Ensaio: Avaliação e Políticas Públicas em Educação
ISSN: 0104-4036
ensaio@cesgranrio.org.br
Fundação Cesgranrio
Brasil

Klein, Ruben
Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências
Ensaio: Avaliação e Políticas Públicas em Educação, vol. 21, núm. 78, enero-marzo, 2013, pp. 35-55
Fundação Cesgranrio
Rio de Janeiro, Brasil

Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências

Ruben Klein*

Resumo

Este artigo trata da importância do erro de medida tanto na Teoria Clássica dos Testes (TCT) como na Teoria de Resposta ao Item (TRI) e de alguns aspectos ligados a estimação das proficiências pelos modelos logísticos da TRI. O artigo mostra que somente no modelo de 3 parâmetros, a consistência no padrão de respostas afeta a estimação das proficiências. O artigo mostra também a importância de se ter testes adequados ao aluno ou a população.

Palavras-chave: Teoria Clássica dos Testes (TCT). Teoria de Resposta ao Item (TRI). Erro de medida. Proficiência. Estimador de Máxima Verossimilhança (EMV). Estimador Expected a Posteriori (EAP). Padrão de respostas.

Aspects of the Item Response Theory related to the estimation to the proficiencies

Abstract

This paper deals with the importance of the error measurement both in the Classical Test Theory and in the Item Response Theory (IRT) and about some aspects related to the estimation of the proficiencies under the IRT logistic models. The paper shows that only in the 3 parameter model, the consistency of the response pattern affects the estimation of the proficiencies. The paper also shows the importance of having adequate tests to the student or population.

Keywords: *Classical Test Theory. Item Response Theory (IRT). Measurement error. Proficiency. Maximum Likelihood Estimator (MLE). Expected a Posteriori Estimator (EAP). Response pattern.*

* Doutor em Matemática, Massachussetts Institute of Technology, EUA. Fundação Cesgranrio.
E-mail: ruben@cesgranrio.org.br

Algunos aspectos de la Teoría de Respuesta al Ítem (TRI) sobre la estimación de las proficiencias

Resumen

Este artículo analiza la importancia del error de medida en la teoría Clásica de los Tests (TCT) y en la Teoría de Respuesta al Ítem y de algunos aspectos relacionados con la estimación de las proficiencias por los modelos logísticos de la TRI. El artículo muestra que la consistencia en el patrón de respuestas afecta la estimación de las proficiencias sólo en el modelo de 3 parámetros. También presenta la importancia de poseer tests adecuados al alumno o a la población.

Palabras clave: *Teoría Clásica de los Tests (TCT). Teoría de Respuesta al Ítem (TRI). Error de medida. Proficiencia. Estimador de Máxima Verosimilitud (EMV). Estimador Expected a Posteriori (EAP). Patrón de respuestas.*

Introdução

Neste artigo, faremos uma curta revisão da Teoria Clássica de Testes (TCT) e da Teoria Resposta ao Item (TRI), procurando explicar alguns aspectos menos conhecidos e com vistas a uma melhor construção de testes e comparabilidade de resultados. Em particular, chamaremos a atenção para o erro de medida. Trataremos neste artigo somente de testes de múltipla escolha.

Um teste de múltipla escolha consiste em n itens ou questões, em geral, com 4 ou 5 alternativas de respostas, sendo uma, e somente uma, correta. O resultado é dado pelo número de acertos ou pelo percentual de acertos em uma escala própria de 0 (zero) a 100 (cem). Em geral, não há julgamento sobre o que uma determinada nota significa, a não ser que notas acima de 70 costumam ser consideradas adequadas.

No entanto é provável que, se o mesmo aluno repetir a mesma prova em outra ocasião ele obtenha um resultado diferente, pois o resultado da prova depende de seu bem-estar e ânimo no dia da prova, de seu nervosismo, da chance de acertar uma questão na qual tenha dúvidas, de alguma distração, etc.

Também não é costume associar-se a nota à dificuldade da prova e, por exemplo, perceber-se que uma média 50 em uma prova aplicada a um grupo de alunos pode significar mais que uma média 80 em outra prova aplicada a outro grupo de alunos. Da mesma maneira, uma nota 100 em uma prova difícil significa mais que uma nota 100 em uma prova fácil. Para resolver esse problema é necessário que os resultados dessas duas provas sejam colocados em uma mesma escala.

Para lidar com esses problemas, precisamos de modelos. Na seção 1, mostraremos o modelo da TCT, e na seção 2, o modelo da TRI, e enfatizaremos o erro da medida.

Teoria Clássica dos Testes (TCT)

A Teoria Clássica de Testes tem um modelo para a habilidade (escore verdadeiro) no qual o erro não depende da habilidade do aluno. A habilidade não é observada diretamente e é estimada pelo número de acertos. Uma boa referência para a TCT é Lord e Novick (1968).

O escore verdadeiro T é o escore esperado no teste, uma variável não observável. Pode ser pensado como a "média dos escores em sucessivas testagens do mesmo teste nas mesmas condições".

O escore observado X é o escore obtido por um estudante em um teste.

O erro do escore E é a diferença entre o escore verdadeiro e o escore observado. Logo
X = T + E
Hipóteses adicionais:
1) Média(E|T) = 0. A média do erro de escore é zero;
2) Cor(T,E) = 0. O escore verdadeiro e o erro de escore são não correlacionados..;
3) Cor(E_1,E_2) = 0. Os erros de escore são não correlacionados em duas aplicações repetidas;

Como consequência de (2), temos que a variância do escore observado é igual à soma das variâncias do escore verdadeiro e do erro do escore.

$$V(X) = V(T) + V(E)$$

A covariância entre X e T é:

$$Cov(X,T) = Cov(T+E,T) = V(T) + Cov(E,T) = V(T)$$

Um conceito importante é o da **confiabilidade do teste (*test reliability*)**, R, definido como sendo o quadrado da correlação entre o escore observado e o escore verdadeiro.

$$R = Cor^2(X,T) = \frac{Cov^2(X,T)}{V(X)V(T)} = \frac{V(T)}{V(X)}$$

Ou seja,

$$R = \frac{V(T)}{V(T)+V(E)} = 1 - \frac{V(E)}{V(X)}$$

As expressões acima indicam que a confiabilidade pode ser interpretada também como a proporção da variância do escore observado explicada pela variância do escore verdadeiro ou ainda por 1 menos a razão entre a variância do erro e a variância do escore observado.

Uma confiabilidade alta indica que o escore observado é uma boa estimativa do escore verdadeiro. A confiabilidade de um teste pode ser estimada através da construção e aplicação de duas formas de teste "paralelas". Diz-se que duas formas de teste *(X,T,E)* e *(X',T',E')* são paralelas se *T = T'* e:

$$X = T + E$$

$$X' = T + E'$$

$$Cov(E,E') = 0,\ Cov(X,E') = 0,\ Cov(X',E) = 0 \text{ e } V(E) = V(E')$$

Logo, *V(X) = V(X')* e as confiabilidades dos dois testes, *X* e *X'* são iguais.

A confiabilidade é estimada pela correlação, *Cor(X,X')*, entre as duas formas do teste, pois

$$Cor(X,X') = \frac{Cov(T+E,\ T+E')}{\sqrt{V(X)V(X')}} = \frac{V(T)}{V(X)} = R$$

Logo, temos também estimativas da variância do escore verdadeiro e do erro:

$$V(T) = V(X)\ Cor(X,X')$$

$$V(E) = V(X) - V(T) = V(X)\ (1 - cor(X,X'))$$

O erro padrão de medida (standard error measurement – SEM) é dado pela raiz quadrada de V(E), logo, se s(X) é o desvio padrão do escore observado X,

$$SEM = s(X)\sqrt{(1-R)}$$

A dificuldade dessa estimação é a necessidade de haver duas formas paralelas do teste aplicadas à mesma população.

Uma estimativa usual da confiabilidade é dada pelo estimador chamado de Alpha de Cronbach que no caso, onde todos os itens são de múltipla escolha, reduz-se a fórmula Kuder-Richardson 20 (KR20). Na realidade essas estimativas são limites inferiores para a confiabilidade.

Se $X = \sum_{i=1}^{n} X_i$, n é o número de itens do teste e p_i é a proporção de acerto no item i, temos:

$$\alpha = \frac{n}{n-1} * \left(1 - \frac{\sum_{i=1}^{n} V(X_i)}{V(X)}\right)$$

$$KR20 = \frac{n}{n-1} * \left(1 - \frac{\sum_{i=1}^{n} p_i * (1 - p_i)}{V(X)}\right)$$

A confiabilidade depende do tamanho do teste, de todos os itens do teste conjuntamente e cresce quando se acrescentam itens ao teste, Lord e Novick (1968). A confiabilidade depende também dos alunos testados. Ao se acrescentarem ou se substituirem itens, todo o processo de estimação tem de ser refeito.

Itens muito difíceis ou muito fáceis e itens com correlação entre acerto e número de acertos na prova (coeficiente ponto bisserial) baixo ou negativo acrescentam pouco à confiabilidade e devem ser substituídos.

Isso ocorre, pois, ao se acrescentar um item ao teste, para aumentar "razoavelmente" o Alpha de Cronbach ou o KR20, V(X) tem de crescer "bem" mais que a variância do item.

No caso de um item em que todos erram ou todos acertam, V(X) não se modifica e a variância do item é zero. Se o item for muito difícil ou fácil, V(X) vai crescer "pouco", alterando pouco o Alpha de Cronbach ou o KR20.

No caso do coeficiente ponto-bisserial baixo ou negativo, V(X) cresce "pouco" também, pois o acréscimo de um ponto de acerto não tendo a consistência esperada tende a diminuir ou acrescer pouco à variância.

O quadro 1 mostra os valores de SEM para alguns valores de R, desvio padrão de X igual a 1. Logo para uma confiabilidade de 0.90, o SEM é de cerca de 30% do desvio padrão de X, o que evidencia o problema do erro de medida e a necessidade de uma confiabilidade alta.

Quadro 1 - Alguns valores de SEM em função da confiabilidade

dp =1	
R	SEM
0.80	0.45
0.85	0.39
0.90	0.32
0.95	0.22

Fonte: Autor (2012).

Listam-se a seguir as principais limitações da Teoria Clássica dos Testes:

- As estatísticas que descrevem os itens de teste dependem do grupo de estudantes que fazem o teste.
- Os escores de teste que descrevem o desempenho dos alunos dependem dos itens apresentados aos alunos.
- A Teoria Clássica dos Testes só pode ser utilizada em situações nas quais todos os alunos fazem o mesmo teste (ou formas "paralelas" de teste).
- A Teoria Clássica dos Testes não fornece um modelo do desempenho de um aluno em um item.
- A maioria das aplicações da Teoria Clássica dos Testes assume incorretamente que os erros de medida têm a mesma variabilidade para todos os alunos.

Teoria de Resposta ao Item (TRI)

A introdução da TRI na avaliação brasileira trouxe muitas vantagens sobre o método tradicional de avaliação. Colocando os itens em uma mesma escala, a TRI permite estimar e comparar os resultados dos alunos, mesmo que eles respondam a itens diferentes.

A TRI é um conjunto de modelos estatísticos onde a probabilidade de resposta a um item é modelada como função da proficiência (habilidade) do aluno (variável não observável) e de parâmetros que expressam certas propriedades dos itens, com a propriedade de que quanto maior a proficiência do aluno, maior a probabilidade de ele acertar o item. Podemos citar as seguintes referências: Lord e Novick (1968), Hambleton, Swaminathan e Rogers (1991), Baker (1992), Hambleton (1993), Andrade e Klein (1999), Andrade, Tavares e Valle (2000) e Klein (2003).

Os modelos utilizados hoje, para itens de múltipla escolha (certo ou errado) baseiam-se na família de locação-escala da função logística (que é uma curva de crescimento com valores de 0 a 1).

O modelo logístico de 3 parâmetros é definido por:

$$P(x = 1 \mid \theta, a, b, c) = c + \frac{(1 - c)}{1 + \exp[-Da(\theta - b)]}$$

Onde

x é a resposta ao item, $\begin{cases} = 1 \text{ se correta} \\ = 0 \text{ se errada,} \end{cases}$

a onde a>0 é o parâmetro de inclinação do item, também chamado de parâmetro de discriminação do item

b é o parâmetro de dificuldade (ou de posição) do item, e
c onde 0<c<1 é o parâmetro da assíntota inferior do item, refletindo as chances de um estudante de proficiência muito baixa selecionar a opção de resposta correta.

O modelo logístico de 2 parâmetros é obtido com c = 0, e se além disso todos os a's são iguais, obtém-se o modelo de 1 parâmetro. Nesse caso, pode-se considerá-lo igual a 1. Este modelo é conhecido como o modelo de Rasch. Este modelo foi obtido através de algumas hipóteses, Rasch (1960), independentemente do desenvolvimento da TRI.

O parâmetro D, na fórmula, assume os valores D=1, para a métrica logística e D=1,7 para a métrica normal. O uso da métrica normal vem do fato de que os primeiros modelos utilizavam a função ogiva normal (função de distribuição cumulativa da distribuição normal) e de que a função de distribuição cumulativa normal com média 0 e desvio padrão 1 é bem aproximada pela função logística com parâmetro b=0 e parâmetro a=1,7, no sentido de que o máximo da diferença pontual entre as duas funções é sempre menor que 0,01.

Uma propriedade importante da TRI é que os parâmetros dos itens obtidos de grupos diferentes de alunos testados são invariantes, exceto pela escolha de origem e escala. Outra propriedade importante é que os parâmetros de dificuldade dos itens e as proficiências estão na mesma escala.

Na prática, os parâmetros dos itens de um teste têm que ser estimados e as proficiências também. Nessa estimação, os parâmetros de todos os itens estão na mesma escala. Com planejamento e outras testagens, pode-se colocar outros itens na mesma escala que os itens originais. Dessa maneira, pode-se construir um banco de itens na mesma escala.

Observa-se que mesmo supondo-se os parâmetros dos itens de um teste conhecido, a proficiência de um aluno, estimada a partir de suas respostas aos itens, tem um erro de medida que depende do valor da proficiência, o que como já dito não ocorre na Teoria Clássica dos Testes.

As proficiências estimadas de alunos que respondem a subconjuntos de itens diferentes de um mesmo banco de itens estão na mesma escala e podem ser comparadas. Como escrito em Hambleton (1993, p. 3), a proficiência de um aluno é a mesma não dependendo do particular subconjunto de itens utilizado, mas suas estimativas variam por causa do erro de medida e algumas estimativas serão melhores que outras por causa do uso de itens mais ou menos apropriados nos testes para ele.

Por isso, avaliações podem utilizar um grande número de itens na mesma série e utilizar um conjunto de provas diferentes com itens comuns entre elas, de modo que um aluno não responde a todos os itens, mas todos os itens são calibrados na mesma escala e as estimativas das proficiências também.

Geralmente 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, em avaliações de várias séries e em vários anos, 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.

Se utilizarmos o método de máxima verossimilhança para estimar a proficiência, supondo os parâmetros dos itens conhecidos, o erro padrão (standard error) é dado por

$$se(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

onde I(θ) é a informação do teste na proficiência θ. Quanto maior a informação em θ, menor é o erro padrão, logo maior é a precisão da estimativa.

Pode-se mostrar que se P(θ) é a probabilidade de acerto de um item em função da proficiência θ e P'(θ) sua derivada, então a informação do item é dada por:

$$I(\theta) = \frac{[P'(\theta)]^2}{P(\theta)(1 - P(\theta))}$$

Como P(θ) é uma curva de crescimento, P'(θ) converge para 0 (zero) quando θ converge para $\pm\infty$. Logo a informação I(θ) converge para 0 (zero) quando θ converge para $\pm\infty$ e o erro padrão se(θ) converge para $+\infty$.

A TRI apresenta a propriedade de que a informação do teste é a soma das informações dos itens (um item é um teste com somente um item). Como toda informação é positiva, a informação do teste aumenta com o número de itens e, conseqüentemente, o erro padrão da estimativa de q decresce. Ao contrário da TCT, o erro padrão da estimativa de proficiência depende da proficiência q e sabe-se também o que acontece com a informação do teste na substituição ou acréscimo de itens.

Itens muito fáceis ou muito difíceis para um aluno fornecem pouca informação para a estimativa de sua proficiência. Itens com parâmetro de dificuldade "b" próximos da proficiência do aluno fornecem mais informação. O aumento do parâmetro "c" diminui a informação do item, pois aumenta a chance do acerto casual.

Esse resultado permite construir testes adaptados a populações diferentes, desde que se tenha informações *a priori* sobre a população e se tenha um banco de itens calibrados. É desejável um teste que tenha informação alta em uma boa extensão.

É possível também construir testes adaptativos no computador (CAT – Computer Adaptive Testing) para um aluno que, além de garantir precisão na estimativa, pode também diminuir o tamanho do teste. Não discutiremos estratégias de seleção de itens em um CAT neste artigo. Veldkamp (2013), nesta revista trata desse assunto.

Hoje, há uma tendência cada vez maior de estimar proficiências dos alunos, o que não era objetivo no início.

Esses comentários indicam que, para conseguir esse objetivo, há necessidade de se construirem bons bancos de itens, com estratégias de manutenção e reposição de itens, de se construirem critérios de seleção de testes adequados a uma população e de se desenvolverem plataformas de CAT. Esse é o caminho do futuro.

Nesse artigo, vamos supor os parâmetros dos itens conhecidos e investigar aspectos relacionados às estimativas das proficiências.

Em primeiro lugar, vale a pena ressaltar as diferentes implicações nas estimativas das proficiências quando se usam os modelos de 1, 2 ou 3 parâmetros, supondo os parâmetros dos itens conhecidos. Para isso utilizaremos as equações de máxima verossimilhança para os diversos modelos.

Seja $\Psi(\mathrm{x}) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$ a função logística e

seja $u_j = \begin{cases} = 1 \text{ se resposta ao item } j \text{ é correta} \\ = 0 \text{ se resposta ao item } j \text{ é errada} \end{cases}$

A equação de máxima verossimilhança para o modelo de 1 parâmetro é dada por:

$$\sum_{j=1}^{m} \Psi(\mathrm{D}(\theta - b_j) = \sum_{j=1}^{m} u_j$$

Portanto, dado um conjunto de itens, a proficiência estimada só depende do número de acertos, não importando os parâmetros de dificuldade dos itens, o estimador de máxima verossimilhança (EMV) é uma função não linear do número de acertos e de uma função de todos os parâmetros (de dificuldade) dos itens.

A equação de máxima verossimilhança para o modelo de 2 parâmetros é dada por:

$$\sum_{j=1}^{m} a_j \Psi(\mathrm{D}a_j(\theta - b_j)) = \sum_{j=1}^{m} a_j u_j$$

No modelo de 2 parâmetros, dado um conjunto de itens, a proficiência estimada depende dos parâmetros de discriminação dos itens acertados, não importando os parâmetros de dificuldade dos itens e de uma função de todos os parâmetros dos itens. Acertos em itens de maiores parâmetros de discriminação implicam maiores estimativas de proficiências. Dessa maneira o EMV da proficiência depende do padrão de respostas e não somente do número de acertos como no caso do modelo de 1 parâmetro.

A equação de máxima verossimilhança para o modelo de 3 parâmetros é dada por:

$$\sum_{j=1}^{m} a_j \Psi(\mathrm{D}a_j(\theta - b_j)) = \sum_{j=1}^{m} \frac{\omega_j(\theta)}{D} u_j$$

onde

$$\omega_j(\theta) = Da_j \Psi(\mathrm{D}a_j(\theta - b_j) - \log(c_j)) = \frac{Da_j}{1 + c_j e^{-\mathrm{D}a_j(\theta - b_j)}}$$

No modelo de 3 parâmetros, dado um conjunto de itens, o EMV da proficiência depende dos 3 parâmetros dos itens acertados através da função "peso" $\omega_j(\theta)$. Pode-se ver que a função peso:

- ✓ cresce com o aumento do parâmetro de discriminação "a" e, portanto, aumenta a estimativa de proficiência,
- ✓ decresce com o aumento do parâmetro "c" e, portanto, diminui a estimativa da proficiência,
- ✓ decresce com o aumento do parâmetro "b" e, portanto, diminui a estimativa da proficiência.

Dessa maneira, importa a consistência do padrão de resposta. Para "a" e "c" iguais, um acerto em um item mais fácil vale mais.

Somente no modelo de 3 parâmetros, o acerto ao acaso, "chute", é penalizado. Nesse modelo, é melhor o aluno acertar os itens mais fáceis que os mais difíceis. Mas, mesmo assim, é melhor tentar o acerto do que deixar em branco, o que corresponde a um não acerto ou erro. Um acerto sempre aumenta a estimativa da proficiência.

Pode-se ver, nos três casos, que o EMV não existe quando todos os itens são acertados (EMV = $+\infty$) ou quando todos são errados (EMV = $-\infty$). Nesse caso o erro padrão seria $+\infty$. Isso faz sentido, pois aluno com tudo errado não tem sua proficiência bem estimada, pois não se mede o que ele sabe. Da mesma maneira, aluno que acerta tudo também não tem sua proficiência bem estimada, pois não se mede até onde suas habilidades vão.

As estimativas de proficiência dependem de todos os itens no teste. Se algum parâmetro "b" aumenta, a função a esquerda da equação decresce, implicando um aumento da estimativa da proficiência.

Para os modelos de 1 e 2 parâmetros, se somarmos uma quantidade $d > 0$ a todos os parâmetros b's, isto é, tornarmos os itens mais difíceis deslocando suas curvas para a direita, então o EMV também aumenta de d. Da mesma maneira, se subtrairmos d de todos os b's, os itens ficam mais fáceis e o EMV diminui de d.

O segundo aspecto importante nas estimativas das proficiências é o estimador a ser utilizado. Em geral, se utiliza um estimador bayesiano, para que todos os padrões de resposta tenham estimativas de proficiência, mesmo os todos certos e os todos errados. Dessa maneira, mesmo quem tem uma nota 0 (nenhum acerto) ou uma nota 100 (tudo certo) tem uma proficiência estimada, que depende de todos os itens do teste e da priori. É claro que os erros de estimativa nesses casos são grandes.

O estimador bayesiano mais utilizado é o EAP (expected a posteriori), que é a média da distribuição a posteriori e o erro padrão é o desvio padrão a posteriori. A média da distribuição a posteriori é uma média ponderada do EMV e da média da distribuição a priori. Quanto maior a variância da distribuição a priori, o EAP fica mais próximo da média da priori, e quanto menor a variância, mais próximo do EMV. O EAP tem um efeito de encolhimento (shrinking) do EMV em relação à média da priori.

O fato de o estimador EAP depender da média da priori cria um problema de comparação. Se usarmos prioris diferentes para alunos que façam a mesma prova, teremos estimativas de proficiência diferentes. Usando a mesma priori, garantimos a mesma estimativa.

Esse problema aparece em avaliações de grupos diferentes, como, por exemplo o SAEB (Sistema Nacional de Avaliação da Educação Básica), avaliação brasileira de Leitura e Matemática nos 5º e 9º anos do Ensino Fundamental (EF) de 9 anos (antigas 4ª e 8ª séries do EF de 8 anos) e na 3ª série do Ensino Médio (EM), realizada a cada dois anos, a partir de 1995. O grupo de referência é a 8ª série da avaliação de 1997. Utiliza-se a distribuição normal com média 0 e desvio padrão 1 como a distribuição a priori para o EAP de todos os grupos e anos de avaliação. Garante-se a comparação mencionada anteriormente, mas criam-se vieses nas estimativas, especialmente no 5º ano/4ª série do EF e na 3ª série do EM. Para garantir a comparabilidade ao longo dos anos, não se pode modificar essa distribuição a *priori*.

Por outro lado, se aplicarmos uma prova de 5º ano a um aluno do 7º ano, usamos a mesma priori. Caso contrário, teríamos que decidir que priori usar.

O método de equalização dos itens utilizado no SAEB (estimação dos parâmetros dos itens na escala SAEB), descrito em Klein (2003) não utiliza as estimativas das proficiências, de modo que a priori utilizada para o cálculo das proficiências não tem influência.

Ilustração

Nesta seção, ilustraremos alguns aspectos do exposto na seção anterior com alguns cálculos de proficiência utilizando o método EAP com a priori da distribuição normal com média zero e variância 1.

Trataremos de dois casos, modelo logístico de 1 parâmetro (ML1P, Rasch) e modelo logístico de 3 parâmetros (ML3P). Para facilitar comparações, manteremos o parâmetro de discriminação "a" fixo igual a 1 no modo normal ou equivalentemente a 1.7 no modo logístico. No modelo de 3 parâmetros, fixaremos o parâmetro "c" igual a 0.2.

Consideraremos 6 situações ML1P_27, com 27 itens, ML1P_31 com 31 itens, ML3P_27 com 27 itens, ML3P_31 com 31 itens, ML3P_53 e ML3P_61.

Em cada situação consideraremos 3 modelos, que diferem pela variação do parâmetro "b". A notação x(z)y significa a sequência de números de x a y com incremento z. Por exemplo, 1.0(0.1)1.5 é a sequência 1.0, 1.1, 1.2, 1.3, 1.4, 1.5.

O quadro ao lado lista os 6 casos com os 3 modelos.

Quadro 2 - Descrição dos modelos

Situação	Modelo 1	Modelo 2	Modelo 3
1. ML1P_27	b = -2.0(0.1)0.6	b = -1.3(0.1)1.3	b = -0.6(0.1)2.0
2. ML1P_31	b = -2.0(0.1)1.0	b = -1.3(0.1)1.7	b = -0.6(0.1)2.4
3. ML3P_27	b = -2.0(0.1)0.6	b = -1.3(0.1)1.3	b = -0.6(0.1)2.0
4. ML3P_31	b = -2.0(0.1)1.0	b = -1.3(0.1)1.7	b = -0.6(0.1)2.4
5. ML3P_53	b = -2.0(0.05)0.6	b = -1.3(0.05)1.3	b = -0.6(0.05)2.0
6. ML3P_61	b = -2.0(0.05)1.0	b = -1.3(0.05)1.7	b = -0.6(0.05)2.4

Fonte: Autor (2012).

Nas situações 1, 3 e 5, os itens com parâmetros "b" variando de -0.6 a 0.6 são comuns aos 3 modelos e nas situações 2, 4 e 6, isso ocorre com os itens com parâmetros "b" variando de -0.6 a 1.0.

Desse modo, em cada situação, todas as curvas são paralelas e a dificuldade do item fica determinada pelo parâmetro "b". A dificuldade aumenta do modelo 1 para o modelo 3.

Observa-se que nas situações 2 e 4 acrescentaram-se em cada modelo 4 itens em relação às situações 1 e 3. Na situação 6, acrescentaram-se 8 itens em relação à situação 5.

As Tabelas de 1 a 6 mostram os resultados do cálculo das proficiências em alguns casos para as situações de 1 a 6. Na primeira parte das Tabelas, as proficiências estão na escala original (média 0 (zero), desvio padrão 1 (um)) e na segunda parte foi feita a seguinte transformação para média 250, desvio padrão 50:

$$Proft = 50*Prof + 250$$
$$dpt = 50*dp$$

Essa transformação é análoga à da escala SAEB, para melhor compreensão.

A primeira linha das Tabelas mostra as proficiências calculadas para o caso de o aluno ter errado todos os itens. Lembramos que, nesse caso, não existe o estimador de máxima verossimilhança (EMV). Pode-se ver que a estimativa EAP da proficiência aumenta do modelo 1 para o modelo 3, e que o desvio padrão a posteriori dp é grande, cerca de meio desvio padrão.

A segunda linha apresenta o caso do aluno que acertou tudo. Novamente o EMV não existe, a estimativa EAP da proficiência aumenta do modelo 1 para o modelo 3 e o desvio padrão é grande.

Observa-se que as estimativas das proficiências de "tudo errado" na Tabela 2 são ligeiramente menores que na Tabela 1, enquanto o aumento de acerto em 4 itens mais difíceis aumentam as estimativas das proficiências do "tudo certo".

As linhas 3 e 4 apresentam as estimativas das proficiências quando somente o 1º item (o mais fácil) foi acertado e quando somente o último item (o mais difícil) foi acertado. Pode-se ver que nas situações ML1P_27 e ML1P_31, como predito na seção 2, a proficiência estimada é a mesma em cada modelo e que também aumenta do modelo 1 para o modelo 3.

Na Tabela 1, nas linhas 5, 6 e 7, apresentamos as proficiências estimadas para os alunos que acertaram os 13 itens com parâmetro "b", variando de -0.6 a 0.6, e que acertaram todos os itens mais fáceis em seus modelos e não acertaram nenhum item mais difícil. Assim no modelo 1, o aluno acertou 27 itens, no modelo 2, 20 itens e no modelo 3 somente esses 13 itens. Observa-se que no modelo 1 esses 13 itens são os mais difíceis. Acrescentaram-se 4 itens mais difíceis para se ver como a proficiência é modificada no modelo 1 (supondo-se que o aluno errou esses 4 itens).

No modelo 1, ML1P_27, o aluno acerta todos os itens e o dp é grande. No modelo 1, ML1P_31, acrescentados 4 itens mais difíceis que ele erra, a proficiência estimada cai muito, cerca de 0.8 desvio padrão da escala. O dp também cai. Nos modelos 2 e 3, ML1P_27, onde aumenta o número de erros dos alunos, as proficiências estimadas caem, e o dp, também. Pode-se ver que as proficiências estimadas mudam pouco para os respectivos modelos 2 e 3, ML1P_31, especialmente no modelo 3.

Os resultados parecem indicar que o "ideal" para estimar a proficiência de um aluno é um teste no qual ele acerte metade e erre metade. Isso indica que se aumentássemos o número de itens mais difíceis no modelo 2, a proficiência e o dp cairiam mais. Esse é um forte argumento para os testes adaptativos computadorizados

Finalmente na linha 8, mostramos no modelos 3, ML1P_27 e ML1P_31, que, como o previsto, a proficiência estimada não se altera, se em vez dos 13 itens mais fáceis pegarmos os 13 itens mais difíceis do modelo 3, ML1P_27, e supusermos que o aluno errou todos os outros (um padrão não consistente e improvável).

Tabela 1 - ML1P_27

							Escala Transformada					
	modelo 1		modelo 2		modelo 3		modelo 1		modelo 2		modelo 3	
	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	sp
tudo errado	-2.818	0.455	-2.292	0.504	-1.790	0.554	109.1	22.8	135.4	25.2	160.5	27.7
tudo certo	1.790	0.554	2.292	0.504	2.818	0.455	339.5	27.7	364.6	25.2	390.9	22.8
Só 1º certo	-2.506	0.402	-1.927	0.427	-1.363	0.455	124.7	20.1	153.7	21.3	181.9	22.8
Só 27º certo	-2.506	0.402	-1.927	0.427	-1.363	0.455	124.7	20.1	153.7	21.3	181.9	22.8
Itens 1 a 27	1.790	0.554					339.5	27.7				
Itens 1 a 20			0.792	0.283					289.6	14.1		
Itens 1 a 13					0.595	0.259					279.7	13.0
itens 15 a 27					0.595	0.259					279.7	13.0

Fonte: Autor (2012).

Tabela 2 - ML1P_31

							Escala Transformada					
	modelo 1		modelo 2		modelo 3		modelo 1		modelo 2		modelo 3	
	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp
tudo errado	-2.822	0.455	-2.295	0.503	-1.793	0.553	108.9	22.7	135.3	25.2	160.4	27.7
tudo certo	2.076	0.524	2.594	0.476	3.124	0.420	353.8	26.2	379.7	23.8	406.2	21.0
Só 1º certo	-2.510	0.402	-1.930	0.426	-1.366	0.455	124.5	20.1	153.5	21.3	181.7	22.7
Só 31º certo	-2.510	0.402	-1.930	0.426	-1.366	0.455	124.5	20.1	153.5	21.3	181.7	22.7
Itens 1 a 27	0.986	0.323					299.3	16.2				
Itens 1 a 20			0.685	0.257					284.3	12.9		
Itens 1 a 13					0.568	0.254					278.4	12.7
itens 15 a 27					0.568	0.254					278.4	12.7

Fonte: Autor (2012).

Nas Tabelas 3 e 4, passamos para o modelo logístico TRI de 3 parâmetros, fixando c= 0.2. Mantemos "c" em um valor fixo para facilitar as comparações.

Comparando com as Tabelas 1 e 2, vemos que as proficiências estimadas no caso de "tudo errado" não se alteram em nenhum caso, pois o parâmetro "c" não aparece no lado esquerdo da equação de verossimilhança.

Mas como o parâmetro "c" diminui a informação, as proficiências estimadas no caso "tudo certo" diminuem.

Nas linhas 3 e 4 vemos nas duas tabelas, que como o previsto, em todos os modelos, a proficiência estimada é maior para quem acerta somente o item mais fácil do que para quem acerta somente o item mais difícil.

Na linha 8 das Tabelas 3 e 4, vemos as proficiências estimadas para o modelos 3 para os itens do 15 a 27 (os mais difíceis no modelo 3, ML3P_27), supondo que os alunos só acertaram esses itens, podemos comparar com a linha 7 onde estão as proficiências estimadas para os 13 itens mais fáceis. A queda da proficiência é dramática, cerca de 2 desvios padrões, de cerca de 270 para cerca de 174 no modelo 3. Observa-se que, nesses modelos, as proficiências estimadas para "tudo errado" são cerca de 160.5.

Aqui observa-se uma enorme diferença de resultados quando se utiliza o modelo de 3 parâmetros, em vez do de 1 parâmetro.

As linhas 5, 6 e 7 apresentam os casos equivalentes do modelo logístico de 1 parâmetro para o de 3 parâmetros e o comportamento observado é semelhante. Para maior precisão da estimativa da proficiência, é importante que existam itens "difíceis" para ele, nos quais o aluno erre.

Tabela 3 - ML3P_27

	Escala Original						Escala Transformada					
	modelo 1		modelo 2		modelo 3		modelo 1		modelo 2		modelo 3	
	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp
tudo errado	-2.818	0.455	-2.292	0.504	-1.790	0.554	109.1	22.8	135.4	25.2	160.5	27.7
tudo certo	1.671	0.576	2.156	0.528	2.668	0.484	333.6	28.8	357.8	26.4	383.4	24.2
Só 1º certo	-2.624	0.450	-2.086	0.496	-1.579	0.551	118.8	22.5	145.7	24.8	171.0	27.5
Só 27º certo	-2.812	0.456	-2.286	0.505	-1.785	0.554	109.4	22.8	135.7	25.2	160.7	27.7
Itens 1 a 27	1.671	0.576					333.6	28.8				
Itens 1 a 20			0.644	0.312					282.2	15.6		
Itens 1 a 13					0.424	0.307					271.2	15.3
Itens 15 a 27					-1.510	0.606					174.5	30.3

Fonte: Autor (2012).

Tabela 4 - ML3P_31

	Escala Original						Escala Transformada					
	modelo 1		modelo 2		modelo 3		modelo 1		modelo 2		modelo 3	
	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp
tudo errado	-2.822	0.455	-2.295	0.503	-1.793	0.553	108.9	22.7	135.3	25.2	160.4	27.7
tudo certo	1.948	0.547	2.450	0.502	2.970	0.454	347.4	27.3	372.5	25.1	398.5	22.7
Só 1º certo	-2.628	0.449	-2.090	0.496	-1.583	0.550	118.6	22.5	145.5	24.8	170.9	27.5
Só 31º certo	-2.819	0.455	-2.292	0.504	-1.790	0.553	109.1	22.8	135.4	25.2	160.5	27.7
Itens 1 a 27	0.843	0.347					292.2	17.4				
Itens 1 a 20			0.536	0.291					276.8	14.5		
Itens 1 a 13					0.395	0.303					269.8	15.1
itens 15 a 27					-1.515	0.604					174.2	30.2

Fonte: Autor (2012).

A seguir, somente para o modelo de 3 parâmetros, aumentamos o número de itens, reduzindo o intervalo de variação entre os b´s de 0.1 para a metade 0.05. Mantivemos os b´s iniciais e finais nos 3 modelos. Assim, no 1º caso, ficamos com 53 itens em vez de 27, Tabela 5, e, no 2º caso, 61 itens, ampliando em 8 itens no final, Tabela 6. Os 13 itens comuns passam a 25 itens comuns nos 3 modelos. Os itens consistentes passam a 53 no modelo 1, a 39 no modelo 2 e a 25 no modelo 3.

Observa-se que o erro da estimativa diminui como o esperado em todos os casos, e a estimativa de "tudo errado" cai e a de "tudo certo" aumenta. Fora o caso do tudo certo (53 itens) no modelo 1, na Tabela 5, as estimativas de proficiência nesses casos "consistentes" (com 53 itens ou 61 itens) são praticamente iguais aos das Tabelas 3 e 4, embora com erros menores.

Tabela 5 - ML3P_53

	Escala Original						Escala Transformada					
	modelo 1		modelo 2		modelo 3		modelo 1		modelo 2		modelo 3	
	Prof	Dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp
tudo errado	-3.155	0.406	-2.617	0.465	-2.091	0.514	92.3	20.3	119.2	23.2	145.4	25.7
tudo certo	1.980	0.530	2.495	0.482	3.029	0.429	349.0	26.5	374.8	24.1	401.5	21.4
Só 1º certo	-3.034	0.410	-2.487	0.464	-1.959	0.516	98.3	20.5	125.6	23.2	152.0	25.8
Só 53º certo	-3.152	0.406	-2.614	0.465	-2.089	0.515	92.4	20.3	119.3	23.2	145.6	25.7
Itens 1 a 53	1.980	0.530					349.0	26.5				
Itens 1 a 39			0.648	0.214					282.4	10.7		
Itens 1 a 25					0.428	0.231					271.4	11.5
Itens 29 a 53					-1.828	0.553					158.6	27.6

Fonte: Autor (2012).

Tabela 6 - ML3P_61

	Escala Original						Escala Transformada					
	modelo 1		modelo 2		modelo 3		modelo 1		modelo 2		modelo 3	
	Prof	Dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp	Prof	dp
tudo errado	-3.158	0.405	-2.620	0.464	-2.094	0.514	92.1	20.3	119.0	23.2	145.3	25.7
tudo certo	2.274	0.502	2.803	0.453	3.331	0.385	363.7	25.1	390.2	22.6	416.5	19.3
Só 1º certo	-3.038	0.410	-2.491	0.464	-1.963	0.515	98.1	20.5	125.5	23.2	151.9	25.8
Só 61º certo	-3.157	0.406	-2.618	0.464	-2.093	0.514	92.2	20.3	119.1	23.2	145.4	25.7
Itens 1 a 53	0.860	0.257					293.0	12.8	250.0			
Itens 1 a 39			0.548	0.204					277.4	10.2		
Itens 1 a 25					0.395	0.230					269.7	11.5
Itens 29 a 53					-1.834	0.552					158.3	27.6

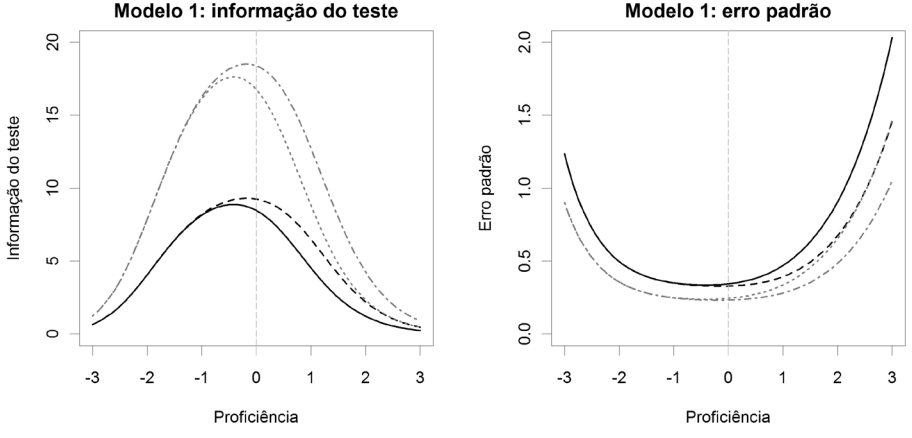
Fonte: Autor (2012).

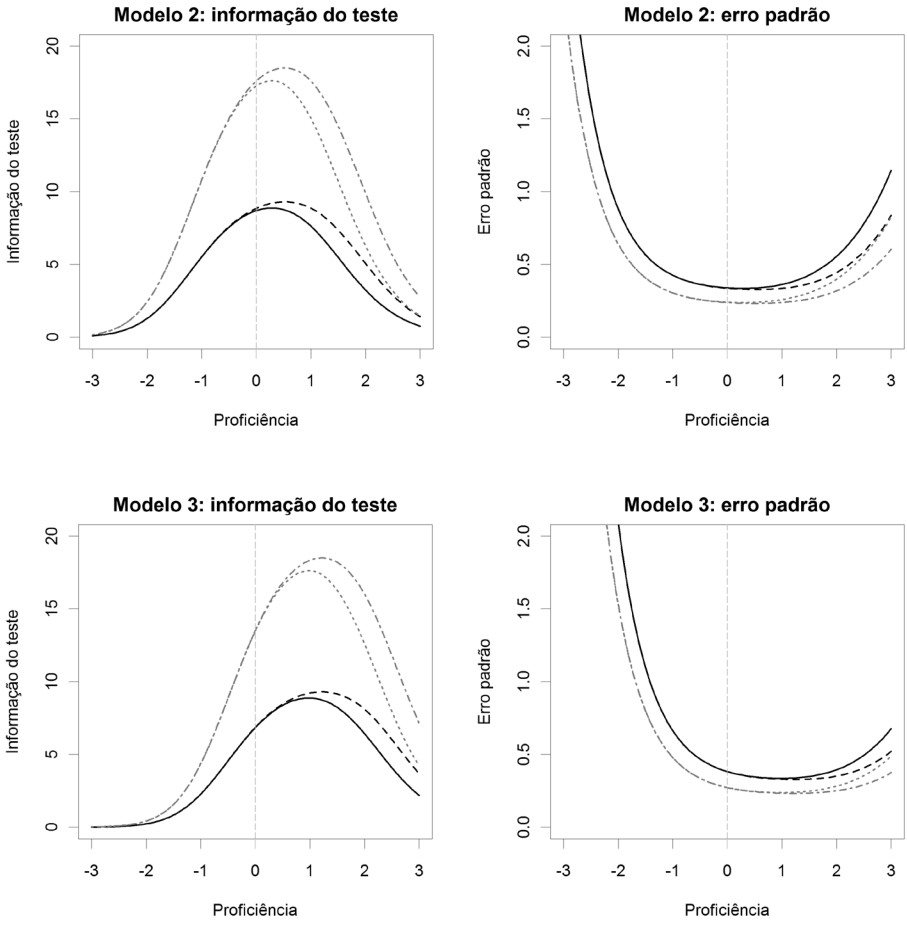
A Figura 1 mostra as curvas de informação e dos erros padrão para os 3 modelos para as 4 situações ML3P_27, ML3P_31, ML3P_53, ML3P_61, correspondentes às Tabelas de 3 a 6.

Pode-se ver que os modelos com 31 e 61 itens têm mais informação que os modelos com 27 e 53 itens, respectivamente. As razões entre as curvas de informação para os modelos das Tabelas 5 e 6 para os modelos correspondentes das Tabelas 3 e 4 variam de 1.92 a 1.99, indicando que têm forma semelhante.

Essa análise mostra que a precisão aumenta com o número de itens, mas que as estimativas dependem muito da forma da curva de informação.

Figura 1 - Curvas de informação e dos erros padrão para os 3 modelos com o modelo de 3 parâmetros





Fonte: Autor (2012).

Conclusão

A TRI é um grande avanço em relação à Teoria Clássica dos Testes, pois permite que, em qualquer prova com itens calibrados, as estimativas de proficiência de um aluno sejam colocadas em uma mesma escala, mas a precisão da estimativa depende da adequabilidade da prova ao aluno e da proficiência do aluno. Como visto, para maior precisão, é preciso que o aluno acerte e erre alguns itens.

A TRI permite um maior número de questões apresentadas à população e que alunos façam provas diferentes.

O modelo de 3 parâmetros (ML3P)apresenta uma propriedade nova que os modelos de 1 parâmetro (ML1P, Rasch) e dois parâmetros (ML2P) não possuem, que é a de penalizar respostas não consistentes aos itens. Essa diferença, como mostrado, tem grande impacto na estimação das proficiências.

Quando o objetivo é estimar proficiência de aluno e fazer comparações ao longo das diversas edições de uma "Avaliação", ou quando há diferentes cadernos (provas), os cadernos precisam ser "equivalentes", no sentido de que os mínimos e máximos possíveis e as curvas de informação sejam semelhantes.

A estimação das proficiências dos alunos começa também a ser comum nas avaliações estaduais e municipais e mesmo na Prova Brasil com a identificação do aluno. As formas (cadernos) de teste devem ser "equivalentes".

Os exemplos na seção 4 ilustram essas afirmações. Para se estimarem com mais precisão as proficiências dos alunos, são necessárias provas adequadas aos alunos e uma quantidade razoável de itens. Para manter melhor comparabilidade entre diversas avaliações, é necessário que as formas das curvas de informação dos testes sejam semelhantes.

Então para permitir o uso melhor da TRI com a construção de cadernos equivalentes e adequadas à população é necessário um banco de itens calibrados com muitos itens e satisfazendo os critérios desejados. Um dos objetivos dessa construção de testes é minimizar os erros de medida para a população e garantir uma melhor comparabilidade.

Finalmente, como observado na seção 4, testes adaptativos computadorizados podem ser utilizados para construir testes adequados a um aluno e, assim, estimar melhor sua proficiência. Mas, mesmo nesse caso, é preciso tomar decisões criteriosas sobre o número mínimo de itens a ser apresentado ao aluno, sobre os critérios de seleção dos itens e sobre o método de estimativa da proficiência.

Referências

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, Helitan Ribeiro; VALLE, Raquel da Cunha. *Teoria da resposta ao Item:* conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

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

HAMBLETON, R. K. Principles and selected applications of item response theory. In: LIMA, Rober (Ed.). *Educational Measurement*. 3.ed. [S.l.]: American Council of Education, 1993.

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

KLEIN, R. Utilização da Teoria de Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (SAEB). *Ensaio*: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 11, n. 40, p. 283-296, 2003.

LORD, F. M.; Novick, M. R. *Statistical theories of mental test Score*. Reading, MA: Addison Wesley, 1968.

RASCH, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.

VELDKAMP, B. P. Bayesian computerized adaptive testing. *Ensaio*: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 21, n. 78, 2013. No prelo.

Recebido em: 19/12/2012
Aceito para publicação em: 18/04/2013