AVALIAÇÃO DA APRENDIZAGEM: UMA ANÁLISE DESCRITIVA SEGUNDO A TEORIA DE RESPOSTAS AO ITEM (TRI)

(LEARING EVOLUTION: A DESCRIPTIVE ANALYSIS USING THE ITEM RESPONSE THEORY)

WAGNER BANDEIRA ANDRIOLA

INTRODUÇÃO

Dentro do campo educacional, o conceito de avaliação tem gerado inúmeros debates e continua a ser fruto de muitas polêmicas (Andriola & Barreto, 1997). Analisando as inúmeras definições, Silva (1992) revela a existência de alguns pontos comuns:

a) o termo *avaliação* difere semanticamente de *medida*, porém a inclui como condição indispensável à sua objetividade e precisão;

b) a avaliação realiza-se em função de objetivos claramente definidos;

c) a avaliação é um processo amplo, influenciado por diferentes aspectos da situação educacional;

d) a avaliação fornece informações úteis para a tomada de decisões com relação a alunos;

e) a avaliação é uma atividade que deve ser contínua, sistemática e científica.

Apesar das polêmicas, pode-se postular uma definição de avaliação que possibilite um grau maior de concordância entre os estudiosos do assunto (Barreto, 1993). A que mais se aproxima dessa desejada concordância foi apresentada por Popham (1977). Diz ele:

> *... embora venham ocorrendo algumas diferenças de opinião através dos anos, a maioria dos educadores concebe a avaliação educacional como operação na qual a qualidade de uma iniciativa educacional é julgada. Em outras palavras, para a maior parte dos educadores o termo avaliação significa o julgamento do valor de uma iniciativa educacional* (pág. 11).

Referindo-se a sala de aula, Tyler (1981) afirma que:

> *... o processo de avaliação da aprendizagem consiste, essencialmente, em determinar em que medida os objetivos educacionais estão sendo realmente alcançados pelo programa planejado* (p. 98).

É claro que, para a avaliação do desempenho escolar, o professor pode e deve usar todas as informações ao seu dispor. Entretanto, nas avaliações envolvendo um número muito grande de alunos, existem limitações de ordem prática (por exemplo: restrições financeiras e dificuldades operacionais), obrigando a utilização de testes objetivos e possibilitando, ainda assim, uma medida válida e precisa do rendimento (Andriola & Barreto, 1997).

Tratando-se de uma atividade científica, que utiliza medidas objetivas (testes), a avaliação da aprendizagem pode estar fundamentada em modelos matemáticos. Os dois mais utilizados são conhecidos sob os nomes de "Teoria Clássica dos Testes (TCT)" e "Teoria de Resposta ao Item (TRI)".

CONSIDERAÇÕES SOBRE A TEORIA CLÁSSICA DOS TESTES

O nascimento formal da Teoria Clássica dos Testes (TCT) ocorreu com os primeiros trabalhos de *Sir* Charles Spearman no início do século, entre os anos 1904-1913, sobre a caracterização e avaliação da inteligência. O objetivo central era encontrar o modelo estatístico que fundamentasse as pontuações nos testes, permitindo, assim, a estimação dos erros que estão associados a todo processo de mensuração (Muñiz & Hambleton, 1992).

A TCT é um conjunto de teorias e técnicas, com graus distintos de formalização, que tem como núcleo a Teoria da Pontuação Verdadeira, que, por sua vez, se baseia no conceito de fidedignidade (Gavíria, 1995). Segundo tal conceito, a pontuação

e os escores forem diferentes, então, não se poderá ter confiança no instrumento porque não haverá consistência nas medidas.

Estatisticamente pode-se dizer que, quanto maior o erro, menor a fidedignidade do instrumento. A fidedignidade é pois, a correlação entre a variância do escore verdadeiro e a variância do escore obtido. Sua formulação matemática é expressa por:

$$rxx = \frac{Sv^2}{So^2}$$; onde:

rxx = coeficiente de fidedignidade;

Sv^2 = variância do escore verdadeiro;

So^2 = variância do escore obtido;

Diversos métodos foram desenvolvidos para estimar a quantidade de erro presente no instrumento de medida, dentre os quais destacam-se: teste-reteste (coeficiente de estabilidade), formas paralelas (coeficiente de consistência interna), bi-partição (coeficiente de consistência interna), Sperman-Brown (coeficiente de consistência interna), Kuder-Richardson - fórmulas KR20 e KR21 (coeficiente de consistência interna), *Alpha* de Cronbach (coeficiente de consistência interna);

b) Validade: diz respeito ao grau com que um teste, efetivamente, mede aquilo que se propõe. Há diversos tipos de validade, segundo os objetivos do teste. Por conseguinte, há diversas maneiras de se determinar a validade. Será discutida aqui a determinação através do coeficiente de correlação de Pearson , do erro-padrão da estimativa ou da análise fatorial.

Empregando-se o coeficiente de correlação de Pearson, obtém-se o grau de associação entre os escores de um teste (X) e os escores da variável critério (Y), cuja representação padrão é rxy. De acordo com Silva (1992), *se as duas variáveis referirem-se a mesma coisa, espera-se encontrar, entre seus resultados, uma correlação positiva, isto é, que os indivíduos bem sucedidos em uma sejam também bem sucedidos na outra, da mesma forma que os que obtiverem os escores mais baixos numa delas deverão também obter os mais baixos na outra* (p. 123).

O erro-padrão da estimativa estabelece os limites dentro dos quais se situa o escore verdadeiro. É definido por Vianna (1982), como *a diferença entre o escore verdadeiro do examinando no critério e o escore estimado para esse mesmo critério, e resulta de erros casuais e de diferenças entre o teste e o critério* (p. 176), sendo portanto, o desvio-padrão das diferenças. Seu cálculo é feito através da fórmula:

$$Syx = S_y \sqrt{1-(rxy)^2}$$; onde:

S_yx = erro-padrão da estimativa;

S_y = desvio-padrão do critério;

rxy = coeficiente de validade.

Já a análise fatorial corresponde a um conjunto de técnicas estatísticas, que analisa as inter-relações existentes entre um conjunto de variáveis visando *resumir as relações entre variáveis de forma concisa, mas acurada com o propósito de facilitar a sua conceituação* (Dias,1997, p. 2). A síntese de um grande número de variáveis em uma quantidade menor.

Bryman e Cramer (1992), destacam alguns objetivos da análise fatorial, dentre os quais:

a) avaliar a validade das questões componentes de um instrumento de medida, informando até que ponto elas estão medindo os mesmos conceitos ou variáveis;

b) reduzir um grande número de variáveis que podem ser explicadas por agrupamentos ou fatores (*data reduction*).

Os conceitos básicos da análise fatorial são: fator e carga fatorial. O termo fator refere-se a uma dimensão ou construto, constituído da junção de diversas afirmações entre um conjunto de variáveis da mesma natureza. Já o termo carga fatorial refere-se à correlação entre uma variável e seu fator correspondente (Dias, 1997).

PARÂMETROS MÉTRICOS DOS ITENS

O objetivo central da análise dos itens é obter informações sobre sua pertinência aos objetivos dos testes. Segundo Fernández (1990), *a análise de itens é o estudo daquelas propriedades que estão diretamente relacionadas com as propriedades do teste.*

A análise dos itens pode ser feita através de procedimentos racionais ou teóricos e também por procedimentos estatísticos, que é o que interessa no momento discutir. A análise estatística dos itens é realizada através do cálculo de índices que definem as propriedades de um item.

Os índices mais relevantes são:

a) dificuldade: refere-se a proporção de sujeitos que respondem corretamente ao item. Seu valor varia de 0 a 1, e quanto mais próximo de 1 mais fácil o item. O valor do índice está diretamente relacionado à média do teste.

empírica de um sujeito em um teste, consta de dois componentes aditivos: a verdadeira pontuação e o erro que, inevitavelmente, está associado à medição. Assim, a TCT busca estimar a quantidade do erro que afeta a pontuação obtida em um teste, ou seja, procura estabelecer a fidedignidade do mesmo (Fernández, 1990). Segundo Andriola (1998), os erros que afetam a pontuação dos sujeitos em um teste são aleatórios, podendo ocorrer por variadas fontes: devido ao respondente (por conta da fadiga, desinteresse, incompreensão da tarefa solicitada, etc.), devido ao ambiente externo (iluminação inadequada, barulho excessivo, etc.), devido ao instrumento de medida (instruções pouco claras, itens ambíguos, tempo insuficiente para a resolução dos itens, excessivo número de itens, etc.) e devido ao próprio ato de aplicação (pouco controle sobre os respondentes, variação de humor dos aplicadores, etc.).

De acordo com Gavíria (1995), os pressupostos da TCT são:

O erro é inerente ao processo de mensuração

A formulação matemática é dada por:

X = T + E; onde:

X = escore bruto do sujeito (escore empírico);

T = escore verdadeiro do sujeito (escore teórico);

E = erro aleatório.

A medida do desempenho de um sujeito (X), obtida através de um teste, é igual a soma da medida verdadeira do sujeito (T) e o erro de medida (E) cometido no processo de mensuração;

O erro de medida tem uma distribuição normal.

A sua formulação matemática é dada por:

E = N (0; 1); onde:

E = erro aleatório;

N = distribuição normal reduzida.

O erro (E) é uma variável aleatória cuja distribuição é normal (N), com média 0 (zero) e uma variância finita, ainda que desconhecida;

A correlação entre a medida verdadeira e o erro é igual a zero.

A sua formulação matemática é dada por:

RTE = 0; onde:

RTE = correlação (R) entre o escore verdadeiro do sujeito (T) e o erro aleatório cometido (E).

A suposição é que não há relação entre a pontuação do sujeito e o erro cometido no processo de mensuração, pois o erro depende de vários fatores, já enumerados;

A correlação entre os erros cometidos em duas medidas distintas é zero.

A sua formulação matemática é dada por:

RE1 E2 = 0; onde:

RE1 E2 = correlação (R) entre o erro cometido numa primeira aplicação (E1) e numa segunda aplicação (E2).

Dessa forma, supõe-se que não existe relação entre o erro cometido durante a aplicação de um teste no momento um (E1) e o erro cometido, durante a aplicação do mesmo teste, no momento dois (E2);

A correlação entre o erro cometido em duas aplicações independentes e distintas quanto à temporalidade é zero.

A sua formulação matemática é dada por:

RE1 T2 = 0; onde:

RE1 T2 = correlação (R) entre o erro cometido numa aplicação (E1) e o escore verdadeiro resultante de uma segunda aplicação (T2) distante temporalmente da primeira.

Assim, o erro de medida que se comete no momento um (E1) é independente da pontuação verdadeira obtida no momento dois (E2).

Os pressupostos da TCT são direcionados às medidas psicológicas e educacionais, que por sua vez resultam do uso de instrumentos de medida: os testes. Assim, é interessante ressaltar os conceitos dos parâmetros métricos relativos ao teste e aos itens, além das suas respectivas formulações matemáticas.

PARÂMETROS MÉTRICOS DO TESTE

Os parâmetros de um teste dizem respeito a duas características almejadas pelo elaborador, de forma a garantir a inquestionável qualidade e utilidade do instrumento de medida. São determinadas através dos coeficientes de fidedignidade e validade.

a) Fidedignidade: também conhecida como precisão, é o conceito mais importante da TCT. Segundo Vianna (1982), *a fidedignidade de um teste* (r_{xx})*refere-se à estabilidade dos seus resultados, ou seja, ao grau de consistência dos escores* (p.145). Se um teste é aplicado em um mesmo grupo, um grande número de vezes, conservando-se as mesmas condições,

Segundo Fernández (1990), uma séria limitação do índice de dificuldade é sua dependência em relação aos sujeitos que responderam ao teste. Se os sujeitos tiverem domínio do assunto tratado no teste, o item poderá ser fácil, do contrário, poderá ser difícil;

b) discriminação: o índice de discriminação é definido como a correlação entre as pontuações dos sujeitos no item e sua pontuação no teste (Fernández, 1990). Assim, um item é considerado discriminador quando diferencia os respondentes que conseguem sair-se melhor, daqueles que não conseguem resultados satisfatórios.

No entanto, cumpre salientar que, para a TCT, o objeto de análise é o teste. Assim, os parâmetros dos itens servem para demonstrar os pontos fortes e fracos do teste. De acordo com Gavíria (1995), os principais problemas da TCT são:

a) inexistência de invariância nas medições: uma mesma característica medida com distintos testes proporciona distintas pontuações. A escala de medida da variável varia de um teste a outro, logo, existe uma dependência entre a variável e o instrumento utilizado. Isto significa que é necessário igualar as pontuações obtidas em distintos testes através de um processo de equalização, fazendo-se transformações lineares dos escores brutos através dos métodos da normatização (percentil, T, escore z, etc.), tendo-se assim, uma compreensão mais realista a respeito da medida realizada pelo instrumento;

b) inexistência de invariância das propriedades métricas dos itens e, por conseguinte, do instrumento: o grau de dificuldade de um item e o seu poder discriminativo, referem-se a um determinado grupo de respondentes, examinado num certo momento e sob a influência de um conjunto específico de circunstâncias. Logo, as propriedades métricas do instrumento variam de acordo com a amostra de sujeitos utilizada para a sua determinação;

c) tratamento dos itens como "réplicas" uns dos outros: a TCT não proporciona uma análise do item baseada nele mesmo. A discriminação de um item depende da pontuação final do teste, portanto, um mesmo item pode obter distintos índices de discriminação em função do conjunto de itens ao qual se apresenta correlacionado. O mesmo se repete para o índice de dificuldade.

A representação do desempenho de sujeitos sobre bases tão oscilantes interfere na confiança dos resultados. Todas essas deficiências deram espaço para o surgimento de outra teoria, que incorporou alguns dos pressupostos da TCT e ainda a complementa. A nova teoria chama-se Teoria da Resposta ao Item (TRI), ainda conhecida pelos nomes de Teoria de Características Latentes e Teoria das Curvas Características.

CONSIDERAÇÕES SOBRE A TEORIA DE RESPOSTA AO ITEM (TRI)

A Teoria de Resposta ao Item (TRI) tem origem, segundo Fernández (1990), nos trabalhos pioneiros de Richardson em 1936; Lawley em 1943; Tucker em 1946; Lord por volta de 1952-53 e Birbaum por volta de 1957-58; ampliando-se, sobretudo, com as contribuições de Birbaum, Lord e Novick em 1968. A elaboração de programas de computadores que realizam análises baseadas na TRI incrementou ainda mais a utilização desta teoria, principalmente por facilitar os cálculos que seriam muito complexos se fossem realizados manualmente. Conquistou inúmeros adeptos nas áreas de psicologia e educação devido ao fato de *oferecer recursos mais flexíveis e eficazes na confecção, análise e apresentação dos resultados de prova que quaisquer outros recursos equivalentes, derivados a partir da teoria clássica* (Fletcher, 1994, p.21).

O centro desta teoria está

> *na relação que estabelece entre as características dos itens e as características operacionais da prova composta desses itens. Em sua essência, esta relação é invariável e permanente, não dependendo do número de itens da prova, do subconjunto de itens selecionados ou das habilidades das pessoas que participam da prova. Quaisquer outros resultados comparáveis advindos da teoria clássica servem apenas para testes hipotéticos de uma infinidade de itens aplicados a uma única população padrão* (Fletcher, 1994, p. 23).

O processo de medida feito através da TRI parte da suposição de que existe no sujeito um traço (uma característica individual que determina sua forma de responder ao teste), que possui uma relação probabilística com cada um dos itens utilizados. *O modelo de resposta que fundamenta a teoria da resposta ao item permite ao analista prever a probabilidade de acerto de uma pessoa com determinada habilidade representada pelo resultado da prova* (Fletcher, 1994, p. 23).

Os parâmetros de cada item não dependem, em absoluto, dos outros itens do teste, ao contrário, a pontuação do teste se faz em função das respostas do sujei-

to a cada item e dos parâmetros de cada item. Coloca-se assim, itens e pessoas na mesma escala de desempenho. Assim, pode-se afirmar se as pessoas são mais ou menos hábeis e se os itens são mais ou menos difíceis, na mesma escala de uma característica latente subjacente.

A relação entre os valores da variável que é medida pelo item e a probabilidade de acertá-lo é descrita por uma função matemática denominada *Curva Característica do Item (CCI)*. A CCI relaciona a probabilidade de êxito em um item com a capacidade medida pelo conjunto de itens que o teste contém (Gavíria, 1995). São as propriedades desta função que fazem a diferença entre a TRI e TCT.

São os seguintes os pressupostos da TRI, segundo Gavíria (1995):

a) unidimensionalidade: o grupo de itens deve medir uma mesma variável. Assim, ao elaborar-se um teste, define-se, previamente, os traços que se deseja avaliar, e se supõe que estes possuem as dimensões necessárias para descrever a característica estudada. Geralmente espera-se que um só traço seja necessário para explicar ou dar conta da atuação do indivíduo no teste. Para comprovação da unidimensionalidade utiliza-se, geralmente, a análise fatorial;

b) independência local dos itens e dos sujeitos: supõe-se que a resposta de um sujeito a um item não é influenciada pelas respostas fornecidas a outros itens. Segundo Hambleton, Swaminathan e Rogers (1991), se a unidimensionalidade é comprovada, disto deriva, matematicamente, a independência local entre os itens, dado que os dois conceitos são equivalentes.

A independência local entre sujeitos dá-se quando o rendimento de um sujeito que responde a um teste é independente do rendimento dos outros (Gavíria, 1995).

CURVA CARACTERÍSTICA DO ITEM (CCI)

A Curva Característica do Item (CCI) representa os parâmetros típicos do mesmo, fornecendo-lhe uma identidade própria. Apresenta-se em três modelos diferentes, nos quais podem ser observados parâmetros que caracterizam as qualidades técnicas dos itens, independentes da população investigada. Esses parâmetros representam a dificuldade, o poder discriminativo e a proporção de acertos casuais.

O tipo de CCI utilizado na TRI é do tipo "S", que representa o modelo idealizado. No eixo das abcissas está indicado o nível do sujeito na variável observada (traço latente) designada pela letra grega θ (theta), cujo valor pode variar de $-\infty$ a $+\infty$. A probabilidade de responder corretamente ao item, dado por P(θ), está indicada no eixo das ordenadas.

Os três parâmetros representados em uma CCI, correspondem aos seguintes aspectos:

1. discriminação (Parâmetro *a*): determinado pelo ponto máximo da inclinação da reta, seu valor é proporcional a esta pendência, e quanto maior a inclinação, maior será o índice de discriminação;

2. dificuldade (Parâmetro *b*): é um parâmetro de posição do item na escala de θ, informando em que parte da escala se encontra o ponto de inclinação máxima da curva. Também informa em qual parte da escala de θ se dá a transição desde uma maior probabilidade de responder incorretamente ao item, a uma maior probabilidade de respondê-lo corretamente (Gavíria, 1995);

3. acerto ao acaso (Parâmetro *c*): representa a probabilidade de acertar o item ao acaso, isto é, "quando não se tem certeza da resposta certa". Gavíria (1995), diz que os examinandos, na incerteza, buscam indícios indiretos que podem orientá-los na localização da opção correta.

DESCRIÇÃO DOS MODELOS DE CCI's

As informações contidas nas CCI's, a respeito dos parâmetros métricos dos itens, dependem do modelo teórico escolhido. O mais simples foi sugerido por G. Rasch em 1960 e recebeu o nome de *modelo logístico de um parâmetro*. Contém o presuposto de que a probabilidade de acerto de um item é influenciada pela sua dificuldade. Sua formulação matemática é:

$$P_1(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$, onde:

$P_1(\theta)$: probabilidade de acertar o item $_i$ para um determinado valor de θ;

θ: valor da variável medida;

b_i: índice de dificuldade do item i;

e: base dos logaritmos neperianos (2,72);

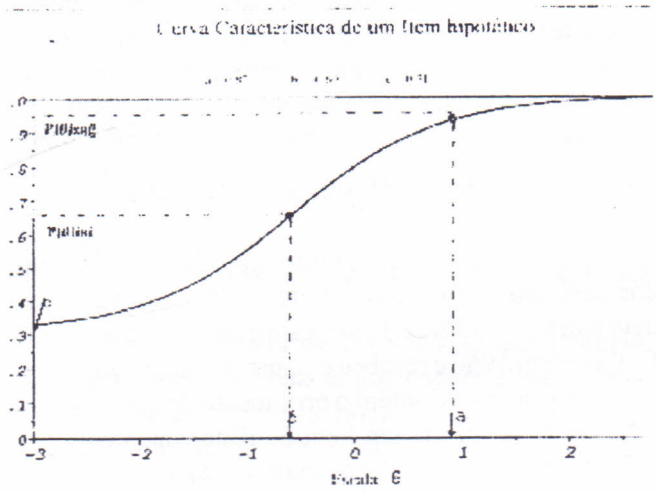
D: constante de valor 1,7 (com este valor, a função logística aproxima-se notavelmente da curva normal padronizada).

O segundo tipo, denominado *modelo logístico de dois parâmetros*, foi apresentado, por volta de 1968, por A. Birnbaum. Neste modelo, a probabilidade de acerto de um item é influenciada pela sua dificuldade e discriminação. Sua definição matemática é:

$$P_2(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$, onde $P_2(\theta)$, θ, b_i, e, D assumem o mesmo significado do modelo de um parâmetro. Sua diferença está no aparecimento, na sua formulação, do índice de discriminação do item (a_i).

Por último, *o modelo logístico de três parâmetros,* também desenvolvido a partir dos trabalhos de A. Birnbaum. Assume que a probabilidade de acerto de um item é influenciada pela sua dificuldade, discriminação e probabilidade de acerto ao acaso. Em termos matemáticos, o modelo é expresso por:

$$P_3(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$, onde $P_3(\theta)$, θ, a_i, b_i, e, D possuem o mesmo significado dos modelos aqui mencionados, e c_i indica a probabilidade de acerto ao acaso.



Para ilustrar o exposto, é mostrada a representação gráfica dos três parâmetros de um item hipotético, descritos através de sua CCI.

A CCI informa que:

- a dificuldade (parâmetro a) tem valor 0,87;
- a discriminação (parâmetro b) tem valor -0,60;
- a probabilidade de acerto ao acaso (parâmetro c) tem valor 0,31;
- a partir do valor de b a probabilidade de acerto ao item (P(θ)) aumenta proporcionalmente ao nível de aprendizagem (θ), assim, os itens não são considerados réplicas uns dos outros;
- o item consegue discriminação máxima para os valores de θ entre -0,60 (valor de b) e 0,87 (valor de a), ou seja, é útil para discriminar os sujeitos com nível de aprendizagem (θ) neste intervalo;
- independentemente dos níveis de aprendizagem dos respondentes (valores de θ) a probabilidade de acerto é a mesma, ou seja, para um respondente que tenha θ=2,0 a probabilidade de acerto ao item, dada por P(θ), está no intervalo entre 0,67 [P(θ)inf] e 0,95 [P(θ)sup], que é o mesmo intervalo para um sujeito que tenha θ=-2,0. Ou seja, os parâmetros métricos dos itens são invariantes, não dependendo dos níveis de aprendizagem (θ) dos respondentes.

OBJETIVO DO TRABALHO

Efetivar, através de estudo de casos, uma análise comparativa entre a TCT e a TRI considerando:

- o desempenho dos respondentes;
- os parâmetros métricos dos itens.

METODOLOGIA

Os dados que originaram este estudo são resultantes da Avaliação da Qualidade do Ensino das Escolas Públicas do Estado do Ceará, realizada em 1996, através da Fundação Cearense de Amparo à Pesquisa (FCPC), sob o financiamento da Secretaria Estadual de Educação e Cultura (SEDUC) do Estado do Ceará.

Amostra

Foi constituída por oito alunos da 8ª série, sendo quatro homens e quatro mulheres, com idades entre 14 e 21 anos. Quanto ao critério de escolha, foram selecionados quatro respondentes com desempenhos dois desvios padrões acima da média (representantes do grupo superior-GS) e quatro com desempenhos dois desvios padrões abaixo da média (grupo inferior-GI).

Instrumento

Utilizou-se um teste de português,[1] destinado aos alunos da 8ª série, composto por 25 questões fechadas seguidas, cada uma, de cinco opções de resposta. O tempo limite destinado à resolução foi 90 minutos.

[1] O teste foi elaborado com base na proposta curricular da SEDUC para a disciplina português (8a série). A avaliação dos processos cognitivos exigidos em cada uma das 25 questões foi baseada na Taxonomia de B. Bloom.

Procedimento

A coleta de dados efetivou-se através da aplicação coletiva do teste de português em 7 576 estudantes de escolas públicas do Estado do Ceará. Concomitantemente foram aplicados, nos citados alunos, um teste de matemática e um questionário sobre os professores dessas disciplinas.

Resultados

Como se trata de um trabalho que visa comparar a interpretação dos resultados obtidos pelos respondentes e dos parâmetros métricos dos itens, segundo TCT e a TRI, é apresentada, inicialmente, a matriz de respostas dos oito sujeitos às 25 questões do teste de português.

Quadro 2: Síntese do desempenho dos respondentes segundo a TCT

RESPONDENTE	DESEMPENHO (TCT)
A	3
B	3
C	3
D	18
E	18
F	18
G	2
H	18

Utilizando, inicialmente, a TCT para interpretar os resultados, pode-se descrever o desempenho dos respondentes:

Quadro 1 - Matriz de respostas dicotomizadas dos oito sujeitos

	QUESTÕES																								
S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
A	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
C	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
D	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1	1	0	1	0	1	1	1
E	1	1	0	0	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	0	1	1
F	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	1	0	1
G	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
H	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0	0	0	1	0

Legenda: S=sujeito.

A matriz fornece os acertos, representados pelo algarismo um (1), e os erros, pelo zero (0). As análises, a seguir, objetivam a comparação da TCT com a TRI.

ANÁLISE DO DESEMPENHO DOS RESPONDENTES

A síntese do desempenho[2] dos respondendentes é apresentada no quadro abaixo.

- D, E, F e H, obtiveram 18 acertos. São os representantes do GS;
- A, B, e C, obtiveram três acertos. Já o respondente G obteve apenas dois acertos. São os representantes do GI.

Após a descrição, pode-se hierarquizar os respondentes, segundo os desempenhos individuais. Assim, os sujeitos D, E, F e H encontram-se no mesmo nível de aprendizagem, porém, superior aos sujeitos A, B e C. Por sua vez, tais sujeitos encontram-se, também, no mesmo nível de aprendizagem, porém superior ao respondente G.

[2] Na TCT o desempenho é resultante do somatório dos acertos obtidos no conjunto de itens (teste).

De acordo com o procedimento adotado para a descrição do nível de aprendizagem, em cada que cada acerto equivale a um ponto e o somatório dos acertos revela o desempenho dos respondentes no teste, apresenta-se o primeiro problema da TCT, já mencionado anteriormente:

• *o tratamento dos itens como "réplicas" uns dos outros.*

O acerto aos itens obtém sempre um mesmo valor numérico. Ora, teoricamente, cada item exige um determinado nível de aprendizagem para que seja resolvido a contento, ou seja, exige processos cognitivos diferenciados. Se a afirmação é verdadeira, por que quando são acertados recebem o mesmo valor numérico? Ou em outros termos: é pertinente dar o mesmo valor numérico para um item que exige um processo cognitivo simples e, igualmente, para outro que exige um processo mais complexo?

Para continuar com o raciocínio, tome-se outro exemplo, desta vez considerando o nível de dificuldade do item. Na TCT a dificuldade de um item é determinada pelo percentual de acerto que o mesmo obtém. Assim:

• os itens 1 e 8 obtiveram 75% de acerto entre os oito respondentes. De acordo com esse valor percentual, pode-se afirmar que são itens fáceis.[3]

Porém, ao separar-se os respondentes em grupos, GS e GI, os resultados modificam-se:

• para o GI, o item 1 passa a ser de dificuldade média, já que 50% dos seus componentes o acertaram (sujeitos C e G). O item 8 continua a ser fácil, já que foi acertado por 75% dos sujeitos (A, B e C);

• para o GS, os itens 1 e 8 são fáceis, pois todos os seus componentes os acertaram.

Essa interpretação, quanto à dificuldade dos itens, revela o segundo problema da TCT, também mencionado anteriormente:

• *inexistência de invariância das propriedades métricas dos itens e, por conseguinte, do instrumento.*

É a prova empírica de que as características dos itens e, por conseguinte, do teste, dependem da amostra utilizada para a sua determinação.

A partir de agora a análise dos resultados é feita considerando-se a TRI. Para tanto, é apresentada a síntese do desempenho[4] dos respondentes.

Quadro 3: Síntese do desempenho dos respondentes segundo a TRI

RESPONDENTE	DESEMPENHO (TRI)
A	-1,37
B	-1,68
C	-1,33
D	2,21
E	2,66
F	1,66
G	-1,55
H	2,41

De acordo com o desempenho individual, determinado através do uso da TRI, pode-se afirmar que:

• D, E, F e H obtiveram os mais elevados desempenhos e continuam representando o GS;

• A, B, C e G obtiveram os mais baixos desempenhos e continuam a representar o GI.

Hierarquizando os respondentes, segundo o desempenho individual, obtém-se a seguinte distribuição:

• E, H e D possuem os desempenhos mais elevados e estão no mesmo nível de aprendizagem, isto é, possuem $\theta > 2,0$;

• F está num nível inferior aos sujeitos E, H e D, isto é, possui $1,5 < \theta < 2,0$;

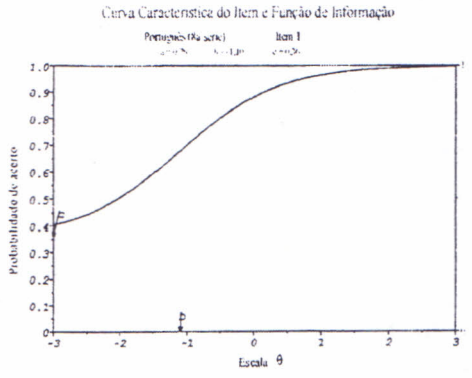
• C, A, G e B possuem os mais baixos desempenhos, isto é, $\theta < -1,0$.

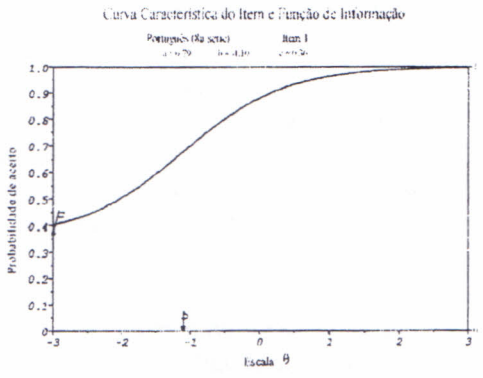
Observa-se que F possui nível de aprendizagem (θ) inferior aos sujeitos E, H e D. No entanto, acertou o mesmo número de itens: 18! Como explicar essa aparente contradição entre tais informações? Há que se recorrer às CCI's. É pertinente esclarecer que as CCI's foram determinadas utilizando-se os 7576 participantes da Avaliação da Qualidade do Ensino das Escolas Públicas do Ceará, pois é impensável determiná-las utilizando as respostas de apenas oito sujeitos.

[3] Foram adotados os seguintes níveis de dificuldade: entre 0% e 30% de acerto o item é difícil; entre 31% e 70% o item é de dificuldade média e entre 71% e 100% o item é fácil.

[4] O desempenho dos respondentes é obtido através da probabilidade de acertar o item ($P(\theta)$), que, por sua vez, é influenciada pelos parâmetros dos itens (a, b e c), de acordo com o modelo que se adequa à distribuição dos dados.

A seguir são apresentadas as CCI's dos itens 1 e 4.





Como dito, os valores dos parâmetros métricos dos itens (a, b e c) e o desempenho dos sujeitos (isto é, o nível de aprendizagem ou θ) estão numa mesma escala. Assim, a CCI do item 1 informa que a discriminação (b) possui valor -1,10, isto é, apenas os indivíduos que possuam níveis de aprendizagem (θ) próximos a este valor poderão acertá-lo. De fato, de acordo com a matriz de respostas dicotomizadas (quadro 1), pode-se observar que apenas os sujeitos A e B não conseguiram acertá-lo, já que possuem θ<-1,10 (Quadro 3). Ou seja, têm conhecimento inferior ao exigido pelo referido item.

Prosseguindo com o raciocínio, pode-se fazer a seguinte indagação: por que, então, os sujeitos C e G conseguiram acertá-lo, já que também possuem θ<-1,10? Para responder a contento tal questão, deve-se considerar o valor do parâmetro c (probabilidade de acerto ao acaso). Observa-se que o valor de c foi 0,36, isto é, os sujeitos que possuem nível de aprendizagem inferior ao exigido pelo item (θ em torno de -1,10) têm 36% de chances de acertá-lo, através de respostas aleatórias ou "no chute".

Retornando ao caso do respondente F, muito provavelmente, acertou o item 4 ao acaso, visto que seu nível de aprendizagem (θ=1,66) é inferior ao valor do parâmetro de discriminação (b=2,05). Além disso, o valor do parâmetro c (probabilidade de acerto ao acaso) é bastante elevado (c=0,38) no comentado item.

Sobre os parâmetros métricos dos itens, determinados através da TRI e representados pelas CCI's, pode-se comentar o seguinte:

- o modelo que melhor se ajustou à distribuição empírica dos resultados foi o de três parâmetros e, por conseguinte, para todos os 25 itens do teste foram determinadas a dificuldade (parâmetro a), a discriminação (parâmetro b) e a probabilidade de acerto ao acaso (parâmetro c);
- os valores de a e b estão na mesma escala de θ, permitindo, dessa forma, uma análise qualitativa acurada do desempenho, ou seja , do nível de aprendizagem (θ);
- a determinação empírica da probabilidade de acerto ao acaso (parâmetro c) enriquece as análises qualitativas baseadas no valor de θ;
- não há tratamento dos itens como réplicas uns dos outros, visto que, para cada item há uma determinada probabilidade de acerto (P(θ)), que, por sua vez, é estabelecida em função da dificuldade, discriminação e probabilidade de acerto ao acaso;
- há invariância das propriedades métricas dos itens, isto é, um item exige sempre o mesmo valor de θ para ser acertado, que é expresso por P(θ), independendo, dessa maneira, do nível de aprendizagem do respondente (ver comentário feito ao final da seção 3.2 deste trabalho).

CONSIDERAÇÕES FINAIS

A adoção do modelo TRI para a criação de bancos de itens, a partir da determinação dos seus parâmetros métricos, é uma tendência universal em áreas como a educação e a psicologia (Hambleton, 1990). Apesar dessa constatação, os psicometristas e pedagogos brasileiros ainda "engatinham" na atividade de utilização do mencionado modelo (Pasquali, 1997).

Diante disso, nada mais adequado que apresentar algumas vantagens de organizarem-se bancos de itens utilizando o modelo TRI. De acordo com Fletcher (1994):

> *... talvez o aspecto mais importante da nova teoria é a promessa de fornecer medidas invariantes do desempenho cognitivo, que não dependem dos itens que compõem a prova ou das pessoas investigadas na amostra.*
> *... A calibração fornece a cada item, parâmetros que caracterizam suas qualidades técnicas, independentes da população investigada. ... Sendo invariantes, eles não dependem da amostra selecionada para fins de calibração. Sendo invariantes, podem ser aplicados a qualquer outra população, proporcionando resultados na mesma escala de habilidade* (p. 24).

Para finalizar, são citadas algumas palavras de Pasquali (1997) a respeito do uso da TRI:

> *Uma das conseqüências mais radicais da TRI no campo dos testes consiste em que o objetivo básico nesta área não reside em elaborar e validar testes ou instrumentos, como se fazia tradicionalmente, mas consiste em elaborar e validar tarefas, itens... Assim, o objetivo final deste modo de pensar em instrumentação consiste na criação de bancos de itens para cada traço latente e, a partir desse banco, construir os testes adaptados a cada sujeito respondente. Assim, a tarefa do psicometrista já não será mais de validar e normatizar testes e sim de parametrizar tarefas ou itens. Com isso se quer dizer que a tarefa consiste em redigir a carteira de identidade de cada item, contendo os seus parâmetros distintivos, tais como o seu coeficiente de validade (a carga no traço latente), seu índice de discriminação, nível de dificuldade, seu índice de disfunção cultural (DIF), e outros... é de se prever que esta será a tecnologia do futuro na área dos testes. Conseqüentemente é nela que o país deve investir, o que concretamente significa em investir na elaboração de bancos de itens* (pp. 59-60).

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRIOLA, Wagner B. Inteligência, Aprendizagem e Rendimento Escolar segundo a Teoria Triárquica da Inteligência (TTI). *Educação em Debate, 35 (1),* p. 75-80, 1998.

ANDRIOLA, Wagner B. & BARRETO, José A. E. Análise métrica de instrumento de medida da aprendizagem através da Teoria de Resposta ao Item (TRI). *Ensaio, 5,* jan-mar, 1997.

BARRETO, José A. E. Avaliação: mitos e armadilhas. *Ensaio, 1 (1),* p. 46-48, out-dez, 1993.

BRYMAN, Alan & CRAMER, Duncan. *Análise de dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS.* Oeiras: Celta Editora, 1992.

DIAS, Mardonio R. 1997. *Análise fatorial: uma introdução.* João Pessoa: Universidade Federal da Paráiba. Manuscrito não publicado.

FERNÁNDEZ, José M. *Teoría de Respuesta a los Ítems. Un nuevo enfoque en la evolución psicológica y educativa.* Madrid: Ediciónes Pirámide S.A., 1990.

FLETCHER, Philip R. A Teoria da Resposta ao item: medidas invariantes do desempenho escolar. *Ensaio. Avaliação e Políticas Públicas em Educação, 2 (1),* p. 21-28, jan-mar. 1994.

GAVÍRIA, José L. *Breve introducción a la Psicometria. Principales Teorias.* Madrid: Universidad Complutense de Madrid, 1995. Manuscrito não publicado.

HAMBLETON, R. K. Item response theory: a broad psychometric framework for measurement advances. *Psicothema, 6 (3),* p. 535-556, 1994.

HAMBLETON, R. K., SWAMINATHAN, H. & ROGERS, H. J. *Fundamentals of Item Response Theory.* North Caroline: Sage Publications, 1991.

MUÑIZ, J. & HAMBLETON, R.K. Medio siglo de Teoría de Respuestas a los Ítems. *Anuario de Psicologia, 52,* p. 41-66, 1992.

PASQUALI, Luiz. O investimento em Testes Psicológicos. *Anais do I Congresso Ibero-Americano de Avaliação Psicológica (p. 59-60).* Porto Alegre: PUCRS, 1997.

POPHAM, William J. *Manual de avaliação:* regras práticas para o avaliador educacional. Petrópolis, Vozes, 1977.

SILVA, Céres Santos. *Medidas e avaliação em educação.* Petropólis, Vozes, 1992.

TYLER, Leona E. *Testes e medidas.* Rio de Janeiro, Zahar Editores, 1981.

VIANNA, Heraldo M. *Testes em Educação.* São Paulo: IBRASA, 1982.