

Fundamentos da Teoria da Resposta ao Item –TRI

Basic Theory of Item Response Theory – IRT

Luiz Pasquali

Universidade de Brasília

Ricardo Primi

Universidade de São Francisco

Resumo

O artigo apresenta o histórico e a lógica básica da nova teoria em psicometria, a teoria de resposta ao item, abreviadamente a TRI. A teoria tem suas origens nos anos 1930, mas foi axiomatizada nos anos 1960, tomando conta de grande parte da psicometria nos anos 1980. A TRI constitui uma teoria dentro das teorias da modelagem latente que surgiram nos anos 1930. Estas teorias postulam que o comportamento humano é consequência de processos hipotéticos chamados de traços latentes. A TRI assume este modelo e expressa a relação entre os comportamentos (chamadas variáveis observáveis) e os traços latentes (as variáveis hipotéticas) através uma equação matemática chamada de equação logística. Esta produz uma curva ou ogiva conhecida como a curva característica do item, a CCI. A CCI define os parâmetros dos comportamentos, ditos itens (dificuldade, discriminação) em função do tamanho do traço latente, expresso como teta (θ). Discutem-se também as vantagens que a TRI apresenta frente à teoria tradicional em psicometria, a Teoria Clássica dos Testes (TCT).

Palavras chave: Teoria de resposta ao item, TRI, psicometria.

Abstract

The paper offers a historical view and the basic rationale of the modern theory in psychometrics: item response theory (IRT). This theory has its roots in the 1930, but it was fully developed only in the 1950's, and became the standard theory in psychometrics in the 1980's. The IRT is one of the latent trait modeling theories that appeared in the 1930's. Latent trait modeling theories assume that the human behavior, called observable variables, is dependent and caused by latent traits, the hypothetical variables. The IRT assume this modeling and expresses the relationship between these two types of variables through a mathematical equation called the logistic equation. This equation produces a curve called the item characteristic curve (ICC). This curve defines the items parameters (difficulty, discrimination) in terms of the level of the latent trait, symbolized as theta (θ). The paper also discusses the advantages that IRT offers over the traditional theory in psychometrics, that is, the classical theory of tests (CTT).

Key words: Item response theory, IRT, psychometrics.

Fundamentos da Teoria da Resposta ao Item –TRI

No contexto teórico da Psicometria, era inevitável que alguma teoria alternativa à teoria clássica dos testes (TCT) surgisse para resolver alguns gra-

ves problemas da medida em Psicologia. Um dos problemas a que nos referimos já vinha sendo frustrantemente assinalado por Thurstone antes mesmo dos anos 30. Dizia Thurstone (1928/1959, p. 547):

“Um instrumento de medida, na sua função de medir, não pode ser seriamente afetado pelo objeto de medida. Na extensão em que sua função de medir for assim afetada, a validade do instrumento é prejudicada ou limitada. Se um metro mede diferentemente pelo fato de estar

medindo um tapete, uma pintura ou um pedaço de papel, então nesta mesma extensão a confiança neste metro como instrumento de medida é prejudicada. Dentro dos limites de objetos para os quais o instrumento de medida foi produzido, sua função deve ser independente da medida do objeto”.

A Psicometria, hoje em dia chamada de clássica (Teoria Clássica dos Testes – TCT), estava bastante bem axiomatizada já nos anos 50, sobretudo com os trabalhos de Guilford (1936/1954) e Gulliksen (1950). Ela, contudo, continha o grave problema que Thurstone menciona acima, isto é, o instrumento construído dependia intrinsecamente do objeto medido. De fato, os testes psicológicos elaborados dentro da Psicometria Clássica são dependentes dos itens que os compõem (*test-dependent*). Assim, por exemplo, ao se querer medir a inteligência de um sujeito, o resultado vai depender muito do instrumento utilizado, obviamente um instrumento que mede inteligência. Se utilizar, digamos, as Matrizes Progressivas de Raven (SPM), obtenho um resultado; se utilizar o *Wechsler Adult Intelligence Scales* (WAIS), obtenho outro. Qual dos dois resultados é o correto? A resposta fica ligada ao instrumento utilizado, de tal forma que o objeto medido, a inteligência no caso, afeta diretamente o instrumento utilizado; aliás, ela é definida pelo instrumento utilizado. Você talvez se lembra como Binet (Binet & Simon, 1908) definiu a inteligência: Inteligência é o que o meu teste mede! Seria como dizer em Física que o comprimento do objeto é o que o metro mede ou a massa o que a balança mede.

Embora Thurstone tenha percebido aguçadamente este problema da medida em Psicologia, ele não conseguiu encontrar uma solução para o mesmo. Foi somente após os anos 50 que os psicometristas começaram a descobrir a solução para o problema, baseados na teoria do traço latente de Lazarsfeld (1959) e nos trabalhos de Lord (1952) e do dinamarquês Rasch (1960), os quais se tornaram as bases da moderna Teoria da Resposta ao Item - TRI (*Item Response Theory* - IRT), inclusive conhecida como a Teoria do Traço Latente, esboçada por Lord em 1952 e finalmente axiomatizada por Birnbaum em 1968 e por Lord em 1980.

O que pode parecer estranho nesta história é o fato de que o problema levantado por Thurstone tenha sido detectado já nos anos 30 e que a resposta já tinha sido dada nos anos 60. Por que então a nova teoria somente veio a ser utilizada nos anos 80? A

resposta está no fato de que a solução dada ao problema da independência do instrumento de medida em relação ao objeto de medida que a Teoria da Resposta ao Item propôs apresentava algoritmos matemáticos de tal complexidade que a tecnologia computacional da época era incapaz de resolver de uma maneira útil e prática. Com o avanço da tecnologia da informática (microcomputadores) e da disponibilidade de softwares apropriados, este problema foi solucionado e a TRI entrou em moda. Aliás, o primeiro software para as análises da TRI surgiu somente em 1979 com o BICAL de Wright, Mead e Bell, seguidos depois pelo LOGIST (Wingersky, Barton, & Lord, 1982) e pelo BILOG (Mislevy & Bock, 1984).

Problemas da Psicometria Clássica

Além do grave problema mencionado na introdução deste capítulo, outros problemas são salientados contra a Psicometria Clássica. Um deles afirma que os parâmetros dos itens de um teste dependem da amostra de sujeitos em que eles foram calculados. Assim, um item qualquer se torna mais difícil ou mais fácil dependendo da amostra ser composta de sujeitos mais inteligentes ou menos inteligentes. Desta forma, o parâmetro de *difficuldade do item* vai variar de pesquisa para pesquisa em função da amostra de sujeitos; isto é, este parâmetro é dependente dos sujeitos utilizados na pesquisa (*subject-dependent*). Esta crítica é válida, mas ela se refere mais a um problema de amostragem e não tanto da análise que se faz da dificuldade do item em termos de número de acertos. Se a mostra for aleatória e representativa, este problema não existe na Psicometria Clássica. No entanto, uma grande vantagem da TRI é que mesmo em amostras não representativas os parâmetros poderão ser estimados corretamente (Embretson & Reise, 2000).

Um problema mais grave ocorre com o cálculo do parâmetro de *discriminação do item*. Esta análise, dentro da Psicometria Clássica, é feita baseada no score total de um teste, seja utilizando grupos-critério ou coeficientes de correlação (Pasquali, 1996). Tal procedimento incorre numa incongruência lógica, pois a discriminação de cada item é testada contra o score total que é constituído por todos os itens do teste, inclusive o item que se está analisando. Isto supõe que os outros itens, pelo menos, sejam adequados. Mas se o são, então por que se fazer a análise? E se não o são, então a análise está simplesmente falha, errada.

Outro problema grave na Psicometria Clássica se situa no cálculo da *fidedignidade* de um teste. Esta é definida comumente em termos de formas paralelas de um teste. Estas formas precisam ser estritamente paralelas, isto é, elas precisam produzir um escore verdadeiro idêntico e variâncias também iguais. Obviamente, a obtenção de formas assim paralelas de um mesmo teste é algo difícil de ser conseguido. Ademais, os sujeitos, ao tomar uma forma paralela, nunca serão exatamente os mesmos, pois há os problemas de maturação que Campbell e Stanley (1973) discutem no contexto do delineamento de pesquisa, tais como a aprendizagem, o cansaço, a motivação, etc. que muda da aplicação de uma forma para outra, inclusive diferencialmente para diferentes sujeitos, tornando a comparação entre as duas formas não mais paralela.

Um outro problema presente na teoria tradicional em Psicometria consiste na suposição que nela se faz de que a variância dos *erros de medida* é a mesma para todos os testandos, suposição de difícil sustentação, pois parece óbvio que alguns testandos realizam a tarefa mais consistentemente que outros e que a consistência varia em função da habilidade dos sujeitos (Hambleton & Swaminathan, 1985). Por exemplo um teste com itens medianamente fáceis poderá diferenciar mais os sujeitos com habilidade média mas não irá diferenciar da mesma maneira os sujeitos com habilidade superior que provavelmente obterão escores perto dos mais altos. Conseqüentemente, o erro de medida neste segundo grupo será maior que no primeiro (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985).

Há, além disso, a condição típica dos testes de aptidão construídos dentro dos moldes da teoria clássica da Psicometria. Os testes são elaborados para avaliar maximamente os sujeitos de habilidades medianas, sendo, por isso, bem menos apropriados e válidos para avaliar sujeitos com habilidades superiores ou de pouca habilidade. De fato, a validade de um teste se maximiza na medida em que o nível de dificuldade do mesmo se aproxima do nível de habilidade do sujeito (Lord, 1980; Weiss, 1983). De sorte que aplicando testes de dificuldade média diferente a sujeitos de diferentes níveis de aptidão irá produzir resultados nem sempre comparáveis, pois é óbvio que obter 50 num teste fácil não é a mesma que obter 50 num teste mais difícil que meça a mesma aptidão. A tarefa de comparar os sujeitos em tais situações é de difícil manejo dentro dos modelos tradicionais de análise.

A Teoria da Resposta ao Item

Um pouco de história

Como já assinalamos, a TRI foi sendo elaborada aos poucos desde os anos 50 por vários autores, embora suas raízes remontem há mais de uma década anterior. Entre estes precursores se encontram os trabalhos de Richardson (1936), comparando os parâmetros dos itens obtidos pela teoria clássica da Psicometria com os moldes que hoje usa a TRI; os trabalhos de Lawley (1943, 1944), indicando alguns métodos para estimar os parâmetros dos itens, os quais se afastavam da teoria clássica e os trabalhos de Tucker (1946), que parece ter sido o primeiro a utilizar a expressão curva característica do item - *Item Characteristic Curve*, ICC - que constitui um conceito chave na TRI. Também deve ser mencionada a contribuição de Lazarsfeld (1950), que introduziu o conceito de traço latente, ainda que no contexto da medida das atitudes, conceito novamente que se constituiu num parâmetro chave da nova TRI.

Entretanto, o responsável mais direto que deu origem à TRI moderna, é Frederic Lord (1952, 1953) por ter elaborado, não somente um modelo teórico, mas ainda métodos para estimar os parâmetros dos itens dentro da nova teoria, utilizando o modelo da ogiva normal. Os modelos elaborados por Lord se aplicam a testes onde as respostas são dicotômicas, isto é, certo e errado, ou seja, testes de aptidão. Mais tarde, Samejima (1969, 1972) elaborou modelos para tratar respostas politômicas e mesmo para dados contínuos, como é caso em testes de personalidade. Outro passo importante na história da TRI foi dado por Birnbaum (1957) ao substituir as curvas de ogiva por curvas logísticas, isto é, baseadas nos logaritmos, tornando o tratamento matemático dos dados bem mais fácil.

Um dos fatores que concretamente mais contribuíram para o uso generalizado da TRI hoje em dia foi o avanço da informática. Como a complexidade matemática no campo da TRI é enorme, o progresso vertiginoso nas máquinas de processamento (microcomputadores) possibilitou a viabilização dos cálculos que o modelo TRI exige em Psicometria. Com este progresso das máquinas, foi possível também, nos anos 80, o desenvolvimento de softwares apropriados para os tais cálculos.

A Teoria Básica da Resposta ao Item

A Teoria da Resposta ao Item é uma teoria do traço latente aplicada primariamente a testes de ha-

bilidade ou de desempenho. O termo *teoria do traço latente* se refere a uma família de modelos matemáticos que relaciona *variáveis observáveis* (itens de um teste, por exemplo) e *traços hipotéticos não-observáveis* ou aptidões, estes responsáveis pelo aparecimento das variáveis observáveis ou, melhor, das respostas ou comportamentos emitidos pelo sujeito que são as variáveis observáveis. Assim, temos um estímulo (item) que é apresentado ao sujeito e este responde a ele. A resposta que o sujeito dá ao item depende do nível que o sujeito possui no traço latente ou aptidão. Desta forma, o traço latente é a causa e a resposta do sujeito é o efeito. Agora, para se poder estimar, a partir da resposta dada pelo sujeito, o seu nível no traço latente, é preciso que se hipotizem relações entre as respostas observadas do sujeito e o seu nível neste mesmo traço latente. Quando estas relações são expressas numa equação matemática, constando de variáveis e de constantes, temos um modelo ou teoria do traço latente. Como tanto as variáveis e constantes que entram numa tal equação, quanto as formas matemáticas que as curvas, que expressam a relação hipotetizada, podem ser as mais variadas, segue que, em princípio, existe um número sem fim de tais equações possíveis. A TRI se decidiu por algumas destas equações que achou mais adequadas ou produtivas, como veremos mais adiante.

De qualquer forma, o fundamental da teoria do traço latente consiste em expressar numa fórmula matemática a relação existente entre variáveis observadas e variáveis hipotéticas, chamadas estas de traços latentes. Assim, se conhecemos as características das variáveis observadas (como os itens de um teste), estas se tornam constantes na equação e esta se torna solucionável, permitindo que se estime então o nível do traço latente ou a aptidão do sujeito e vice-versa, isto é, se for conhecido o nível do traço latente é possível serem estimadas as características dos itens respondidos por este sujeito.

Assim, a TRI faz dois postulados básicos, a saber:

- 1) o desempenho do sujeito numa tarefa (item de um teste) pode ser predito a partir de um conjunto de fatores ou variáveis hipotéticas, ditos aptidões ou traços latentes (identificados na TRI com a letra grega teta: θ); o teta sendo a causa e o desempenho o efeito. Trata-se de modelagem latente (*latent trait modeling*). Ou seja, comportamento = função (traço latente)
- 2) a relação entre o desempenho e os traços

latentes pode ser descrita por uma equação matemática monotônica crescente, chamada de Curva Característica do Item – CCI (veja Figura 1).

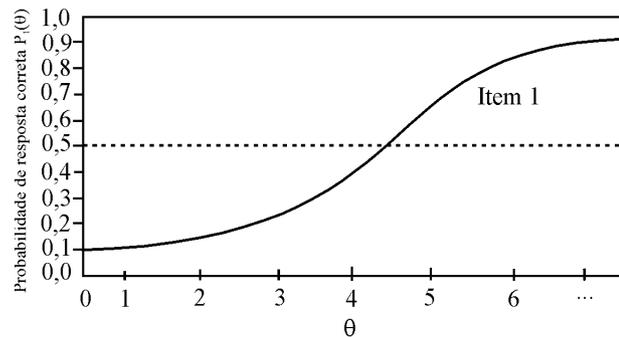


Figura 1 – ICC para item 1

A Figura 1 mostra que, à medida que aumenta o teta, aumenta também a probabilidade de acertar o item (relação monotônica crescente entre aptidão e probabilidade de acerto)¹.

Vantagens da TRI

As restrições que a TRI faz sobre os dados, como veremos, são muito fortes. Segundo o princípio geral da parcimônia em ciência, isto é, de que se deve privilegiar a explicação e o modelo mais simples, a TRI seria um modelo bem mais complexo do que o modelo da Psicometria Clássica e, conseqüentemente, deveria ser preterido em favor desta. Isto seria válido se o modelo mais complexo não explicasse mais coisas do que o mais simples. Então, pergunta-se: a TRI oferece vantagens importantes em relação ao modelo tradicional da Psicometria ou não? Quais são estas vantagens que torna a TRI preferível?

Primeiramente, é preciso alertar que a TRI não veio para substituir toda a Psicometria Clássica, mas apenas partes dela, particularmente na análise dos itens e no tema da fidedignidade da medida; mas sua contribuição na temática da validade dos instrumentos psicológicos, ainda que importante, não parece revolucionária e esta temática é a mais relevante em Psicometria. Evidentemente, é de se esperar que logo se consiga desenvolver a Psicometria pura e simplesmente em Psicologia, sem distinção entre Clássica e Moderna, mas utilizando os avanços definitivos de ambas. Dito isto, então

¹ Os modelos atuais da TRI completos incluem outros parâmetros na equação, como veremos ao falarmos dos mesmos; contudo, a lógica fundamental da teoria do traço latente consiste em expressar o comportamento em termos de traços latentes.

quais são os avanços que a TRI trouxe no campo da Psicometria? Hambleton, Swaminathan e Rogers (1991) apresentam cinco grandes avanços que a TRI trouxe, sendo os três primeiros particularmente importantes. Vejamos:

- a) o *cálculo do nível de aptidão* do sujeito independe da amostra de itens utilizados: diz-se que a habilidade do sujeito é independente do teste (*not test-dependent*). Na Psicometria Clássica, o escore do sujeito dependia e variava segundo o teste aplicado fosse mais fácil ou mais difícil, ou produzisse maiores ou menores erros. Assim, tais escores não eram comparáveis e, mesmo aplicando ajustes, os escores ainda continuavam não comparáveis sobretudo porque os testes produziam diferenças nas variâncias de erros de medida. No caso da TRI, não importa que itens ou conjunto de itens que você utilize, obviamente que estejam medindo o mesmo traço latente, irão produzir o mesmo nível de aptidão do sujeito, dentro, é óbvio, dos sempre presentes erros de medida em qualquer ramo da ciência;
- b) o *cálculo dos parâmetros dos itens* (dificuldade e discriminação) independe da amostra de sujeitos utilizada: diz-se que os parâmetros são independentes dos sujeitos (*not group-dependent*). Na clássica, os parâmetros dependiam muito dos sujeitos amostrados possuírem maior ou menor aptidão;
- c) a TRI permite *emparelhar itens com a aptidão* do sujeito. Isto quer dizer que se avalia a aptidão de um sujeito, utilizando itens com dificuldade tal que se situam em torno do tamanho da aptidão do sujeito, sendo, assim, possível utilizar itens mais fáceis para sujeitos com habilidades inferiores e itens mais difíceis para sujeitos mais aptos, produzindo escores comparáveis em ambos os casos. Na psicometria clássica sempre era aplicado o mesmo teste, hermeticamente fechado, para todos os sujeitos, de sorte que, se o teste fosse fácil, avaliaria bem sujeitos de aptidão menor e mal sujeitos de aptidão superior e, se o teste fosse difícil, faria o contrário. Esta história ficará mais clara ao falarmos de testes adaptativos e montagem de testes otimizados;
- d) a TRI constitui um modelo que não precisa fazer suposições que aparentam serem im-

- prováveis, tais como os erros de medida serem iguais para todos os testandos;
- e) a TRI não necessita trabalhar com testes estritamente paralelos como exige a psicometria clássica.

Podem parecer que as duas primeiras vantagens são fictícias pois se utilizássemos amostras representativas, tanto na construção quanto na padronização dos testes, teríamos medidas independentes do teste utilizado e parâmetros dos itens independentes da amostra de sujeitos. Considere, por exemplo, dois testes, um mais fácil e um mais difícil, padronizados com amostras representativas de sujeitos. Neste caso poder-se-ia supor que, resultados semelhantes na escala padronizada, obtidos por meio do teste fácil ou pelo teste difícil, significariam a mesma coisa. Mas isso não é verdade pois as distâncias entre os sujeitos com habilidade superior à média, obtidas caso aplicássemos o teste difícil, seriam maiores do que se aplicássemos o teste mais fácil uma vez que estes sujeitos, com alta habilidade, teriam notas muito próximas das mais altas e, portanto, mais próximas no segundo teste. Portanto a medida é afetada pelo instrumento utilizado. Na TRI isto não ocorre porque ela trabalha com o escore teta. Como mostra a Figura 1 o escore teta tem uma relação não linear com a probabilidade de acerto associada ao escore total e, por isso, corrige as distorções do escore total. Nos extremos, uma diferença de pequena magnitude na probabilidade de acerto corresponde a uma diferença de magnitude muito dos escores teta, ainda que, com precisão muito mais reduzida.

Suposições da TRI

Entre as características ou pré-requisitos da TRI, duas são de especial relevância: a unidimensionalidade e a independência local, que representam suposições básicas dos modelos mais correntes desta teoria psicométrica. A história das suposições é frustrante para muita gente, porque elas não podem ser empiricamente demonstradas e nem possuem bases lógicas; elas são simplesmente aceitas ou não aceitas, isto é, elas são gratuitas, elas são hipóteses. Mas, observe o seguinte: (1) qualquer modelo matemático, para poder funcionar e ser útil, precisa fazer algumas suposições entre o modelo e os dados empíricos, inclusive especificando as relações que existem entre as variáveis hipotéticas do modelo e as variáveis observáveis ou empíricas; (2) embora as suposições não possam ser provadas diretamente, elas podem sê-lo indiretamente (em suas conseqüências), isto é, verificando se a sua violação produz resultados incon-

gruentes no estudo da realidade empírica; quer dizer que os resultados práticos irão determinar se as suposições foram ou não úteis ou adequadas.

A Unidimensionalidade

As teorias do traço latente (que vêm dos anos 30) afirmam que existe um conjunto de traços latentes que estão por detrás de um desempenho comportamental qualquer. Assim, um sujeito qualquer se situa num espaço de n dimensões (agora chamadas de teta – q) e seu desempenho depende do tamanho que ele possui em cada um desses tetas, de sorte que o seu comportamento pode ser expresso como um vetor de escores ou pesos nos vários tetas, isto é,

Desempenho = $f(q_1, q_2, q_3, \dots, q_n)$.

Entretanto, em sua maioria, os modelos da TRI postulam que há apenas uma aptidão (teta) responsável pela realização de um conjunto de tarefas (itens de um teste). Parece pacífico que qualquer desempenho humano é sempre multideterminado ou multimotivado, dado que mais de um traço latente entra na execução de qualquer tarefa, como o modelo geral do traço latente afirma. Contudo, para satisfazer o postulado da unidimensionalidade é suficiente admitir que haja uma aptidão *dominante* (um fator ou traço dominante) responsável pelo desempenho num conjunto de itens de um teste. Este fator é o que se supõe estar sendo medido pelo teste. O postulado da unidimensionalidade continua importante dado que a TRI ainda não possui soluções adequadas para modelos multidimensionais, embora estes estejam sendo pesquisados já de longa data (Mulaik, 1972; Samejima, 1974; Bock & Aitkin, 1981; Reckase, 1985). Isso também não quer dizer que os autores se entendem completamente sobre o que seja unidimensionalidade e como ela deve ser estimada.

A Independência Local

Este postulado afirma que, mantidas constantes as aptidões que afetam o teste, menos o teta dominante, as respostas dos sujeitos a quaisquer dois itens são estatisticamente independentes. Isto implica em que o desempenho do sujeito num item não afeta o desempenho em outro item: cada item é respondido exclusivamente em função do tamanho do seu teta dominante². Esta suposição é importante e útil, porque, sendo ela verdadeira, então a seqüência de respostas do sujeito a uma série de itens será o produto das probabilidades de cada item individual. As probabilidades de cada item dependem das ca-

racterísticas do mesmo (dificuldade, discriminação) e do tamanho do teta do sujeito respondente, como veremos mais adiante. Acompanhe o seguinte raciocínio, para entender o que seja independência local:

Seja

- q a aptidão dominante que afeta um conjunto de itens,
- U_i a resposta de um sujeito ao item i ($i = 1, 2, \dots, n$); $U_i = 1$ se o sujeito acerta o item e $U_i = 0$ se ele erra;
- $P(U_i=1/2q)$ a probabilidade de resposta do sujeito j com aptidão q ao item i . $P(U_i=1/2q)$ significa a probabilidade de uma resposta correta dada ao item i , levando em conta o teta do sujeito e $P(U_i=0/2q)$ a probabilidade de uma resposta errada. Estas duas instâncias são abreviadamente expressas como P_i e Q_i (sendo $Q_i = 1 - P_i$).

Com tais informações, a independência local pode ser matematicamente afirmada como

$$\begin{aligned} \text{Prob}(U_1, U_2, \dots, U_n/2q) &= P(U_1/2q) \\ &P(U_2/2q) \dots P(U_n/2q) \\ &= \prod_{i=1}^n P(U_i|\theta) \end{aligned}$$

onde a última expressão significa o produtório das probabilidades dos n itens do teste.

A independência local significa que, para examinando com uma aptidão dada, a probabilidade de resposta a um conjunto de itens é igual aos produtos das probabilidades das respostas do examinando a cada item individual. Assim, se um sujeito acertou os itens 1 e 2 e errou o 3, a configuração ou padrão de suas respostas, isto é, o padrão de resposta é $U_1=1, U_2=1, U_3=0$, ou seja, 1 1 0 e a independência local implica que

$$\begin{aligned} P(U_1=1, U_2=1, U_3=0/2q) &= P(U_1=1/2q) \\ &P(U_2=1/2q) P(U_3=0/2q) \\ &= P_1 P_2 Q_3. \end{aligned}$$

Embora pareça improvável que os comportamentos (respostas) de um mesmo sujeito não estejam

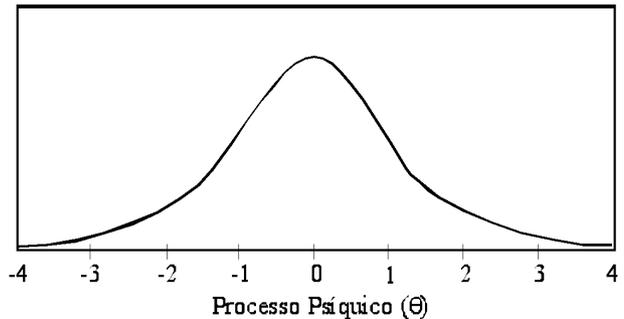
² Os matemáticos complicam esta história dizendo que $P(X_i = 1|X_j = 1, x_k, q_s) = P(X_i = 1|x_k, q_s)$, isto é, a probabilidade de que o item i seja acertado ($X_i = 1$), dependendo ou condicionado ao fato de que o item j seja acertado ($X_j = 1$) e dos parâmetros dos itens (dificuldade, discriminação, etc.: x_k) e do teta do sujeito respondente (q_s) é a mesma que resolvendo o item i sozinho. Em outras palavras, o que acontece com o item j é irrelevante com respeito à história do item i .

correlacionados, a independência local afirma que, se houver correlação, esta se deve à influência de fatores outros que não o fator dominante. Se estes outros fatores forem controlados (mantidos constantes), o fator dominante será a única fonte de variação e as respostas se tornam independentes, porque o sujeito responde exclusivamente em função do tamanho do seu teta. Assim, a independência local implica também a unidimensionalidade, já que a única causa da resposta do sujeito é este suposto teta dominante (Lord, 1980; Lord & Novick, 1968).

Também é preciso anotar que a independência local afirma a independência das respostas do sujeito, o que não significa que os itens do teste não possam estar correlacionados. Como assim? A correlação entre os itens, se existe, é definida sobre um grupo de sujeitos, que responderam do mesmo jeito, enquanto a independência local diz que cada sujeito individualmente deu respostas independentes para cada item do teste. Aliás, se os itens estão medindo o mesmo traço latente, é de se esperar que estejam correlacionados.

Quando um sujeito responde a uma série de itens, ele produz o que se chama de um padrão de respostas, composto de acertos (valor 1) e erros (va-

Supõe-se que os processos psíquicos se distribuem normalmente entre a população. Isto quer di-



zer que um processo psíquico se distribui dentro da população como uma curva normal, implicando que alguns sujeitos da população possuem um nível baixo de tal processo, outros um nível alto, e a maioria um nível médio, como mostra a Figura 2.

Figura 2 – Distribuição normal dos processos psíquicos

Para expressar a distribuição da totalidade dos elementos de uma população utiliza-se a métrica dos escore padrão, que é ancorada na média (valor 0) e que vai de $-\infty$ a $+\infty$, mas que na prática vai de -3 a +3,

Tabela 1 – Padrões de resposta em 4 itens para 10 sujeitos

Sujeitos	Padrão de resposta para item				Probabilidade do padrão
	1	2	3	4	
1	0	0	0	0	$Q_1 Q_2 Q_3 Q_4 = 0,8 \times 0,6 \times 0,5 \times 0,2 = 0,048$
2	0	0	0	1	$Q_1 Q_2 Q_3 P_4 = 0,8 \times 0,6 \times 0,5 \times 0,8 = 0,192$
3	0	0	1	0	$Q_1 Q_2 P_3 Q_4 = 0,8 \times 0,6 \times 0,5 \times 0,2 = 0,048$
4	0	0	1	1	$Q_1 Q_2 P_3 P_4 = 0,8 \times 0,6 \times 0,5 \times 0,8 = 0,192$
5	0	1	0	0	$Q_1 P_2 Q_3 Q_4 = 0,8 \times 0,4 \times 0,5 \times 0,2 = 0,032$
6	0	1	0	1	$Q_1 P_2 Q_3 P_4 = 0,8 \times 0,4 \times 0,5 \times 0,8 = 0,128$
7	0	1	1	1	$Q_1 P_2 P_3 P_4 = 0,8 \times 0,4 \times 0,5 \times 0,8 = 0,128$
8	1	0	0	0	$P_1 Q_2 Q_3 Q_4 = 0,2 \times 0,6 \times 0,5 \times 0,2 = 0,012$
9	1	0	1	1	$P_1 Q_2 P_3 P_4 = 0,2 \times 0,6 \times 0,5 \times 0,8 = 0,048$
10	1	1	1	1	$P_1 P_2 P_3 P_4 = 0,2 \times 0,4 \times 0,5 \times 0,8 = 0,032$

lor 0). Veja os exemplos de padrões de resposta e sua probabilidade de ocorrência na Tabela 1 para 10 sujeitos em 4 itens, cuja probabilidade individual de acerto $[P_i(q_j)]$ destes últimos é de, respectivamente, 0,2, 0,4, 0,5, 0,8.

A Curva Característica do Item – CCI

Na visão da teoria do traço latente, os itens de um teste constituem a expressão comportamental ou comportamentos, que o sujeito expressa como resposta a um ou mais traços latentes. Assim, qualquer comportamento seria a expressão física ou comportamental de um processo psíquico, sendo este a causa daquele.

porque entre estes dois pontos cai 99,97% de todos os sujeitos de uma população. A TRI faz uso desta métrica. O processo psíquico é chamado de teta (θ). Supõe-se, assim, que cada sujeito da população possui um tamanho ou nível de teta que o posiciona nesta escala de -3 a +3. Como é que se vai descobrir o tamanho do teta de cada sujeito? Resposta: Através do seu comportamento. Como assim? A forma usual para medir um processo latente (teta) consiste em desenvolver um teste com uma série de questões ou itens, os quais expressam algum aspecto do traço latente. Em se tratando de um teste de aptidão, então o tamanho de teta do sujeito é expresso pelo nú-

mero de itens que ele acertou. Isto na Teoria Clássica dos Testes (TCT). No caso da Teoria de Resposta ao Item (TRI), entretanto, não se pergunta quantos itens o sujeito acertou e, sim, por que ele acertou ou errou cada item individual. Desta forma, a TRI está interessada em descobrir qual é o tamanho de teta que o sujeito deve ter para poder acertar o item, cada item individualmente. Você já pode adivinhar, então, que em teoria basta até um único item para se poder descobrir o tamanho do teta do sujeito. O problema que fica para resolver consiste em se saber como o item sinaliza o tamanho do teta do sujeito. A resposta ficará clara no decorrer desta exposição, mas a resposta sucinta é a seguinte: através da CCI. Então vamos ver o que é Teesta CCI.

Parece razoável se supor que um sujeito que tem maior aptidão, isto é, que possui um nível mais elevado do processo latente que um dado item mede, terá uma probabilidade maior de acertar este item do que um sujeito com nível inferior de aptidão. Se o processo latente é expresso como q , então esta probabilidade de acerto é definida como $p_i(q)$, que se lê como: a probabilidade (p) de acertar o item (i) dado um tamanho tal de teta (q). Assim, o sujeito com menor habilidade terá uma $p_i(q)$ pequena, enquanto um de aptidão superior terá tal probabilidade bem maior. Desta forma, a $p_i(q)$ de acertar um dado item vai de 0 a 1, onde ela será 0 para o sujeito que não tiver nenhuma aptidão que o item mede e 1 para o sujeito que tem uma aptidão teta ótima. Esta situação faz com que, à medida que cresce o tamanho do teta, vai crescendo também a $p_i(q)$, provocando visualmente uma curva de tipo S na escala de aptidão, como mostra a Figura 3.

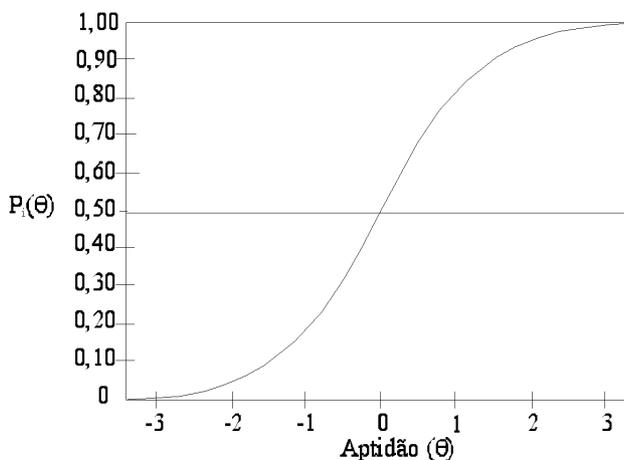


Figura 3 – A CCI de um item em função do θ

Esta curva expressa tudo que a TRI pode descobrir sobre o item. Ela é chamada de curva logística e é caracterizada por duas características, a saber, a dificuldade e a discriminação. Vamos examinar um pouco a história da curva logística e as duas características.

A função logística

A função logística se expressa numa ogiva, que já era utilizada desde o século 19 em biologia para expressar o crescimento dos seres vivos e ela entrou em voga na Psicometria com Birnbaum nos anos 1950. Os pioneiros da TRI não trabalhavam com a função da ogiva logística e sim com a da ogiva normal (veja Lord, 1952; Lord & Novick, 1968). A expressão da função em termos logísticos evita trabalhar com integrais, o que permite tratamento matemático mais simples, segundo os estatísticos. Essa história do modelo da ogiva normal e do modelo logístico é um pouco complicada para estudiosos que não são estatísticos. Vamos ver brevemente esta história: o modelo da ogiva normal se baseia na chamada função phi-gamma que trabalha com os mínimos quadrados. Phi-gamma vem das letras gregas que compõem a fórmula: $p = f(g)$, onde o gama (g) é o desvio em relação à média, mais ou menos como é o caso dos desvios-padrão z . A Equação que trabalha a ogiva normal é complexa e é a seguinte: $p = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$. Os estatísticos preferem trabalhar com outra função, que produz uma curva bem parecida com a ogiva normal e que tipicamente se ajusta melhor aos dados. Esta função é a função logística, a qual trabalha com o método da máxima verossimilhança em lugar dos mínimos quadrados e é matematicamente mais fácil de processar. A função logística é a seguinte: $Q = \frac{1}{1 + e^{-\alpha - \beta x}}$ (veja Guilford, 1954). O expoente do denominador é chamado de logit, ou seja, o desvio logístico. Aplicado ao caso da TRI, esta função é a seguinte:

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$

onde, e = constante que equivale a 2,7182818...

(geralmente expresso como 2,72)

b = dificuldade do item

a = discriminação do item

θ = aptidão ou traço latente

$a(\theta - b)$ = logit.

A conceituação e a métrica do q foi explicada acima (veja figura 2). A conceituação e a métrica dos parâmetros de dificuldade e de discriminação serão explicados a seguir. Quando se fala de métrica, quer-se referir ao tipo de escala utilizada para medir um dado fenômeno. Assim, quando se diz que alguém obteve 9 numa prova de desempenho, sendo um desempenho excelente, está-se supondo que a métrica utilizada é uma escala que vai de 0 a 10. Se a escala utilizada fosse de 0 a 100, então a nota 9 representaria péssimo desempenho. Desta forma, você vê que é de fundamental importância, quando se dão valores numéricos aos fenômenos, se saber qual é a métrica (o metro) utilizada e, assim, se poder entender o significado do valor atribuído.

As características do item

O parâmetro de dificuldade, chamado de b_i , é o ponto na escala de aptidão no qual a probabilidade de uma resposta correta é de 50% (ou seja, 0,5). A métrica teórica deste parâmetro vai de $-\infty$ a $+\infty$, mas na prática vai de -3 a +3.

O parâmetro da discriminação, chamado de a_i , é expresso pela inclinação da CCI no momento da inflexão, isto é, quando a curva CCI corta a linha que corresponde à probabilidade de 0,5 de resposta correta., o que acontece sempre quando $q = b$. A métrica teórica deste parâmetro também vai de $-\infty$ a $+\infty$, mas valores negativos ficam estranhos porque eles diriam que sujeitos de maior ap-

tidão tendem a errar o item enquanto sujeitos de menor habilidade tendem a acertar. Na prática, a métrica deste parâmetro vai de 0 a 3, onde 0 significa nenhuma discriminação e 3, discriminação praticamente perfeita. Veja estes dois parâmetros expressos na Figura 4.

Na figura 4, o parâmetro b_i do item corresponde ao ponto na escala de aptidão q onde a probabilidade de resposta é 0,5. Quanto maior for o b_i , maior deve ser o nível de aptidão exigido para que o examinando tenha a chance de 50% de acertar o item. Transformando a escala da aptidão em escores padrões, com média = 0 e desvio padrão = 1, os valores de b_i tipicamente se situam entre -3 (itens fáceis) e +3 (itens difíceis); assim, conforme figura 4, o item 1 exige aptidão de mais ou menos -1,6 e o item 2 aptidão de 0,60, sendo este último mais difícil que o item 1.

O parâmetro de discriminação do item (isto é, o a_i) é representado pela inclinação da curva no ponto de inflexão, onde a probabilidade de resposta correta é 0,5. Na ilustração da figura 1-4, o item 2 é mais difícil (parâmetro b_2) que o item 1 (b_1), mas menos discriminativo, pois a inclinação da curva dele é menor que a do item 1 (parâmetro a_2 é menos íngreme que a_1 , isto é, seu ângulo de incidência no ponto de inflexão é mais agudo que do item 1). Veja na Figura 5 por que o ângulo de inclinação da curva característica expressa a discriminação do item.

De fato, para distinguir um sujeito que tem a chance de 70% de acertar o item com relação a um

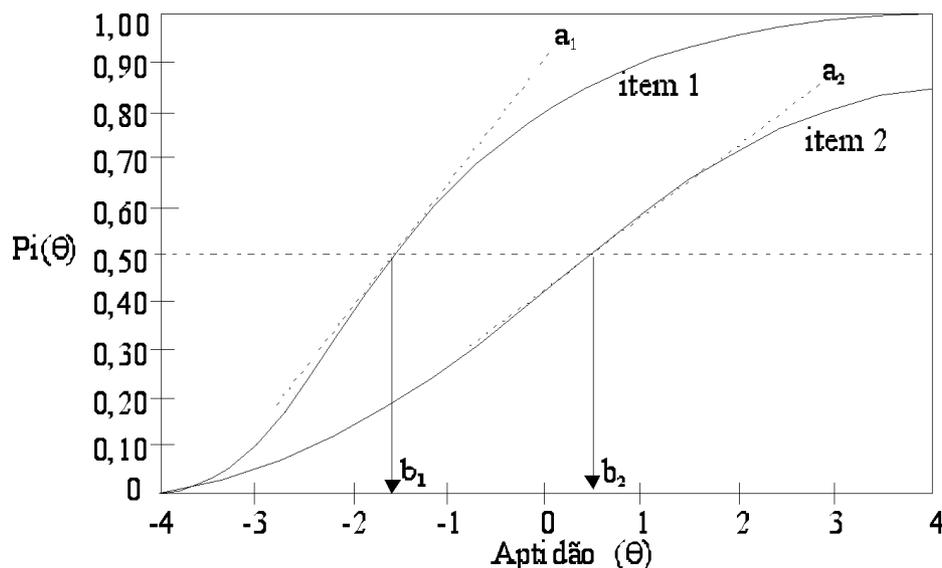


Figura 4 – Parâmetros de dificuldades (b) e discriminação (a) de dois itens

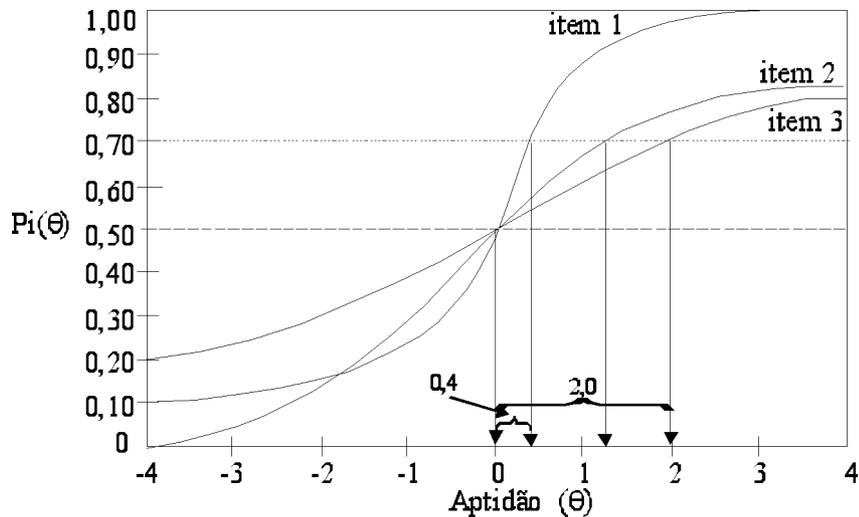


Figura 5 – Parâmetros de discriminação de três itens

que tem a chance de 50%, o item 1 precisa apenas de uma diferença de 0,40 desvios padrões de teta, enquanto o item 3 já precisa uma diferença de 2 desvios padrões de teta. Isto é, um sujeito, para aumentar suas chances de 50% para 70% de acertar o item, basta aumentar a sua aptidão (o teta) em 0,4 no caso do item 1 e em 2,0 no caso do item 3. Assim, o item 1 diferencia tamanhos de teta bem mais próximos do que o item 3 ou o item 2, isto é, reage a pequenas diferenças de teta e, conseqüentemente, ele é mais discriminativo de diferenças de níveis do teta do que estes dois últimos. Note também que todos os três itens têm a mesma dificuldade, ou seja, $b = 0$, mas seus índices de discriminação são muito

diferentes. Assim, se um item apresenta uma discriminação perfeita, então o ângulo de incidência da curva seria de 90 graus, ou seja, uma perpendicular. Neste caso, o item é capaz de discriminar diferenças infinitesimamente mínimas de níveis de teta. Veja isto na Figura 6.

Neste caso da Figura 6, os sujeitos à esquerda da perpendicular teriam chance 0 de acertar o item, enquanto os da direita teriam probabilidade 1 de acertar. Observe-se que em inglês estes dois parâmetros dos itens vêm indicados com diferentes nomes, como segue:

- Dificuldade: b , location, threshold, position;
- Discriminação: a , slope, inclination, dispersion.

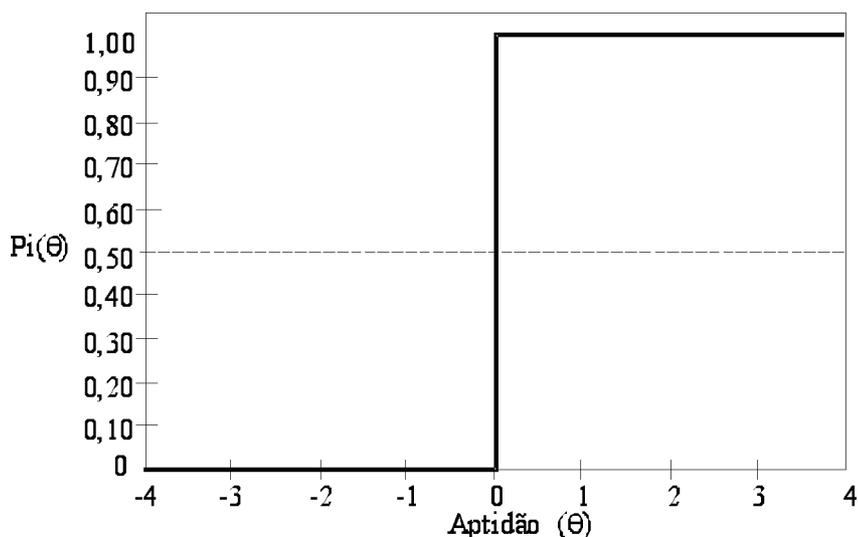


Figura 6 – Item com discriminação perfeita em $\theta = 0$

Referências

- Baker, F.B. (2001). *The basics of item response theory*. Washington, DC: ERIC.
- Binet, A. & Simon, Th. (1908). Le développement de l'intelligence chez les enfants. *Année Psychologique*, 14, 1-94.
- Birnbaum, (1957). Efficient design and use of test of a mental ability for various decision-making problems. (Series Report No. 58-16). Washington, DC: USAF School of Aviation Medicine.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring and examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, ch. 17-20.
- Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Campbell, D.T. & Stanley, J. (1973). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum.
- Guilford, J.P. (1936, 1954). *Psychometric methods*. New York: McGraw-Hill.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory. Principles and applications*. Boston, MA: Kluwer Nijhoff Publishing.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.
- Lawley, D.N. (1943). Richardson, M.W. (1936). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, Series A*, 61, 273-287.
- Lawley, D.N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh*, 62-A, 74-82.
- Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stauffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (Eds.), *Measurement and prediction*. Princeton, NJ: Princeton University Press, 1950.
- Lazarsfeld, P.F. (1959). Latent structure analysis. In S.Koch, *Psychology: A study of science*. New York: McGraw-Hill, 476-543.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lord, F.M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society.
- Lord, F.M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 17, 181-194.
- Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13, 517-549.
- Mislevy, R.J. & Bock, R.D. (1984). *BILOG: Maximum likelihood item analysis and test scoring logistic models*. Mooresville, IN: Scientific Software.
- Mulaik, S.A. (1972). *The Foundations of factor analysis*. New York: McGraw-Hill.
- Pasquali, L. (1996). Teoria da resposta ao item - IRT: uma introdução. In L. Pasquali (Org.), *Teoria e métodos de medida em ciências do comportamento*. Brasília: INEP, 173-195.
- Rasch, G. (1960, 1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.
- Reckcase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401-412.
- Richardson, M.W. (1936). Notes on the rationale of item analysis. *Psychometrika*, 1, 69-76.
- Richardson, M.W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1, 33-49.
- Samejima, F. (1972). *A general model for tree-response data* (Psychometric Monograph, No. 18). Psychometric Society.
- Samejima, F. (1974). Normal ogive model on the continuous response level in the multi-dimensional latent space. *Psychometrika*, 39, 111-121.
- Samejima, R. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Psychometric Society.
- Samejima, R. (1973). Homogeneous case of the continuous response model. *Psychometrika*, 38, 202-219.
- Thurstone, L.L. (1959). *The measurement of values*. Chicago, IL: Chicago University Press.

- Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529-554.
- Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11, 1-13.
- Weiss, D.J. (1983). Introduction. In David J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press, 1-8.
- Weiss, D.J. (Ed. - 1983), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.

Recebido em 17/12/2002

Aceito em 03/04/2003