Utilização da Teoria da Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (SAEB)

Ruben Klein

Resumo

O artigo introduz a Teoria da Resposta ao Item (TRI) em sua forma usual de um único grupo e para grupos múltiplos e explica como a TRI para grupos múltiplos está sendo utilizada no Sistema Nacional de Avaliação da Educação Básica (SAEB) para a calibração dos itens e para a obtenção de uma escala única, por disciplina, para as proficiências dos alunos das 4ª e 8ª séries do Ensino Fundamental e para a 3ª série do Ensino Médio, para os SAEBs a partir de 1995. O artigo apresenta o procedimento adotado de equalização entre séries e entre anos e introduz um método novo para equalizar avaliações que utilizam alguns itens do SAEB com o SAEB. Finalmente o artigo explica como as escalas do SAEB são interpretadas.

Palavras-chave: *TRI - grupos múltiplos – SAEB – equalização - interpretação de escala - proficiência- estimação Bayesiana marginal.*

Ruben Klein
Doutor em Matemática, Massachussets Institute of Technology, EUA. Pesquisador do Laboratório Nacional de Computação Científica - LNCC/CNPq e Consultor da Fundação Cesgranrio.

1.Introdução

A Teoria de Resposta ao Item (TRI), ver Lord e Novick (1968); Hambleton, Swaminathan e Rogers (c1991) e Baker (1992), surge da necessidade de superar as limitações da apresentação de resultados somente através de percentuais de acertos ou escores dos testes e ainda da dificuldade de comparar resultados de diferentes testes em diversas situações. Referências recentes, em Português, sobre a TRI são Andrade, Tavares e Valle (2000) e Andrade e Klein (1999).

Na Teoria Clássica dos Testes, os resultados dependem do particular conjunto de questões que compõem a prova e dos indivíduos que a fizeram, ou seja, as análises e interpretações estão sempre associadas à prova como um todo e ao grupo de indivíduos. Assim, a comparação entre indivíduos ou grupos de indivíduos somente é possível quando eles são submetidos às mesmas provas ou, pelo menos, ao que se denomina de provas paralelas, quase sempre difíceis de serem

construídas. Desta maneira, fica muito difícil fazer comparações quando diferentes indivíduos fazem provas diferentes. Isto é importante em avaliações de larga escala, quando se quer avaliar uma grande parte de um currículo de uma determinada disciplina e série e, para tal, é necessário apresentar um grande número de itens aos alunos, maior do que aquele que eles poderiam responder em 1 ou 2 horas de prova. Além disso, nas avaliações ao longo dos anos, torna-se difícil comparar resultados quando as provas não são as mesmas. Este é o caso do Sistema Nacional de Avaliação da Educação Básica (SAEB), ver Fontanive e Klein (2000).

A TRI muda o foco de análise da prova como um todo para a análise de cada item. A TRI é um conjunto de modelos matemáticos onde a probabilidade de resposta a um item é modelada como função da proficiência (habilidade) do aluno (variável latente, não observável) e de parâmetros que expressam certas propriedades dos itens. Quanto maior a proficiência do aluno, maior a probabilidade de ele acertar o item.

Uma das propriedades importantes da TRI é o fato de os parâmetros dos itens e as proficiências dos indivíduos serem invariantes. Tanto os parâmetros dos itens obtidos de grupos diferentes de alunos testados quanto os parâmetros de proficiência baseados em grupos diferentes de itens são invariantes, exceto pela escolha de origem e escala.

Graças a essa propriedade, a TRI, associada a outros procedimentos estatísticos, permite comparar alunos, estimar a distribuição de proficiências da população e subpopulações e ainda a monitorar os progressos de um sistema educacional.

Um dos procedimentos mais utilizados para colocar indivíduos que fazem provas diferentes em uma mesma escala de proficiência é a utilização de itens comuns nas provas. Desta maneira, os desempenhos dos indivíduos podem ser comparados.

Por exemplo, no SAEB, o uso de itens comuns entre séries e entre anos permite que os alunos de todas as séries e de todos os anos sejam postos em uma mesma escala de proficiência de modo que seus desempenhos possam ser comparados.

Como a escala de proficiências obtida pela TRI não tem uma origem e unidade de medida absoluta, estas têm que ser arbitradas. É necessário, portanto, que esta escala seja interpretada com o intuito de se saber o que alunos ou indivíduos sabem e são capazes de fazer em determinados níveis da escala.

Na realidade, esta interpretação deve ser feita em qualquer escala, como, por exemplo, a escala dos percentuais de acerto, pois, dependendo da dificuldade da prova, um 3 (três) pode significar mais que um 7 (sete).

As proficiências dos indivíduos e os parâmetros dos itens do teste precisam ser estimados a partir das respostas em cada item do teste. O significado de calibrar um item é justamente o de estimar seus parâmetros.

O emprego desta tecnologia permite a elaboração de escalas de proficiências comuns a todas as séries e na mesma escala obtida em anos anteriores. Isso permite fazer comparações e acompanhar a evolução do sistema de ensino tanto entre séries, como ao longo dos anos.

Os modelos utilizados nos SAEB/95, SAEB/97, SAEB/99 e SAEB/2001são indicados a seguir:

i) Para os itens de múltipla escolha, é utilizado o modelo da logística de 3 parâmetros definido por:

$$P(x_j = 1 \mid \theta, a_j, b_j, c_j) = c_j + \frac{(1 - c_j)}{1 + \exp\left[-D a_j (\theta - b_j)\right]} \equiv P_{j1}(\theta)$$

onde $\begin{cases} = 1 \text{ se correta} \\ = 0 \text{ caso contrário} \end{cases}$

x_j é a resposta ao item j,

a_j onde $a_j > 0$ é o parâmetro de inclinação do item j, também chamado de parâmetro de discriminação do item

b_j é o parâmetro de posição (ou de dificuldade) do item, e

c_j onde $0 < c_j < 1$ é o parâmetro da assíntota inferior do item j, refletindo as chances de um estudante de proficiência muito baixa selecionar a opção de resposta correta.

D=1 métrica logística; D=1,7 métrica normal

Denota-se a probabilidade de resposta errada ao item de:

$$P_{j0}(\theta) = P(x_j = 0 \mid \theta, a_j, b_j, c_j) = 1 - P_{j1}(\theta)$$

i) Para os itens de resposta construída pelos alunos (itens abertos) corrigidos como certos ou errados, utiliza-se o modelo da logística de 2 parâmetros, obtida da logística de 3 parâmetros, com c_j igual a 0 (zero).

ii) Para os itens de resposta construída corrigidos com graduação de 0 a 2 ou 0 a 3 (SAEB/95), foi utilizado o modelo de crédito parcial generalizado, ver Muraki (1992), definidos por

$$P_j(\theta) = P(x_j = k \mid \theta, a_j, b_{jm}, m = 0, \ldots, m_{j-1}) = \frac{\exp\left[\sum_{v=0}^{k} D a_j (\theta - b_{ju})\right]}{\sum_{c=0}^{m_{j-1}} \exp\left[\sum_{v=0}^{c} D a_j (\theta - b_{jv})\right]}$$

onde

m_j o número de categorias na resposta ao item j

x_j a resposta ao item j, com possibilidades $0,2,\ldots,m_{j-1}$

a_j o parâmetro de inclinação.

b_{jk}, $k=1,\ldots,m_{j-1}$ parâmetros de posição,

sendo que b_{jk} é o ponto onde as curvas

$P_{j,k-1}$ (q) e P_{jk}(q) se interceptam.

D=1 métrica logística; D=1,7 métrica normal

A indeterminação nos parâmetros do modelo é resolvida colocando-se $b_{j0}=0$.

O uso da TRI pressupõe as seguintes hipóteses:

i) unidimensionalidade da proficiência;

ii) independência condicional das respostas de um aluno a um conjunto de itens, dado a proficiência do aluno. Isto é:

$$P(x_i = (x_{i1},\cdots x_{in}) \mid \theta_i, \text{parâmetros dos itens}) = \prod_{j=1}^{n}\prod_{k=0}^{m_{j-1}} P_{jk}(\theta_i)^{u_{ijk}}$$

onde q_i é a proficiência do aluno i, $x_i = (x_{ik},...,x_{in})$ são as respostas aos itens, $P_{ik}(q_i)$ é a probabilidade do aluno i responder o nível k de resposta do item ($m_j=2$ para os itens de múltipla escolha) e

$$u_{ijk} = \begin{cases} 1 \text{ se } \mathbf{x}_{ij} = k \\ 0 \text{ caso contrário} \end{cases}$$

Essas duas condições são equivalentes, isto é, se a proficiência é unidimensional, então vale a independência condicional. Inversamente, se vale a independência condicional com uma única variável, então a proficiência é unidimensional.

Generalizando, a dimensão da proficiência é igual ao número de variáveis necessárias para a independência condicional.

Os parâmetros dos itens no SAEB/2001 foram estimados conjuntamente para todos os itens (que são todos de múltipla escolha) de todas as séries de uma mesma disciplina, pelo programa BILOG-MG, que implementa uma extensão da TRI a grupos múltiplos de respondentes (BOCK ; ZIMOWSKI, 1996; ZIMOWSKI et al. 1996), utilizando o procedimento de estimação Bayesiana marginal (cf. MISLEVY 1986; BAKER 1992).

Esta extensão da TRI de uma população para várias subpopulações ou grupos não-equivalentes é feita considerando-se distribuições *a priori* diferentes para cada subpopulação. Para eliminar a indeterminação do modelo, arbitra-se como na TRI de um único grupo a distribuição *a priori* do grupo de referência. Desta maneira, estimam-se conjuntamente os parâmetros das distribuições *a priori* dos outros grupos e dos parâmetros dos itens. Este procedimento permite que se faça simultaneamente a equalização entre séries e, quando for o caso, a equalização entre anos.

Nos SAEBs, as respostas faltantes de um aluno no fim de um bloco são consideradas como "não alcançadas" e tratadas como se não tivessem sido apresentadas ao aluno. As respostas faltantes aos itens de múltipla escolha anteriores ao último item respondido são consideradas omissões intencionais e suas respostas são consideradas erradas.

Na seção 2, apresenta-se uma breve descrição técnica sobre os procedimentos de estimação dos parâmetros dos itens. Os leitores, que preferirem, podem passar para as seções 3 a 5, que descrevem a equalização entre séries e entre anos feita nos SAEBs desde 1995. Finalmente a seção 6 descreve os procedimentos de interpretação da escala, um dos conceitos mais importantes introduzidos no país na análise do SAEB 95.

2. Estimação dos parâmetros dos itens

No caso de um grupo, a estimação dos parâmetros dos itens, sem considerar as distribuições *a priori* para estes parâmetros, é feita pelo método de máxima verossimilhança marginal, isto é, os estimadores dos parâmetros são os valores que maximizam a verossimilhança marginal.

Seja ξ_j' o vetor de parâmetros do item j e $\xi' = (\xi_1', \ldots, \xi_m')$ o vetor de parâmetros de todos os itens.

Seja $x' = (x_1, \ldots, x_m)$ um vetor de respostas, $x_j = 1$ ou 0, de um indivíduo com proficiência θ. Então a probabilidade de ocorrência deste vetor de resposta é:

$$P(x/\theta,\xi) = \prod_{j=1}^{m} P_{j_1}^{x_j}\left(1 - P_{j_1}\right)^{(1-x_j)} = \prod_{j=1}^{m} P_{j_1}^{x_j} P_{j_0}^{(1-x_j)}$$

Supondo-se que a população de respondentes pertença a uma população em que a proficiência θ tenha uma distribuição de probabilidade contínua, com densidade $g(\theta)$ média e variância finitas, tem-se que a probabilidade marginal é dada por:

$$P(x/\xi) = \int P(x/\theta,\xi) g(\theta) d\theta$$

e a função de verossimilhança marginal de ξ é dada por:

$$L = P(x_1, \ldots, x_n/\xi) = \prod_{i=1}^{n} P(x_i/\xi),$$

onde n é o número de indivíduos.

Em geral, a integral

$$\int P(x/\theta,\xi) g(\theta) d\theta$$

não pode ser expressa em forma fechada, mas pode ser estimada pela fórmula de quadratura gaussiana:

$$P(x/\xi) \approx \sum_{q=1}^{Q} P\left(x/X_q \xi\right) A\left(X_q\right)$$

onde X_q é um ponto de quadratura e $A(X_q)$ é um peso positivo correspondendo a função de densidade $g(\theta)$ no ponto X_q.

Maximizar **L** é equivalente a maximizar **logL** que por sua vez é equivalente a obter a solução do sistema

$$\frac{\partial \log L}{\partial \xi_{j\ell}} = \sum_{q=1}^{Q} \sum_{k=0}^{1} \bar{r}_{jkg} \frac{1}{P_{jk}\left(X_q, \xi_j\right)} \frac{\partial P_{jk}\left(X_q, \xi_j\right)}{\partial \xi_{j\ell}}$$

$$j = 1, \ldots m; \ell = 1, \ldots, \ell_j$$

onde **m** é o número de itens, ℓ_i é o número de parâmetros do item **j**

$$\tilde{r}_{jkq} = \sum_{i=1}^{n} x_{ijk} P\left(X_q/x_i, \xi\right) \Delta X_q$$

é o número esperado de respostas na categoria **k=0** ou **1**, do item **j** por indivíduos com proficiência no intervalo

$$\left(X_q - \Delta X_q/2, X_q + \Delta X_q/2\right)$$

e $\tilde{N}_{jkq} = \sum_{k=0}^{1} \tilde{r}_{jkq}$ é o número esperado de indivíduos que responderam ao item **j** com proficiência no intervalo

$\left(X_q - \Delta X_q/2, X_q + \Delta X_q/2\right)$.

Observa-se também que

$$p\left(X_q/x_i,\xi\right)\Delta X_q = \frac{P\left(x_i/X_q,\xi\right)A\left(X_q\right)}{\sum_{q=1}^{Q} P\left(x_i/X_q,\xi\right)A\left(X_q\right)}$$

é a probabilidade *a posteriori* de a proficiência estar no intervalo

$\left(X_q - \Delta X_q/2, X_q + \Delta X_q/2\right)$

dado o vetor de resposta x_i e o vetor de parâmetros ξ.

A solução do sistema

$$\frac{\partial \log L}{\partial \xi_{j\ell}} = 0 \qquad j = 1,\ldots m; \ell = 1,\ldots,\ell_j$$

pode ser obtida pelo algoritmo **E-M (Esperança-Maximização)**, que é executado em dois passos.

Passo E. Dados valores de $\xi_1,\ldots,\xi_m$, calcula-se $\tilde{r}_{jkq}$ e $\tilde{N}_{jkq}$.

Passo M. Dados $\tilde{r}_{jkq}$ e $\tilde{N}_{jkq}$, obter ξ que resolve o sistema ou equivalentemente maximiza a verossimilhança.

Após cada passo **M**, comparam-se os novos estimadores de ξ com os anteriores e o ciclo **E-M** é repetido até que se obtenham estimadores estáveis ξ.

A distribuição da proficiência θ pode ser qualquer, mas em geral é razoável supor-se a distribuição normal. Como a distribuição é latente (pois a proficiência não é observável diretamente), a escala em que θ é medida é arbitrária.

Resolve-se a indeterminação especificando-se a origem e a unidade de escala. Para fins computacionais, é conveniente assumir que o parâmetro de locação (origem) e o parâmetro de escala (desvio padrão) da distribuição sejam respectivamente 0 (zero) e 1 (um).

A estimação pelo método de máxima verossimilhança marginal pode apresentar problemas de convergência, em particular, quando se usa o modelo logístico de 3 parâmetros. Por isto, utiliza-se o método Bayesiano marginal, ver Mislevy (1986) e Baker (1992).

Neste procedimento, considera-se uma distribuição contínua *a priori* para os parâmetros ξ dos itens, dada pela densidade $g(\xi)$.

O estimador Bayesiano é o valor de ξ que maximiza a densidade da distribuição *a posteriori*

$$g\left(\xi/x_1,\ldots,x_n\right) = \frac{P\left(x_1,\ldots,x_n/\xi\right)g\left(\xi\right)}{P\left(x_1,\ldots,x_n\right)}$$

onde $P\left(x_1,\ldots,x_n/\xi\right) = L$ é a função de verossimilhança marginal de ξ.

Então maximizar $g\left(\xi/x_1,\ldots,x_n\right)$ é equivalente a maximizar $\log L + \log g\left(\xi\right)$, o que permite que se utilize o procedimento anterior somente acrescentando os termos referentes às distribuições *a priori* dos parâmetros.

Na Teoria de Resposta ao Item (TRI) com grupos múltiplos, tem-se uma distribuição contínua com densidade $g_k(\theta)$ média e variância finitas para cada grupo. Em geral, supõe-se a distribuição normal para todos os grupos, mas com médias e variâncias diferentes.

Resolve-se a indeterminação do modelo, arbitrando-se a média e a variância de um dos grupos chamado de referência. Como no caso de um único grupo, é conveniente arbitrar-se o valor 0 (zero) para a média e o valor 1 (um) para a variância no grupo de referência.

A verossimilhança marginal neste caso é dada por:

$$L = \prod_{k=1}^{K} \prod_{i=1}^{n_k} P\left(x_{ki}/\xi,\eta_k\right) =$$

$$= \prod_{k=1}^{K} \prod_{i=1}^{n_k} \int P\left(x_{ki}/\theta,\xi\right) g_k(\theta)\, d\theta$$

onde K é o número de grupos, n_k o número de indivíduos no grupo k e η_k é o vetor de parâmetros da distribuição do grupo k.

Os métodos de estimação dos parâmetros dos itens são os mesmos no caso de um grupo. De maneira análoga ao caso de um grupo, pode-se utilizar o método Bayesiano marginal para a estimação dos parâmetros dos itens. No caso de grupos múltiplos, as médias e variâncias das distribuições dos outros grupos precisam ser estimadas conjuntamente com os parâmetros dos itens.

3. Equalização entre séries e entre anos.

A equalização entre séries no SAEB é feita automaticamente, da maneira indicada anteriormente, com o uso da extensão da TRI a grupos múltiplos não equivalentes, onde as diferentes séries seriam os grupos. Dessa maneira, tem-se uma estimativa única dos parâmetros, não importando a série, e evita-se a propagação de erros na equalização quando se estimam os parâmetros série a série.

A escala comum entre séries e entre anos do SAEB está baseada no SAEB/97, quando a média e o desvio padrão da distribuição de proficiências da 8ª série desse ano foram arbitrados, respectivamente, em 250 e 50.

A atual escala do SAEB é baseada na calibração conjunta dos itens de todas as séries dos SAEBs 95 e 97, utilizando a 8ª série de 1997 como grupo de referência e omitindo-se os itens de resposta construída que admitem mais graduações, além de certo e errado. Utilizaram-se aqui **todas as respostas individuais de todos os alunos** das 3 séries nos SAEBs 95 e 97.

No SAEB/99, a equalização entre anos para a mesma disciplina e a obtenção dos parâmetros dos itens do SAEB/99 na mesma escala do SAEB/97 foram obtidos conjuntamente com a calibração entre séries, incluindo-se na estimação **todas as respostas individuais de todos os alunos da amostra de 1997** com os parâmetros dos itens de 1997 considerados conhecidos (SAEB/99, Relatório Téc-

nico). A amostra dos alunos da 8ª série de 1997 serviu como grupo de referência, como já o tinha sido na estimação conjunta dos itens do SAEB/97 e SAEB/95. A inclusão da amostra de 1997 (ou de uma subamostra) é necessária para que se fixem os parâmetros, já conhecidos dos itens de 1997, os quais estão relacionados com esta população. Utiliza-se o procedimento Bayesiano e os parâmetros são "fixados" através de distribuições *a priori* com variância muito pequena. Desta maneira, obtêm-se estimações dos parâmetros dos itens novos na mesma escala. As médias e desvios padrões dos grupos correspondentes aos alunos da 4ª série e 3ª série do Ensino Médio, assim como os parâmetros dos itens de 97, são recuperados, isto é, são re-estimados e são praticamente iguais aos valores anteriores.

O uso de uma subamostra dos alunos precisa ser estudado. A subamostra precisa ser retirada criteriosamente, levando em conta o plano amostral do SAEB.

No SAEB/2001, a equalização entre anos para a mesma disciplina e a obtenção dos parâmetros dos itens do SAEB/2001 na mesma escala do SAEB/97 foram reliazadas conjuntamente com a calibração entre séries incluindo-se na estimação todas as respostas individuais dos alunos das 3 séries da amostra de 1999, com os parâmetros dos itens de 1999 considerados conhecidos. O SAEB/99 foi utilizado, pois o SAEB/2001 tem itens comuns com o SAEB/99 em todas as séries. O grupo dos alunos da 8ª série de 1999 serviu como grupo de referência. Como se supõe que a média e o desvio padrão da distribuição de proficiências do grupo de referência são, respectivamente 0 (zero) e 1 (um), e isto não é verdade para a distribuição de proficiências da 8ª série de 1999, foi feita uma transformação linear para que isto aconteça. Os parâmetros dos itens de todas as séries de 1999 sofreram a transformação correspondente. Os parâmetros desta transformação são baseados na média e desvio padrão do grupo dos alunos de 8ª série de 1999, estimado no processo de calibração dos parâmetros dos itens. Utilizou-se o *default* do programa, a distribuição normal com média 0 e variância 1 como a distribuição *a priori* das proficiências do grupo de referência e também dos outros. Após a calibração, faz-se a transformação linear inversa, de modo que os parâmetros dos itens de 1999 retornem a seus valores originais e os parâmetros dos itens novos de 2001 fiquem na escala comum do SAEB (SAEB/2001, Relatório Técnico).

Uma outra opção é utilizar, como distribuição *a priori* para o grupo de referência, os alunos da 8ª série de 1999, a distribuição empírica estimada para a 8ª série no SAEB/99 transformada, para que tenha média 0 e variância 1. Pode-se, da mesma maneira, utilizar as distribuições empíricas estimadas no SAEB/99 para os grupos definidos pela 4ª série do Ensino Fundamental e pela 3ª série do Ensino Médio transformadas como as distribuições *a priori* para estes grupos. Observa-se que estas distribuições são re-estimadas pelo programa a cada iteração.

4. Estimação das proficiências dos alunos

No procedimento de estimação por máxima verossimilhança marginal ou Bayesiana marginal, pressupõe-se uma

distribuição *a priori* para as proficiências dos alunos de cada grupo. Em geral, esta distribuição *a priori* é atualizada nos diferentes passos do processo de estimação. O programa chama a última atualização da *priori* de distribuição empírica.

No procedimento usado para se estimar os parâmetros dos itens, obtém-se para cada aluno a sua distribuição *a posteriori* de proficiência, condicional a suas respostas no teste.

O estimador da proficiência do aluno mais utilizado é um estimador Bayesiano, que é a média dessa distribuição *a posteriori* (o chamado estimador EAP), e o estimador da precisão é o desvio padrão dessa distribuição. Neste procedimento, é necessário fornecer a distribuição *a priori* da distribuição de proficiências de cada grupo. No SAEB a distribuição *a priori* que está sendo utilizada para todos os grupos é a distribuição normal com média 0 e desvio padrão 1, N(0,1), distribuição *default* de muitos programas.

Por isto, no SAEB/2001, as proficiências dos alunos de 2001 são estimadas com os parâmetros dos itens de todas as séries transformados para a escala SAEB original (antes da transformação para média 250 e desvio padrão 50 para a 8ª série de 1997) e a *priori* N(0,1) para todos as séries. Para isto, roda-se de novo o programa supondo-se os parâmetros conhecidos.

Observa-se que com esta *priori*, a média e desvio padrão das proficiências estimadas de um grupo não coincidem com a média e desvio padrão do grupo estimado pelo programa na fase de calibração, mesmo após a transformação linear inversa destes. Utilizando-se a distribuição empírica de cada grupo, obtida pelo programa como a distribuição *a priori* dos indivíduos do grupo, obtém-se que a média das proficiências estimadas é igual à média do grupo estimado na fase de calibração, mas o desvio padrão é diferente.

No primeiro procedimento (uso da priori N(0,1)), em uma prova composta somente com itens do SAEB, para estimar as proficiências dos alunos basta o conhecimento dos parâmetros dos itens. No segundo procedimento, é necessário estimar a distribuição de proficiências *a priori* da nova população de alunos. Esta estimação está descrita na seção seguinte.

5. Equalização em testes que utilizam itens do SAEB

A maneira mais comum de se fazer a equalização entre um teste novo com itens novos e itens comuns com o SAEB é calibrar separadamente os itens para o novo teste em relação à população testada e depois aplicar a propriedade de invariância da escala, utilizando algum método para obtenção da transformação linear que levará à escala SAEB. O problema deste método é que o erro da equalização costuma ser grande. Chama-se a atenção de que como no modelo de 3 parâmetros, o parâmetro "c" não é transformado, os parâmetros "c" conhecidos dos itens comuns devem ser fixados.

Como o SAEB tem utilizado o primeiro procedimento descrito na seção 4, para obtenção das proficiências na escala SAEB, **mesmo nesta forma de equa-**

lização, é necessário estimar estas rodando o programa com os parâmetros na escala SAEB e a priori N(0,1).

Se o SAEB tivesse utilizado o segundo procedimento (proficiências estimadas utilizando-se as distribuições empíricas), bastaria aplicar a mesma transformação linear às proficiências estimadas na calibração independente. O mais correto aqui seria também estimar estas proficiências na calibração independente (1 grupo) com a distribuição empírica estimada e não com a distribuição *default* N(0,1).

O **novo procedimento proposto aqui** é baseado na equalização realizada no SAEB e na TRI para grupos múltiplos. Este procedimento utiliza a amostra (ou subamostra) de alunos da série considerada do SAEB, do qual os itens comuns provêm. Utiliza também todos **os dados individuais de resposta a todos os itens da amostra** e **todos os parâmetros de todos os itens da série considerada são fixados**. Esta amostra (subamostra) do SAEB deve ser utilizada como o grupo de referência. Os parâmetros dos itens do SAEB devem ser transformados para a escala do grupo de referência (média 0 e desvio padrão 1) como explicado no item anterior.

Após a calibração dos novos itens, fixando-se os parâmetros transformados, volta-se à escala SAEB e se estimam as proficiências dos alunos com o primeiro procedimento descrito acima.

Recomenda-se este procedimento para a equalização com a escala SAEB de avaliações estaduais ou municipais que utilizam itens do SAEB.

Este procedimento pode ser utilizado para estimar simultaneamente a distribuição empírica da nova população de alunos quando todos os itens utilizados são do SAEB.

Observa-se que um teste pode conter itens de vários SAEBs, mas no seu planejamento deveria conter uma quantidade razoável de itens da mesma série e ano do SAEB, para que os alunos desta série e ano SAEB sirvam como grupo de referência. Os parâmetros dos outros itens SAEB devem ser transformados da mesma maneira.

6. Interpretação da escala

As escalas obtidas ordenam os desempenhos dos alunos (do nível mais baixo ao mais alto) em um *continuum*. Interpretar a escala significa escolher alguns pontos ou níveis da escala e descrever os conhecimentos e habilidades que os alunos demonstraram possuir quando situados em torno desses pontos.

A metodologia para interpretação das escalas inclui dois procedimentos principais: identificação de itens âncoras e a apresentação desses itens a um painel de especialistas.

A metodologia de escolha dos níveis âncora e a identificação dos itens âncora, utilizadas no SAEB 99 e 2001, foram diferentes das utilizadas nos SAEBs 95 e 97.

Nos SAEBs 95 e 97, utilizou-se a metodologia descrita em Beaton e Allen (1992). Nesta metodologia, itens âncora são itens que caracterizam os pontos ou

níveis das escalas, no sentido de que a grande maioria dos alunos situados em cada um dos níveis acerta o item, enquanto menos da metade dos alunos situados no nível imediatamente inferior também o acerta. Para o item ser considerado âncora, por exemplo, no nível 250, ele deve satisfazer ao seguinte critério:

· que 65% ou mais dos respondentes em torno do nível 250 acertem o item, que menos de 50% dos alunos posicionados no nível anterior acertem o item e que a diferença entre os percentuais dos que acertaram seja maior que 30%.

A seleção dos itens âncora para os demais níveis das escalas obedece ao mesmo critério.

Os problemas observados com este procedimento foram:

i) para se ter muitos itens âncora, os níveis selecionados têm que ser bem espaçados;

ii) para se ter mais itens para auxiliar a interpretação é necessário ampliar o conceito de nível âncora para quase-âncora;

iii) dificuldade dos membros dos painéis de especialistas de utilizarem o conceito de discriminação entre níveis, só usando praticamente itens considerados dominados pelos alunos em torno do nível;

iv) alguns descritores cobertos por alguns itens não aparecerem na descrição, pois, mesmo sendo bons itens, não foram classificados como itens âncora.

A solução para este último problema foi identificar estes itens e verificar em que níveis os alunos estariam dominando estes assuntos, através do conhecimento dos percentuais de acerto em diversas partes da escala.

A modificação introduzida em 1999 e mantida em 2001 foi selecionar os pontos da escala de 100 a 425, em intervalos de 25, o que inclui o ponto 250, média arbitrada da distribuição de proficiências da 8ª série nas disciplinas de língua portuguesa e matemática.

Para cada um destes pontos ou níveis, foi considerado um intervalo de comprimento 25 e centrado no ponto.

Para cada item foi estimado o percentual de acerto dos alunos em cada nível, calculando-se o percentual de acerto dos alunos com proficiência no intervalo que contém o nível.

Nesta nova abordagem, um item é dito "âncora" em um nível se:

i) O número de alunos no nível que respondeu ao item é maior que 50.

ii) O percentual de acerto do item nos níveis anteriores é menor que 65%.

iii) O percentual de acerto do item no nível considerado e nos níveis acima é maior que 65%.

iv) O ajuste da curva é bom.

Desta maneira, todo item considerado bom é utilizado, pois vai ser "âncora" em algum nível, a não ser que seja muito difícil, e dá flexibilidade para o painel de especialistas escolher alguns níveis para resumir a interpretação da escala.

De acordo com o segundo procedimento, a análise dos itens âncora, visando buscar a explicação do significado das respostas dadas pelos alunos, foi realizada por Painéis de Especialistas. No Painel, os especialistas recebem os itens separados por níveis, cada item com seu enunciado completo, suas estatísticas clássicas e um gráfico com sua curva característica proveniente da calibração da TRI, os percentuais esperados de acerto e os percentuais empíricos de acerto. No SAEB/2001, acrescentou-se um gráfico com os percentuais empíricos por alternativa de resposta.

Uma maneira equivalente de descrever os níveis da escala é pensar nos intervalos ou faixas entre dois níveis interpretados, por exemplo, 175 e 200. A interpretação deste intervalo ou faixa é a interpretação no nível correspondente ao ponto inicial do intervalo ou faixa. Desta maneira, garante-se que o percentual estimado de acerto nos itens âncora em cada ponto do intervalo ou faixa é maior que 65%.

Recebido em: 26/11/2002
Aceito para publicação em: 01/08/2003

ABSTRACT

Use of Item Response Theory at the Brazilian Assessment System for Basic Education.

The paper introduces Item Response Theory (IRT) in its usual form of one group and for multiple groups and explains how IRT for multiple groups is being used in the Brazilian Assesment System for Basic education (SAEB) for item calibration and for obtaining a unique scale, by discipline, for student proficiencies of the 4th, 8th and 11th grade of Basic Education, for all SAEB´s since 1995. The paper presents the adopted equating procedure among grades and among years and introduces a new method for equating assessments that use some SAEB items with SAEB. Finally, the paper explains how SAEB scales are interpreted.

Keywords: *IRT - multiple groups – SAEB – equating - scale interpretation – proficiency - Bayes marginal estimation.*

RESUMEN

Utilización de a Teoría de Respuesta al Ítem en el Sistema Nacional de Evaluación de la Educación Básica

El artículo introduce la Teoría de la Respuesta al Ítem (TRI) en su forma usual de un único grupo y para grupos múltiples y explica como la TRI para grupos múltiples está siendo utilizada en el tema Nacional de Evaluación de la Educación Básica (SAEB) para la calibración de los ítems y para la obtención de una escala única, por disciplina, para las proficiencias de los alumnos de 4º y 8º curso de la Enseñanza Fundamental y para 3 curso de la Enseñanza Media, para los SAEB´s a partir de 1995. El artículo presenta el procedimiento adoptado de ecualización entre cursos y entre años e introduce un método nuevo para ecualizar evaluaciones que utilizan algunos ítems del SAEB con el SAEB. Finalmente el artículo explica como las escalas do SAEB son interpretadas.

Palabras clave: *TRI - grupos múltiples – SAEB – ecualización - interpretación de escala - proficiencia- estimación Bayesiana marginal.*

Referências Bibliográficas:

ANDRADE, D. F. ; KLEIN, R. Métodos estatísticos para avaliação educacional: Teoria da Resposta ao Item. *Boletim da ABE*, ano 15, n. 43, p. 21-28, 1999.

ANDRADE, D. F. ; TAVARES, H. R. ; VALLE, R. C. Teoria da Resposta ao Item: conceitos e aplicações. SINAPE, 14., 2000, Associação Brasileira de Estatística, 2000.

BAKER, F. B. *Item Response Theory*: parameter estimation techniques. New York : Marcel Dekker, 1992.

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

BOCK, R.D. ; ZIMOWSKI, M.F. Multiple group IRT. In: LINDEN, W.J. van der ; HAMBLETON, R.K. (Ed.). Handbook of modern Item Response Theory. Newbury Park. :[s.n.], 1996.

HAMBLETON, R.K.; SWAMINATHAN, H. ; ROGERS, H.J. *Fundamentals of Item Response Theory*. Newbury Park: Sage, c.1991. (Measurement methods for the social sciences series; v. 3).

FONTANIVE, N.S. ; KLEIN, R. Uma visão sobre o Sistema de Avaliação da Educação Básica do Brasil – SAEB. *Ensaio: Avaliação e Políticas Públicas em Educação,* Rio de Janeiro, v.8, n.29, p.409-442, out./dez., 2000.

LORD, F.M. ; NOVICK, M.R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

MISLEVY, R.J. Bayes modal estimation in Item Response Models. *Psychometrika,* 51, p.177-195, 1986.

MURAKI, E. A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, v.16, n.2, p.159-176,1992.

Sistema Nacional de Avaliação da Educação Básica: SAEB 95, Relatório Técnico (1996). Fundação Carlos Chagas/Fundação Cesgranrio, São Paulo/Rio de Janeiro. Enviado ao MEC/INEP/DAEB.

Sistema Nacional de Avaliação da Educação Básica: SAEB 97, Relatório Técnico (1998). Fundação Cesgranrio, Rio de Janeiro. Enviado ao MEC/INEP/DAEB.

Sistema Nacional de Avaliação da Educação Básica: SAEB 99, Relatório Técnico. (2000). Consórcio Fundação Cesgranrio/Fundação Carlos Chagas, Rio de Janeiro. Enviado ao MEC/INEP/DAEB.

Sistema Nacional de Avaliação da Educação Básica: SAEB 2001, Relatório Técnico. (2002). Consórcio Fundação Cesgranrio/Fundação Carlos Chagas, Rio de Janeiro.

ZIMOWSKI, M.F.; MURAKI, E.; MISLEVY, R. ; BOCK, R.D. *BILOG-MG*: Multiple-group IRT Analysis and Test Maintenance for Binary Items. Scientific Software International, 1996.

Correspondência: deptri@cesgranrio.org.br