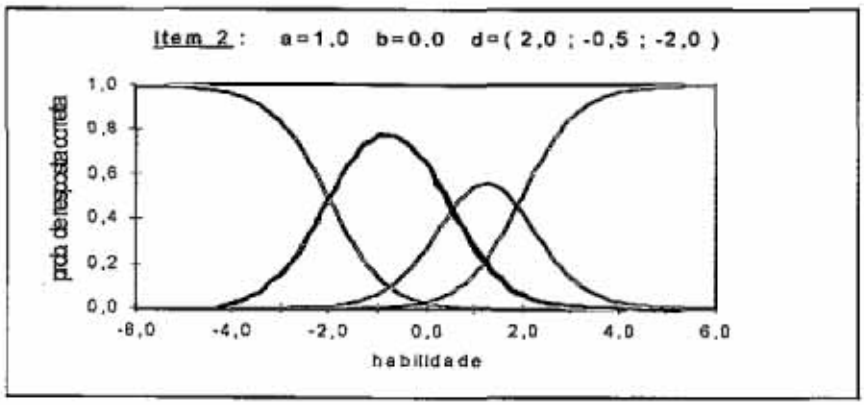
Cabe ressaltar que os parâmetros de categoria d_k não dependem do item, isto é, são comuns a todos os itens do teste. Logo, se os itens que compõem a prova tiverem suas próprias categorias de resposta, que podem diferir no número, então este modelo não é adequado.

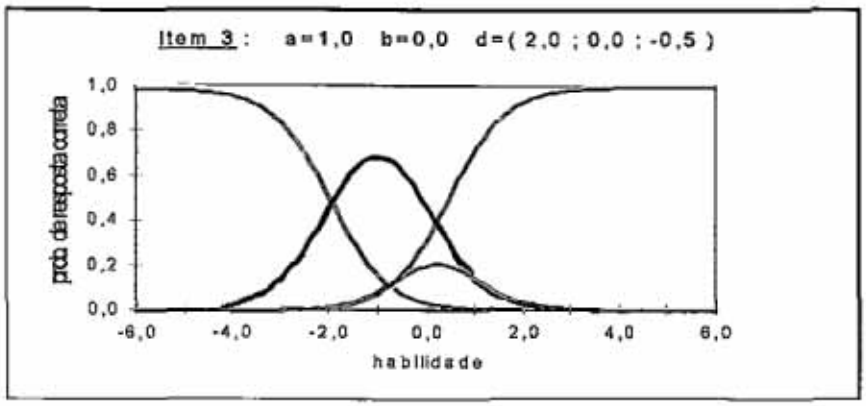
Assim, em teste compostos por itens com (m+1) categorias de resposta cada um, m parâmetros de categoria necessitam ser estimados, além dos parâmetros de inclinação e de locação de cada item. Logo, se tivermos um teste com I itens, teremos então [I x 2 + m] parâmetros de item a serem estimados.

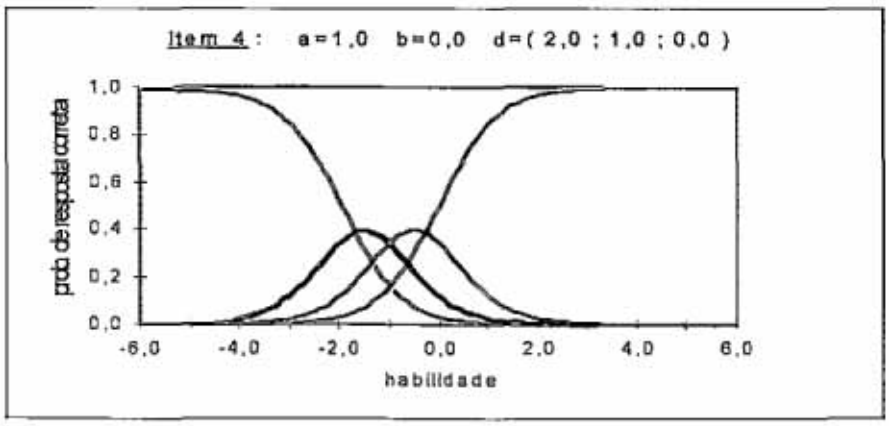
A seguir, faremos a representação gráfica do modelo de escala gradual e do modelo de resposta gradual para alguns itens com 4 categorias de resposta.

Figura 2.3 Representação gráfica dos modelos de escala gradual e de resposta gradual









A figura acima ilustra o significado dos parâmetros a_i, b_i e d_k. Em todos os itens, os parâmetros a_i e b_i foram mantidos fixos (em 1,0 e 0,0, respectivamente). Dessa maneira, podemos verificar a importância dos

parâmetros de categoria, d_k. Os itens 1 e 4, por terem os parâmetros de categoria igualmente espaçados, podem então ser representantes do modelo de escala gradual. Já o modelo de resposta gradual poderia ser representado por qualquer um dos itens acima.

Observando o item 1, podemos notar que indivíduos com habilidade até -2,0 têm maior probabilidade de responder apenas à categoria 0. Já indivíduos com habilidade entre -2,0 e 0,0, têm mais chance de alcançarem a categoria 1. Para habilidades entre 0,0 e 2,0, a maior probabilidade é que os indivíduos respondam até a categoria 2. Finalmente, indivíduos com habilidade acima de 2,0, devem alcançar a última categoria de resposta (que deverá representar o acerto total).

Note que do item 1 para o 2, a distância entre d_2 e d_3 tornou-se menor. A conseqüência disto é que aumenta a faixa de habilidade em que os indivíduos deverão responder somente até a categoria 1: de -2,0 a 0,0 no item 1 para -2,0 a 0,5 no item 2. Em outras palavras, a categoria 2 ficou mais difícil de ser alcançada, uma vez que no item 1 indivíduos com habilidade entre 0,0 e 2,0 tinham maior probabilidade de conseguir responder à essa categoria do que indivíduos com habilidade entre 0,5 e 2,0 no item 2.

No item 3, praticamente não há chance dos indivíduos responderem até a categoria 2: indivíduos com habilidade entre -2,0 e 0,0 têm mais chance de conseguir responder somente à categoria 1, enquanto que os indivíduos com habilidade maior do que esse valor já têm maior probabilidade de atingir a última categoria do item.

Finalmente, o item 4 é um exemplo de item onde a maioria dos indivíduos ou responde somente à primeira categoria, ou consegue chegar até a última. Apenas indivíduos com habilidade entre -2,0 e 0,0 apresentam uma chance maior de responderem somente às categorias 1 e 2.

2.1.2.4 Modelo de Crédito Parcial (Partial Credit Model)

O **modelo de crédito parcial** foi desenvolvido por Masters (1982) e é também um modelo para análise de respostas obtidas de duas ou mais categorias ordenadas. Nesse sentido, esse modelo é utilizado com os mesmos propósitos que outros já citados, inclusive o modelo de resposta gradual. O modelo de crédito parcial difere do gradual, entretanto, por pertencer à família de modelos de Rasch. Na verdade, o modelo de

crédito parcial é uma extensão do modelo de Rasch para itens dicotômicos. Logo, todos os parâmetros no modelo são de locação, sendo que o poder de discriminação é assumido ser comum para todos os itens.

Supondo que o item i tem (m_i+1) categorias de resposta ordenáveis $(k=0,1,...,m_i)$, temos que o modelo de crédito parcial é dado por:

$$P_{i,k}(\theta_j) = \frac{\exp[\sum_{u=0}^{k}(\theta_j - b_{i,u})]}{\sum_{u=0}^{m_i}\exp[\sum_{v=0}^{u}(\theta_j - b_{i,v})]}$$

com $i=1,2,...,I$; $j=1,2,...,n$; $k=0,1,...,m_i$ e $\sum_{u=0}^{0}(\theta_j - b_{i,u}) \equiv 0$,

onde:

$P_{i,k}(\theta_j)$ probabilidade de um indivíduo com habilidade θ_j escolher a categoria k, dentre as (m_i+1) categorias do item i.

$b_{i,k}$ parâmetro de item que regula a probabilidade de escolher a categoria k em vez da categoria adjacente (k-1) no item i . Cada parâmetro $b_{i,k}$ corresponde ao valor de habilidade em que o indivíduo tem a mesma probabilidade de responder à categoria k e à categoria (k-1), isto é, onde $P_{i,k}(\theta_j) = P_{i,k-1}(\theta_j)$.

Assim, para itens com (m_i+1) categorias de resposta, será necessário estimar m_i parâmetros de item. Note que para itens com apenas 2 categorias de resposta, este modelo fica análogo ao modelo de Rasch para itens dicotômicos.

2.1.2.5 Modelo de Crédito Parcial Generalizado (Generalized Partial Credit Model)

O **modelo de crédito parcial generalizado** – MCPG foi formulado por Muraki (1992), que se baseou no modelo de créditos parciais de Masters, relaxando a hipótese de poder de discriminação uniforme para todos os itens. O modelo de crédito parcial generalizado é dado por:

$$P_{i,k}(\theta_j) = \frac{\exp[\sum_{u=0}^{k} Da_i(\theta_j - b_{iu})]}{\sum_{u=0}^{m_i} \exp[\sum_{v=0}^{u} Da_i(\theta_j - b_{iv})]}$$

com $i=1,2,...,I$; $j=1,2,...,n$; $k=0,1,...,m_i$

Se o número de categorias de respostas é (m_i+1), somente m_i parâmetros de categoria do item podem ser identificados. Qualquer um dos (m_i+1) parâmetros de dificuldade das categorias pode ser arbitrariamente definido com qualquer valor. A razão é que o termo incluso no parâmetro é cancelado no numerador e no denominador do modelo. Em geral, definimos $b_{i,1} \equiv 0$.

Os parâmetros de categoria do item, $b_{i,k}$, são os pontos na escala de θ_j em as curvas de $P_{i,k-1}(\theta_j)$ e $P_{i,k}(\theta_j)$ se interceptam. Essas duas funções se interceptam somente uma vez, e a intersecção pode ocorrer em qualquer ponto da escala θ_j. Então, sob a hipótese de que $a_i > 0$,

se $\theta_j = b_{i,k}$ então $P_{i,k}(\theta_j) = P_{i,k-1}(\theta_j)$

se $\theta_j > b_{i,k}$ então $P_{i,k}(\theta_j) > P_{i,k-1}(\theta_j)$

se $\theta_j < b_{i,k}$ então $P_{i,k}(\theta_j) < P_{i,k-1}(\theta_j)$

Da mesma maneira como no modelo de escala gradual, no MCPG o parâmetro $b_{i,k}$ pode ser decomposto como:

$$b_{i,k} = b_i - d_k$$

Mas, é importante ressaltar que, diferentemente do modelo de escala gradual, aqui os valores de d_k **não** são necessariamente ordenados seqüencialmente dentro de um item. O parâmetro d_k é interpretado como a dificuldade relativa da categoria k em comparação com as outras categorias do item ou o desvio da dificuldade de cada categoria em relação à locação do item, b_i.

Assim, em teste compostos por itens com (m_i+1) categorias de resposta, m_i parâmetros de categoria necessitam ser estimados, além dos parâmetros de inclinação e de locação de cada item. Logo, se tivermos um teste com I itens, teremos então $[\sum_{i=1}^{I} m_i + I \times 2]$ parâmetros de item a serem estimados.

2.2 Modelos envolvendo dois ou mais grupos

Alguns modelos já foram desenvolvidos para serem aplicados quando um teste envolve mais de uma população, sendo basicamente extensões dos modelos até aqui apresentados. No entanto, um dos poucos modelos que já se encontram implementados computacionalmente e que, portanto, já estão sendo utilizados na prática, quando um teste é aplicado a dois ou mais grupos de respondentes, é uma generalização dos modelos logísticos unidimensionais de 1, 2 e 3 parâmetros, que foi recentemente proposta por Bock e Zimowski (1997). O modelo é dado por:

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

com $i = 1, 2, ..., I$, $j = 1, 2, ..., n_k$, $k = 1, 2, ..., g$, onde:

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

θ_{jk} representa a habilidade (traço latente) do j-ésimo indivíduo da população k.

$P(X_{ijk} = 1 \mid \theta_{jk})$ é a probabilidade de um indivíduo j da população k, com habilidade θ_{jk}, responder corretamente ao item i.

Os demais parâmetros já foram descritos anteriormente.

Em geral, indivíduos pertencentes a diferentes populações não são submetidos todos aos mesmos itens. Mas, para que seja possível a comparação entre populações, é necessário que haja pelo menos alguns itens comuns entre elas. Assim, I representa o número total de itens distintos apresentados.

A recente implementação computacional desse modelo para mais de um grupo de respondentes foi um dos maiores avanços da TRI nos últimos anos. Através dele a comparação de indivíduos de grupos distintos, submetidos a provas diferentes mas com itens comuns, passou a ser feita de uma maneira bem mais eficiente do que era feita até então, uma vez que diminui possíveis erros de modelagem que a metodologia anterior poderia vir a ter. Algumas das questões mais importantes envolvendo a comparação de duas ou mais populações serão detalhadamente discutidas no Item 4.

3.0 ESTIMAÇÃO

Uma das etapas mais importantes da TRI é a estimação dos parâmetros dos itens e das habilidades dos respondentes. Como foi visto no tópico anterior, a probabilidade de uma resposta correta num determinado item depende da habilidade do indivíduo e dos parâmetros que caracterizam o item. Mas, em geral, ambos são desconhecidos. O que é conhecido são as respostas dos indivíduos aos itens do teste.

Assim, nos modelos de resposta ao item temos um problema de estimação que envolve dois tipos de parâmetros: os parâmetro dos itens e as habilidades dos indivíduos. Então, do ponto de vista teórico, podemos dividir o problema em três situações: quando *já* conhecemos os parâmetros dos itens, temos apenas que estimar as habilidades; se já conhecemos as habilidades dos respondentes, estaremos interessados apenas na estimação dos parâmetros dos itens e, por fim, a situação mais comum, em que desejamos estimar os parâmetros dos itens e as habilidades dos indivíduos simultaneamente. Na TRI, o processo de estimação dos parâmetros dos itens é conhecido como **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 indivíduos de uma ou mais populações, estimam-se os parâmetros dos itens e/ou as habilidades a partir de métodos de máxima verossimilhança ou de procedimentos bayesianos. Ambas as soluções exigem procedimentos iterativos que envolvem cálculos bastante complexos e, consequentemente, programas de computador específicos. É importante ressaltar que, em qualquer um desses casos, 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 padrão 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.

Nesse tópico, vamos abordar cada um dos 3 problemas de estimação citados. Para ilustrar os métodos de estimação estaremos sempre utilizando o modelo logístico de 3 parâmetros, por ser um dos

modelos mais populares da TRI, e considerando que os itens envolvidos são dicotômicos. Procedimentos de estimação para itens não dicotômicos ou envolvendo outros modelos podem ser encontradas, por exemplo, em Baker (1992).

Cabe também ressaltar que estaremos sempre nos referindo aos modelos unidimensionais, ou seja, aqueles envolvendo um único traço latente ou habilidade. Além disso, estaremos considerando os modelos para um único grupo de respondentes, com o intuito de facilitar as explicações, uma vez que nosso objetivo é dar uma visão geral das diferentes técnicas de estimação mais utilizadas. Porém, no final do tópico, descreveremos um dos processos de estimação para modelos de duas ou mais populações.

3.1 Estimação da habilidade

Inicialmente vamos considerar a situação em que os parâmetros dos itens são conhecidos e portanto, o problema de estimação se resume em estimar as habilidades dos indivíduos submetidos ao teste. Na prática, esta situação é bastante frequente, uma vez que submeter indivíduos a itens já calibrados, apenas visando a estimação de suas habilidades com objetivos de classificação ou seleção, é uma das vantagens da TRI e que já vem sendo bastante explorada. Uma prova disso é o crescente interesse na criação de bancos de itens.

Cabe ressaltar que estaremos tratando da estimação da habilidade de um único respondente, digamos o indivíduo j, e que para estimar as habilidades de n indivíduos, o processo pode ser repetido independentemente n vezes.

Seja X_{ij} a variável aleatória dicotômica (1=acerto e 0=erro) que representa a resposta do j-ésimo indivíduo (com j=1, 2, ..., n) ao i-ésimo item (com i= 1, 2, ..., I). Assim, $X_j = (X_{1j}, X_{2j}, ..., X_{Ij})^t$ é o vetor aleatório (I x 1) que representa as respostas do j-ésimo indivíduo a todos os itens.

Seja também $P_{ij} = P(X_{ij} = 1 \mid \theta_j)$ o modelo logístico unidimensional de 3 parâmetros apresentado no Tópico 2 e, além disso, seja $Q_{ij} = 1 - P_{ij}$.

3.1.1 Estimação por Máxima Verossimilhança

Sob a suposição de independência local, a probabilidade do vetor de respostas X_j do indivíduo j, condicionado na sua habilidade θ_j, é dada por:

$$P_j(X_j|\theta_j) = \prod_{i=1}^{I} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}}$$

e então, o logaritmo natural da função de verossimilhança, baseada nas respostas desse j-ésimo indivíduo, pode ser escrita como:

$$\ln L(x_j|\theta_j) = \sum_{i=1}^{I} [\, x_{ij} \ln P_{ij} + (1 - x_{ij}) \ln Q_{ij} \,]$$

Portanto, o estimador de máxima verossimilhança para a habilidade do j-ésimo indivíduo é o valor de θ que maximiza a função acima. Ou seja, devemos solucionar a equação:

$$\frac{\partial \ln L(x_j|\theta_j)}{\partial \theta_j} = 0$$

isto é,

$$\sum_{i=1}^{I} \left[\frac{x_{ij} - P_{ij}}{P_{ij} Q_{ij}} \times \frac{\partial P_{ij}}{\partial \theta_j} \right] = 0$$

No entanto, a equação anterior não pode ser resolvida diretamente, e então são necessários métodos iterativos. Em geral, utiliza-se o método de Newton-Raphson, que pode ser encontrado, por exemplo, em Hambleton e Swaminathan (1985).

Além disso, essa equação pode não ter solução finita para certos padrões peculiares de resposta como, por exemplo, em situações de acerto total ou erro total. Nesses casos, o estimador de máxima verossimilhança não está definido, e a solução seria buscar outros métodos de estimação.

Como todo estimador de máxima verossimilhança, $\hat{\theta}_{j\,MV}$ – o estimador de θ_j – é normalmente distribuído com média θ_j, no caso de testes com um número suficientemente grande de itens. O erro padrão de $\hat{\theta}_{j\,MV}$ é dado por:

$$EP(\hat{\theta}_{j\,MV}) = \frac{1}{\sqrt{I(\theta_j)}}$$

onde $I(\theta_j)$ é a função de informação do teste, definida em 2.1.1.1 .

Assim, podemos observar que diferentemente do erro padrão de medida utilizado na Teoria Clássica, que é constante para todos os respondentes, o erro padrão fornecido pela TRI varia de acordo com a escala de habilidades. Ele é tipicamente menor no centro da escala, onde geralmente há mais itens, e maior nos extremos da escala, onde há poucos itens.

3.1.2 Estimação por Métodos Bayesianos

O problema da não existência de estimadores de máxima verossimilhança em algumas situações pode ser resolvido se procedimentos de estimação bayesianos são usados. A idéia básica é modificar a função de verossimilhança de modo a incorporar qualquer informação a priori que se possa ter sobre o parâmetro de habilidade. Por exemplo, podemos dizer que θ é normalmente distribuído com média μ e desvio padrão σ . Nesse caso, a informação a priori pode ser expressa na forma de uma função densidade, denotada por $g(\theta)$.

3.1.2.1 Estimador EAP

O estimador bayesiano EAP (Esperança A Posteriori) é baseado na seguinte forma do teorema de Bayes:

$$g(\theta_j|x_j) = \frac{P_j(x_j|\theta_j)\, g(\theta)}{P(x_j)}$$

No nosso caso, temos que:

$P_j(x_j|\theta_j) = \prod_{i=1}^{I} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}}$, que é a verossimilhança condicionada em θ_j;

$g(\theta)$ é distribuição a priori das habilidades da população à qual pertence θ_j . (Utiliza-se a mesma priori para todos os θ_j) e

$P(x_j) = \int_{-\infty}^{+\infty} P(x_j|\theta)\, g(\theta)\, d\theta$, que é a probabilidade marginal de x_j.

O estimador EAP de Bayes é a média da distribuição a posteriori de θ, dado o vetor de respostas observado x_j, e essa esperança é dada por:

$$\hat{\theta}_{j\,EAP} = E(\theta_j | x_j) = \frac{\int_{-\infty}^{+\infty} \theta_j \, g(\theta) \, P_j(x_j|\theta_j) \, d\theta}{\int_{-\infty}^{+\infty} g(\theta) \, P_j(x_j|\theta_j) \, d\theta}$$

Mas, os cálculos envolvidos na resolução da equação acima são bastante complexos, pois em geral essas integrais não podem ser expressas de forma fechada. Assim, a solução encontrada é a utilização de uma aproximação para a distribuição de $g(\theta)$, através de pontos de quadratura, desenvolvidos por Hermite-Gauss. Esse procedimento consiste em aproximar a densidade $g(\theta)$ por um histograma definido num intervalo finito, isto é, formar uma distribuição de freqüência com q valores de θ.

A idéia de utilizar uma aproximação por pontos de quadratura é substituir o problema de encontrar a área sob uma curva contínua por um problema mais simples, de encontrar a soma das áreas de um número finito de retângulos que aproximem essa área. O ponto médio de cada retângulo na escala de habilidade será denotado por X_k (k=1, 2, ..., q), e cada um desses pontos têm um peso $A(X_k)$ associado. Esse peso leva em conta a altura da função densidade $g(\theta)$ nas vizinhanças de X_k e a largura dos retângulos. Os valores de X_k e $A(X_k)$ são encontrados resolvendo-se um conjunto de equações que envolvem a distribuição contínua a ser aproximada e o número desejado de pontos de quadratura. Tabelas contendo pontos de quadraturas e seus pesos correspondentes podem ser encontradas para várias escolhas de $g(\theta)$ (veja, por exemplo, Stroud e Sechrest, 1966).

Assim, no nosso caso teremos:

$$\hat{\theta}_{j\,EAP} \equiv \frac{\sum_{k=1}^{q} X_k \, P_j(x_j|X_k) \, A(X_k)}{\sum_{k=1}^{q} P_j(x_j|X_k) \, A(X_k)}$$

onde X_k é o k-ésimo ponto de quadratura, e $A(X_k)$ é o peso correspondente à função densidade $g(X_k)$.

Uma medida da precisão desse estimador é o erro padrão a posteriori, aproximado por:

$$EP\,(\hat{\theta}_{j\,EAP}) \equiv \frac{\sum_{k=1}^{q} (X_k - \hat{\theta}_{j\,EAP})^2 \; P_j(x_j|X_k)\, A(X_k)}{\sum_{k=1}^{q} P_j(x_j|X_k)\, A(X_k)}$$

As vantagens do estimador EAP são que ele está definido para qualquer padrão de resposta e tem um erro médio menor do que qualquer outro estimador, inclusive o de máxima verossimilhança. No entanto, ele é, em geral, um estimador viciado, embora o viés seja pequeno quando o erro padrão a posteriori for pequeno.

Embora a média amostral das estimativas EAP seja um estimador não viciado da média da população em estudo, o desvio padrão amostral é, em geral, menor do que o da população. Esse vício não representa um problema sério se todos os indivíduos forem medidos com o mesmo erro padrão a posteriori. No entanto, pode ser um problema se compararmos indivíduos usando testes que têm desvios padrão a posteriori muito diferentes. Assim, deve-se evitar fazer comparações entre indivíduos que tiveram suas habilidades estimadas com precisões muito distintas, ou seja, *indivíduos que foram submetidos a testes diferentes, que foram calibrados* separadamente e que tiveram desvios padrão a posteriori significativamente diferentes.

3.1.2.2 Estimador MAP

O estimador modal de Bayes ou MAP (Máximo A Posteriori) é baseado na seguinte forma do teorema de Bayes:

$$g\,(\theta_j|x_j) \;\propto\; L(x_j|\theta_j)\; g(\theta)$$

onde:

$L(x_j|\theta_j) = \prod_{i=1}^{I} P_{ij}^{x_{ij}}\, Q_{ij}^{1-x_{ij}}$ é a verossimilhança condicionada em θ_j e

$g(\theta)$ é distribuição a priori de θ_j .

Então, tomando o logaritmo natural da equação acima, temos:

$$\ln g\,(\theta_j|x_j) \;\propto\; \ln L_j(x_j|\theta_j) \;+\; \ln g(\theta)$$

ou seja,

$$\ln g\,(\theta_j|x_j) \;\propto\; \sum_{i=1}^{I} [\, x_{ij}\, \ln P_{ij} \;+\; (1-x_{ij})\, \ln Q_{ij}\,] \;+\; \ln g(\theta)$$

Portanto, o estimador MAP para a habilidade do j-ésimo indivíduo é o valor de θ que maximiza a função acima. Ou seja, devemos solucionar a equação:

$$\frac{\partial \ln g(\theta_j | x_j)}{\partial \theta_j} = 0$$

isto é,

$$\sum_{i=1}^{I} \left[\frac{x_{ij} - P_{ij}}{P_{ij} Q_{ij}} \times \frac{\partial P_{ij}}{\partial \theta_j} \right] + \frac{\partial \ln g(\theta_j | x_j)}{\partial \theta_j} = 0$$

Como podemos notar, a equação a ser resolvida é bastante semelhante à equação do estimador de máxima verossimilhança. Na verdade, o primeiro termo dessa equação é exatamente a equação resultante naquele caso. Logo, os mesmos procedimentos iterativos utilizados para obter-se o estimador de máxima verossimilhança são necessários aqui e então, também, pode-se utilizar o método de Newton-Raphson.

O erro padrão do estimador MAP é dado por:

$$EP(\hat{\theta}_{j\,MAP}) = \frac{1}{\sqrt{J(\theta_j)}}$$

onde:

$J(\theta_j) = I(\theta_j) - \dfrac{\partial^2 \ln g(\theta_j | x_j)}{\partial \theta_j^{\,2}}$ é a informação a posteriori.

Apesar das semelhanças com o estimador de máxima verossimilhança nos procedimentos para sua obtenção, em termos de propriedades o estimador MAP é mais parecido com o EAP, pois está definido para qualquer padrão de resposta e é, em geral, um estimador viciado.

3.2 Estimação dos parâmetros de item

Vamos agora considerar a situação em que conhecemos as habilidades dos indivíduos que responderam a um determinado teste e estamos interessados em estimar os parâmetros dos itens que compõem esse teste. Na prática, essa situação não ocorre, mas o estudo desse caso é

importante para o desenvolvimento do caso mais complexo, que virá a seguir, onde teremos que estimar os parâmetros dos itens e das habilidades simultaneamente.

Estimar os parâmetros de <u>todos</u> os itens de um teste simultaneamente é um problema que envolve recursos computacionais bastante dispendiosos. Assim, na maioria dos procedimentos utilizados na TRI, os parâmetros dos itens são estimados item a item. Portanto, estaremos tratando agora das técnicas de estimação de um <u>único</u> item, digamos o item i. Para obter os parâmetros dos I itens que compõem o teste, o processo deverá ser repetido independentemente I vezes.

Suponhamos que sejam conhecidas as habilidades de n indivíduos, e que esses indivíduos sejam submetidos a um dado item i. Suponhamos também que podemos dividir esses indivíduos em <u>k</u> grupos, de acordo com suas habilidades. Então, teremos k grupos, cada um com f_j sujeitos com habilidade θ_j, onde j=1, 2, ..., k e $\sum_{j=1}^{k} f_j = n$.

Seja r_j o número de indivíduos que responde corretamente ao item i, dentre os f_j indivíduos com habilidade θ_j. Conseqüentemente, $f_j - r_j$ é o número de indivíduos dentro desse grupo que não acerta o item i. Então, sejam $R_i = (r_1, r_2, ..., r_k)$ o vetor observado do número de respostas corretas ao item i e a_i, b_i e c_i os parâmetros do item i, como descritos no tópico anterior.

3.2.1 Estimação por Máxima Verossimilhança

A probabilidade de observarmos um dado vetor $R_i = (r_1, r_2, ..., r_k)$ para o item i é dada pela função de verossimilhança:

$$L(R_i \mid a_i , b_i , c_i) = P(R_i) = \prod_{j=1}^{k} \frac{f_j !}{r_j ! \ (f_j - r_j)!} P_{ij}^{r_j} \ Q_{ij}^{f_j - r_j}$$

e, então, o logaritmo natural dessa função de verossimilhança pode ser escrito como:

$$\ln L(R_i | a_i, b_i, c_i) = \text{const.} + \sum_{j=1}^{k} [r_j \ln P_{ij} + (f_j - r_j) \ln Q_{ij}]$$

Portanto, o estimador de máxima verossimilhança dos parâmetros a_i, b_i e c_i são os valores que maximizam a superfície tri-dimensional dada pela função anterior. Ou seja, devemos solucionar simultaneamente as equações:

$$\frac{\partial \ln L(R_i \mid a_i, b_i, c_i)}{\partial a_i} = 0$$

$$\frac{\partial \ln L(R_i \mid a_i, b_i, c_i)}{\partial b_i} = 0$$

$$\frac{\partial \ln L(R_i \mid a_i, b_i, c_i)}{\partial c_i} = 0$$

No entanto, a solução das equações acima em geral não pode ser obtida diretamente, e então são necessários métodos iterativos, como, por exemplo, o método de Newton-Raphson, na sua forma multivariada.

Como podemos observar, a estimação dos parâmetros de item, supondo as habilidades dos respondentes conhecidas é bastante semelhante ao processo inverso, explicado na seção 3.1. A diferença é que a função de verossimilhança para um item é multidimensional nos parâmetros do item, o que certamente torna os cálculos mais complexos do que no caso anterior.

Mas, certamente, as semelhanças não param por aqui. Assim como ocorrem dificuldades nos processos de estimação em 3.1, neste caso também podem ocorrer os mesmos problemas. No entanto, deixaremos a abordagem de métodos alternativos de estimação para a seção 3.3, pois, como já foi dito, a situação aqui estudada não acontece na prática. Cabe porém ressaltar que tanto essa situação como a anterior podem ser vistas como "blocos" do processo de estimação por máxima verossimilhança conjunta, que será abordado no próximo caso.

3.3 Estimação conjunta dos parâmetros dos itens e das habilidades

Sem dúvida alguma, a situação mais comum na prática, é esta: um teste, composto de I itens, é aplicado a um grupo de n indivíduos, e estamos interessados em estimar tanto os 3I parâmetros dos itens (no caso de assumirmos um modelo com 3 parâmetros), como as habilidades dos n sujeitos.

Cabe ressaltar que apesar de os processos de estimação envolverem a estimação das habilidades de cada um dos respondentes, em geral estamos preocupados com a obtenção dos parâmetros populacionais – e não individuais –, ou seja, estamos interessados na média e na variabilidade das habilidades da população de respondentes. Essas estimativas populacionais dos parâmetros podem ser obtidas ao final de

alguns dos processos de estimação, como por exemplo, o método EAP, descrito na seção 3.1.2.1. Obviamente, outra alternativa seria calcular a média e o desvio padrão das estimativas das habilidades individuais obtidas. No entanto, apesar da média amostral ser um estimador não viciado da média da distribuição da população de respondentes, o desvio padrão amostral é um estimador viciado para a variabilidade populacional (Mislevy, 1991).

Sejam $X = | X_{ij} |$ a matriz (n x I) das respostas dos n indivíduos aos I itens, $\theta = (\theta_1, \theta_2, ..., \theta_n)^t$ o vetor (n x 1) das suas respectivas habilidades, $a=(a_1, a_2, ..., a_I)^t$ o vetor (I x 1) dos parâmetros de discriminação, $b = (b_1, b_2, ..., b_I)^t$ o vetor (I x 1) dos parâmetros de dificuldade e $c = (c_1, c_2, ..., c_I)^t$ o vetor (I x 1) dos parâmetros de acerto ao acaso.

3.3.1 Estimação por Máxima Verossimilhança Conjunta (MVC)

Talvez por ser o mais natural, este método foi o primeiro a ser utilizado na resolução desse tipo de problema de estimação dos parâmetros da TRI.

Sob a suposição de independência local, a probabilidade do vetor de respostas x_j do indivíduo j, condicionado na sua habilidade θ_j e nos parâmetros dos itens, é dada por:

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

e a função de verossimilhança, baseada nas respostas dos n indivíduos, pode ser escrita como:

$$L(X | \theta, a, b, c) = \prod_{j=1}^{n} P_j(x_j | \theta_j, a, b, c) = \prod_{j=1}^{n} \prod_{i=1}^{I} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}}$$

e, então, tomando o logaritmo natural da função de verossimilhança, temos:

$$\ln (X | \theta, a, b, c) = \sum_{j=1}^{n} \sum_{i=1}^{I} [x_{ij} \ln P_{ij} + (1 - x_{ij}) \ln Q_{ij}]$$

Observe que será necessário derivar, igualar a zero e resolver 3I + n equações simultaneamente, para obtermos as estimativas de máxima verossimilhança de todos os parâmetros envolvidos. Trata-se, portanto, de um problema computacional bastante complexo. Para simplificar o

problema, uma saída é a utilização de um procedimento em 2 estágios onde, num primeiro estágio, os parâmetros dos itens são estimados, assumindo-se que as habilidades dos indivíduos são conhecidas. No segundo estágio, as habilidades são estimadas, considerando-se agora que os parâmetros dos itens são conhecidos.

Para inicializar o processo, é comum utilizar-se o escore padronizado de cada um dos respondentes como o valor "conhecido" de suas respectivas habilidades. O escore padronizado nada mais é do que o escore (número de acertos) obtido na prova, menos o escore médio, dividido pelo desvio padrão dos escores de todos os indivíduos submetidos ao teste. Daí, cada item é então considerado separadamente, de modo bastante semelhante ao apresentado em 3.2, e como já foi citado naquela seção, métodos iterativos do tipo Newton-Raphson são necessários para resolver simultaneamente as 3 equações geradas para cada um dos I itens.

Neste ponto, teremos então um conjunto de 3I estimativas para os parâmetros dos itens. Assim, no segundo estágio do processo, essas estimativas podem ser tratadas como os "verdadeiros" parâmetros dos itens, e caímos, portanto, numa situação bastante semelhante àquela abordada na seção 3.1. Dessa maneira, concluímos um "ciclo" do processo, que é então reinicializado utilizando-se as estimativas das habilidades obtidas no final do ciclo anterior como sendo os verdadeiros valores.

Entretanto, devido à indeterminação associada ao modelo, já apresentada em 2.1.1.1, os valores dos parâmetros que maximizam a função de verossimilhança 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.

Assim, esse problema de indeterminação é solucionado definindo-se uma escala arbitrária para os valores das habilidades ou para os valores dos parâmetros de dificuldade b, uma vez que tanto as habilidades quanto os parâmetros de dificuldade são medidos na mesma escala. Usualmente, estabelece-se os valores 0 para a média e 1 para o desvio padrão dos valores das habilidades e então a transformação linear

$\hat{\theta}_j^* = (\hat{\theta}_j - \bar{\hat{\theta}}) / S_{\hat{\theta}}$ é feita no final de cada ciclo, sendo que $\bar{\hat{\theta}}$ e $S_{\hat{\theta}}$ são a média e o desvio padrão dos valores de $\hat{\theta}_j$, respectivamente. Após essa transformação nos parâmetros de habilidades, também são necessárias as mesmas transformações nos parâmetros dos itens, isto é:

$$\hat{b}_i^* = \frac{\hat{b}_i - \bar{\hat{\theta}}}{S_{\hat{\theta}}} \quad \text{e} \quad \hat{a}_i^* = \hat{a}_i \, S_{\hat{\theta}}$$

lembrando que o parâmetro c não sofre qualquer transformação, uma vez que não depende da escala de medida por ser uma probabilidade. Se o problema de indeterminação não fosse resolvido, os parâmetros poderiam flutuar indefinidamente com o aumento do número de ciclos, e a convergência nunca seria alcançada.

Note que um critério de convergência é necessário para determinar quando um número suficiente de ciclos já foi realizado. Tal critério é arbitrário, mas é usual utilizar a própria função de verossimilhança para determinar o momento de parada. Para tanto, substituem-se as estimativas dos parâmetros obtidas após cada ciclo para avaliar numericamente a função de verossimilhança, e então, a diferença entre duas verossimilhanças sucessivas deverá ser menor do que um valor especificado. Outro critério de parada seria simplesmente pré-estabelecer o número de ciclos do processo.

Como foi visto em 3.1.1, quando obtemos uma estimativa de máxima verossimilhança para os parâmetros de habilidade, sua variância é dada pelo inverso da função de informação do teste. Analogamente, quando obtemos estimativas de máxima verossimilhança para os parâmetros dos itens, a matriz de variância e covariância dessas estimativas é dada pelo inverso da matriz de informação dos parâmetros do item estimados.

Finalmente, cabe ressaltar que as habilidades de indivíduos que acertaram ou erraram todos os itens ou os parâmetros dos itens respondidos correta ou erroneamente por todos os indivíduos não podem ser estimados através deste método. 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 verossimilhança

não se aplicam aqui. Atualmente, este método de estimação tem sido utilizado somente como base para outros procedimentos.

3.3.2 Estimação por Máxima Verossimilhança Marginal (MVM)

Devido às dificuldades de estimar-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. No entanto, ao contrário do procedimento descrito anteriormente, 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.

A essência deste procedimento é a integração em θ, de modo que a função de verossimilhança não dependa dos parâmetros de habilidade. Conseqüentenente, os parâmetros dos itens são estimados na distribuição marginal, e essa estimação não depende mais da estimação das habilidades de cada respondente, mas sim das distribuições dessas habilidades.

A probabilidade marginal do vetor de respostas x_j com respeito aos parâmetros dos itens e à distribuição a priori da habilidade θ_j pode ser escrita como:

$$\int P_j(x_j|\theta_j,a,b,c)\, g(\theta_j|\tau)\, d\theta$$

Logo, as estimativas de máxima verossimilhança para os parâmetros dos itens são obtidas a partir da maximização da função de verossimilhança marginal:

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

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 de 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 de escala da priori. Em geral, utiliza-se como priori a distribuição normal com média 0 e desvio padrão 1.

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 algoritmo EM, produz estimativas consistentes para os parâmetros dos itens e é computacionalmente muito mais simples (Dempster, Laird e Rubin, 1977).

Em linhas gerais, o algoritmo EM é um procedimento iterativo para encontrar estimativas de máxima verossimilhança de parâmetros de modelos de probabilidade na presença de variáveis aleatórias não observáveis, chamadas de variáveis latentes. O **E** representa o passo em que se calcula a Esperança e o **M** representa o passo de Maximização. No caso particular da TRI, nós desejamos encontrar estimativas dos parâmetros de item na presença de um variável aleatória não observável e para fazer inferências sobre essa variável θ, informações observáveis baseadas nas respostas dadas aos itens são usadas.

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 algoritmo, 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 considerados como sendo os verdadeiros valores dos parâmetros, estimam-se as habilidades dos repondentes através de métodos de máxima verossimilhança ou bayesianos, 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.

Cabe ressaltar que 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 ter-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 itens respondidos corretamente ou erroneamente por todos os indivíduos.

3.3.3 Estimação através de Procedimentos Bayesianos

A principal característica dos procedimentos de estimação bayesianos empregados na TRI é que eles utilizam distribuições de probabilidade para as habilidades dos indivíduos e também para os parâmetros dos itens. Em linhas gerais, a idéia é combinar a função de verossimilhança, que é baseada nos dados – que no caso da TRI são as respostas dadas pelos indivíduos –, com informações a priori sobre a distribuição dos parâmetros de interesse. Utilizando uma aplicação do teorema de Bayes, teremos uma distribuição de probabilidade a posteriori, que é proporcional ao produto da função de verossimilhança e da distribuição a priori. Então, as inferências sobre os parâmetros que queremos estimar serão baseadas nessa distribuição a posteriori.

Assim, primeiro podemos escrever a densidade conjunta de todos os parâmetros a priori para os dados. Esses parâmetros são assumidos como sendo variáveis aleatórias contínuas e independentes com distribuição de probabilidade dada por:

$$g(\theta, \tau, a, b, c, \eta) = \prod_{j=1}^{n} g(\theta_j | \tau) \prod_{i=1}^{I} g(a_i, b_i, c_i | \eta)\, g(\tau)\, g(\eta)$$

onde $g(\theta_j|\tau)$ é a distribuição de probabilidade do parâmetro de habilidade θ_j de um indivíduo e é condicional aos parâmetros populacionais da distribuição de habilidade contida no vetor τ. Tipicamente, é assumido que a distribuição a priori da habilidade é uma distribuição normal. Desde que as habilidades são assumidas como sendo independentes e identicamente distribuídas, τ contém a média e a variância comuns μ_θ e σ^2_θ daquelas distribuições de habilidade a priori. Nesse tipo de modelo, os θ_j's são os parâmetros e os parâmetros populacionais μ_θ e σ^2_θ são conhecidos como hiperparâmetros.

Os hiperparâmetros podem também ser tratados como variáveis aleatórias tendo uma distribuição de probabilidade que, nesse caso, é denotada por $g(\tau)$.

$g(a_i, b_i, c_i|\eta)$ é a distribuição de probabilidade para os parâmetros do item i, condicional aos parâmetros populacionais do vetor η. Analogamente ao caso dos parâmetros de habilidade, os a_i, b_i e c_i serão referidos como parâmetros, e η será referido como o vetor contendo os hiperparâmetros do item.

Dada a matriz de resposta dos indivíduos, a distribuição a posteriori de todos os itens, obtida via uma aplicação do teorema de Bayes é:

$$g(\theta, \tau, a, b, c, \eta \mid X) \propto L(X \mid \theta, a, b, c)\, g(\theta|\tau)\, g(\tau)\, g(a, b, c|\eta)\, g(\eta)$$

Daí, para estimarmos os parâmetros dos itens a partir da equação acima, a função de verossimilhança é marginalizada com respeito à habilidade. Além disso, a distribuição dos parâmetros dos itens pode ser removida da equação se integrarmos com respeito a suas distribuições de probabilidade.

Então, integrando a distribuição de probabilidade da habilidade $g(\theta|\tau)$ com respeito a θ e os parâmetros populacionais dos itens $g(\eta)$ com respeito a η, temos a distribuição marginal a posteriori:

$$g(a, b, c, \tau \mid X) \propto \iint L(X \mid a, b, c, \theta)\, g(\theta|\tau)\, g(a, b, c|\eta)\, g(\tau)\, g(\eta)\, d\theta d\eta$$

$$\propto L(X \mid a, b, c, \tau)\, g(a, b, c)\, g(\tau) \qquad (3.3.3)$$

onde $L(X \mid a, b, c, \tau)$ é a verossimilhança marginal, resultante de $L(X \mid a, b, c, \theta)$ ter sido integrada com respeito à habilidade.

Note que, integrando sobre a distribuição populacional da habilidade, eliminamos a dependência das estimativas dos parâmetros dos itens nas estimativas das habilidades individuais. Entretanto, a verossimilhança marginal ainda é condicional aos hiperparâmetros da distribuição populacional das habilidades e, então, $g(\tau)$ e os valores desses hiperparâmetros deverão ser especificados.

Analogamente, a integração sobre a distribuição populacional dos parâmetros dos itens não elimina a necessidade de especificar os valores dos hiperparâmetros η em g(a,b,c).

A distribuição marginal a posteriori dada pela equação (3.3.3) é que será maximizada para a obtenção das estimativas dos parâmetros dos itens.

Devido à hipótese de independência entre os itens, podemos simplificar o processo considerando a estimação de um item de cada vez e repetindo o processo n vezes. Assim, para estimar os parâmetros do item i, as derivadas parciais da equação (3.3.3) são tomadas com respeito aos parâmetros do item e igualadas a zero. Por conveniência, podemos trabalhar com o logaritmo natural dessa equação. Assim, o sistema de equações a ser resolvido será:

$$\frac{\partial}{\partial z_i}[\ \ln L(X \mid a,b,c,\tau)\ ] + \frac{\partial}{\partial z_i}[\ \ln g(a,b,c)\ ] + \frac{\partial}{\partial z_i}[\ \ln g(\tau)\ ] = 0$$

onde z_i representa um dos três parâmetros do item i, isto é, a_i, b_i ou c_i .

Como a distribuição $g(\tau)$ não depende de nenhum dos parâmetros do item, sua derivada com respeito a z_i será sempre zero. Assim, podemos simplificar o sistema de equações a ser resolvido:

$$\frac{\partial}{\partial z_i}[\ \ln L(X \mid a,b,c,\tau)\ ] + \frac{\partial}{\partial z_i}[\ \ln g(a,b,c)\ ] = 0 \qquad (3.3.4)$$

Note que o sistema dado por (3.3.4) é formado por duas componentes: o primeiro termo é a componente da verossimilhança e o segundo é a componente da priori. A verossimilhança é simplesmente dada por

$$L(X \mid a,b,c,\tau) = \prod_{j=1}^{n} P_j(x_j \mid a,b,c,\tau)$$

mas a distribuição da priori g(a,b,c) precisa ser definida. Uma das possibilidades para essas distribuições a priori dos parâmetros dos itens, que se encontra implementada computacionalmente, é utilizar a

distribuição log-normal para o parâmetro a, a distribuição normal para o b e a distribuição beta para o parâmetro c.

A escolha dessas distribuições a priori certamente não é única, mas foi feita de maneira conveniente, de acordo com as características de cada parâmetro. Sob os modelos utilizados pela TRI não são esperados valores negativos para o parâmetro a. Por essa razão, a distribuição log-normal é uma boa opção. O parâmetro b é medido na mesma escala da habilidade dos respondentes. Como a distribuição normal tem uma forma que representa bem essa distribuição das habilidades e, além disso, é uma distribuição de probabilidade que possui propriedades conhecidas, sua escolha é bastante natural. Já o parâmetro c, deve assumir somente valores entre 0 e 1. Assim, foi feita a escolha da distribuição beta.

Uma vez definidas as prioris, teremos que resolver, para cada um dos itens, um sistema formado por 3 equações, determinadas por (3.3.4). Mais detalhes sobre como chegar na forma final desse sistema podem ser encontrados, por exemplo, em Baker (1992).

A solução simultânea dessas 3 equações não é um cálculo trivial. No entanto, podemos notar que nesse ponto, teremos um problema bastante semelhante ao do caso de estimação por Máxima Verossimilhança Marginal e, analogamente àquele caso, o algoritmo EM poderá ser empregado. Além disso, como o sistema de equações a ser resolvido é não linear, um procedimento iterativo como Newton-Raphson deverá ser utilizado no passo M (de Maximização) do algoritmo EM, para a obtenção das estimativas dos parâmetros de item. Essas estimativas serão então usadas na fase E (do cálculo da Esperança) do próximo ciclo do EM, e este processo se repetirá até que algum critério de convergência seja satisfeito.

Também de forma análoga ao caso de estimação por MVM, uma vez estimados os parâmetros dos itens, os parâmetros de habilidade dos respondentes podem ser estimados considerando-se os parâmetros dos itens como conhecidos e escolhendo-se um dos métodos de estimação ilustrados em 3.1 .

Na verdade, no caso da estimação dos parâmetros dos itens, o procedimento de estimação bayesiano modal marginalizado que acabamos de descrever, pode ser considerado como uma extensão do processo de estimação por MVM. Mas a grande vantagem desse procedimento sobre o de MVM e também sobre os anteriores é que as

estimativas para os parâmetros dos itens estarão definidas em todas as situações.

3.4 Estimação para duas ou mais populações

Os problemas de estimação para modelos envolvendo duas ou mais populações, em geral são extensões diretas dos métodos aqui apresentados. Por isso, vamos apenas ilustrar a estimação para dois ou mais grupos através de uma das técnicas possíveis: o método da máxima verossimilhança marginal.

Para que seja possível o processo de estimação conjunta de indivíduos vindos de duas ou mais populações, a única exigência é que eles sejam submetidos a itens comuns. Mas, para facilitar a explicação do processo de estimação, vamos supor que todos os indivíduos dos diversos grupos sejam submetidos ao mesmo teste, ou seja, que todos os I itens que compõem o teste sejam comuns. Vamos então supor que temos g grupos, com n_k indivíduos no k-ésimo grupo, k = 1, 2, ..., g.

Seja $X_{jk} = [\ X_{1jK},\ X_{2jk},\ ...,\ X_{ijk}\ ]^t$ o vetor aleatório (I x 1) que representa o j-ésimo padrão de respostas do grupo k, j = 1, 2, ..., 2^I. Vamos trabalhar com o número de ocorrências distintas de padrões de resposta em cada um dos grupos. Assim, seja r_{jk} o número de ocorrências do j-ésimo padrão de respostas no grupo k. Seja também $s_k \leq \min(n_k, 2^I)$, o número de padrões de resposta com $r_{jk} > 0$. Note que então devemos ter:

$$\sum_{j=1}^{s_k} r_{jk} = n_k$$

A probabilidade marginal do vetor X_{jk} pode ser escrita como:

$$P_M(X_{jk}) = \int_{-\infty}^{+\infty} P(X_j|\theta, a, b, c)\ g(\theta|\tau_k)\ d\theta$$

onde:

$P(X_j|\theta, a, b, c) = \prod_{i=1}^{I} P(X_{ijk} = 1 |\ \theta_{jk})$ sendo que $P(X_{ijk} = 1 |\ \theta_{jk})$ é o modelo logístico unidimensional para dois ou mais grupos apresentado em 2.2 e $g(\theta|\tau_k)$ é a distribuição a priori de θ para o grupo k.

Considerando a independência entre as respostas dos indivíduos nos diferentes grupos, temos que os dados seguem uma distribuição Produto de Multinomiais, dada por:

$$L_M(a,b,c,\tau) = \prod_{k=1}^{g} \left\{ \frac{n_k!}{\prod_{j=1}^{s_k} r_{jk}!} \prod_{j=1}^{s_k} [\, P_M(X_{jk}) \,]^{r_{jk}} \right\}$$

Logo, tomando o logaritmo natural da equação acima, teremos:

$$\ln L_M(a,b,c,\tau) = \sum_{k=1}^{g} \ln \left\{ \frac{n_k!}{\prod_{j=1}^{s_k} r_{jk}!} \right\} + \sum_{k=1}^{g} \sum_{j=1}^{s_k} r_{jk} \ln P_M(X_{jk})$$

Deveremos então tomar a derivada da equação acima e igualá-la a zero com relação aos 3 parâmetros de cada um dos I itens, ou seja, 3I vezes. Serão portanto, 3I equações que não podem ser escritas de forma fechada a serem resolvidas simultaneamente e, por isso, é preciso calculá-las numericamente através de processos iterativos como os já descritos, por exemplo, na seção 3.3.2. Com relação aos parâmetros populacionais τ_k o procedimento é semelhante, mas se a distribuição a priori de θ para o grupo k não for considerada normal, a obtenção dessas estimativas torna-se bem mais complexa.

Há também o problema de indeterminação do modelo que, nesse caso, não se resolve simplesmente ao estabelecer-se distribuições a priori para as habilidades, como no caso de um único grupo, descrito em 3.3.2. Aqui, devido à existência de mais de um grupo de respondentes, é necessário também que se defina uma das populações como a população de referência. Dessa maneira, seus parâmetros populacionais já estarão pré-determinados e, portanto, não precisarão ser estimados. E, então, as demais populações serão posicionadas com relação a essa população de referência. Em geral, utiliza-se a mesma priori para todos os grupos – que para facilitar os processos de estimação – usualmente é a distribuição normal com média 0 e desvio padrão 1.

Maiores detalhes sobre esse processo de estimação podem ser encontrados em Bock e Zimowski (1997). Outros problemas envolvendo dois ou mais grupos serão descritos no Tópico 4.

3.5 Sumário

Para finalizar, iremos listar de forma resumida as principais características, vantagens e desvantagens dos métodos de estimação

apresentados nas seções 3.1 e 3.3, que são as duas situações que iremos encontrar nos problemas práticos.

3.5.1 Estimação da habilidade

MV	⇒	**maximizar a função de verossimilhança**
	⇑	para testes "longos" produz estimadores não viciados
	⇓	não está definido para alguns padrões de resposta

EAP	⇒	**esperança da distribuição a posteriori de θ**
	⇑	definido para qualquer padrão de resposta
	⇑	possui o menor erro médio
	⇓	viciado
	⇓	exige cálculos mais complexos do que o método de MV
	⇓	necessidade de uma distribuição a priori para θ

MAP	⇒	**máximo da distribuição a posteriori de θ**
	⇑	definido para qualquer padrão de resposta
	⇓	viciado
	⇓	exige cálculos mais complexos do que o método de MV
	⇓	necessidade de uma distribuição a priori para θ

3.5.2 Estimação dos parâmetros de item e das habilidades

MVC	⇒	**maximizar a função de verossimilhança**
	⇑	serve como base para outros procedimentos
	⇓	não está definido para alguns padrões de resposta
	⇓	viciado
	⇓	apresenta problemas de indeterminação
	⇓	não possui propriedades assintóticas, pois o aumento do número de respondentes aumenta o número de parâmetros a serem estimados

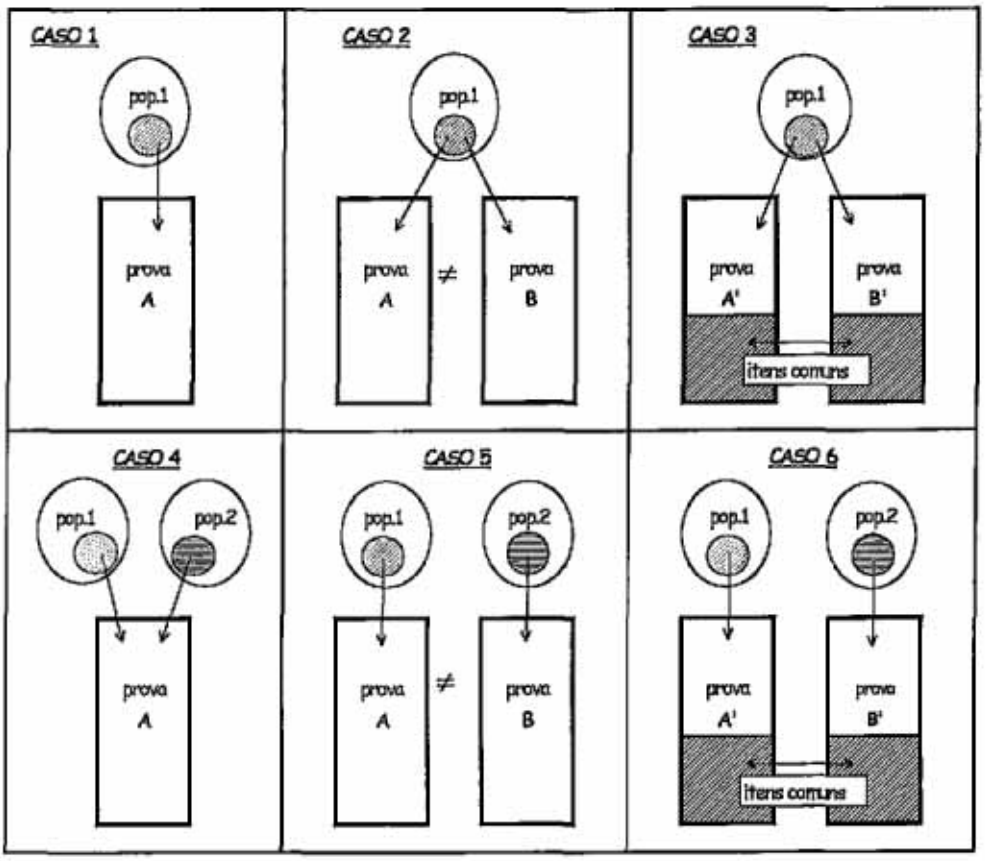
MVM	⇒ **maximizar a função de verossimilhança marginal** ⇑ possui propriedades assintóticas: se o número de respondentes aumenta, as estimativas dos parâmetros a, b e c são consistentes. ⇑ uma vez estimados os parâmetros dos itens, pode-se estimar as habilidades através de um dos métodos em 3.4.1 ⇓ não está definido para itens com acerto total ou erro total ⇓ mais trabalhoso computacionalmente do que o método de MVC ⇓ necessidade de uma distribuição a priori para θ ⇓ apresenta problemas na estimação do parâmetro c em alguns casos; deveria ser usado somente com um número suficientemente grande de respondentes

bayesiano	⇒ **maximizar a distribuição marginal a posteriori** ⇑ definido para qualquer padrão de resposta ⇑ uma vez estimados os parâmetros dos itens, pode-se estimar as habilidades através de um dos métodos em 3.4.1 ⇓ é bastante trabalhoso computacionalmente ⇓ necessidade de distribuições a priori para as habilidades e para os parâmetros dos itens

4.0 EQUALIZAÇÃO

No tópico anterior, apresentamos os métodos de estimação mais utilizados quando todos os parâmetros dos itens de uma única prova devem ser estimados. No entanto, esta é apenas uma das possíveis situações que na prática iremos encontrar. A seguir, listaremos os 6 casos possíveis, quanto ao número de grupos e de tipos de prova envolvidos. Esses casos estão esquematizados na Figura 4.1.

Figura 4.1 Representação gráfica das 6 situações que serão abordadas quanto ao número de grupos e de tipos de provas



1. Um único grupo fazendo um único tipo de prova.
2. Um único grupo fazendo 2 tipos de prova, totalmente distintos (nenhum item comum).
3. Um único grupo fazendo 2 tipos de prova, apenas parcialmente distintos, ou seja, com alguns itens comuns.
4. Dois grupos fazendo um único tipo de prova.
5. Dois grupos fazendo 2 tipos de prova, totalmente distintos (nenhum item comum).
6. Dois grupos fazendo 2 tipos de prova, apenas parcialmente distintos, ou seja, com alguns itens comuns.

Note que para simplificar, listamos os casos acima utilizando apenas 2 tipos de provas e 2 populações, mas as situações envolvendo um número maior de provas e/ou de populações são análogas.

Além disso, os problemas de estimação também podem diferir dependendo do conjunto de itens que necessita ser estimado, ou seja, se nosso conjunto de itens é composto de:

(a) apenas itens novos (ou seja, itens que ainda não foram calibrados);

(b) apenas itens já calibrados;

(c) itens novos e itens calibrados.

Em primeiro lugar, é importante definir o conceito de **Equalização** (ver Kolen e Brennan (1995), por exemplo), que é um dos mais importantes da TRI e um dos grandes objetivos das Avaliações Educacionais. Equalizar significa equiparar, tornar comparável, o que no caso da TRI significa colocar parâmetros de itens vindos de provas distintas ou habilidades de respondentes de diferentes grupos, na mesma métrica, isto é, numa escala comum, tornando os itens e/ou as habilidades comparáveis.

Existem dois tipos de equalização: a equalização **via população** e a equalização **via itens comuns**. Isto significa que há duas maneiras de colocar parâmetros, tanto de itens quanto de habilidades, numa mesma métrica: na primeira usamos o fato de que se um único grupo de respondentes é submetido a provas distintas, basta que todos os itens sejam calibrados conjuntamente para termos a garantia de que todos estarão na mesma métrica. Já na equalização via itens comuns, a garantia de que as populações envolvidas terão seus parâmetros em uma única escala será dada pelos itens comuns entre as populações, que servirão de ligação entre elas.

4.1 Diferentes tipos de situações encontradas em avaliações educacionais, quanto ao número de grupos e de tipos de provas

Uma vez listadas as diversas situações que podemos ter, vamos agora discutir cada uma delas. Obviamente, podemos ter os casos 1 a 6 combinados com as situações (a) a (c). Mas, mais uma vez para facilitar a explicação, trataremos inicialmente dos casos 1 a 6 considerando sempre a situação mais simples, ou seja, o caso (a).

Cabe ainda ressaltar que todas as análises e comentários deste tópico serão feitos considerando-se o modelo logístico unidimensional de 3 parâmetros.

4.1.1 Um único grupo fazendo um único tipo de prova

Este é o caso trivial, em que se aplicam diretamente os modelos matemáticos e os métodos de estimação descritos nos tópicos anteriores. Foi o caso considerado até agora, nos tópicos 2 e 3, e pela própria natureza do problema não é necessário nenhum tipo de equalização.

Um exemplo para ilustrar este caso seria uma prova de 30 itens aplicada à 4ª série diurna do Ensino Fundamental da rede pública estadual de São Paulo.

4.1.2 Um único grupo fazendo 2 tipos de prova, totalmente distintos

Este é um caso clássico do que chamamos de equalização via população. Para resolvê-lo, basta que todos os itens de ambas as provas sejam calibrados simultaneamente. O fato de todos os indivíduos representarem uma amostra aleatória de uma mesma população é que garante que todos os parâmetros envolvidos estarão na mesma escala.

Um exemplo para este caso seria quando 2 provas distintas (tipo A e tipo B), com 30 itens cada, são aplicadas, de maneira aleatória, aos alunos da 4ª série diurna do Ensino Fundamental da rede pública estadual de São Paulo. Ao final dos processos de estimação, todos os resultados obtidos serão comparáveis, não importando a que tipo de prova cada aluno foi submetido.

4.1.3 Um único grupo fazendo 2 tipos de prova, apenas parcialmente distintos, ou seja, com alguns itens comuns

Este caso é bastante semelhante ao caso anterior, e aqui também podemos fazer a equalização via população. Assim, valem os mesmos comentários do caso 4.1.2.

Um exemplo dessa situação seria a aplicação de 2 tipos de provas (tipo A e tipo B), com 30 itens cada e com 10 itens comuns entre elas, aos alunos da 4ª série diurna do Ensino Fundamental da rede pública estadual

de São Paulo. Aqui, o número total de itens a serem calibrados seria 50 (= 30 + 30 – 10). Analogamente ao exemplo anterior, ao final dos processos de estimação todos os resultados obtidos serão comparáveis, não importando a que tipo de prova cada aluno foi submetido.

Outro exemplo bastante interessante para este caso, seria a aplicação do SAEB – Sistema Nacional de Avaliação da Educação Básica. Nesse estudo, uma das populações alvo é a 3ª série do Ensino Médio. Como a aplicação é de caráter nacional, alunos de vários estados do país são avaliados, mas todos são considerados como respondentes vindos da mesma população, ou seja, como um único grupo. Além disso, o SAEB procura cobrir a grade curricular de forma completa, e para tanto, é considerado um grande número de itens distintos em cada disciplina. Como seria inviável a aplicação de todos os itens a um único aluno, as provas são montadas segundo um esquema BIB – Blocos Incompletos Balanceados – no qual os itens são divididos em blocos, que por sua vez são reunidos em cadernos, e estes cadernos – que nada mais são do que tipos distintos de provas -, é que são aplicados aos alunos. No caso da 3ª série do Ensino Médio, os itens foram divididos em 13 blocos com 13 itens distintos cada um. Foram então montados 26 cadernos, cada um composto por 3 blocos distintos. Assim, cada aluno responde a 39 itens. É importante notar que diferentes blocos não têm itens comuns entre si, mas que diferentes cadernos podem – ou não – ter itens comuns: basta que tenham algum bloco em comum. Concluindo, desta maneira foram aplicados diferentes tipos de provas – representados pelos 26 cadernos – com itens comuns a um único grupo de respondentes – alunos da 3ª série do Ensino Médio brasileiro.

O SAEB também é um bom exemplo prático do que chamamos de provas com itens não apresentados. Podemos considerar que a prova é composta dos 169 itens, mas que apenas 39 são submetidos a cada aluno. Consequentemente, temos 130 itens que não foram apresentados para cada aluno. Quando temos provas com um número originalmente grande de itens, podemos resolver o problema utilizando esquemas semelhantes ao usado no SAEB. Assim, o que inicialmente poderia ser considerado como um único tipo de prova, pode vir a ser considerado como vários tipos diferentes, se não submetermos todos os itens a todos os alunos.

4.1.4 Dois grupos fazendo um único tipo de prova

Este é um exemplo de equalização via itens comuns (só que no caso, todos). Como as 2 populações fazem exatamente a mesma prova, basta que os itens sejam calibrados utilizando-se as respostas dos respondentes de ambos os grupos simultaneamente. Para tanto, devemos apenas utilizar um modelo para duas populações, como apresentado no tópico 2.

Um exemplo para este caso seria a aplicação de uma única prova, composta de 40 itens, nos períodos diurno (população 1) e noturno (população 2) da 8ª série do Ensino Fundamental da rede pública estadual de São Paulo. Ao final dos processos de estimação, todos os resultados obtidos serão comparáveis, não importando a que população o aluno pertence.

4.1.5 Dois grupos fazendo 2 tipos de prova, totalmente distintos

Este é o único dos 6 casos que não pode ser resolvido pela TRI. Obviamente é possível calibrar separadamente os itens das duas provas, mas o problema é que não podemos fazer nenhum tipo de comparação entre os resultados obtidos, uma vez que eles estarão em métricas diferentes. Neste caso, não faz sentido comparar os resultados destes 2 grupos, assim como não faz sentido comparar diretamente 40° C com 40° F . Assim como essas duas temperaturas estão em escalas de medida diferentes, os parâmetros obtidos nestas 2 provas também estarão. A diferença é que, no caso das temperaturas, há uma relação conhecida entre as duas escalas, e, assim, é possível colocarmos uma das temperaturas na mesma escala que a outra, possibilitando, então, a comparação. Já no caso das provas, não existe nenhuma relação entre elas e nem entre os dois grupos que torne possível a comparação.

Um exemplo que ilustra esta situação seria a elaboração de 2 provas distintas: uma, composta de 30 itens, seria aplicada à 4ª série diurna (população 1) e a outra prova, composta de 40 itens, seria aplicada à 5ª série diurna (população 2) do Ensino Fundamental da rede pública estadual de São Paulo. Estas duas provas poderiam ser calibradas separadamente e seus resultados poderiam ser interpretados isoladamente dentro de cada série, mas não poderíamos comparar os resultados dos itens e nem das habilidades estimadas para os indivíduos das duas séries.

4.1.6 Dois grupos fazendo 2 tipos de prova, apenas parcialmente distintos, ou seja, com alguns itens comuns

Finalmente, vamos comentar o último caso, em que 2 grupos são submetidos a 2 tipos de provas diferentes, mas que têm alguns itens comuns. Assim como o caso 4.1.4, este também é um exemplo de equalização via itens comuns. Este caso representa o melhor exemplo do uso e da importância da equalização e, sem dúvida, ilustra o maior avanço da TRI sobre a Teoria Clássica. O uso de itens comuns entre provas distintas aplicadas a populações distintas permite que todos os parâmetros estejam na mesma escala ao final dos processos de estimação, possibilitando comparações e a construção de "escalas do conhecimento" interpretáveis, que são de grande importância na área educacional. A resolução deste caso é bastante semelhante ao que foi descrito no caso 4.1.4, com a diferença que aqui apenas alguns dos itens (e não a prova toda) fazem a ligação entre as 2 populações envolvidas. Este caso será abordado mais detalhadamente através de um exemplo prático, que será apresentado no tópico 6.

Um exemplo que ilustra bem esta situação seria a aplicação de uma prova com 30 itens à 3ª série diurna (população 1) e de outra prova, também com 30 itens, à 4ª série diurna da rede pública estadual de São Paulo (população 2). Entre elas poderiam haver 10 itens comuns (por exemplo, 10 itens da matriz curricular da 3ª série). Desta maneira, no final do processo de estimação teríamos todos os 50 itens numa mesma métrica, possibilitando comparações entre alunos de 3ª e 4ª séries, e também possibilitando a criação de uma "escala de conhecimento" da 3ª e da 4ª série nesta dada disciplina. Como veremos no tópico 6, esta escala possibilitaria a verificação dos conteúdos que os alunos destas duas séries dominam, dos conteúdos onde há falhas, acompanhar a "evolução do conhecimento" de uma série para outra, etc.

4.2 Diferentes problemas de estimação encontrados em avaliações educacionais, quanto ao conjunto de itens a ser estimado

Vamos agora considerar outro ponto bastante importante na TRI: o conjunto de itens a ser calibrado. Vamos comentar inicialmente os casos

(a) a (c), considerando-se o caso 1, ou seja, o caso em que uma única prova foi aplicada a um único grupo de respondentes.

4.2.1 Quando todos os itens são novos

Neste caso, todos os itens são considerados "novos", ou seja, deseja-se calibrar o conjunto completo de itens. Este é o caso trivial, que foi considerado até agora. Para resolver este problema basta utilizar alguma das técnicas de estimação descritas no tópico anterior.

Trata-se exatamente da mesma situação descrita em 4.1.1 e, portanto, poderíamos utilizar o mesmo exemplo: a aplicação de uma prova, composta de 30 itens novos (ou seja, com 30 itens que desejamos calibrar), aos alunos da 4ª série diurna da rede pública estadual de São Paulo.

4.2.2 Quando todos os itens já estão calibrados

Este é o caso em que todos os itens já foram calibrados anteriormente, ou seja, quando não desejamos calibrar nenhum dos itens e estamos interessados apenas em estimar as habilidades dos respondentes. Este é um caso também bastante freqüente na TRI, devido ao impulso que esta teoria deu na criação de bancos de itens. Tais bancos são formados por conjuntos de itens que já foram testados e calibrados a partir de um número significativo de indivíduos de uma dada população. Desta maneira, assumimos que os parâmetros desses itens já são "conhecidos", ou seja, assumimos que conhecemos os verdadeiros valores dos parâmetros desses itens e, assim, sempre que desejarmos, podemos aplicar novamente alguns desses itens do banco a outros indivíduos (ou até mesmo a um único indivíduo) e poderemos então estimar apenas suas habilidades, que estarão sempre na mesma métrica das habilidades do grupo de indivíduos utilizado na calibração inicial.

A questão da métrica é um ponto que deve ser considerado com bastante cuidado numa situação como esta. Quando se "constrói" um banco de itens, uma informação fundamental é a escala em que aqueles itens foram calibrados. Isto porque as habilidades de indivíduos que serão estimadas futuramente a partir daqueles itens estarão nesta mesma métrica e, portanto, quaisquer comparações diretas só poderão ser feitas com outros sujeitos que também tenham suas habilidades nesta escala.

Assim, para resolver este problema, basta utilizar um dos processos de estimação das habilidades dos indivíduos quando os parâmetros dos itens já são conhecidos, que foram descritos na seção 3.1 do tópico 3.

Um exemplo para este tipo de situação seria a aplicação de uma prova, composta de 30 itens que já foram calibrados numa aplicação anterior (por exemplo, numa aplicação de nível nacional como o SAEB), feita a uma amostra de alunos da 4ª série diurna da rede pública estadual brasileira, aos alunos da 4ª série diurna da rede pública estadual de São Paulo. Este tipo de procedimento é bastante comum e, nesse caso, o objetivo seria comparar a rede pública paulista com o desempenho nacional.

4.2.3 Quando alguns itens são novos e outros já estão calibrados

Neste caso, temos itens "novos" e itens já calibrados, ou seja, desejamos calibrar alguns itens e manter os parâmetros de outros, que já foram calibrados anteriormente. Este também é uma situação que está tipicamente ligada à criação de bancos de itens. Isto porque um banco de itens está continuamente em formação, ou seja, é bastante comum estarmos interessados em acrescentar novos itens ao conjunto que já se encontra no banco (assim como também é comum a retirada de itens do banco). Neste caso, o problema fundamental é garantir que os itens novos sejam calibrados na mesma métrica em que estão os outros itens do banco.

Na prática, este é um problema de solução mais complexa do que possa parecer a princípio. Isto porque é indispensável o uso de programas computacionais especificamente desenvolvidos para a análise de itens via TRI e esses programas ainda apresentam algumas dificuldades com relação a situações como essa. Vamos comentar especificamente os problemas que podem surgir em casos como esse, quando utilizamos o BILOG ou o BILOG-MG, no próximo tópico.

Um exemplo para esse caso seria a aplicação de uma prova, composta de 30 itens, aos alunos da 4ª série diurna da rede pública estadual de São Paulo. Desses 30 itens, 15 são itens novos e 15 são itens que já foram calibrados numa aplicação de nível nacional do SAEB, utilizando-se uma amostra de alunos da 4ª série diurna da rede pública brasileira. Na prática, esta é uma situação bastante comum, pois quando

são feitas avaliações regionais, por um lado há o interesse em criar e aplicar itens novos, mas por outro lado, há também o interesse em que os resultados obtidos possam ser comparados aos resultados nacionais.

4.3 Considerações finais sobre equalização

Ilustramos até aqui os casos (a), (b) e (c) considerando-se a situação 1. As outras situações onde tratamos apenas de uma população (situações 2 e 3), são análogas. No entanto, quando temos 2 (ou mais) populações envolvidas (situações 4 e 6), e desejamos estimar itens novos e manter fixos os parâmetros dos itens já calibrados (caso (c)), poderemos ter problemas com a métrica. Os casos (a) e (b) não apresentam problemas, sendo análogos à situação anterior.

Sempre que há mais de uma população envolvida nos processos de estimação, como já foi comentado anteriormente, existem problemas de indeterminação de escala. Para resolver este problema, devemos definir uma das populações como sendo a referência e, então, as demais populações serão posicionadas com relação a ela.

Este tipo de problema sempre irá ocorrer quando fazemos a equalização entre 2 ou mais populações durante o processo de estimação dos itens. No entanto, há uma outra maneira de solucionarmos o problema de maneira adequada: poderíamos optar por fazer uma equalização a posteriori, que será discutida a seguir.

4.3.1 Equalização a posteriori

Até aqui discutimos formas de equalização entre 2 ou mais populações feitas durante o próprio processo de estimação dos parâmetros. Mas, também, é possível fazer a equalização a posteriori, isto é, depois de terminado o processo de calibração dos itens. Basicamente, a equalização a posteriori é feita da seguinte maneira: calibram-se separadamente os dois conjuntos de itens, que foram submetidos às duas populações de interesse. Obviamente, a condição necessária é que hajam itens comuns entre os dois conjuntos. Assim, para os itens comuns, teremos dois conjuntos de estimativas, cada uma na métrica de suas respectivas populações. Daí, através dessas duas estimativas para os itens comuns estabelece-se algum tipo de relação que permita colocarmos os parâmetros de um dos conjuntos de itens na escala do outro. Com todos

os itens na mesma métrica, pode-se então estimar as habilidades de todos os respondentes, que então estarão também na mesma escala.

Pela propriedade de invariância, já discutida no tópico 2, dado que o modelo é adequado aos dados, os parâmetros **a** e **b** de um certo item apresentado a 2 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 colocados na mesma escala das estimativas do grupo 1.

Vários métodos, que se baseiam nessas relações lineares existentes entre os parâmetros de um mesmo item medidos em escalas diferentes, poderiam ser então utilizados para determinar os coeficientes α e β. A solução mais natural – pelo próprio tipo de relação existente entre os parâmetros – seria determinar esses coeficientes através de uma regressão linear simples. No entanto, a crítica feita à utilização desse método é que ele não é simétrico, ou seja, uma regressão de x por y é diferente de uma regressão de y por x.

Um dos métodos de equalização a posteriori existentes que não apresenta esse problema, ou seja, é invariante (simétrico) em relação às variáveis utilizadas, é denominado **Média-Desvio** (Mean-Sigma). O método Média-Desvio 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 1 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. Maiores detalhes sobre este e outros métodos de equalização podem ser encontrados em Kolen e Brennan (1995), por exemplo.

Exemplificando, uma avaliação feita no estado do Rio Grande do Norte (Fundação Carlos Chagas, 1997) utilizou alguns itens do SAEB 95, com o intuito de colocar os resultados obtidos na mesma métrica do SAEB. Os gráficos a seguir mostram as relações entre as estimativas dos parâmetros **a** e **b** nas duas avaliações, para a disciplina Língua Portuguesa da 8ª série do Ensino Fundamental.

Figura 4.2 Gráfico de dispersão das estimativas do parâmetro de dificuldade – b dos itens comuns da prova de Língua Portuguesa da 8ª série entre o RN e o SAEB

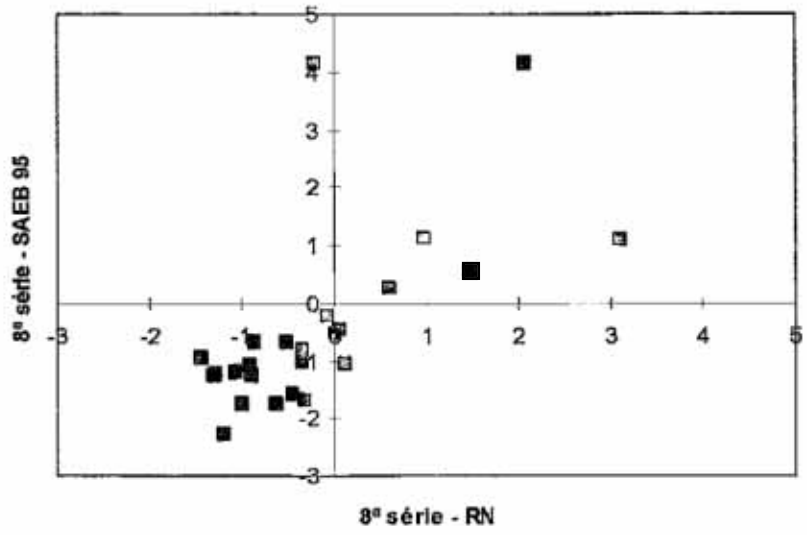
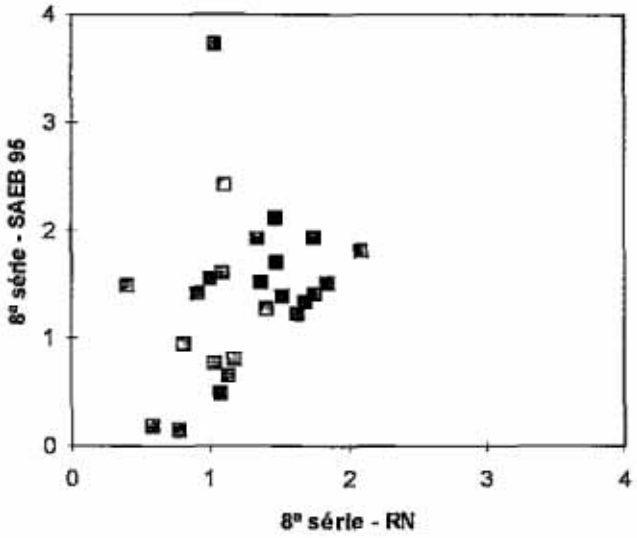


Figura 4.3 Gráfico de dispersão das estimativas do parâmetro de discriminação – a dos itens comuns da prova de Língua Portuguesa da 8ª série entre o RN e o SAEB



Utilizando o método Média-Desvio, os coeficientes α e β obtidos foram:

$$\alpha = \frac{S_{SAEB}}{S_{RN}} = \frac{1{,}614}{1{,}104} = 1{,}462$$

$$\beta = M_{SAEB} - \alpha M_{RN} = -0{,}363 - 1{,}462 \times -0{,}162 = -0{,}126$$

Logo, as estimativas dos parâmetros obtidas na avaliação feita com os alunos do Rio Grande do Norte foram colocadas na mesma métrica do SAEB 95 através das seguintes expressões:

$$a_{RN}^{novo} = \frac{1}{\alpha}\, a_{RN} = \frac{1}{1{,}462}\, a_{RN}$$

$$b_{RN}^{novo} = \alpha\, b_{RN} + \beta = 1{,}462\, b_{RN} - 0{,}126$$

$$\theta_{RN}^{novo} = \alpha\, \theta_{RN} + \beta = 1{,}462\, \theta_{RN} - 0{,}126$$

Uma última observação sobre equalização deve ser feita com relação à quantidade de itens comuns. Certamente, quanto maior o número de itens comuns, melhor será a qualidade da equalização. Assim, o melhor caso de equalização entre 2 grupos distintos é a situação do caso 4, ou seja, quando se trata exatamente da mesma prova. No entanto, já sabemos que não é necessário que todos os itens sejam comuns. O número mínimo de itens comuns necessários para uma boa equalização entre 2 populações depende basicamente de dois fatores: do tipo de equalização que será feita e da "qualidade" desses itens comuns.

Equalizações feitas durante o processo de calibração são mais "eficazes" e portanto, exigem um número menor de itens comuns do que equalizações feitas a posteriori. Além disso, se os itens comuns utilizados na equalização tiverem níveis de dificuldade baixos ou altos demais com relação às populações envolvidas ou, então, se apresentarem baixo poder de discriminação, haverá necessidade de um número maior de itens.

Alguns autores têm sugerido pelo menos 6 itens comuns entre 2 provas de 30 itens, quando a equalização é feita durante a calibração. Um estudo de simulação considerando diferentes situações de equalização pode ser encontrado em Andrade (1999).

4.4 Construção e interpretação de escalas de habilidade

Uma vez que todos os parâmetros dos itens e que todas as habilidades dos respondentes – tanto individuais como populacionais – de todos os grupos avaliados estão numa mesma métrica, ou seja, quando todos os parâmetros envolvidos são comparáveis, pode-se então construir escalas de conhecimento interpretáveis.

Devido à natureza arbitrária das estimativas dos parâmetros dos itens e das habilidades, já comentada anteriormente, sabemos que podemos comparar entre si as habilidades obtidas para os diferentes respondentes, mas que, no entanto, elas não possuem "de per si" qualquer significado prático em termos pedagógicos. Assim, a menos que se efetue uma ligação desses valores com os conteúdos envolvidos na avaliação, pode-se dizer apenas que um indivíduo com habilidade 1,80 na escala (0,1) deve possuir um conhecimento muito maior do conteúdo avaliado do que um indivíduo com habilidade –0,50, e também que o primeiro indivíduo tem uma habilidade 1,80 desvios padrão acima da média da população avaliada enquanto que o segundo tem habilidade 0,50 desvios

padrão abaixo da média dessa mesma população. Por outro lado, não podemos afirmar nada a respeito do que o indivíduo com habilidade 1,80 sabe a mais do que aquele com habilidade –0,50.

Estes fatos motivaram então a criação de escalas de conhecimento – também chamadas de escalas de habilidade –, que tornam possível a interpretação pedagógica dos valores das habilidades. Essas escalas são definidas por níveis âncora, que por sua vez são caracterizados por conjuntos de itens denominados itens âncora. Níveis âncora são pontos selecionados pelo analista na escala da habilidade para serem interpretados pedagogicamente. Já os itens âncora são itens selecionados, segundo a definição dada abaixo, para cada um dos níveis âncora.

Definição de item âncora: Considere dois níveis âncora consecutivos Y e Z com $Y < Z$. Dizemos que um determinado item é âncora para o nível Z se e somente se as 3 condições abaixo forem satisfeitas simultaneamente:

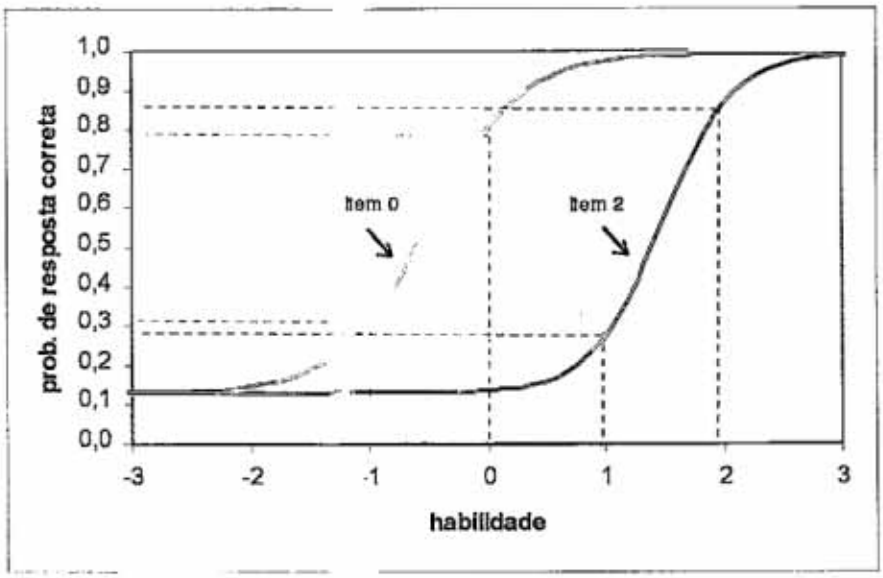
1. $P(X = 1 \mid \theta = Z) \geq 0{,}65$ e
2. $P(X = 1 \mid \theta = Y) < 0{,}50$ e
3. $P(X = 1 \mid \theta = Z) - P(X = 1 \mid \theta = Y) \geq 0{,}30$.

Em outras palavras, para um item ser âncora em um determinado nível âncora da escala, ele precisa ser respondido corretamente por uma grande proporção de indivíduos (pelo menos 65%) com este nível de habilidade e por uma pequena proporção de indivíduos (no máximo 50%) com o nível de habilidade imediatamente anterior. Além disso, a diferença entre a proporção de indivíduos com esses níveis de habilidade que acertam a esse item deve ser de pelo menos 30%. Assim, para um item ser âncora ele deve ser um item "típico" daquele nível, ou seja, bastante acertado por indivíduos com aquele nível de habilidade e pouco acertado por indivíduos com um nível de habilidade imediatamente inferior.

No gráfico a seguir são apresentados em uma escala de habilidade com níveis âncora -3, -2, -1, 0, 1, 2, e 3, exemplos de 2 itens âncora (item0 e item2) para os níveis âncora 0 e 2, respectivamente. Os parâmetros dos itens são:

$$a_0 = 1{,}52 \quad , \quad b_0 = -0{,}47 \quad \text{e} \quad c_0 = 0{,}13 \ ,$$
$$a_2 = 1{,}97 \quad , \quad b_2 = 1{,}50 \quad \text{e} \quad c_2 = 0{,}13 \ .$$

Figura 4.4 Exemplo de 2 itens âncora



A partir das expressões abaixo, pode-se verificar que os dois itens satisfazem a definição de item âncora:

(i) $P(X_0 = 1 \mid \theta = 0) = 0,80 \quad \geq 0,65$
(ii) $P(X_0 = 1 \mid \theta = -1) = 0,31 \quad \leq 0,50$
(iii) $P(X_0 = 1 \mid \theta = 0) - P(X_0 = 1 \mid \theta = -1) = 0,80 - 0,31 = 0,49 \geq 0,30$
e
(i) $P(X_2 = 1 \mid \theta = 2) = 0,86 \quad \geq 0,65$
(ii) $P(X_2 = 1 \mid \theta = 1) = 0,27 \quad \leq 0,50$
(iii) $P(X_2 = 1 \mid \theta = 2) - P(X_2 = 1 \mid \theta = 1) = 0,86 - 0,27 = 0,59 \quad \geq 0,30$

A priori, não se pode ter certeza de quantos itens âncora serão selecionados para cada nível âncora e nem se existirão no teste aplicado itens âncora para todos os níveis âncora determinados. Por isto, é fundamental que os níveis âncora sejam escolhidos não muito próximos uns dos outros e também que o número de itens aplicados seja bastante grande de modo a possibilitar a construção e interpretação da escala de habilidade. No SAEB, por exemplo, foram aplicados 130 itens para cada uma das disciplinas avaliadas na 4ª série do Ensino Fundamental, e 169 itens de cada uma das disciplinas da 8ª série do Ensino Fundamental e da 3ª série do Ensino Médio. Como já foi comentado anteriormente, essa

quantidade de itens foi aplicada visando cobrir amplamente a grade curricular de cada uma das séries nas disciplinas avaliadas, e também propiciou a identificação e caracterização de diversos níveis âncora para a construção das escalas de habilidade. Maiores detalhes sobre construção e interpretação de escalas de habilidade poderão ser encontrados em Beaton e Allen (1992).

Um último comentário, é que antes da construção das escalas é bastante comum fazer uma transformação linear em todos os parâmetros envolvidos. Tal procedimento tem como único objetivo facilitar a construção e utilização da escala, uma vez que procura transformar valores negativos ou decimais em números positivos e inteiros.

No próximo tópico, discutiremos alguns dos recursos computacionais disponíveis para a análise de dados via TRI. Em particular, descreveremos o desempenho de dois programas computacionais frente aos diferentes tipos de equalização abordados neste tópico.

5.0 RECURSOS COMPUTACIONAIS

Sem dúvida alguma, o crescimento e a divulgação da TRI sempre estiveram intimamente ligados ao desenvolvimento paralelo de recursos computacionais que viabilizassem sua utilização. Isto porque as ferramentas matemáticas necessárias para sua aplicação são muito mais complexas do que as técnicas empregadas na Teoria Clássica de Medidas.

Desde suas primeiras aplicações, alguns pesquisadores desenvolveram seus próprios programas computacionais, mas é certo que sua utilização em larga escala depende diretamente da disponibilidade de programas computacionais comerciais no mercado.

Na Europa e nos Estados Unidos, desde a década de 70, foram lançados alguns programas específicos para análises via TRI. Aqui no Brasil, onde a utilização da TRI é bem mais recente, há uma variedade bem menor de programas computacionais comerciais sendo usados.

Neste tópico, vamos comentar alguns programas computacionais comerciais que são os mais usados atualmente – no Brasil – e que se propõem a resolver, na prática, muitos dos problemas abordados pela TRI e que foram descritos nos tópicos anteriores.

O TESTFACT (Wilson et al., 1991) é um programa que produz várias estatísticas descritivas para os itens de um teste, inclusive algumas

das utilizadas pela Teoria Clássica, mas que também tem recursos importantes para a TRI, usados na verificação da dimensionalidade dos testes: técnicas de análise fatorial específicas para serem aplicadas em itens. Dois tipos especiais de análise fatorial, que foram elaboradas para variáveis dicotômicas (como é o caso dos itens, quando são considerados como certo ou errado), estão implementados neste programa. Uma delas é a análise fatorial feita a partir da matriz de correlação tetracórica, que é um tipo especial de correlação, utilizada quando as variáveis assumem apenas os valores 0 ou 1 (ver Divgi (1979), por exemplo). A outra técnica implementada é a análise fatorial plena, baseada no método de máxima verossimilhança (ver Mislevy (1986), por exemplo).

Para a análise de itens não dicotômicos, podemos citar o programa PARSCALE (Muraki e Bock, 1997), que tem implementados os modelos de Resposta Gradual e de Créditos Parciais, descritos no tópico 2. Em sua versão mais recente, é possível fazer análises para mais de um grupo de respondentes.

Dos programas disponíveis no mercado, os que são atualmente mais utilizados nas análises envolvendo a TRI – aqui no Brasil – são o BILOG (Mislevy e Bock, 1990) e o BILOG-MG (Zimowski et al., 1996). Estes dois programas são específicos para análises via TRI de itens dicotômicos e ambos têm implementados os modelos unidimensionais logísticos de 1, 2 e 3 parâmetros. A diferença básica entre eles é que o BILOG-MG permite a análise de mais de um grupo de respondentes, enquanto que o BILOG permite apenas analisar respondentes considerados como provenientes de uma única população.

Vamos comentar a seguir quais dos métodos de estimação descritos no tópico 3 estão implementados nestes dois programas e também dar uma ênfase especial ao desempenho deles perante as diversas situações que envolvem equalizações, descritas no tópico 4.

5.1 Considerações gerais sobre os programas BILOG for Windows v. 3.09 e BILOG-MG v 1.0

Esses dois programas executam a análise em três etapas, chamadas de fases 1, 2 e 3, que se caracterizam pelo tipo de tarefas realizadas em cada uma delas. Na **fase 1**, que é a fase de **entrada e leitura de dados**, o usuário deve fornecer ao programa basicamente dois tipos de informação: a identificação de cada indivíduo com suas respectivas respostas ao teste e

o gabarito (que é uma sequência contendo as alternativas corretas dos itens que compõem o teste). Também é possível fornecer as respostas já corrigidas, ou seja, já codificadas como 0 ou 1. Nesse caso não há a necessidade do gabarito, pois o programa irá interpretar 1 como acerto e 0 como erro. No caso de esquemas amostrais complexos, pode-se fornecer ao programa pesos diferentes para cada um dos indivíduos. Essas informações devem estar em arquivos do tipo ASCII. Os arquivos de saída, fornecidos ao usuário, também estarão neste formato.

Nessa fase é feita a "correção" da prova de cada sujeito (no caso de ter sido fornecido o arquivo com as respostas originais) e são calculadas algumas estatísticas descritivas, tais como: número de indivíduos submetidos a cada item, número e porcentagem de acerto em cada item e algumas correlações de interesse, como por exemplo, a correlação bisserial (ver Lord e Novick (1968), por exemplo), usada na Análise Clássica.

A importância dessa etapa do processamento, além da verificação de que a leitura dos dados foi feita corretamente, é que estas estatísticas serão utilizadas posteriormente como valores iniciais para os processos de estimação realizados nas fases seguintes. Além disso, estatísticas como a correlação bisserial servem para um diagnóstico preliminar dos itens, servindo, por exemplo, na identificação de itens com problemas no gabarito.

A **fase 2** é a fase da **calibração** dos itens. Nesta fase, são obtidos os parâmetros dos itens, com seus respectivos erros padrão. Os métodos de estimação disponíveis serão comentados na próxima seção. O BILOG fornece ainda gráficos contendo algumas informações de interesse, tais como as curvas características e as curvas de informação de cada item e do teste. No BILOG-MG esses gráficos também podem ser obtidos, mas com uma resolução bastante baixa. Isto se deve ao fato de que o programa BILOG já está disponível para o sistema Windows, enquanto que o BILOG-MG ainda só tem versões para o sistema operacional DOS.

A **fase 3** é a fase da **estimação das habilidades** dos respondentes. Aqui são estimadas as habilidades de cada um dos indivíduos, a partir dos resultados obtidos na fase anterior. Essas habilidades inicialmente são estimadas na escala dos parâmetros dos itens. No entanto, pode-se especificar alguns tipos de mudanças na escala, que serão feitas tanto nas habilidades como nos parâmetros estimados na fase anterior. Maiores

detalhes quanto aos métodos de estimação realizados nesta fase que estão disponíveis nesses programas serão fornecidos na próxima seção.

5.2 O processo de estimação nos programas BILOG e BILOG-MG

5.2.1 Métodos implementados para a calibração dos itens (fase 2)

Como foi dito na seção anterior, esses dois programas realizam inicialmente a calibração dos itens e depois a estimação das habilidades dos respondentes. Dois métodos de estimação para os parâmetros dos itens estão implementados, tanto no BILOG, como no BILOG-MG: máxima verossimilhança marginal e um método bayesiano de estimação por maximização da distribuição marginal a posteriori.

Assim como foi descrito no tópico 3, para que os parâmetros dos itens possam ser estimados através de qualquer um desses dois métodos, é necessária a utilização de distribuições de probabilidade para as habilidades dos respondentes. Esses programas assumem que os respondentes são uma amostra aleatória de uma população de habilidades que pode ser assumida como tendo ou uma distribuição normal padrão, ou uma distribuição discreta arbitrariamente especificada pelo usuário, ou ainda uma distribuição empírica, a ser estimada conjuntamente com os parâmetros dos itens. Esta distribuição empírica é representada na forma de uma distribuição discreta, através de pontos de quadratura.

No caso de mais de um grupo de respondentes, quando usamos o BILOG-MG, ao final do processo de calibração dos itens são fornecidas estimativas da média e desvio padrão da distribuição de habilidades a posteriori para cada grupo.

Também como já foi dito no tópico 3, na estimação por maximização da distribuição marginal a posteriori, distribuições a priori são definidas para os parâmetros dos itens. No caso desses dois programas, o usuário pode especificar prioris normais para o parâmetro de dificuldade, prioris log-normais para os parâmetros de discriminação e prioris beta para o parâmetro de acerto casual.

O BILOG e o BILOG-MG utilizam duas formas de resolver as equações de verossimilhança marginal: o método EM e o Scoring de Fisher.

5.2.2 Métodos implementados para a estimação das habilidades (fase 3)

Uma vez terminada a calibração dos parâmetros, será feita a estimação das habilidades dos respondentes. O BILOG e o BILOG-MG têm implementados os 3 métodos para estimação de habilidades quando os parâmetros dos itens são conhecidos, que foram descritos na seção 3.1, ou seja, os métodos de estimação por máxima verossimilhança, por esperança a posteriori (EAP) e por máximo a posteriori (MAP).

No método da máxima verossimilhança, as estimativas das habilidades dos respondentes são calculadas pelo método de Newton-Raphson, utilizando-se uma transformação linear do logito do percentual de acertos dos indivíduos como valores iniciais. Os problemas já descritos com as estimativas dos respondentes que tiveram erro total ou acerto total são contornados através de um artifício: os alunos que erraram todos os itens "ganham" um meio certo no item mais fácil. Alunos que acertaram todos os itens, "ganham" um meio certo no item mais difícil. Apesar dessas alternativas implementadas pelos dois programas, este método nem sempre fornece boas estimativas nestes casos.

No método EAP, as estimativas para as habilidades são calculadas utilizando-se pontos de quadratura para aproximar a distribuição a priori das habilidades de cada respondente. O número de pontos de quadratura é definido pelo usuário, que pode também escolher entre uma priori que seja normal (e cujos parâmetros podem ser especificados pelo usuário), ou uma distribuição discreta arbitrária (fornecida pelo usuário), ou ainda uma distribuição discreta empírica, através do uso dos pontos de quadratura e de seus respectivos pesos gerados na fase 2.

As estimativas EAP para as habilidades dos respondentes estão sempre definidas, qualquer que seja o padrão de respostas. Além disso, quando utilizamos a estimação por EAP, é fornecida uma estimativa da distribuição de habilidades da população de respondentes, na forma de uma distribuição discreta, dada pelos pontos de quadratura. Esta distribuição é obtida acumulando-se as densidades a posteriori de todos os sujeitos em cada ponto de quadratura. As somas são então normalizadas para obter-se as probabilidades estimadas em cada ponto. Também são fornecidos a média e o desvio padrão para essa distribuição estimada.

No método MAP, as estimativas das habilidades são calculadas pelo método de Newton-Gauss. Este procedimento sempre converge e fornece estimativas para todos os padrões de resposta possíveis. É assumida uma distribuição a priori normal, cujos parâmetros podem ser especificados pelo usuário, sendo que o padrão definido nesses programas é a normal padrão.

5.3 A equalização nos programas BILOG e BILOG-MG

Quando desejamos que a equalização seja feita durante o processo de calibração dos itens, o uso de programas computacionais especificamente desenvolvidos para esse fim são uma ferramenta bastante importante. O BILOG-MG é um bom exemplo de um programa que pode ser utilizado na maioria dos casos descritos no tópico anterior. Nesta seção, vamos então descrever quando é possível sua utilização em cada um daqueles casos e em quais deles o BILOG também pode ser usado.

Os casos 1 a 6 tratam, respectivamente, das situações descritas nas seções 4.1.1 a 4.1.6 do tópico 4. Já os casos (a) a (c) tratam, respectivamente, das situações descritas nas seções 4.2.1 a 4.2.3.

5.3.1 O BILOG e o BILOG-MG frente a grupos e/ou provas distintas

Caso 1: Aqui temos um único grupo fazendo um único tipo de prova.

Por se tratar do caso mais básico, em que não se faz necessário nenhum tipo de equalização, podemos utilizar qualquer um dos programas computacionais disponíveis que tratam da TRI, inclusive o BILOG e o BILOG-MG.

Caso 2: Aqui temos um único grupo fazendo 2 tipos de prova totalmente diferentes.

Por se tratar de um caso de equalização via população, basta que todos os itens de ambas as provas sejam calibrados simultaneamente. Para tanto, devemos fazer apenas uma ligeira alteração nos modelos já propostos, incorporando a informação do tipo de prova a que cada aluno foi submetido, uma vez que a cada tipo de prova está associado um conjunto de itens distintos.

Este também é um caso bastante comum que a maioria dos programas computacionais para análise via TRI é capaz de resolver. No BILOG-MG está situação é tratada sem maiores problemas. Já no BILOG, há uma limitação técnica: as respostas dos alunos devem estar já corrigidas (codificadas com 0 ou 1), para que não haja necessidade de utilizar os gabaritos das provas, uma vez que o programa não consegue ler 2 tipos de gabaritos distintos.

Caso 3: Aqui temos um único grupo fazendo 2 tipos de prova parcialmente diferentes, isto é, com alguns itens comuns.

Este caso é bastante semelhante ao caso anterior, ou seja, a equalização também pode ser feita via população. A única observação que podemos acrescentar é que devemos ter bastante cuidado no tratamento dos itens comuns. É que embora esses itens apareçam nos dois tipos de provas, não podem ser "contados" duas vezes, ou seja, o número total de itens a ser calibrado é o total de itens da prova A, mais o total de itens da prova B, menos o número de itens comuns entre A e B.

Caso 4: Aqui temos 2 grupos fazendo uma mesma prova.

Por se tratar de uma situação onde se faz necessária uma equalização via itens comuns, este caso necessita de programas computacionais para análise via TRI que tenham implementados modelos para mais de um grupo. O BILOG, por exemplo, não comporta esse tipo de problema, enquanto que o BILOG-MG foi especialmente desenvolvido para modelar esse tipo de situação.

Se só dispuséssemos do BILOG, uma alternativa seria calibrar as provas dos dois grupos separadamente, e depois realizar uma equalização a posteriori, como foi descrito no tópico anterior. Nesse caso, como todos os itens são comuns, métodos de equalização a posteriori como o média-desvio produzem resultados bastante satisfatórios, quando comparados à equalização feita durante o processo de calibração dos itens (Andrade, 1999).

Caso 5: Aqui temos 2 grupos fazendo 2 tipos de prova totalmente diferentes.

Como já foi explicado no tópico anterior, não há nenhuma maneira de tornar comparáveis os resultados desses 2 grupos.

Caso 6: Aqui 2 grupos são submetidos a 2 tipos de provas diferentes, mas que têm alguns itens comuns.

Assim como no caso 4, esta é uma situação típica para ser abordada no BILOG-MG, utilizando-se um modelo para mais de uma população e, portanto, não é possível o uso do BILOG. Como já foi citado no caso 3, devemos apenas ter o cuidado de não considerar duas vezes os itens repetidos.

Assim como foi comentado no caso 4, aqui também se pode resolver o problema através de uma equalização a posteriori. No entanto, o desempenho desse tipo de equalização torna-se bastante inferior à equalização feita durante o processo de calibração se o número de itens comuns for pequeno.

5.3.2 O BILOG e o BILOG-MG frente ao conjunto de itens a ser calibrado

Caso (a): todos os itens são "novos".

Quando desejamos calibrar o conjunto completo de itens, temos o problema de estimação mais comum e, portanto, ele pode ser resolvido utilizando-se qualquer um dos programas computacionais disponíveis que tratam da TRI, inclusive o BILOG e o BILOG-MG.

Caso (b): todos os itens já são calibrados.

Se não desejamos calibrar nenhum dos itens, estamos interessados apenas em estimar as habilidades dos respondentes. Este problema pode ser resolvido de maneira relativamente simples através dos programas BILOG e BILOG-MG, sendo necessário apenas fornecermos um arquivo contendo as estimativas dos parâmetros de interesse. No entanto, cabe aqui uma observação: quando se utilizar o BILOG ou o BILOG-MG é sempre recomendável utilizar as estimativas dos parâmetros na escala "original", isto é, como foram fornecidas pelo programa, sem que tenham sofrido nenhum tipo de transformação linear. Isto porque, quando se utilizam os métodos EAP ou MAP para estimar as habilidades dos respondentes, faz-se necessário o uso de uma distribuição a priori para a habilidade de cada um desses respondentes, e o padrão desses programas é utilizar a distribuição normal padrão ou outras distribuições discretas, mas sempre com média e desvio padrão nas vizinhanças dos valores 0 e 1, respectivamente. Assim, se por exemplo, a métrica da população em que

os parâmetros foram estimados tiver sido transformada para (200; 40), haverá problemas na estimação das habilidades dos novos respondentes.

Caso (c): alguns itens são "novos" e outros já estão calibrados.

Neste caso, desejamos calibrar alguns itens e manter os parâmetros de outros, que já foram calibrados anteriormente. Para que possamos fixar parâmetros de alguns itens e calibrar o restante utilizando o BILOG e o BILOG-MG, deveremos necessariamente utilizar um método de estimação bayesiano, uma vez que o único procedimento disponível nesses programas para fixar apenas parte dos itens, é o uso de distribuições a priori convenientes para os parâmetros desses itens. Para os itens novos, que desejamos calibrar, utilizamos as prioris padrão sugeridas pelo programa. Já para os outros itens, definimos prioris cujas médias são os próprios valores dos parâmetros que desejamos fixar e cujos desvios padrão são tão pequenos que a distribuição se torna praticamente degenerada naquele ponto. O que ocorre na prática é que todos os parâmetros são estimados novamente, mas a convergência daqueles itens conhecidos é artificialmente induzida para os valores que desejamos. Pode-se também "reforçar" ainda mais a convergência utilizando-se outro recurso do programa, que é a definição, por parte do usuário, de valores iniciais convenientes.

Mas, o uso deste tipo de procedimento pode acarretar alguns problemas. Por exemplo, se não utilizarmos novamente o mesmo grupo de respondentes da calibração inicial, poderemos ter problemas para obter a convergência nessa segunda calibração. E, na prática, muitas vezes não dispomos do conjunto original de respondentes para juntarmos aos respondentes da nova aplicação. E devemos ressaltar que estamos nos referindo ao caso em que há uma única população sendo submetida a uma única prova. O problema se torna ainda mais complexo, no caso de termos mais de uma população envolvida (comentaremos essa situação a seguir).

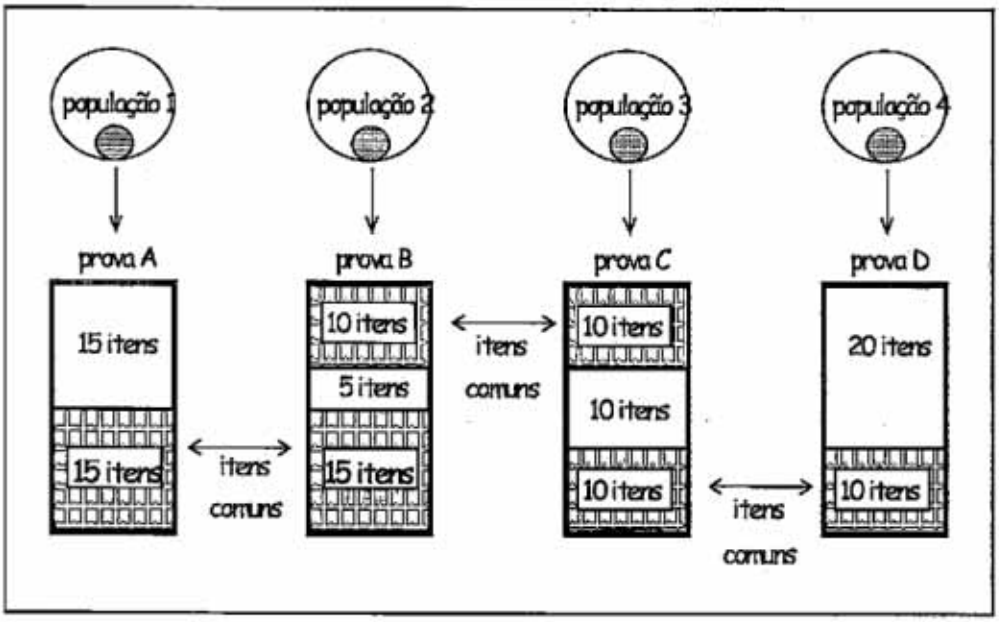
5.3.3 O uso do BILOG-MG quando desejamos fixar parte dos itens e calibrar o restante, e há mais de uma população envolvida.

Quando há duas (ou mais) populações envolvidas (casos 4 e 6), e utilizamos o BILOG-MG para estimar parte do conjunto de itens, fixando os demais (caso (c)), poderemos ter problemas com a métrica. É que como há mais de uma população envolvida nos processos de estimação, para resolver os problemas de indeterminação de escala, o programa pede ao usuário que defina uma das populações como sendo a referência, que será definida como tendo média 0 e desvio padrão 1 e, então, as demais populações serão posicionadas com relação a ela.

Vamos então imaginar a seguinte situação, ilustrada na Figura 5.1: utilizamos amostras das populações 1 e 2 para calibrar um conjunto de itens, provenientes de 2 tipos de provas (tipo A e B). Estas provas tinham 30 itens cada, sendo 15 itens comuns. A população 1 foi utilizada como referência. Ao final do processo, temos um conjunto de 45 itens (= 30 + 30 – 15) calibrados, além das habilidades dos respondentes das 2 populações. Digamos que as estimativas obtidas para os parâmetros populacionais dos 2 grupos foram respectivamente (0; 1) e (2; 2). Desse modo, um item i, cuja estimativa do parâmetro **b** foi 1 está, usando-se como unidade o desvio padrão da população 1, um desvio padrão acima da média da população 1 (e, portanto, é relativamente difícil para este grupo) e um desvio padrão abaixo da média da população 2 (e, portanto, é relativamente fácil para este grupo).

Suponha agora que temos outras 2 provas C e D, que serão submetidas, respectivamente, a amostras das populações 3 e 4. Ambas as provas são compostas de 30 itens, sendo que há 10 itens comuns entre elas. Suponha ainda que, além disso, há 10 itens na prova C que são comuns com a prova B e, portanto, que já foram calibrados anteriormente.

Figura 5.1. Esquematização dos itens comuns entre as provas



Desejamos então fixar os parâmetros desses 10 itens obtidos na calibração anterior e estimar todos os restantes. O motivo para isto seria que, procedendo desta maneira, faríamos uma equalização entre as populações 1, 2, 3 e 4, tornando possível qualquer comparação entre elas.

Mas, o que aconteceria se, para tanto, utilizássemos apenas as populações 3 e 4 ? Para começar, teríamos que definir uma população de referência, digamos a população 3. Logo, essa população será definida como tendo parâmetros (0; 1), para que a população 4 seja posicionada com relação a ela. Supondo que aquele item i, cujo valor de **b** é 1, foi um dos 10 itens que tiveram seus parâmetros fixados, que interpretação deveríamos ter sobre a relação desse item com a população 3 ? A mesma que já tivemos com relação à população 1: que ele está um desvio padrão acima da média da população 3 e, portanto, é relativamente difícil para este grupo.

O fato de termos as populações 1 e 3 necessariamente com a mesma distribuição de probabilidade é um problema, pois sabemos que se tratam de populações diferentes. Suponhamos que essas populações sejam, respectivamente, a 3ª, a 4ª, a 5ª e a 6ª séries do Ensino Fundamental. Seria perfeitamente razoável esperarmos que as médias das distribuições de habilidades destas populações mantivessem uma relação crescente de ordem. Assim, se a 3ª série fosse fixada como tendo

parâmetros (0; 1) e a 4ª série tivesse então seus parâmetros estimados em (2; 2), esperaríamos ter uma média maior do que 2 para a 5ª série, e não (0; 1)! Desta maneira, aquele item i, cujo parâmetro de dificuldade foi estimado em 1, deveria estar necessariamente abaixo da média da 5ª série.

Há pelo menos 2 maneiras de solucionarmos este problema. A primeira, que nem sempre é possível, é utilizarmos novamente os respondentes utilizados nas provas A e B no processo da calibração das provas C e D. Fixaríamos todos os itens das provas A e B enquanto calibraríamos os itens novos das provas C e D. Desta maneira, poderíamos definir novamente a população 1 como sendo a referência e, então, não haveria mais problemas no posicionamento das populações 3 e 4. Mas, como já foi dito, nem sempre é possível proceder desta maneira, pois poderíamos não dispor dos respondentes utilizados na primeira calibração. Uma outra maneira de solucionar o problema de maneira adequada seria fazer uma equalização a posteriori, que já foi comentada na seção 4.3.1 do tópico anterior.

No próximo tópico, descreveremos um exemplo completo de um estudo em avaliação educacional onde foi necessário fazer uma equalização e no qual foi utilizado o BILOG-MG.

6.0 UMA APLICAÇÃO PRÁTICA

Nas últimas décadas, a TRI vem tornando-se a técnica predominante no campo dos testes, em países de todo o mundo. Aqui no Brasil, nos últimos anos, vários estudos na área educacional já foram feitos, cuja técnica estatística empregada foi a TRI. Em Andrade e Klein (1999) podemos encontrar uma lista atualizada das principais aplicações brasileiras. Dentre elas, podemos destacar:

> Sistema Nacional de Avaliação da Educação Básica – SAEB. Aplicado de 2 em 2 anos, desde 1995, em todo o território nacional, em amostras de escolas públicas e privadas, em alunos da 4ª e 8ª séries do Ensino Fundamental e da 3ª série do Ensino Médio.
>
> Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo – SARESP. Aplicado em 1996 (3ª e 7ª séries do Ensino Fundamental), em 1997 (4ª e 8ª séries do Ensino Fundamental) e em 1998 (5ª série do Ensino Fundamental e 1ª série do Ensino

Médio) em todas as escolas públicas estaduais do Estado de São Paulo.

Avaliação das Escolas Estaduais do Estado do Rio Grande do Norte. Aplicada em 1996, seguindo os mesmos moldes do SAEB. A inclusão de itens do SAEB 95, nas provas, permitiu a comparação do rendimento das escolas do Rio Grande do Norte com o rendimento nacional, medido pelo SAEB 95.

Neste tópico vamos descrever uma aplicação prática da TRI na área de Avaliação Educacional, utilizando uma dessas avaliações.

6.1 O Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo – SARESP

A Secretaria de Estado da Educação de São Paulo – SEE/SP implantou, em 1996, o Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo – SARESP, visando alcançar dois objetivos. O primeiro seria ampliar o conhecimento do perfil de realização dos alunos, fornecendo aos professores informações sobre o desempenho dos alunos de modo a subsidiar o trabalho a ser desenvolvido em sala de aula. Assim, os docentes poderiam identificar, no começo do ano escolar, os pontos fortes e fracos do desempenho dos alunos e, a partir desse diagnóstico, adotar estratégias pedagógicas apropriadas.

O segundo seria fornecer informações essenciais para a melhoria da gestão do sistema educacional, na medida em que identifica os pontos críticos do ensino e possibilita à SEE, por meio de seus órgãos centrais e das Delegacias de Ensino, apoiar as escolas e os educadores com recursos, serviços e orientações.

6.1.1 As características da aplicação

As provas do SARESP são elaboradas a partir de matrizes curriculares, ou seja, tabelas de especificação de conteúdos e objetivos, que indicam os temas e metas do currículo a serem desenvolvidos em cada série e disciplina. Esses parâmetros fundamentam-se nas Propostas Curriculares elaboradas pela Coordenadoria de Estudos e Normas Pedagógicas – CENP e, desde 1997, os itens que compõem as provas vêm sendo construídos pelos professores da Rede Estadual de Ensino.

Até o momento o SARESP foi realizado em 3 anos consecutivos, e a aplicação das provas foi feita segundo o quadro a seguir.

Tabela 6.1 Esquema da aplicação das Provas do SARESP

Ano de Aplicação	Séries e períodos Avaliados	Provas compostas Pelas disciplinas
1996	3ª série diurna do Ensino Fundamental	1-Língua Portuguesa e 2-Matemática
	7.ª série diurna e noturna do Ensino Fundamental	1-Língua Portuguesa, 2-Matemática, 3-Ciências e 4-História e Geografia
1997	4ª série diurna do Ensino Fundamental	1-Língua Portuguesa e 2-Matemática
	8.ª série diurna e noturna do Ensino Fundamental	1-Língua Portuguesa, 2-Matemática, 3-Ciências e 4-História e Geografia
1998	5ª série diurna e noturna do Ensino Fundamental	1-Língua Portuguesa e 2-Matemática
	1ª série diurna e noturna do Ensino Médio	1-Língua Portuguesa, 2-Matemática, 3-Ciências e 4-História e Geografia

Como as avaliações são sempre realizadas no início do ano letivo, as provas de cada uma das séries-alvo são baseadas em conteúdos abordados no ano anterior. Exemplificando, em 1996, as provas dos alunos da 3ª e 7ª séries foram elaboradas com base nos conteúdos relativos ao Ciclo Básico e à 6ª série, respectivamente.

Em todos os anos foram avaliados todos os alunos que frequentavam as séries envolvidas: trata-se, portanto, de uma avaliação de caráter censitário. Cada aluno, entretanto, é avaliado em apenas uma disciplina, ou seja, na 3ª e 4ª séries metade dos alunos responde à prova de Língua Portuguesa e a outra metade à de Matemática. Essa divisão é feita de maneira aleatória. Nas demais séries, os alunos são divididos, também aleatoriamente, em 4 partes e então cada uma delas é submetida a um tipo de prova: Língua Portuguesa, Matemática, Ciências ou História e Geografia. Essa última prova é a única onde aparecem duas disciplinas. No entanto, em termos de análise, as duas disciplinas são obviamente consideradas separadamente.

6.1.2 O tipo de resultados alcançados

A cada ano de aplicação, todos os itens que compõem as provas de cada uma das disciplinas consideradas são cuidadosamente avaliados e interpretados, dentro de cada série envolvida. Para tanto, são considerados

tanto seus parâmetros obtidos através da TRI como também algumas estatísticas fornecidas pela Teoria Clássica. A partir dessas informações, um conjunto de especialistas em cada uma das disciplinas faz um diagnóstico completo de cada item (assunto abordado, grau de dificuldade, erros mais freqüentes, etc.), e também da prova como um todo. Com base nessas informações, pode-se ter uma visão geral do desempenho dos alunos e verificar quais as principais deficiências da série naquele ano.

No entanto, essa avaliação isolada feita ano a ano em cada série, não nos permite comparar o desempenho dos alunos de um ano para o outro, ou seja, verificar se houve realmente um ganho no conhecimento de uma série para a seguinte. Para responder a esta questão, seria necessário que os itens de duas séries consecutivas fossem comparáveis, ou seja, estivessem na mesma métrica. E isto poderia ser conseguido através de uma Equalização.

No entanto, as provas de um ano para outro não apresentavam itens comuns. Como fazer então uma equalização entre duas populações que foram submetidas a provas totalmente diferentes? A solução encontrada foi a criação de uma prova adicional, que serviria de "ligação", uma vez que seria composta de itens que haviam sido submetidos a essas duas populações.

Exemplificando, as provas aplicadas em 1997, na 4ª e 8ª séries, não tinham itens comuns com as provas aplicadas no ano anterior, na 3ª e 7ª séries. Assim, foram montadas duas provas de ligação: a primeira, composta de itens que haviam sido submetidos à 3ª série e à 4ª série e a segunda composta de itens que haviam sido submetidos à 7ª e à 8ª séries. Essas duas provas adicionais foram aplicadas no final do ano de 1997, a uma amostra de alunos da 3ª e da 7ª séries, respectivamente. Cabe ressaltar, que estes dois grupos adicionais foram introduzidos no estudo com o único objetivo de possibilitar a equalização, não havendo nenhum interesse em estudar o desempenho destas populações.

A partir destas provas de ligação foi possível a criação de uma escala única para as séries consecutivas, permitindo assim a comparação dos resultados e a criação de escalas de conhecimento interpretáveis. No SARESP essas escalas foram construídas para as disciplinas Língua Portuguesa e Matemática, por serem as únicas disciplinas avaliadas em todas as séries, todos os anos.

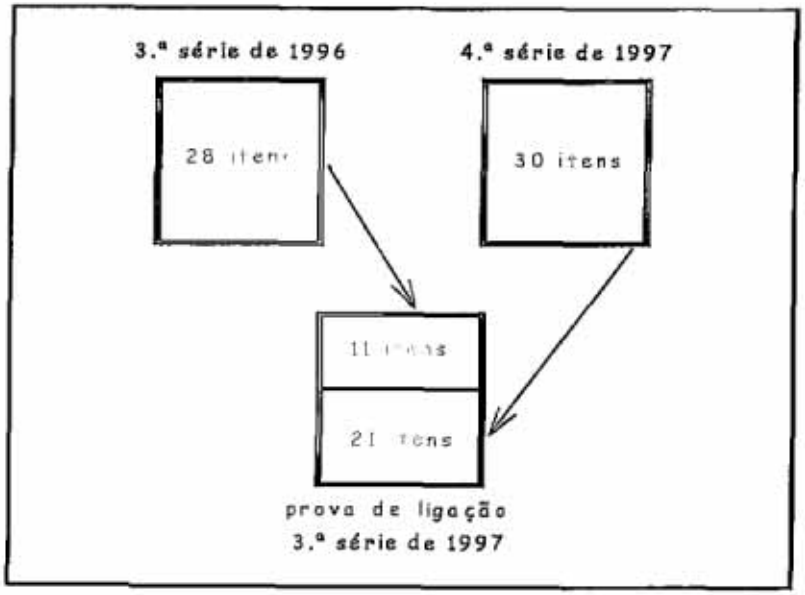
Vamos descrever mais detalhadamente esse processo, usando como exemplo as provas de Língua Portuguesa da 3ª e 4ª séries.

6.1.3 Um exemplo: a Língua Portuguesa na 3ª e 4ª séries

Em 1996 foi aplicada uma prova de 28 itens de Língua Portuguesa aos alunos da 3ª série. Em 1997, os alunos da 4ª série foram avaliados nessa disciplina através de uma prova composta de 30 itens, totalmente distinta da prova aplicada no ano anterior.

Num primeiro momento, cada uma destas provas teve seus itens calibrados e interpretados dentro de suas respectivas séries. Mas, para que a equalização entre as duas séries pudesse ser possível, foi criada uma prova de ligação, composta de 32 itens, sendo 11 provenientes da prova da 3ª série e 21 da prova da 4ª série, como mostra a Figura 6.1. Esta prova foi então submetida a uma amostra de alunos que cursavam a 3ª série, no final de 1997. Esta nova população foi introduzida no estudo apenas para possibilitar a equalização.

Figura 6.1 Esquema da composição da prova de ligação



Cabe ressaltar que a prova de ligação foi composta de mais itens da prova da 4ª série do que da prova da 3ª, pois houve a preocupação de montar-se uma prova com diferentes graus de dificuldade e com um bom nível de discriminação. Uma vez que as provas de 96 e 97 já haviam sido analisadas separadamente através da TRI, foram selecionados os itens com tais características e a prova da 4ª série de 97 apresentou um número maior deles. Também é importante notar que a população escolhida para fazer a prova de ligação foi a 3ª série de 1997, pois como já foi dito, os itens das provas da 3ª série de 96 e da 4ª série de 97 foram elaboradas com base nos conteúdos dos anos anteriores, ou seja, eram referentes aos conteúdos do Ciclo Básico e da 3ª série, respectivamente. Como a prova de ligação foi aplicada no final do ano letivo de 1997, a série mais indicada para ser submetida a tal prova era, portanto, a 3ª série.

Todos os 58 itens, respondidos pelos alunos das 3 populações envolvidas foram então calibrados simultaneamente. O programa computacional utilizado foi o BILOG-MG. Foram utilizados procedimentos bayesianos para a estimação dos parâmetros dos itens e das habilidades. Assim, foram consideradas distribuições a priori para cada um dos parâmetros dos itens e também distribuições normais padrão a priori, para cada uma das populações envolvidas. O grupo submetido à prova de ligação (3ª série de 97) foi considerado a população de referência. Portanto, as outras séries foram posicionadas em relação à ela. No final do processo de estimação, foram fornecidas as estimativas das distribuições a posteriori, para cada uma das populações.

Cabe ressaltar novamente que não havia interesse em estudar o desempenho dos alunos submetidos à prova de ligação, ou seja, ao grupo da 3ª série de 97. O número de alunos que fizeram essa prova foi apenas o suficiente para atender às exigências da TRI, no que se refere ao número mínimo de sujeitos necessários para obter-se boas estimativas dos parâmetros dos itens. Os gráficos a seguir ilustram a forma dessas distribuições, obtidas para as duas populações de interesse.

Figura 6.2 Representação gráfica da distribuição a posteriori das habilidades em Língua Portuguesa dos alunos da 3ª série

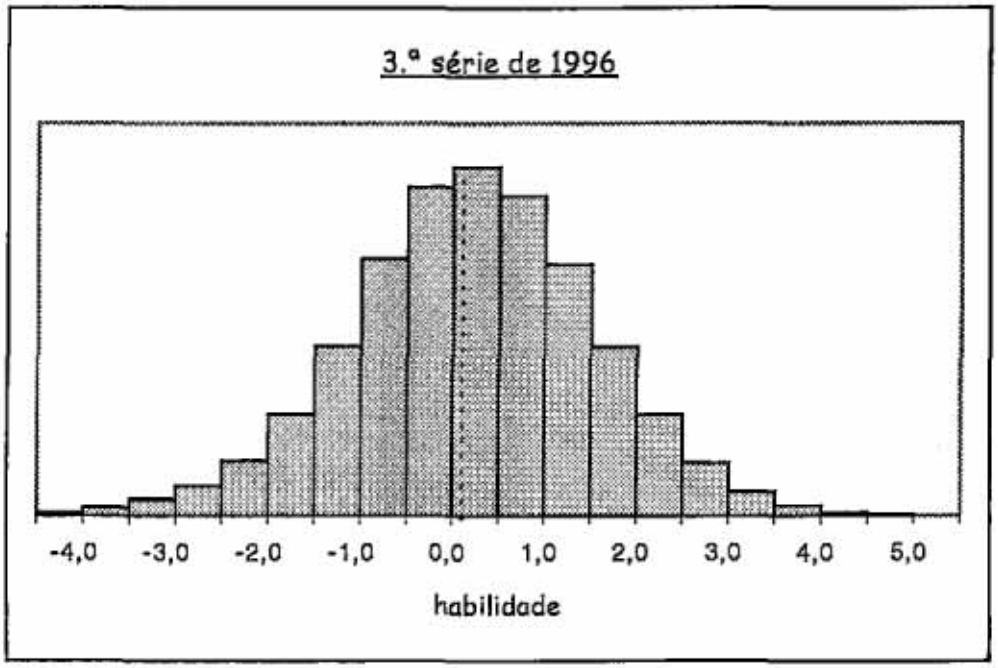
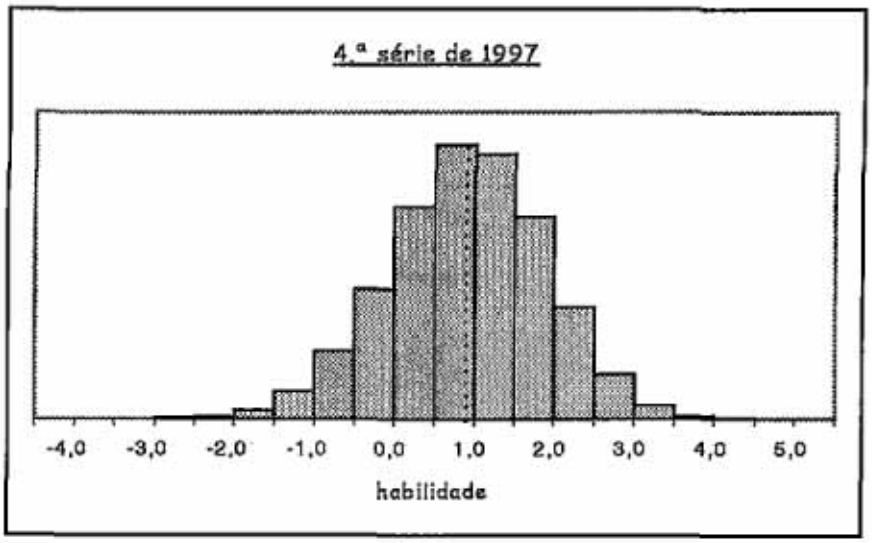


Figura 6.3 Representação gráfica da distribuição a posteriori das habilidades em Língua Portuguesa dos alunos da 4ª série



6.1.3.1 Interpretação dos resultados

Podemos observar que o gráfico da 4ª série encontra-se deslocado para a direita, com relação ao gráfico da 3ª série. Houve um aumento na média da 4ª série em relação à 3ª (representada pela linha tracejada). Também podemos observar que os alunos da 4ª série parecem ser mais homogêneos do que os alunos da série anterior, com relação à habilidade em Língua Portuguesa.

Foi feita uma transformação linear nas estimativas dos parâmetros dos itens e das habilidades dos alunos, visando um melhor entendimento dos resultados. Após essa transformação, a média e o desvio padrão das habilidades dos alunos da 3ª série de 1996 em Língua Portuguesa foram fixados em 50 e 16, respectivamente. Para a 4ª série os valores obtidos foram 62 e 13.

Com todos os 58 itens na mesma métrica, o próximo passo foi a identificação de níveis âncora – conforme descrito na seção 4.4 – , que pudessem caracterizar a escala de conhecimento em Língua Portuguesa da 3ª e 4ª séries.

Assim, foi possível a caracterização de 5 níveis âncora (nos pontos 5, 30, 45, 60 e 75) na escala de habilidades de Língua Portuguesa da 3ª e 4ª séries. Cada um desses níveis âncora é formado por um conjunto de itens, que caracterizam esse ponto na escala de habilidades, de acordo com a natureza e o grau de conhecimentos que eles exigem.

Após a identificação dos níveis âncora, um grupo de especialistas analisa e interpreta o conjunto de itens que o compõem, a fim de caracterizar cada ponto da escala. A seguir, exemplificamos como ficou a caracterização de um determinado nível âncora da escala de habilidades em Língua Portuguesa da 3ª e 4ª séries do SARESP:

Nível 60 – Língua Portuguesa

Neste nível, os alunos são capazes de identificar o narrador e revelam ter noções relativas ao papel geral que este assume na história. Com relação ao uso e interpretação da Língua Portuguesa, reconhecem a função do sinal de interrogação no texto. Nos textos narrativos-descritivos, identificam os diferentes elementos que estruturam o texto, discernindo ou reconstituindo a seqüência lógica dos fatos narrados. Em texto de correspondência (bilhete), conseguem interpretar o sentido da

mensagem, percebendo implicações lógicas entre as informações contidas no texto.

Demonstram, ainda, certa familiaridade com a leitura de histórias em quadrinhos, fazendo a leitura de imagens e inferindo o significado atribuído a uma expressão onomatopaica como, por exemplo, "PLOFT", identificado como o barulho de um livro ao ser fechado.

Além da interpretação de cada ponto que caracteriza a escala de habilidades, também foi calculada a porcentagem de alunos em cada série que dominavam os assuntos descritos em cada nível, visando avaliar os ganhos, em termos de conhecimentos, de um ano para outro. Por exemplo, para o nível 60, descrito anteriormente, chegamos aos seguintes resultados:

Em 1996, a porcentagem de estudantes que respondiam questões desse nível era de 26,6%. Em 1997, essa porcentagem passa a ser de 55,8%. Ou seja, houve um ganho de 29,2% (pontos percentuais) da 3ª para a 4ª série.

Por fim, foi estimada a habilidade média (e respectivo erro padrão) em Língua Portuguesa, para cada escola. Assim, cada uma delas recebeu um boletim, indicando o desempenho médio da escola, da delegacia da qual ela faz parte e também o resultado médio geral (ou seja, da população toda, que no caso, são todas as escolas públicas estaduais de São Paulo). Com base nessas informações, cada instituição de ensino pode verificar qual sua situação em relação às demais, além de avaliar os ganhos de seus alunos de um ano para outro, e de ter indicações sobre quais os assuntos em que seus alunos ainda estão deficientes.

Obviamente, todos os resultados obtidos são também enviados para as Delegacias de Ensino e para a Secretaria de Estado da Educação de São Paulo. Assim, a partir das informações fornecidas pelo SARESP, as ações podem ser tomadas tanto a nível de cada instituição de ensino, quanto em proporções estaduais.

Dando prosseguimento ao estudo, em 1998 uma das séries avaliadas pelo SARESP foi a 5ª série do Ensino Fundamental, nos períodos diurno e noturno. Para cada uma das disciplinas avaliadas dois tipos de provas, com alguns itens comuns, foram aplicados em cada uma das populações – diurna e noturna. Novamente, as provas aplicadas não tinham itens comuns com as provas dos anos anteriores.

Mais uma vez, foi montada uma prova de ligação, composta de itens utilizados nas provas de 3 das 4 populações de interesse: 4ª série de

1997, 5ª série diurna de 1998 e 5ª série noturna de 1998. Essa prova foi aplicada então a uma amostra de alunos que cursavam a 4ª série em 1998. Essa população adicional também foi introduzida no estudo apenas com o objetivo de possibilitar a equalização.

Cabe ressaltar que a meta agora é colocar os alunos da 3ª série de 96, 4ª série de 1997 e 5ª séries diurna e noturna de 98, todos na mesma escala. Nessa nova equalização, os itens da 3ª série não precisaram mais entrar na prova de ligação, pois a 3ª e a 4ª séries já haviam sido colocadas na mesma métrica. Na verdade, agora é como se fossemos apenas "colar" a 5ª série nas séries anteriores. Assim, essa segunda equalização foi realizada de uma maneira bastante distinta da primeira. Os itens calibrados no ano anterior foram mantidos fixos durante o processo de estimação e apenas os itens aplicados à 5ª série foram calibrados, resultando ao final do processo, num conjunto de itens de 3ª à 5ª séries, todos na mesma escala. Dessa maneira, a escala de habilidades da 3ª e da 4ª séries pode ser ampliada com a entrada da 5ª série e interpretada para todo esse conjunto de alunos.

Concluindo, esse estudo, além de avaliar o desempenho da rede estadual de São Paulo ano a ano, também vem fornecendo indicadores quantitativos de como as intervenções no ensino público têm afetado o conhecimento dos alunos de uma série para outra, e esse tipo de questão só pode ser respondida através das ferramentas fornecidas pela TRI.

No próximo tópico serão feitas as considerações finais sobre este trabalho e algumas sugestões para futuras pesquisas e aplicações.

7.0 CONCLUSÕES E SUGESTÕES

Neste trabalho procuramos apresentar os principais conceitos, modelos, métodos de estimação e aplicações da Teoria da Resposta ao Item, com o objetivo de mostrar o grande potencial da sua aplicação na área de avaliação educacional, em especial quando há a necessidade da comparação do desempenho de duas ou mais populações de indivíduos.

Apesar desta teoria ter quase 50 anos, somente nos últimos 15 é que ela vem sendo aplicada em larga escala nas principais avaliações educacionais de diferentes países. Atribui-se este fato à complexidade matemática dos métodos envolvidos, praticamente inviáveis sem o auxílio do computador. O que temos observado é que a teoria vem sendo desenvolvida num ritmo que ainda não vem sendo acompanhado pelo

desenvolvimento de programas computacionais eficientes, que viabilizem sua utilização em maior escala.

Além disso, a aplicação apropriada desta teoria exige necessariamente o envolvimento de especialistas em educação e em estatística. Nesse sentido, faz-se imprescindível a elaboração de grupos de avaliação que possibilitem a integração de profissionais de ambas as áreas.

Justamente pelo fato da TRI ter sido ainda tão pouco explorada, vários pontos têm sido levantados na literatura sobre sua adequação. Alguns deles ainda permanecem em aberto. Podemos citar, por exemplo, a questão da dimensionalidade do espaço de traços latentes envolvidos na avaliação. Todos os modelos que vêm sendo efetivamente utilizados pressupõem que o conhecimento que se deseja medir pode ser representado por uma única habilidade. Alguns autores têm defendido a tese de que os modelos unidimensionais têm fornecido bons resultados, mesmo em situações multidimensionais, desde que uma das dimensões possa ser considerada predominante. Mais recentemente, modelos para mais de uma dimensão têm sido propostos, mas ainda não têm sido aplicados devido a não disponibilidade de recursos computacionais e também à sua maior dificuldade de interpretação.

A questão da equalização entre diferentes populações também sempre foi um ponto bastante discutido na literatura. Conforme comentamos neste trabalho, a proposta recente de modelos para múltiplos grupos de Bock e Zimowski (1997), que viabilizam a equalização durante o processo de calibração, deu um novo rumo à solução desta questão, 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.

Outro ponto que poderíamos citar, foi levantado por Mislevy (1991) e diz respeito à qualidade da estimação da distribuição das habilidades dos elementos de uma população. O autor discute a possibilidade de se obterem melhores estimativas da variabilidade das habilidades, utilizando-se também outras informações dos respondentes que possam estar associadas com suas habilidades. Exemplos dessas informações seriam o grau de escolaridade dos pais, o hábito de leitura do respondente, a condição sócio-econômica da família, etc. Esta

metodologia é baseada no conceito de imputação múltipla de dados faltantes e os valores obtidos para as habilidades são denominados de "valores plausíveis". Mas ainda existem alguns fatores que dificultam a aplicação desta metodologia e o principal deles, como sempre, é a não existência comercial, até o presente momento, de programas computacionais apropriados. Além disso, há também a dificuldade da obtenção de informações adicionais relevantes ao problema que sejam fidedignas e a inclusão dessas mesmas informações no modelo.

Há também muitos pontos que ainda não foram explorados, como por exemplo, a implementação de modelos longitudinais, ou seja, de modelos que permitam o estudo de indivíduos avaliados mais de uma vez ao longo do tempo.

Apesar de termos abordado nesse trabalho as aplicações da TRI na área educacional – que atualmente é onde verificamos suas maiores contribuições –, é importante ressaltar que vem crescendo o interesse na utilização dos modelos propostos por esta teoria em outras áreas, como por exemplo, em medicina, psicologia e marketing.

Para finalizar, gostaríamos de ressaltar que uma maior disseminação do uso da TRI 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. A primeira aplicação da TRI no Brasil foi na análise do SAEB 95. Desde então, os órgãos governamentais, através do MEC, vêm valorizando e incentivando o uso dessa teoria nas avaliações educacionais brasileiras. No entanto, o mercado de trabalho ainda está bastante deficiente de profissionais com tais qualificações.

Referências Bibliográficas

ANDRADE, D. F. (1999). **Comparando o Desempenho de Grupos (Populações) de Respondentes Através da Teoria da Resposta ao Item.** Tese apresentada ao Departamento de Estatística e Matemática Aplicada da UFC para o Concurso de Professor Titular.

ANDRADE, D. F. e KLEIN, R. (1999). Métodos estatísticos para avaliação educacional : teoria da resposta ao item. **Boletim da ABE, 43,** 21-28.

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

ANDRICH, D. (1978). A rating formulation for ordered response categories. **Psychometrika, 43**, 561-573.

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

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

BIRNBAUM, A. (1968). **Some latent trait models and their use in inferring an examinee's ability.** In Statistical Theories of Mental Test Scores, F. M. Lord and M. R. Novick (Eds.). Reading, M. A. : Addison-Wesley.

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

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

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

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

FUNDAÇÃO CARLOS CHAGAS (1997). **Avaliação das Escolas Estaduais de Ensino Fundamental e Ensino Médio do Rio Grande do Norte, 4v.** São Paulo : Fundação Carlos Chagas.

FUNDAÇÃO CARLOS CHAGAS (1998). **Programa de Aceleração da Aprendizagem: avaliação final, avaliação do material didático e apêndice, 3v.** São Paulo : Fundação Carlos Chagas / Instituto Ayrton Senna.

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

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

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

HAMBLETON, R. K.; COOK, L. L. (1997). Latent trait models and their use in the analysis of educational test data. **Journal of Educational Measurement, 14**, 75-96.

HAMBLETON, R. K.; SWAMINATHAN, H. (1985). **Item Response Theory : Principles and Applications.** Boston : Kluwer.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. (1991). **Fundamentals of Item Response Theory.** Newbury Park : Sage Publications.

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

KOLEN, M. J.; BRENNAN, R. L. (1995). **Test Equating – Methods and Practices.** New York : Springer.

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

LORD, F. M. (1952). A theory of test scores. **Psychometric Monograph, 7.**

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

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

MASTERS, G. N. (1982). A Rasch model for partial credit scoring. **Psychometrika, 47**, 149-174.

MINISTÉRIO da EDUCAÇÃO e do DESPORTO (1996). **Sistema Nacional de Avaliação da Educação Básica : SAEB 95 – relatório técnico.** São Paulo/Rio de Janeiro : Fundação Carlos Chagas / Fundação Cesgranrio.

MINISTÉRIO da EDUCAÇÃO e do DESPORTO (1998). **Sistema Nacional de Avaliação da Educação Básica : SAEB 97 – relatório técnico.** Rio de Janeiro : Fundação Cesgranrio.

MISLEVY, R. J. (1986). Recent developments in the factor analysis of categorical variables. **Journal of Educational Statistics, 11**, 3-31.

MISLEVY, R. J. (1991). Randomization-based inference about latent variables from complex samples. **Psychometrika, 56**, 177-196.

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

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

MURAKI, E. (1992). A generalized partial credit model : Application of an EM algorithm. **Applied Psychological Measurement, 16,** 159-176.

MURAKI, E.; BOCK, R. D. (1997). **PARSCALE : IRT Based Test Scoring and Item Analysis for Graded Open-Ended Exercices and Performance Tasks.** Chicago : Scientific Software, Inc.

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

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

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

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

SOARES, J. F.; MARTINS, M. I.; ASSUNÇÃO, C. N. B. (1998). Heterogeneidade acadêmica dos alunos admitidos na UFMG e PUC-MG. **Estudos em Avaliação Educacional, 17,** 61-72. São Paulo : Fundação Carlos Chagas.

STROUD, A. H.; SECREST, D. (1966). **Gaussian Quadrature Formulas.** Englewood Cliffs, New Jersey : Prentice-Hall.

VIANNA, H. M. (1987). **Testes em Educação.** São Paulo : IBRASA

WILSON, D. T.; WOOD, R.; DOWNS, P. K.; GIBBONS, R. (1991). **TESTFACT : Test Scoring, Item Statistics and Item Factor Analysis.** Chicago : Scientific Software, Inc.

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

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

ZWICK, R. (1987). Assessing the dimensionality of NAEP reading data. **Journal of Educational Measurement, 24,** 293-308.