Avanços na Interpretação de Escalas com a Aplicação da Teoria de Resposta ao Item

Advances in Scale Interpretation with the Application of Item Response Theory

Ricardo Primi
Universidade São Francisco

Resumo

Um dos aspectos mais importantes de qualquer instrumento de avaliação refere-se ao significado dos escores, isto é, à interpretação atribuída aos diferentes níveis de desempenho. Tradicionalmente, a inteligibilidade dessas notas pode ser conferida por três procedimentos: referência à norma, referência ao conteúdo, referência ao critério. O mais utilizado é a referência à norma informando a posição relativa que um escore ocupa frente ao desempenho de um grupo de referência. Contudo, a principal limitação deste procedimento é que ele carece de informações exatas sobre o que a pessoa é capaz de realizar. O Escalonamento comportamental, definido por John B. Carroll, é um procedimento baseado na Teoria de Resposta ao Item (TRI) que procura superar essa limitação. Por meio desse método pode-se afirmar, em termos comportamentais, as implicações dos resultados dos testes com respeito ao que o examinado conhece ou pode realizar. Esse artigo discute este procedimento e ilustra sua aplicação na avaliação da compreensão em leitura e inteligência fluida

Palavras chave: Teoria de Resposta ao Item, interpretação referenciada em critério, interpretação de escalas e métodos psicométricos.

Abstract

One of key points in psychological tests is related to the meaning of test scores, that is, the interpretation of different levels of achievement. Usually the scores intelligibility is accomplished by the use of three procedures: (a) norm reference, content reference and criterion reference. The most common method, norm reference, informs the relative standing of a specific score in relation to a group of reference. Although, the main limitation of this procedure is the lack of information about what kind of attainments a person is capable to achieve. Behavioural scaling, defined by John B. Carroll, is a procedure based on Item Response Theory which overcame this limitation. Using this method it is possible to state, in behavioural terms, the implications of test results in respect to what the subject knows or is capable to realize. This paper discusses this procedure and illustrates its application in the assessment of reading comprehension and fluid intelligence.

Key Words: Item Response Theory, criterion referenced measures, scale interpretation and psychometric methods

Notas do autor:

1) Os trabalhos do autor foram financiados pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e também pelo Conselho Nacional de Pesquisa Científica (CNPq).

2) Endereço: Ricardo Primi, Universidade São Francisco, Laboratório de Avaliação Psicológica e Educacional, LabAPE, Programa de Pós Graduação Stricto Sensu em Psicologia, Rua Alexandre Rodrigues Barbosa, 45, CEP 13251-900, Itatiba – SP, correio eletrônico: ricardo.primi@saofrancisco.edu.br

Avanços na Interpretação de Escalas com a Aplicação da Teoria de Resposta ao Item

Um dos aspectos mais importantes de qualquer instrumento de avaliação refere-se ao significado dos escores, isto é, à interpretação atribuída aos diferentes níveis de desempenho. Tradicionalmente, a inteligibilidade dessas notas pode ser conferida por três procedimentos: referência à norma, referência ao conteúdo, referência ao critério (Cronbach, 1996).

A referência à norma compara os escores obtidos por um sujeito com os escores obtidos por um grupo de referência (grupo normativo) e indica a posição relativa desse escore frente ao grupo. A referência ao conteúdo é utilizada quando o conjunto de problemas presente no instrumento pode ser considerado uma amostra representativa do universo de problemas de um determinado conteúdo ou domínio. Nessas condições interpreta-se o escore nas tarefas (amostra) diretamente como uma estimativa do escore que o sujeito teria se respondesse a todos os problemas do universo (população). A referência ao critério confere significado ao escore relacionando-o a alguma outra medida que se queira prever, chamada critério externo. Se existe uma correlação significativa entre as duas medidas pode-se conferir significado ao primeiro escore, indicando, para cada nível, qual a expectativa de desempenho no critério externo.

Dos três procedimentos, um dos mais empregados é a interpretação referenciada no desempenho de um grupo normativo, utilizado, por exemplo, no Exame Nacional de Cursos - ENC, no Exame Nacional do Ensino Médio - ENEM (Instituto Nacional de Estudos e Pesquisas Educacionais 1997, 1998a) e na grande maioria dos testes psicológicos (Anastasi & Urbina, 1997). Embora esse procedimento seja a escolha correta quando se deseja saber a posição que uma pessoa ocupa perante os seus pares, ele não possibilita afirmar com exatidão o que a pessoa é capaz de realizar, isto é, ele não implica, pelo menos diretamente, em afirmações precisas sobre o potencial das pessoas. Com o avanço da Psicometria, culminando nos modelos da Teoria de Resposta ao Item – TRI, essa limitação vem sendo superada (Hambleton, Swaminatham & Rogers, 1991).

Carrol (1993) sistematizou um conjunto de procedimentos baseados na TRI visando o estabelecimento de significados mais precisos para o desempenho. Chamou-os de métodos de *escalonamento comportamental*. Em suas palavras: "O escalonamento comportamental de testes refere-se ao processo de afirmar, em termos comportamentais, as implicações dos resultados dos testes com respeito ao que o examinado conhece ou pode realizar. Portanto ele é um aspecto da validade de construto. Ele não necessariamente precisa estender-se aos problemas de validade externa ou preditiva, embora o escalonamento comportamental adequado de testes possa ser de grande ajuda quando trata desse aspecto" (p. 299).

Pretende-se, nesse artigo, apresentar sucintamente este procedimento, ilustrando algumas aplicações práticas que vêm sendo implementadas em procedimentos de avaliação psicológica e educacional.

Idéias Básicas da Teoria de Resposta ao Item

A TRI propõe um modelo matemático que formaliza a relação entre os elementos essenciais da situação na qual uma pessoa responde a um problema. Nessa situação, quanto maior a *habilidade* da pessoa no fator requerido pelo problema maior será a *probabilidade* de que ela responda corretamente. Por outro lado, sendo a habilidade constante, quanto maior for a *dificuldade* do problema menor será a probabilidade de que ela o acerte. O modelo matemático representa essa situação por meio da Curva Característica do Item (CCI), que indica a probabilidade de acerto em função da habilidade das pessoas que o respondem e da dificuldade do problema (dependendo do modelo, podem ser incluídas outras características do item, como a discriminação e a probabilidade de acertos ao acaso).

O modelo de um parâmetro inicialmente criado por Georg Rasch (1980) e popularizado por Wright e Stone (1979) caracteriza o item somente pela sua dificuldade e é equacionado:

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

Onde,

$P_{ij}(\theta_j)$ = probabilidade de que a pessoa *j*, com habilidade θ_j acerte o item *i*;

θ_j = a habilidade da pessoa *j*,

b_i = dificuldade do item *i*,

D = constante de ajuste igual à 1,7,

e = constante matemática igual à 2,72.

P (

Na Figura 1 apresenta-se a representação gráfica dessa função para dois itens: Item 1, cujo *b* = -1,8,

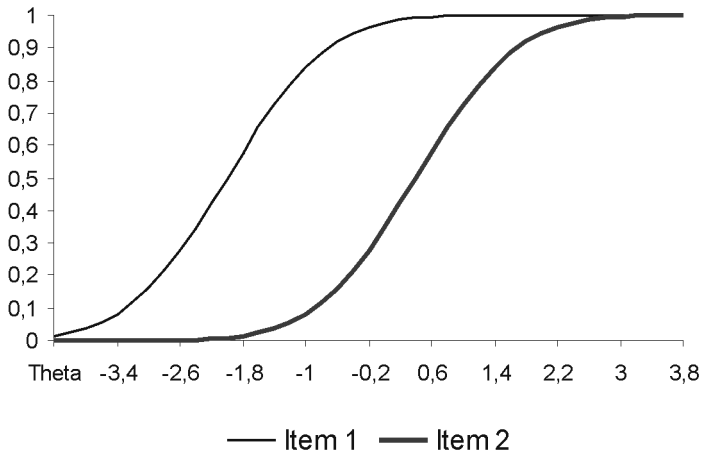

Figura 1. Curvas características de dois itens segundo o modelo de Rasch de um parâmetro (Item 1, b=-1,8; Item 2, b=0,6)

e o Item 2, mais difícil, cujo $b = 0,6$. Este índice indica o valor na escala da habilidade, em que a probabilidade de acerto corresponde a 0,50. Observa-se que, à medida que a habilidade aumenta, aumenta também a probabilidade de se acertar o item. O Item 2 é mais difícil pois, para atingir 50% de acerto, é necessário ter uma habilidade igual a 0,6, em comparação ao –1,8 suficiente para atingir o mesmo nível no Item 1.

Em uma situação de avaliação aplica-se um conjunto de itens previamente calibrados, isto é, com parâmetros conhecidos, a uma pessoa cuja habilidade se pretende conhecer. Após a correção das respostas, tem-se duas informações: a probabilidade de acerto, isto é, o padrão de acertos e erros nos itens aplicados, e as dificuldades desses itens (obtidas previamente nos estudos de calibração). A atribuição do escore ao sujeito é feita comparando-se o perfil de acertos de uma pessoa com a dificuldade dos problemas respondidos, para atribuição de um valor numérico (theta) que indica a habilidade do sujeito. Em síntese, este procedimento encontra o valor da habilidade mais condizente com o padrão observado de respostas, considerando, para isso, a dificuldade dos itens.

Dizendo de outro modo, por meio da curva característica do item se estabelece uma relação que tem, de um lado, a probabilidade de acerto $[P_{ij}(\theta_j)]$, e do outro, uma comparação entre habilidade do sujeito e dificuldade do item $(\theta_j\text{-}b_i)$. Nesse processo de comparação, se o sujeito acertou o item é porque sua habilidade excedeu à dificuldade do item $(\theta_j > b_i)$. Reversamente, se errou, é porque sua habilidade foi inferior à dificuldade do item $(\theta_j < b_i)$. A estimação da habilidade é feita considerando a probabilidade de acerto ou erro e a dificuldade do item, e encontrando o valor da habilidade.

Uma implicação importante da aplicação da TRI é que, uma vez conhecida a habilidade de uma pessoa, pode-se estabelecer as expectativas de acerto nos itens que avaliam aquela habilidade. Por exemplo, quando uma pessoa tem habilidade igual ao índice de dificuldade do item, as chances são de 50% de que ela o acerte. À medida em que sua habilidade aumente em relação à dificuldade do item, suas chances de acertá-lo serão maiores do que 50%. À medida que sua habilidade seja menor do que a dificuldade do item, suas chances de acertá-lo serão menores do que 50%. Portanto, tendo em mãos o escore de uma pessoa, pode-se prever quais itens ela terá mais chances de acertar ou errar, informando-se o domínio que a pessoa possui do que foi avaliado.

Carrol (1993) propôs que esta informação poderia ser visualizada com maior clareza utilizando a Curva Característica da Pessoa (CCP). A CCP representa o decréscimo da probabilidade de acerto de uma pessoa com habilidade q à medida que a dificuldade dos itens aumenta. Na Figura 2 exemplifica-se essa curva para duas pessoas: a pessoa A, com habilidade igual a –1,2 e a a pessoa B com habilidade igual a 1,8. Como pode ser observado, essas curvas demonstram qual a expectativa de acertos em itens dos diferentes ní-

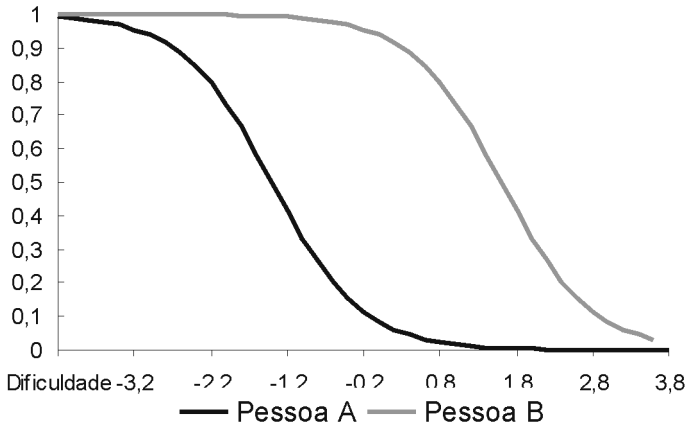

Figura 2. Curvas características de duas pessoas, demonstrando a expectativa de acertos em itens de diferentes níveis de dificuldade (a=1 e c=0).

veis de dificuldade (obviamente, com discriminação e possibilidade de acertos ao acaso semelhantes, e que nesse caso foram, respectivamente, um e zero). Diferentemente da Figura 1, as curvas representam pessoas e o eixo x a dificuldade dos itens. Como pode ser observado, essas curvas informam com clareza quais itens a pessoa consegue resolver com maior êxito.

O Escalonamento Comportamental

O escalonamento comportamental ocorre quando é possível vincular informações mais detalhadas aos diferentes níveis de dificuldade dos itens. Carrol (1993, p.303) classificou quatro maneiras pelas quais essas informações podem ser consolidadas em afirmações comportamentais: (a) Tipo I: apresentando ilustrações de itens ou tarefas com diferentes níveis de dificuldade, exemplificando os vários níveis da escala de dificuldade/habilidade com situações concretas; (b) Tipo II: apresentando descrições verbais dos comportamentos típicos esperados ao longo dos diferentes níveis de dificuldade, essas afirmações são fundamentadas na análise das competências necessárias à resolução de certos itens, representantes dos diferentes níveis de dificuldade, os quais são denominados itens-âncoras (Beaton & Allen, 1992); (c) Tipo III: vinculando informações paramétricas que descrevem como os diferentes níveis de dificuldade das tarefas presentes no teste se relacionam com outros atributos ou universo de tarefas (esse procedimento se assemelha à referência ao critério e à referência ao conteúdo, apresentadas anteriormente); (d) Tipo IV: descrevendo o tipo ou nível de processamento cognitivo necessário para acertar itens de diferentes níveis de dificuldade.

O escalonamento comportamental ocorre porque, por um lado, explica-se com mais exatidão quais processos cognitivos são necessários para o desempenho nos problemas com diferentes níveis de dificuldade e, por outro, porque a função característica da pessoa permite descobrir a expectativa de acerto para os vários níveis de dificuldade. Como tais níveis estão vinculados às informações sobre processos cognitivos pode-se generalizar sobre os processos nos quais a pessoa terá maior êxito. Observe que a interpretação, portanto, descreve exatamente do que a pessoa é capaz e no que terá mais dificuldade.

A interpretação proveniente desse procedimento supera aquela proveniente da referência à norma que, a partir do desempenho de uma pessoa, informa quantas pessoas do grupo de referência foram superadas por ela. Obviamente, os escores obtidos via TRI podem ser padronizados e organizados de forma a fornecer, além das informações sobre a proficiência, também informações normativas.

Exemplos de aplicação

Algumas aplicações do escalonamento comportamental já podem ser observadas na literatura nacional voltada à avaliação educacional e psicológica.

O Sistema de Avaliação do Ensino Básico – SAEB (Instituto Nacional de Estudos e Pesquisas Educacionais, 1998b) seguiu o procedimento proposto por Beaton e Allen (1992) e efetuou o escalonamento do Tipo II, definindo a proficiência, associada aos pontos da escala, com base na análise, por especialistas, dos conteúdos e processos necessários à resolução de itens típicos desses pontos, chamados iten- âncora.

Santos, Primi, Taxa e Vendramini (1999) vêm desenvolvendo um estudo da aplicação desses procedimentos na análise de instrumentos de avaliação da compreensão em leitura baseados na Técnica de Cloze, que consiste na seleção de um texto de aproximadamente 200 vocábulos, dos quais se omite sempre o quinto vocábulo. Os examinandos devem preencher a lacuna com a palavra que julgarem ser a mais apropriada para a constituição de uma mensagem coerente e compreensível.

Alguns estudos como o de Abraham e Chapelle (1992) têm demonstrado que o tipo de habilidade requerida para o preenchimento das lacunas depende da natureza da palavra eliminada. Santos et al. (1999) demonstraram que a complexidade da lacuna depende da classe gramatical das palavras a serem descobertas. A categoria gramatical das palavras explicou 38,4% da complexidade. Do maior nível de complexidade ao menor observou-se: adjetivos, verbos, advérbios, substantivos, pronome, preposição e artigo. Na Tabela 1 apresenta-se as expectativas de acerto para diferentes níveis de habilidade, em lacunas formadas por palavras de diferentes classes gramaticais.

A partir desse estudo, Santos et al. (1999) inferiram que os escores devem ser interpretados como a extensão e profundidade do conhecimento semântico lexical. Como pode ser observado esse estudo segue o escalonamento comportamental do Tipo IV.

Um outro estudo seguindo o escalonamento do Tipo IV é o de Primi (1998), que desenvolveu um instrumento de raciocínio analógico indutivo com figuras geométricas. Nesse estudo, a interpretação dos escores foi sistematizada em vários níveis de exigência da memória de trabalho e gerenciamento metacognitivo, aspectos centrais da inteligência fluida (Gf) medida pelo instrumento.

Tabela 1. Expecatativa Média de Descoberta de Palavras de Diferentes Classes Gramaticais para Pessoas com Diferentes Níveis de Habilidade.

Theta	**-3,00**	**-2,50**	**-2,00**	**-1,50**	**-1,00**	**-0,50**	**0,00**	**0,50**	**1,00**	**1,50**	**2,00**	**2,50**	**3,00**
Adjetivos	0,01	0,02	0,03	0,04	0,06	0,08	0,11	0,15	0,20	0,27	0,36	0,48	0,62
Verbos	0,05	0,06	0,09	0,11	0,14	0,17	0,21	0,25	0,32	0,40	0,50	0,61	0,72
Advérbios	0,03	0,04	0,08	0,13	0,19	0,28	0,37	0,45	0,54	0,62	0,70	0,77	0,84
Substantivos	0,09	0,15	0,23	0,32	0,41	0,48	0,54	0,59	0,64	0,69	0,74	0,79	0,85
Pronomes	0,09	0,13	0,18	0,24	0,33	0,42	0,52	0,62	0,71	0,79	0,85	0,90	0,93
Preposições	0,14	0,24	0,36	0,49	0,60	0,69	0,75	0,80	0,84	0,87	0,90	0,92	0,94
Artigos	0,16	0,27	0,39	0,52	0,63	0,73	0,80	0,87	0,91	0,95	0,97	0,98	0,99

Como pode ser observado o escalonamento comportamental pressupõe um avanço teórico na compreensão do construto medido pelo teste para que seja possível explicar os processos cognitivos vinculados aos diferentes níveis de capacidade. Essa é principal vantagem desse procedimento. Os procedimentos tradicionais referenciados à norma informam a posição relativa ocupada pela pessoa a partir de processos de comparações inter individuais. O escalonamento comportamental descreve a capacidade de uma pessoa a partir de uma comparação intra-individual entre a habilidade da pessoa e a complexidade da tarefa que por sua vez deve estar claramente definida apoiando-se na compreensão do construto.

Pode-se concluir que o escalonamento comportamental supera a limitação dos procedimentos tradicionais quanto ao fornecimento de

afirmações precisas sobre o potencial das pessoas e por isso sugere-se que ele seja utilizado nos procedimentos de avaliação. Obviamente isso não significa que os procedimentos tradicionais devem ser abandonados. Na verdade a melhor combinação seria o uso do escalonamento comportamental em conjunto com informações normativas.

Referências

Abraham, R. G. & Chapelle, C. A. (1992). The meaning of cloze test scores: an item difficulty perspective. *The Modern Language Journal*, 76(4), 468-479

Anastasi, A & Urbina, S. (1997) *Psychological Testing*. Upper Saddle River, New Jersey: Prentice Hall.

Beaton, A. E. & Allen, N. L. (1992) Interpreting scales through scale anchoring. *Journal of Educational Measurement*; 17(2), 191-204.

Carrol, J. B. (1993). Test theory and the behavioral scaling of test performance. In: Frederiksen, N., Mislevy, R. J. & Bejar, I. I. (Eds.*). Test theory for a new generation of tests*. (pp. 297-322) Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1996) *Fundamentos da testagem psicológica*. Porto Alegre: Artes Médicas.

Hambleton, H. K., Swaminatham, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Instituto Nacional de Estudos e Pesquisas Educacionais (1998a). *Exame Nacional do Ensino Médio: Relatório Final*. Brasília: INEP.

Instituto Nacional de Estudos e Pesquisas Educacionais (1998b). *SAEB/95 relatório final*. Brasília: INEP.

Instituto Nacional de Estudos e Pesquisas Educacionais (1997). *Exame Nacional de Cursos: Relatório Síntese*. Brasília: INEP.

Rasch, G. (1980). *Probablistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Primi, R. (1998). *Desenvolvimento de um instrumento informatizado para avaliação do raciocínio analítico*. Tese de Doutorado, Instituto de Psicologia, Universidade de São Paulo, São Paulo.

Santos, A. A., Primi, R., Taxa, F. O. S., Vendramini, C. M. M. (1999*). Análise de um instrumento para avaliação da compreensão em leitura com base na Teoria de Resposta ao Item*. Manuscrito em preparação.

Wright, B. D. & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA.