INTRODUÇÃO À TEORIA DA RESPOSTA AO ITEM: conceitos e aplicações

DALTON FRANCISCO DE ANDRADE[1]
RAQUEL DA CUNHA VALLE[2]

1. Introdução

Resultados obtidos em provas, expressos apenas por seus escores brutos ou padronizados, têm sido freqüentemente 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, a comparação entre indivíduos somente é possível quando eles são submetidos às mesmas provas ou, pelo menos, ao que se denomina de provas paralelas. O leitor encontrará maiores detalhes sobre esta metodologia, incluindo a sua fundamentação matemática, em Gulliksen (1950), Lord e Novick (1968) e Vianna (1987), entre outros.

Atualmente, na área educacional, vem crescendo o interesse pela aplicação de técnicas derivadas da Teoria de Resposta ao Item - TRI que propõem modelos de variáveis latentes para representar a relação entre a probabilidade de um aluno responder corretamente a um item e seus traços latentes ou habilidades na área do conhecimento avaliada, os quais não são observados diretamente. Tendo como elemento central os itens e não a prova como um todo, a TRI permite, por exemplo, a comparação entre populações distintas submetidas a provas diferentes mas com alguns itens comuns ou, ainda, a comparação entre indivíduos da mesma população que tenham sido submetidos a diferentes provas, com ou sem itens comuns.

---

[1] Ph.D. em Bioestatística pela University of North Carolina at Chapel Hill, Professor Assistente do Departamento de Estatística do Instituto de Matemática e Estatística da USP, Pesquisador Senior do Departamento de Pesquisas Educacionais da Fundação Carlos Chagas.

[2] Bacharel em Estatística, Mestranda em Estatística do Departamento de Estatística do Instituto de Matemática e Estatística da USP, Estatístico do Departamento de Pesquisas Educacionais da Fundação Carlos Chagas.

O objetivo principal deste trabalho é apresentar os conceitos básicos envolvidos na TRI e suas principais aplicações em avaliações educacionais brasileiras. Em Lord (1980) e Hambleton, Swaminathan e Rogers (1991) por exemplo, o leitor encontrará maiores detalhes sobre os fundamentos e aplicações desta teoria. Na seção 2 discutem-se os modelos matemáticos com suas suposições básicas; na seção 3, o processo de estimação dos parâmetros dos itens e das habilidades dos respondentes; na seção 4, é introduzido o conceito de equalização ("equating"), a partir do qual pode-se efetuar as comparações entre populações e indivíduos descritas acima e na seção 5 discutem-se a construção e interpretação de escalas de habilidades por meio desta teoria. Na seção 6, apresenta-se uma aplicação da TRI na análise de parte dos dados obtidos pelo Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo - SARESP nos anos de 1996 e 1997 (ver Secretaria da Educação (1996,1997). Finalmente, apresentam-se conclusões e sugestões na seção 7.

2. Modelos matemáticos

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

$$P(X_i = 1|\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

onde

X_i é uma variável dicotômica que assume os valores 1 (quando o indivíduo responde corretamente ao item) ou 0 (quando o indivíduo não responde corretamente ao item),

θ representa a habilidade ou proficiência do indivíduo,

$P(X_i = 1 | \theta)$ é a probabilidade de um indivíduo com habilidade igual a θ responder corretamente ao item,

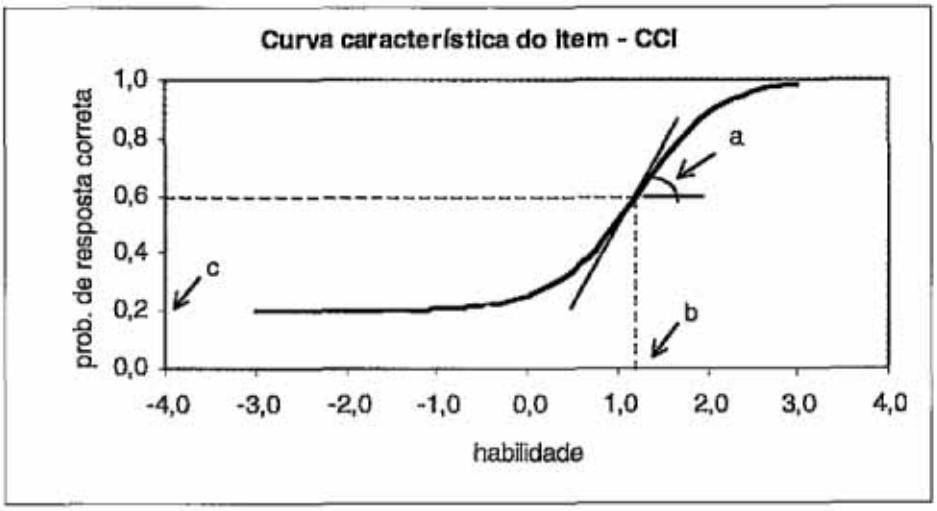
D é um fator de escala constante conhecido, igual a 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal,

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

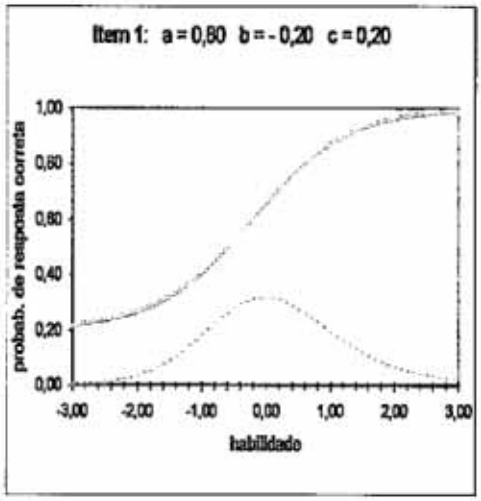
a_i é o parâmetro de discriminação (ou de inclinação) do item, com valor proporcional à inclinação da curva característica do item - CCI no ponto b_i, e

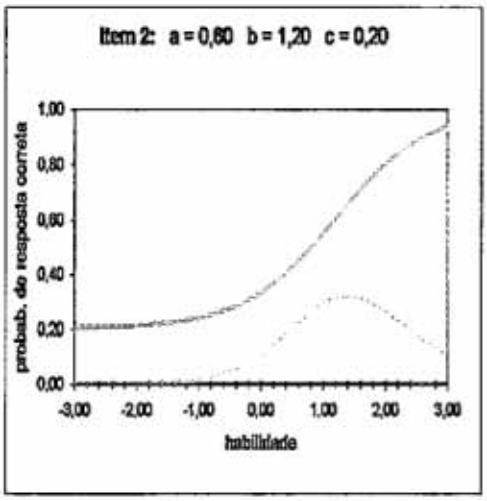
c_i é o parâmetro de acerto ao acaso do item.

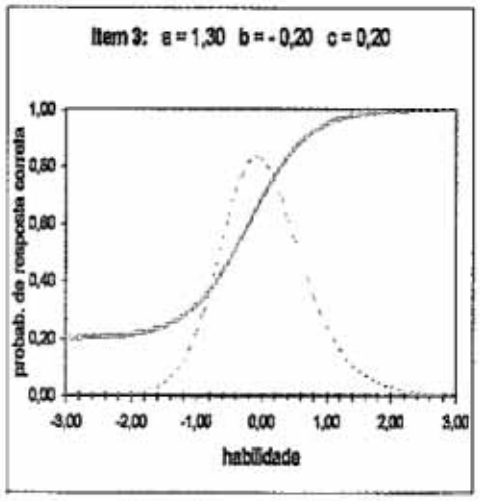
Note que $P(X_i=1|\theta)$ pode ser vista também como a proporção de resposta correta ao item dentre todos os indivíduos da população com a mesma habilidade θ. O gráfico a seguir exemplifica a relação existente entre $P(X_i=1|\theta)$ e os parâmetros do modelo.

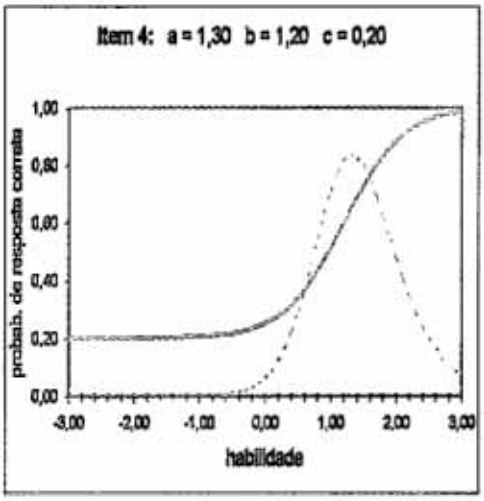


O modelo proposto baseia-se no fato de que indivíduos com maior habilidade possuem maior probabilidade de acertar ao item e que esta relação não é linear. De fato, pode-se perceber a partir do gráfico anterior que a CCI tem forma de "S" com inclinação e deslocamento na escala de habilidade definidos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









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

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

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

2.1 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 ao longo deste trabalho por escala(0,1) . Em termos práticos, não faz a menor diferença estabelecer-se estes valores ou outros quaisquer. O importante são as relações de ordem existentes entre os pontos da escala. Por exemplo, na escala utilizada acima um indivíduo com habilidade 1,20 está 1,20 desvios padrão acima da habilidade média. Este mesmo indivíduo teria a habilidade 92,00, e conseqüentemente estaria também 1,20 desvios padrão acima da habilidade média, se a escala utilizada para esta população fosse a escala(80,10). Isto pode ser visto a partir da transformação de escala

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

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

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

Por exemplo, os valores dos parâmetros **a** e **b** do item 4, mostrado anteriormente, na escala(0,1) são, respectivamente, 1,30 e1,20 e seus correspondentes na escala(80,10) são, respectivamente, 0,13=1,30/10 e 92,00=10x1,20+80. Além disso, um indivíduo com habilidade θ = 1,00 medida na escala(0,1) tem sua habilidade representada por θ^* = 10X1,00+80 = 90,00 na escala(80,10) e

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

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

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

ou seja, a probabilidade de um indivíduo responder corretamente a um certo item é sempre a mesma, independentemente da escala utilizada para medir a sua habilidade ou, ainda, a habilidade de um indivíduo é invariante, independe da escala de medida.

Assim, não faz qualquer sentido querermos analisar itens a partir dos valores de seus parâmetros **a** e **b** sem conhecer a escala na qual eles foram determinados. Na escala(0,1), valores mais apropriados para o parâmetro **a** estão no intervalo [0,60;1,70] e para o parâmetro **b** no intervalo [-2,00;2,00]. É claro que estes valores dependem muito do objetivo da avaliação. Por exemplo, um item com **a** igual a 2,00 serve, basicamente, para discriminar os indivíduos em dois grupos de habilidade, os que tem habilidade menor do que o valor de **b** dos que tem habilidade maior do que o valor de **b**. Note que o valor do parâmetro **c** não se altera com a mudança de escala porque ele mede a probabilidade de acerto para indivíduos com baixa habilidade, qualquer que seja a escala de medida.

2.2 Unidimensionalidade/independência local

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

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

2.3 Outros modelos

Dois outros modelos podem ser facilmente obtidos do modelo apresentado acima. Por exemplo, quando não existe possibilidade de resposta correta ao acaso pode-se considerar $c = 0$ no modelo acima

e tem-se o chamado **modelo logístico unidimensional de 2 parâmetros.** Se além de não existir resposta ao acaso ainda tivermos todos os itens com o mesmo poder de discriminação, tem-se o chamado modelo de 1 parâmetro, o qual possui somente o parâmetro de dificuldade do item. Este modelo é conhecido como modelo de Rasch. Uma generalização desses modelos para 1, 2 e 3 parâmetros, para o caso de duas ou mais populações, foi recentemente proposta por Bock e Zimowski (1997).

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

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

Uma das etapas mais importante da TRI é a estimação dos parâmetros dos itens e/ou das habilidades dos respondentes. Em algumas situações, os parâmetros dos itens já são conhecidos e o que se deseja é estimar as habilidades; já em outras situações menos freqüentes, conhecem-se as habilidades dos respondentes e o que se deseja é a estimação dos parâmetros dos itens. Porém, **as situações mais comuns são aquelas em que se deseja estimar tanto os parâmetros dos itens quanto as habilidades dos respondentes simultaneamente.** O processo de estimação dos parâmetros dos itens é conhecido por **calibração.** Em todas estas situações, assume-se como verdadeiro o modelo proposto e, a partir do conjunto de respostas dadas por um certo número de respondentes selecionados aleatoriamente de uma determinada população, estimam-se os parâmetros e/ou habilidades a partir do método de máxima verossimilhança ou de métodos bayesianos. Ambos os métodos exigem procedimentos iterativos que envolvem cálculos bastante complexos e, conseqüentemente, programas de computador específicos. É importante ressaltar que, em qualquer uma destas situações, os valores das habilidades e dos parâmetros dos itens estarão todos na mesma escala de medida. Vários autores têm sugerido que cada respondente seja submetido a pelo menos 30 itens e que cada item seja submetido a pelo menos 300 respondentes, para que se obtenham estimativas

com erros padrões pequenos. Note que, apesar de estarmos sempre nos referindo à habilidade de um indivíduo, na prática, em geral, o que se deseja é estimar a habilidade média de uma população de indivíduos, por exemplo, a população dos alunos da 3.ª série(*) do ensino fundamental da escola pública estadual de São Paulo. A estimativa da média terá um erro padrão menor do que o erro padrão associado a cada uma das estimativas individuais. O leitor interessado nesses métodos de estimação poderá consultar, entre outros, Baker(1992).

4. Equalização ("Equating")

Uma das grandes vantagens da TRI sobre a Teoria Clássica das Medidas é que ela permite a equalização das habilidades de indivíduos, pertencentes ou não à mesma população, que são submetidos a diferentes provas, possibilitando assim a comparação de seus desempenhos. Um exemplo é o Sistema Nacional de Avaliação do Ensino Básico – SAEB (ver Ministério da Educação e do Desporto (1995)) onde, devido ao grande número de itens considerados por disciplina, é utilizado o plano BIB espiral que possibilita a alocação de subconjuntos de itens a diferentes indivíduos. Um outro exemplo seria a avaliação da escola pública estadual do Estado do Rio Grande do Norte realizada pela Fundação Carlos Chagas em 1997 (ver Fundação Carlos Chagas (1997)), que utilizou alguns itens do SAEB-95 para possibilitar a comparação do desempenho destes alunos em relação ao resto do pais. **A equalização somente é possível devido ao princípio da invariância dos parâmetros dos itens** e é realizada de forma diferente, dependendo dos respondentes pertencerem ou não à mesma população.

Quando os indivíduos avaliados pertencem todos à mesma população, por exemplo, são todos alunos de uma mesma série, a equalização de suas habilidades pode ser feita tendo eles sido submetidos a provas com itens comuns ou não. A idéia básica é que cada prova seja respondida por uma amostra aleatória de alunos de uma mesma população e consequentemente todos pertencem à mesma distribuição de habilidade. Como exemplos podemos citar o SAEB e o

(*) Usou-se a nomenclatura anterior ao estabelecimento da progressão continuada e do ciclo na escola fundamental.

SARESP com os dados analisados por série e o estudo, recentemente publicado, por Soares, Martins e Assunção (1998), que compara o desempenho de alunos que prestaram exames vestibulares para o curso de Direito da UFMG e da PUC-MG em 1995.

A equalização das habilidades de indivíduos de diferentes populações é mais complexa e exige que indivíduos sejam submetidos a provas com itens comuns. Existem duas formas de se realizar tal equalização.

Na primeira forma faz-se a calibração dos itens em separado para cada uma das populações. Como resultado tem-se os parâmetros dos itens, comuns ou não, calibrados nas escalas estabelecidas para cada uma das populações. Utilizando-se o princípio da invariância dos itens, faz-se a equalização dos valores do parâmetro **b** dos itens comuns por meio de técnicas de regressão, por exemplo. Como o valor do parâmetro **b** está na mesma escala de medida da habilidade, este procedimento faz também a equalização das habilidades dos indivíduos das diferentes populações. O leitor poderá encontrar maiores detalhes em Kolen e Brennan (1995), entre outros.

A segunda forma de equalização faz uso do modelo da TRI para duas ou mais populações apresentado por Bock e Zimowisk (1997). Neste caso, a equalização é feita em um único passo tendo em vista que esta formulação permite a estimação simultânea, e, conseqüentemente, na mesma escala de medida, dos parâmetros de todos os itens e das habilidades de todos os indivíduos envolvidos. A equalização dos resultados da 4ª e 8ª séries do ensino fundamental e da 3ª série do ensino médio do SAEB e a equalização dos resultados da avaliação do rendimento da escola pública do Estado do Rio Grande do Norte com o SAEB-95 foram realizadas por meio da primeira forma de equalização. Por outro lado, a equalização dos resultados obtidos no programa de aceleração da aprendizagem do Instituto Ayrton Senna (ver Fundação Carlos Chagas (1998)) com os resultados da 4ª. série do ensino fundamental do SAEB-95 e a equalização dos resultados dos alunos da 3ª série de 1996 com os resultados dos alunos da 4ª série de 1997 da escola pública do Estado de São Paulo foram realizadas por meio da segunda forma. Espera-se que a primeira forma de equalização seja menos precisa do que a segunda (ver Hedges e Vevea (1997), por exemplo) porque na sua aplicação é introduzido um procedimento a mais, o ajuste por regressão, que também está sujeito a erros de modelagem. Conseqüentemente, o número de itens comuns para a equalização

deverá ser maior na primeira forma (pelo menos 15) do que na segunda (pelo menos 6).

5. Construção e interpretação de escalas de habilidade

Como apresentado acima, os valores obtidos (estimativas) das habilidades dos respondentes dependem dos valores fixados para origem e unidade da escala de medida. Estes valores são escolhidos de modo a representarem, respectivamente, a habilidade média e o desvio padrão da distribuição da habilidade da população em estudo. Como estes valores são arbitrários, as habilidades obtidas para os diferentes respondentes são comparáveis entre si mas 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 desvio 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.

A interpretação pedagógica dos valores das habilidades apoia-se nos conceitos abaixo:

Níveis âncora: são pontos selecionados pelo analista na escala da habilidade para serem interpretados pedagogicamente.

Itens âncora: são itens selecionados, segundo a definição dada abaixo, para cada um dos níveis âncora.

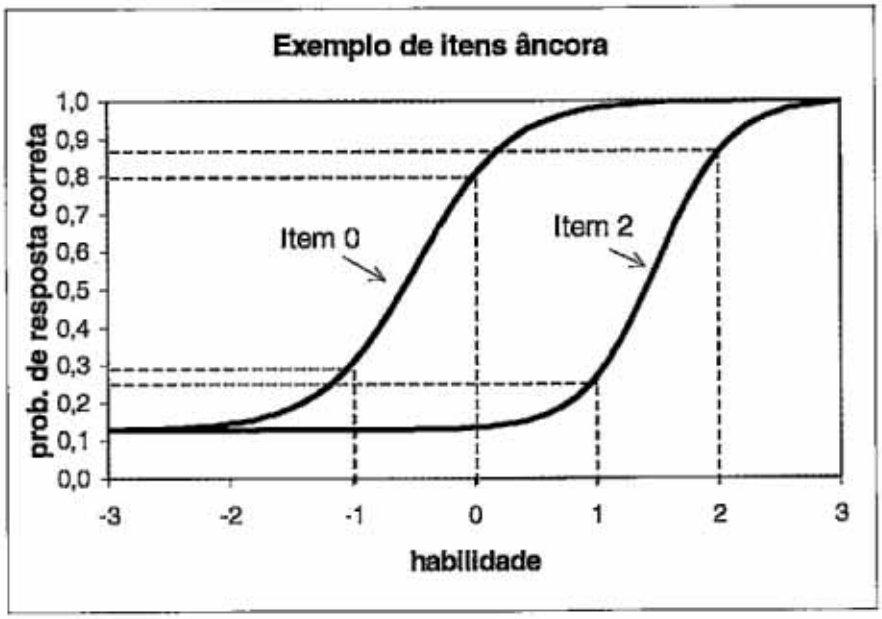
Definição de item âncora: Considere dois níveis âncora consecutivos Y e Z com $Y < Z$.. Dizemos que um determinado item é âncora para o nível Z se e somente se

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 para um determinado nível âncora da escala, ele precisa ser respondido corretamente por uma grande proporção de indivíduos com este nível de habilidade e por uma pequena proporção de indivíduos com o nível de habilidade imediatamente anterior. No gráfico abaixo são apresentados, em uma escala de habilidade com níveis âncora –3, -2, -1, 0, 1, 2, 3, exemplos de itens âncora (item0 e item2) para os níveis âncora 0 e 2, respectivamente. Os parâmetros dos itens são: $a_0 = 1{,}52$, $b_0 = -0{,}47$, $c_0 = 0{,}13$, $a_2 = 1{,}97$, $b_2 = 1{,}50$ e $c_2 = 0{,}13$.



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

$P(X_0=1|\theta=0) = 0,80$ e $P(X_0=1|\theta=-1) = 0,31 \Rightarrow P(X_0=1|\theta=0) - P(X_0=1|\theta=-1) = 0,49$

e

$P(X_2=1|\theta=2) = 0,86$ e $P(X_2=1|\theta=1) = 0,27 \Rightarrow P(X_2=1|\theta=2) - P(X_2=1|\theta=1) = 0,59$

A priori, não se pode ter certeza de quantos itens âncoras serão selecionados para cada nível âncora e nem se existirão no teste aplicado itens âncoras para todos os níveis âncora selecionados. Por isto, é fundamental que os níveis âncoras 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. Por exemplo, no SAEB-95 foram aplicados 169 itens na 3ª série do ensino médio para a construção da escala de matemática. Maiores detalhes sobre construção e interpretação de escalas de habilidade poderão ser encontrados em Beaton e Allen (1992).

6. Aplicação: Equalização dos resultados SARESP96/97

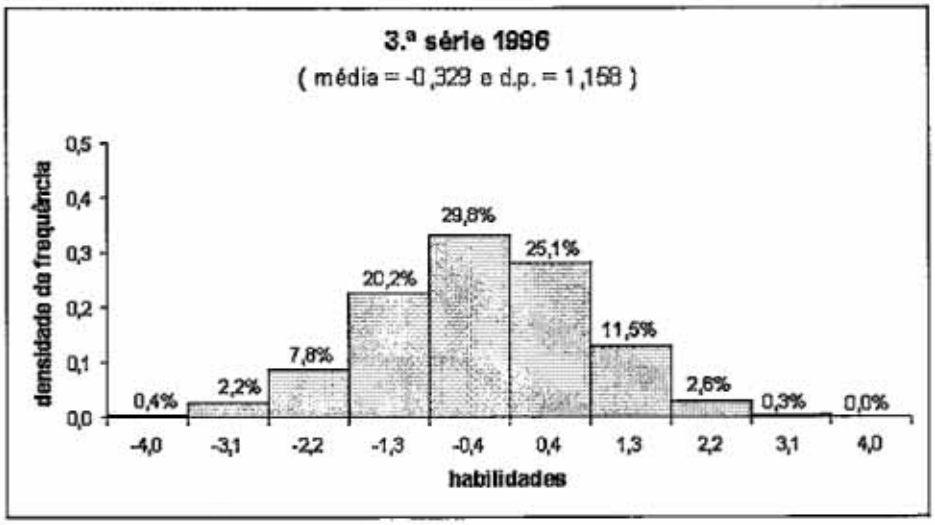
A TRI vem sendo aplicada em diversas avaliações educacionais no Brasil e no exterior. Várias dessas aplicações foram citadas nas seções anteriores. Nesta seção, pretende-se discutir com mais detalhes uma aplicação da TRI, que achamos de relevância, para exemplificar uma das muitas contribuições que a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

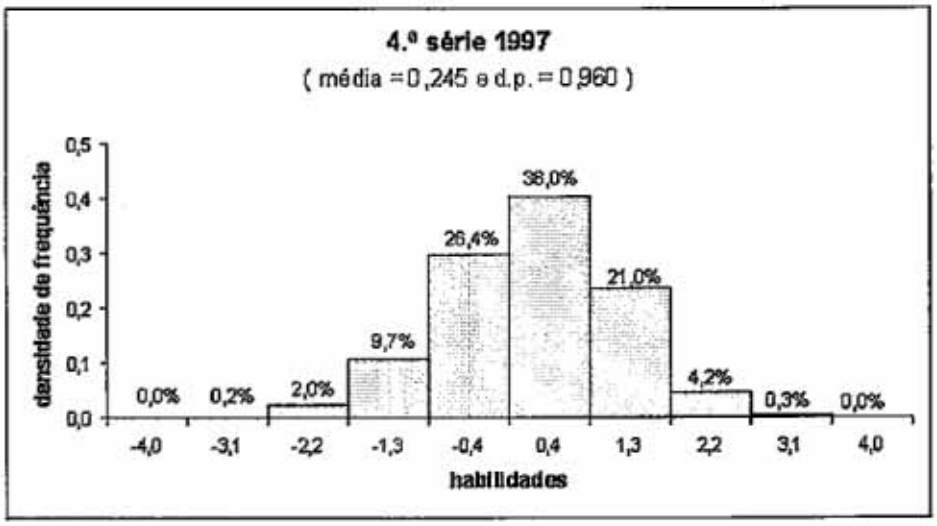
O Sistema de Avaliação do Rendimento das Escolas do Estado de São Paulo -SARESP tem utilizado a TRI na análise dos seus dados dos anos de 1996 (3ª série e 7ª série do ensino fundamental), 1997 (4ª série e 8ª série do ensino fundamental) e 1998 (5ª série do ensino fundamental e 1ª série do ensino médio).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

da conclusão deste artigo a análise dos dados de 1998 estava em andamento.

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

De modo a viabilizar comparações entre as duas populações de interesse, elaborou-se uma terceira prova, chamada de prova de ligação, com 30 itens (11 itens da prova de 1996 e 21 itens da prova de 1997) que foi aplicada em outubro de 1997 a uma amostra de alunos de uma terceira população (3ª série de 1997). A amostra utilizada para a terceira população, chamada de população de ligação, foi definida de modo a ter-se uma boa calibração de todos os 58 itens, em uma mesma escala de habilidade. Não havia qualquer interesse em estimar a distribuição das habilidades dos indivíduos dessa terceira população. O uso de um número diferente de itens de 1996 e de 1997 na terceira prova, deveu-se ao melhor desempenho dos itens da prova de 1997; desempenho este avaliado pela aplicação da TRI nos dados isolados de cada um dos dois anos. Nos dados das tres populações, aplicou-se o modelo logístico unidimensional de tres parâmetros para múltiplos grupos (neste nosso caso tres grupos) que possibilitou, a partir da prova de ligação com itens comuns às provas de 1996 e 1997, a calibração de todos os 60 itens e a estimação das habilidades de todos os respondentes de 1996 e 1997 em uma única escala de medida. Os histogramas a seguir representam as distribuições das habilidades dos alunos da 3ª série de 1996 e da 4ª série de 1997. A habilidade 0 corresponde à habilidade média da população de ligação (3ª série de 1997).





A partir dos histogramas, pode-se concluir que a habilidade média dos alunos da 4ª série de 1997 é maior do que a habilidade média dos alunos da 3ª série de 1996, e também que a variabilidade das habilidades dos alunos da 4ª série de 1997 é menor do que a variabilidade das habilidades dos alunos da 3ª série de 1996. Dos

histogramas pode-se também obter, a partir do cálculo de áreas de retângulos, a porcentagem de alunos em determinado intervalo de habilidade. Por exemplo, enquanto que em 1996 a porcentagem de alunos da 3ª série do ensino fundamental que tinham habilidade maior ou igual a 0 foi de 39,5% (=25,1% + 11,5% + 2,6% + 0,3%), em 1997 a porcentagem de alunos da 4ª série do ensino fundamental que tinham habilidade maior ou igual a 0 foi de 61,5% (=36,0% + 21,0% + 4,2% + 0,3%). Estes valores nos mostram que houve um aumento de 22,0 pontos percentuais na porcentagem de alunos com habilidade igual ou maior do que 0 de 1996 para 1997. Com todos os 60 itens calibrados na mesma escala, pode-se também construir uma interpretação pedagógica para esta escala nos níveis âncora –2, -1, 0, 1, 2. Sugere-se que o leitor consulte Secretaria da Educação (1996,1997) para maiores resultados sobre as escalas construídas e suas interpretações pedagógicas.

Todos os cálculos para calibração e estimação das habilidades foram efetuados com os recursos dos programas computacionais BILOG 3 (ver Mislevy e Bock (1990))., para uma única população, e BILOG-MG (ver Zimowisk e outros (1996)), para mais de uma população. Outros dois programas computacionais bastante utilizados, entre outros, são o TESTFACT (ver Wilson e outros (1991)), para o estudo da dimensionalidade do teste, e o PARSCALE (ver Muraki e Bock (1993)), para a calibração de itens politômicos.

7. Conclusões e sugestões

Nossa principal motivação para escrever este primeiro trabalho sobre a Teoria da Resposta ao Item foi o pouco conhecimento que pesquisadores e profissionais da área de educação e estatística no Brasil têm sobre esta teoria. Por isto, preocupamo-nos muito mais em fornecer uma ampla visão de seus fundamentos básicos e de suas principais aplicações no Brasil, do que apresentar detalhes matemáticos de modelagem. Apesar desta teoria ter mais de 50 anos, somente nos últimos 15 anos é que ela vem sendo aplicada em larga escala nas principais avaliações educacionais em diferentes paises. Atribui-se este fato à complexidade matemática dos métodos envolvidos e também à ausência de programas computacionais eficientes. A aplicação apropriada desta teoria exige necessariamente o

envolvimento de especialistas em educação e em estatística. Sua primeira aplicação no Brasil foi na análise do SAEB95.

Alguns pontos têm sido levantados na literatura sobre a adequação desta teoria. Dois deles que consideramos importantes são a dimensionalidade do espaço de traços latentes envolvidos na avaliação e a equalização de diferentes avaliações. Como exemplos do segundo ponto, destacamos as equalizações dos resultados do SAEB95 e do SAEB97 e de resultados de avaliações estaduais com o SARESP. Com relação ao primeiro ponto, alguns autores têm defendido a tese de que os modelos unidimensionais têm fornecido bons resultados mesmo em situações multidimensionais mas com uma das dimensões predominante. Mais recentemente, modelos para mais de uma dimensão têm sido propostos. Com relação ao problema da equalização, a proposta recente de modelos para múltiplos grupos de Bock e Zimowski (1997) deu um novo rumo à solução deste problema, tendo em vista que os modelos anteriores envolvem outros erros de modelagem além daqueles da própria teoria. Sugerimos a leitura de Goldstein e Wood (1989), Mislevy (1992), Goldstein (1994) e Hedges e Vevea (1997), entre outros, para um melhor entendimento destes problemas e suas soluções. Um terceiro ponto, não menos importante do que os outros dois, é aquele levantado por Mislevy (1991) sobre a 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. Algumas das dificuldades para a aplicação desta metodologia seriam a não existência comercial, até o presente momento, de programas computacionais, a obtenção de informações adicionais fidedignas relevantes ao problema e a inclusão dessas mesmas informações no modelo.

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

departamentos de estatística e de medidas em educação, em algumas de nossas universidades, seria de fundamental importância.

Agradecimentos

Os autores gostariam de agradecer aos professores Heraldo Marelim Vianna e Yara Lúcia Espósito pelas críticas e sugestões efetuadas na primeira versão deste trabalho e a Secretaria de Estado da Educação do Estado de São Paulo pela utilização de parte dos dados do SARESP96 e do SARESP97. Este trabalho foi parcialmente financiado pelo CNPq, pela EMBRAPA através do Projeto Técnico nº 23800.96/021-05-01, pelo PRONEX convênio nº 76.97.1081.00 e pela FAPESP através do Projeto Temático convênio nº 96/01741-7.

Referências bibliográficas

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

BEATON, A.E. e ALLEN, N.L. (1992). *Interpreting Scales Through Scale Anchoring.* **Journal of Educational Statistics, 17** : 191-204.

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

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. e 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., SWAMINATHAN, H. e ROGERS, H.J. (1991). **Fundamentals of Item Response Theory**. Newburry Park: Sage Publications.

HEDGES, L.V. e 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, Michael J. e BRENNAN, Robert L. (1995). ***Test Equating - Methods and Practices***. New York : Springer

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

LORD, F.M. e NORVICK, M.R. (1968). **Statistical Theories of Mental Test Socre**. Reading: Addison-Wesley.

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO (1995). **Sistema Nacional de Avaliação da Educação Básica: SAEB 95 relatório técnico**. São Paulo/Rio de Janeiro: Fundação Carlos Chagas/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. e BOCK, R.D. (1990). **BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models**. Chicago: Scientific Software, Inc.

MURAKI, E. e BOCK, R.D. (1993). **PARSCALE: IRT Based Test Scoring and Item Analysis for Graded Open-Ended Exercices and Performance Tasks.** Chicago: Scientific Software, Inc.

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.** São Paulo: SEE. (em fase de elaboração)

SOARES, J. F., MARTINS, M. I. e 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.

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

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

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