Avanços na Psicometria: Da Teoria Clássica dos Testes à Teoria de Resposta ao Item

Advances in Psychometrics: From Classical Test Theory to Item Response Theory

Laisa Marcorela Andreoli Sartes & Maria Lucia Oliveira de Souza-Formigoni*
Universidade Federal de São Paulo, São Paulo, Brasil

Resumo
No século XX, o desenvolvimento e avaliação das propriedades psicométricas dos testes se embasou principalmente na Teoria Clássica dos Testes (TCT). Muitos testes são longos e redundantes, com medidas influenciáveis pelas características da amostra dos indivíduos avaliados durante seu desenvolvimento, sendo algumas destas limitações consequências do uso da TCT. A Teoria de Resposta ao Item (TRI) surgiu como uma possível solução para algumas limitações da TCT, melhorando a qualidade da avaliação da estrutura dos testes. Neste texto comparamos criticamente as características da TCT e da TRI como métodos para avaliação das propriedades psicométricas dos testes. São discutidas as vantagens e limitações de cada método.
Palavras-chave: Psicometria, teoria de resposta ao item, teoria clássica dos testes.

Abstract
In the 20th century, development and evaluation of psychometric properties of tests was mainly based on the Classical Test Theory (CTT). Many tests are long and redundant, with measures influenced by the characteristics of the sample of the individuals being evaluated. Some of the limitations are a consequence of the use of the CTT. The Item Response Theory (IRT) has been proposed as a solution to some limitations of the CTT, improving the quality of assessment of the tests structure. In this paper we critically compare the characteristics of CTT and IRT methods in determining the psychometric properties of tests. We discuss the advantages and limitations of each method.
Keywords: Psychometrics, Item Response Theory, classical test theory.

O uso de instrumentos e testes psicométricos representa uma importante forma de avaliação objetiva dos fenômenos psicológicos. Embora alguns psicólogos ainda apresentem resistência e exerçam críticas, o uso de testes psicológicos vem ganhando força. Isto se deve ao fato de que eles permitem o estabelecimento de um referencial que pode reduzir os vieses subjetivos da percepção e do julgamento do psicólogo examinador, desde que este tenha completo domínio e conhecimento de sua aplicação e interpretação. Considerando que o trabalho do psicólogo tem se desenvolvido em uma perspectiva multidisciplinar, esses profissionais têm, cada vez mais, atuado em diferentes contextos por solicitação de profissionais de formações variadas, que esperam resultados rápidos e objetivos, que possam contribuir na determinação das intervenções mais adequadas aos sujeitos.

O desenvolvimento de instrumentos de avaliação psicológica se iniciou no século XIX, paralelamente ao avanço da ciência positivista e da necessidade de medidas objetivas e válidas para o desenvolvimento de pesquisas clínicas. Surgiu daí a urgência por métodos que avaliassem as propriedades psicométricas dos instrumentos. As avaliações psicométricas dos testes passaram por diferentes fases, tendo início por volta da década de 1880 com Galton, atravessando as eras de Cattell, na década 1890, de Binet, na década de 1900, a era dos testes de inteligência, entre 1910 e 1930, a da análise fatorial e a era da sistematização, entre 1940 e 1980. Na era de Binet, a partir dos trabalhos de Spearman relacionados à correlação, desenvolveu-se a Teoria Clássica dos Testes (TCT; Pasquali, 1997). Nela se basearam a maioria dos métodos operacionais utilizados para avaliar as duas principais propriedades psicométricas dos instrumentos: a validade e a confiabilidade. A validade de um instrumento pode ser definida como sua capacidade de realmente medir aquilo a que se propõe a medir (Pasquali, 1997; Portney & Watkins, 1993). A validade de

* Endereço para correspondência: Departamento de Psicobiologia, Universidade Federal de São Paulo, Rua Napoleão de Barros, 1038, São Paulo, SP, Brasil 04024-003. E-mail: mlosformigoni@unifesp.br
Agradecimentos*:* Este trabalho contou com o apoio financeiro da *Fundação de Amparo* à Pesquisa do Estado de São Paulo (FAPESP, processo nº 05/60229-5 - bolsa de Doutorado para a primeira autora), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, bolsa de produtividade em pesquisa para a segunda autora) e da Associação Fundo de Incentivo à Pesquisa (AFIP). Agradecemos a valiosa colaboração dos estatísticos Dalton F. Andrade, Adriano Borgatto e Fernando A. B. Colugnati.

construto, mais especificamente, se refere à capacidade do instrumento de efetivamente medir um conceito teórico específico – o construto – seja este um processo psicológico ou uma característica dos indivíduos (Bruscato, 1998; Menezes & Nascimento, 2000; Strauss & Smith, 2009). A confiabilidade de um teste se refere à reprodutibilidade da medida, ou seja, o grau de concordância entre múltiplas medidas de um mesmo sujeito *inter* e *intra* indivíduos (Armstrong, White, & Saracci, 1994).

Segundo Streiner (2010), a avaliação psicométrica sofreu diversas modificações no último século, seguindo tendências observadas em diversos outros contextos nos quais aquilo que é maior é considerado melhor. Desta forma, anteriormente, quanto mais extenso fosse o instrumento para medir um construto, melhor eram suas propriedades de validade. Além disso, um maior número de itens contribuía positivamente no cálculo das fórmulas de confiabilidade, como no caso do conhecido *alpha* de Cronbach (Maroco & Garcia-Marques, 2006; Schmitt, 1996). Entretanto, nas últimas décadas, este conceito tem sido desafiado por uma corrente teórica conhecida como Teoria de Resposta ao Item (TRI). Esta corrente propôs o uso de escalas mais curtas que seriam tão confiáveis quanto as longas, se não mais. Esta era, que se iniciou a partir da década de 1980, é conhecida como a era da psicometria moderna, centrada na TRI.

Em vários países, assim como no Brasil, a TRI tem sido amplamente utilizada na avaliação de testes educacionais padronizados aplicados em grande escala, visando avaliar habilidades e conhecimentos em testes de múltipla escolha. No Brasil, desde o final da década de 90, tem surgido literatura sobre o tema (Pasquali, 1997, 2009; Pasquali & Primi, 2003), além de estudos que utilizaram a TRI para avaliação de testes relacionados à gestão pela qualidade, *marketing*, nutrição, qualidade de vida, psiquiatria, psicologia, raciocínio verbal e compreensão de leitura (Andrade, 2001; Andrade & Valle, 1998; Andriola, 1998, 2001; Bueno et al., 2009; Chachamovich, Fleck, Trentini, Laidlaw, & Power, 2008; Cúri, Singer, & Andrade, 2011; Fernandes, Prieto, & Delgado, 2010; Gabriel, Vasconcelos, Andrade, & Schmitz, 2009; Nunes, Muniz, Nunes, Primi, & Miguel, 2010; Polanczyk et al., 2010; Santos, Primi, Taxa, & Vendramini, 2002; Sisto, Santos, & Noronha, 2007; Soares, 2005; Vendramini, Silva, & Canale, 2004).

Mais recentemente, a TRI tem sido adotada por diversos pesquisadores para avaliação de instrumentos da área de saúde. Além da avaliação psicométrica dos testes, voltada para avaliação dos construtos, destaca-se nesta área o estudo da equivalência de grupos com diferentes características sociodemográficas, por meio do Funcionamento Diferencial do Item (DIF), o desenvolvimento de versões reduzidas dos instrumentos eliminando itens redundantes, e ainda o desenvolvimento de testes adaptativos computadorizados (CAT), que direcionam a apresentação dos itens a partir da identificação do nível de habilidade do indivíduo determinado pelas respostas aos itens anteriormente apresentados (Alterman, Cacciola, Habing, & Lynch, 2007; Cook et al., 2008; Jiang & Hesser, 2009; Kopec et al., 2008; Reise & Waller, 2009; Teresi & Fleishman, 2007; Wainer, 2000).

Para a Psiquiatria e Psicologia, o desenvolvimento da TRI veio ao encontro da tendência atualmente observada, tanto na área de pesquisa clínica, quanto na de assistência, da necessidade de utilização de instrumentos objetivos e breves, mas que possam abranger todo o construto a que se referem. Nesse sentido, a TRI trouxe grandes avanços para avaliação do construto dos instrumentos, proporcionando ainda a possibilidade de otimização de instrumentos que já foram validados por métodos clássicos. Vários pesquisadores utilizaram a TRI nas discussões sobre novos critérios diagnósticos adotados na (5ª) edição do DSM (DSM-V; Martin & Chung, 2008; Schuckit et al., 2005). Lima, Soares e Bacaltchu (2000) discutem a necessidade de pesquisas em Psiquiatria para aproximá-la dos avanços da medicina baseada em evidências, incluindo nisto a qualidade da avaliação dos critérios diagnósticos.

Embora a TRI tenha proporcionado avanços na validação de construto dos testes, esta técnica ainda apresenta certas dificuldades se comparada às técnicas tradicionais da TCT. Neste artigo temos por objetivo apresentar as principais características da TCT e da TRI e realizar uma comparação crítica da utilização de ambas na avaliação da validade de construto e como forma de otimização (ou redução) dos instrumentos.

Teoria Clássica dos Testes

A TCT compreende um conjunto de conceitos e técnicas que tem sido utilizado como base para o desenvolvimento de numerosos instrumentos de medidas e como ponto de referência para abordagens de medidas psicométricas (DeVellis, 2006). Uma vez que seus postulados são pouco exigentes e pouco numerosos, ela pode ser utilizada numa grande variedade de situações (Grégoire & Laveault, 2002). Segundo Grégoire e Laveault (2002), a TCT tem como objetivo fundamental o resultado total obtido por determinada pessoa em um teste, baseando-se em sete postulados, dos quais dois valem ser destacados aqui (para maiores informações sobre os outros postulados ver Grégoire & Laveault, 2002). O primeiro deles faz referência ao fato de que o resultado observado no teste do indivíduo (X) é resultante da soma do resultado verdadeiro do indivíduo (V: valor real), denominado por DeVellis (2006) "escore verdadeiro", com o erro de medida associado a esse mesmo resultado (E: variável aleatória). Uma fórmula simples pode representar o resultado observado: $X=V+E$. Outro postulado estipula que o valor esperado para o resultado observado é o resultado verdadeiro. Isto significa que a precisão de um resultado aumenta com o número de observações. Em resumo, os postulados da TCT assumem que os erros aleatórios da medida devem ser independentes, em qualquer circunstância.

Segundo a TCT, a análise dos itens é um procedimento que visa selecionar os melhores itens de um conjunto de

itens muitas vezes maior que o necessário, avaliando duas características dos itens: a dificuldade e a discriminação. Na TCT, o parâmetro de *dificuldade* do item é dado pela proporção dos indivíduos que respondem afirmativamente ao item, no caso de itens dicotômicos. Quando os itens são politômicos, a dificuldade é determinada pela proporção de respostas a uma categoria de escolha ou pela média das respostas de todos os indivíduos (DeVellis, 2006; Grégoire & Laveault, 2002). O *índice de dificuldade* varia entre 0 e 1, sendo que caso seu valor seja igual a zero, significa que nenhum indivíduo respondeu ao item afirmativamente. Ao contrário, se o índice for igual a 1, significa que todos responderam afirmativamente. O parâmetro de *discriminação* na TCT visa diferenciar o grupo de indivíduos que tiveram alta pontuação total dos que tiveram baixa pontuação total no teste. Alguns estudos utilizam a divisão dos indivíduos em dois grupos, os 27% superiores (que obtiveram os escores mais altos) e os 27% inferiores (que obtiveram os escores mais baixos), proposta por Kelley em 1939 (Grégoire & Laveault, 2002). O item é mais discriminativo quanto maior for o seu valor. O *índice de discriminação* pode assumir qualquer valor entre -1 e +1, correspondendo à diferença entre o *índice de dificuldade* dos indivíduos que obtiveram uma pontuação elevada no escore total do teste e o *índice de dificuldade* dos indivíduos que obtiveram uma pontuação baixa no escore total do teste. Entretanto, a interpretação dos *índices de dificuldade* e de *discriminação* depende das características da amostra de indivíduos utilizada para a normatização do teste. Dois fatores podem influenciar a interpretação do *índice de dificuldade*: o número de respostas omitidas e a probabilidade de um indivíduo responder corretamente ao item devido ao acaso. Sendo assim, se um grande número de indivíduos não responder ao item por falta de tempo, o *índice de dificuldade* não será avaliado corretamente.

Após o desenvolvimento da TCT, a partir dos trabalhos de Thurstone, por volta de 1930 foi desenvolvida a análise fatorial (AF), que é uma análise multivariada cujo principal propósito é avaliar a dimensionalidade dos instrumentos de medida. Ela abrange um conjunto de técnicas estatísticas cujo objetivo é representar (ou descrever) um grande conjunto de variáveis iniciais, utilizando um menor número de variáveis hipotéticas (fatores) obtidas como função das variáveis originais (Reis, 2001). O objetivo da AF é a parcimônia, procurando definir o relacionamento entre as variáveis de modo simples e usando um número de fatores menor que o número original de variáveis. Embora amplamente utilizada, a AF apresenta algumas limitações. Por exemplo, a AF confirmatória permite apenas verificar se os dados são consistentes ou não com a estrutura fatorial postulada. Quando os dados são compatíveis com várias estruturas latentes, a AF não permite determinar qual fator devemos escolher, sendo tal escolha feita em bases teóricas. Além disso, a AF apresenta dificuldades com conjuntos de dados de natureza mista, como ocorre em instrumentos que apresentam respostas do tipo dicotômica, numéricas e que utilizam escalas do tipo Likert ao mesmo tempo. Outra limitação da AF é relativa à suposição de linearidade na relação dos itens com o construto. Isto é, o construto é representado pela soma ponderada das respostas aos itens, sendo os pesos dados pelas cargas fatoriais. Por conseguinte, esta análise fornece informações necessárias, mas não suficientes (Grégoire & Laveault, 2002).

Limitações da Teoria Clássica dos Testes

Ainda que a TCT tenha sido muito útil para o desenvolvimento dos testes psicológicos e continue sendo largamente utilizada, suas limitações têm sido discutidas há muitas décadas (Gulliksen, 1950; Hambleton & Slater, 1997; Lord & Novick, 1968; Rasch, 1961). Atualmente, diversos trabalhos têm apresentado e discutido essas limitações, propondo novas formas de avaliação de medidas psicométricas (Andrade, Tavares, & Valle, 2000; De Champlain, 2010; Embretson & Reise, 2000; Hambleton & Slater, 1997; Pasquali & Primi, 2003; Vendramini et al., 2004). A principal limitação da TCT refere-se ao fato de que todas as medidas são dependentes da amostra dos indivíduos que responderam ao instrumento. Isto significa que as avaliações do teste são válidas somente se a amostra for representativa ou se o instrumento for utilizado em outra amostra com características semelhantes (Embretson & Reise, 2000; Hambleton & Slater, 1997). Além disso, testes diferentes com *índices de dificuldade* e *discriminação* diferentes geram resultados diferentes para os mesmos indivíduos. Uma terceira limitação é que na TCT, se o mesmo construto é medido por dois testes diferentes, os resultados não são expressos na mesma escala, impedindo uma comparação direta. Torna-se necessária a utilização de um método mais complexo de equalização para compará-los.

Outro problema é com relação à avaliação da fidedignidade, a qual exige que dois testes aplicados ao mesmo grupo, devam ter formas estritamente paralelas, ou seja, produzir escores verdadeiros idênticos e variâncias também iguais (Pasquali & Primi, 2003). Por fim, a TCT pressupõe que a variância dos erros de medida de todos os indivíduos é a mesma. Entretanto, deve-se considerar que alguns indivíduos realizam o teste de forma mais consistente do que outros (Hambleton & Swaminathan, 1985).

Outras limitações da TCT foram exploradas por Embretson e Reise (2000) como as suposições de que testes longos são mais confiáveis do que testes curtos e que as propriedades de um intervalo da escala avaliada são obtidas somente se a distribuição de freqüência dos escores for normal. Além disso, instrumentos que contém itens com formatos mistos, conduzem a um escore total dos indivíduos desbalanceado.

Teoria de Resposta ao Item

Considerando as limitações da TCT, após os anos 50 os psicometristas propuseram uma nova abordagem na tentativa de solucionar tais problemas. A TRI deve ser entendida como um conjunto de modelos psicométricos

para desenvolver e refinar medidas psicológicas (Embretson & Reise, 2000). Esta teoria visa superar as limitações da teoria clássica, e não entrar em contradição com seus princípios. Ao contrário, sugere uma nova proposta estatística, a de análise centrada nos itens, apresentando ainda novos recursos tecnológicos para a avaliação psicológica e educacional (Primi, 1998).

A TRI parte da suposição de que existe no indivíduo um traço latente, representado pela letra θ (teta), relativo a uma característica do indivíduo como, por exemplo, a gravidade da dependência de drogas, o nível de depressão, ou o quanto ele sabe sobre matemática. O traço latente pode ser entendido, em outros contextos, como o construto. O nível do traço latente do indivíduo (por exemplo: nível de habilidade em matemática, gravidade da depressão, etc) permite estimar se o indivíduo responderá positivamente ou não a um determinado item, estabelecendo para cada item uma probabilidade de resposta relacionada à presença (ou intensidade) da condição avaliada (Fletcher, 1994). Tal probabilidade é determinada pelas características (parâmetros) de cada item por meio de uma função matemática, tipicamente logística. O primeiro passo para aplicação da TRI é a estimação dos parâmetros dos itens, que pode ser feita por diferentes modelos estatísticos. A utilização dos modelos é baseada no número de populações envolvidas, na natureza das respostas ao item (dicotômico ou não dicotômico), na dimensionalidade do instrumento (Vendramini et al., 2004) e no número de parâmetros a ser avaliados.

Para itens dicotômicos, a função matemática é referida como uma Curva Característica do Item (CCI) e para itens politômicos, a relação entre o traço latente e a resposta a uma das categorias determinada pela CCI é conhecida como Curva de Resposta à Categoria (CRC). Utilizaremos neste texto a denominação CCI para ambos os casos. Um modelo comum aplicado a itens dicotômicos é o modelo logístico de dois parâmetros, o qual avalia as características de *dificuldade* e *discriminação* dos itens. Entretanto, diversos outros modelos são propostos como, por exemplo, o modelo logístico de um parâmetro, que avalia apenas a dificuldade dos itens e o modelo de três parâmetros que avalia ainda a probabilidade de acerto de um item ao acaso (Andrade et al., 2000; Reise & Waller, 2009). Dentre os modelos mais comuns para itens politômicos encontramos o modelo de resposta gradual (Samejima, 1969), que pode ser visto como uma generalização do modelo logístico de dois parâmetros. Além destes modelos paramétricos, podem ser encontrados na literatura vários modelos não paramétricos, dos quais a escala Mokken é o mais conhecido (Ark, 2007). As duas principais suposições para a utilização da maioria dos modelos da TRI são a unidimensionalidade e a independência local (Andrade et al., 2000; Embretson & Reise, 2000). Pasquali e Primi (2003) consideram que as suposições podem ser frustrantes para os pesquisadores porque elas não podem ser empiricamente demonstradas e nem possuem bases lógicas, isto é, elas são hipóteses. A unidimensionalidade é o postulado de que há apenas um traço latente, ou construto, responsável pelas respostas dos indivíduos aos itens. É plausível dizer que o comportamento humano é multideterminado, sendo assim, os modelos que utilizam essa premissa admitem que haja um traço latente dominante que determine as respostas ao teste (Andrade et al., 2000). Este traço é conhecido na análise fatorial como fator dominante. Alguns autores discutem outros métodos de avaliação da dimensionalidade além da análise fatorial clássica, como a análise fatorial tetracórica para itens dicotômicos, a policórica, para itens politômicos e os métodos de máxima verossimilhança (Andrade et al., 2000). Apesar disso, ainda há contradições entre os autores sobre o que seja a unidimensionalidade e como ela deve ser verificada (Pasquali & Primi, 2003). Modelos multidimensionais foram propostos para avaliar instrumentos que apresentam múltiplas dimensões de forma concomitante, sem a necessidade de se replicar um modelo unidimensional diversas vezes. Entretanto, tais modelos são complexos e ainda pouco estudados. Desta forma, para análise de instrumentos multidimensionais a maioria dos autores utiliza múltiplas análises unidimensionais, uma para cada dimensão. Deve-se considerar, que este procedimento tem como suposição básica a independência entre estas dimensões, o que nem sempre é real.

A suposição sobre a independência local postula que, se o traço latente é constante, as respostas do indivíduo aos itens são independentes (Andrade et al., 2000; Hambleton & Swaminathan, 1985; Reeve & Fayers, 2005) e dependem somente do nível do traço latente do indivíduo. Esta suposição é fundamental para a adequada estimação dos parâmetros dos itens e sua violação pode implicar em erro ao escolher os itens melhor estimados para a construção final do instrumento (Reeve & Fayers, 2005).

Segundo Hambleton, Swaminathan e Rogers (1991), a TRI traz ao menos cinco avanços em comparação com a psicometria tradicional. O primeiro deles refere-se ao cálculo do nível de traço latente do sujeito que independe da amostra de itens utilizados. Ao contrário do que ocorria na psicometria clássica, onde o escore do sujeito dependia e variava segundo o grau de dificuldade e precisão do instrumento aplicado, a TRI considera que o indivíduo possui um traço latente "verdadeiro" específico que não irá variar segundo os itens utilizados, desde que os itens estejam medindo o mesmo traço latente. Na TRI, o cálculo dos parâmetros dos itens independe da amostra de sujeitos utilizada, enquanto que na teoria clássica os parâmetros dependiam de os indivíduos da amostra possuírem maior ou menor nível no traço latente. Isto significa que mesmo que a amostra não seja representativa, os parâmetros dos itens serão estimados corretamente pela TRI (Embretson & Reise, 2000). Cabe ressaltar, porém, que a amostra necessária para avaliação dos modelos da TRI necessitam ser grandes, especialmente nos modelos de dois ou mais parâmetros. A TRI permite ainda posicionar os itens (considerando seu parâmetro de dificuldade) e os sujeitos (considerando seus escores) em uma mesma escala do traço latente que está

sendo avaliado. Desta forma, é possível estabelecer uma relação entre itens e níveis do traço latente dos indivíduos, identificando os itens que melhor avaliam cada nível do traço. Por exemplo, se considerarmos que um indivíduo está posicionado no nível 2,5 da escala do traço latente, é mais provável que ele responda positivamente aos itens posicionados neste mesmo nível e aos itens posicionados nos níveis anteriores e menos provável que responda positivamente aos itens posicionados acima desse nível. Assim, itens que são respondidos por um número maior de pessoas avaliam sujeitos com níveis menores do traço e itens mais difíceis discriminam os sujeitos com níveis maiores do traço latente. Outro avanço se refere ao fato de que a TRI constitui um modelo que não precisa fazer suposições improváveis, tais como a de que os erros de medida são iguais para todos os indivíduos. A TRI não exige que se trabalhe com instrumentos estritamente paralelos para avaliar a confiabilidade, como exige a teoria clássica. Além desses cinco avanços citados por Hambleton et al. (1991), a TRI permite que as respostas que não se aplicam a alguns entrevistados não sejam consideradas idênticas à resposta nula, distinguindo-se desta forma dos testes clássicos.

Principais Modelos da Teoria de Resposta ao Item (TRI)

Considerando que os modelos logísticos de dois parâmetros são os mais comumente utilizados e estudados, apresentaremos a seguir suas características. O modelo logístico de dois parâmetros deve ser utilizado quando os itens a serem avaliados são dicotômicos, como por exemplo, respostas do tipo "sim" ou "não" (Andrade et al., 2000; Pasquali & Primi, 2003). Este modelo é definido por

$$P(X_i = 1 \mid \theta, a_i, b_i) = \frac{1}{1 + e^{-1,7.a_i(\theta - b_i)}}$$

Nesta complicada fórmula estatística, o X_i representa a resposta ao item i (se X_i=1 a resposta indica uma resposta positiva, se X_i=0 a resposta é negativa), θ é o traço latente ou a gravidade de problemas, a_i representa o parâmetro de discriminação do item *i* e b_i o parâmetro de dificuldade do item *i*.

$P(X_i = 1 \mid \theta, a_i, b_i)$ é a probabilidade de um indivíduo responder 1 ao item *i*, dado o seu nível de traço latente e valores dos parâmetros *a* e *b* do item. A letra *e* representa uma constante com valor de 2,72.

Este modelo avalia a probabilidade de um indivíduo responder positivamente ("sim") a um item em função do seu traço latente, sendo tal probabilidade influenciada pela capacidade de discriminação do item (parâmetro *a*) e pela sua dificuldade (parâmetro *b*). O parâmetro *a* deve ser entendido como a capacidade do item de diferenciar indivíduos quanto à intensidade do traço latente e o parâmetro *b* representa o grau de dificuldade de o item gerar uma resposta afirmativa. A métrica do parâmetro *b* pode variar entre -∞ e +∞, mas comumente varia entre -4 e +4. Embora, teoricamente o parâmetro *a* também possa variar entre valores infinitos negativos e positivos, na prática, a métrica deste parâmetro é sempre apresentada com valor positivo (Pasquali & Primi, 2003).

Com base nos parâmetros *a* e *b* de cada item, é construída a CCI, um gráfico com uma curva de probabilidade em forma de S (Figura 1), no qual o eixo das abscissas representa a intensidade do traço latente dos indivíduos (nível de gravidade de problemas com álcool, por exemplo) e o eixo das ordenadas é a probabilidade de o indivíduo responder positivamente ao item (variando de 0 a 1). A curva indica que quanto maior o nível de gravidade do indivíduo, maior sua probabilidade de responder afirmativamente ao item. A CCI avalia, portanto, a probabilidade de resposta positiva a um dado item em função da intensidade (ou nível) do traço latente do indivíduo.

O valor de *a* corresponde ao ângulo de inclinação da curva, sendo as curvas mais inclinadas (com tendência a formar um ângulo de 90º) indicativas de maior discriminação, e as menos inclinadas indicativas de menor discriminação, formando curvas mais achatadas (Andrade et al., 2000). A grosso modo, o valor de *b* de um item indica a sua posição na escala do traço latente quando a probabilidade do indivíduo responder positivamente ao item for de 50%, conforme indicado na Figura 1. Neste exemplo, foi utilizada uma escala padronizada na qual a média é igual a zero e o desvio padrão é igual a um. O valor de *b*=0 indica que o item avalia um nível médio (moderado) de intensidade do traço latente. Por outro lado, itens com valores baixos ou negativos de *b* indicam que eles avaliam intensidade do traço latente menor do que a média. Entretanto, esta é uma simplificação didática, pois estes parâmetros podem sofrer a influência da habilidade do grupo e da dificuldade dos itens. Quanto maior for o valor de *b*, maior deve ser o nível de gravidade de problemas do indivíduo para que ele tenha a chance de 50% (0,5) de responder positivamente àquele item. Supõe-se então que um indivíduo que possui maior gravidade de problemas do que outro sujeito terá maior probabilidade de responder positivamente a um dado item que esteja posicionado nos níveis mais altos da escala do que o sujeito que possui menor gravidade de problemas. As respostas a diversos itens, posicionados em diferentes níveis da escala, contribuem para identificar o nível de gravidade do indivíduo. As pessoas com menor traço latente tendem a responder "sim" apenas aos itens com baixos valores de *b* e aquelas com maior traço latente tendem a responder "sim" tanto a itens com valores de *b* negativos ou baixos como a itens com valores de *b* mais elevados.

Diferentemente do que se pode esperar intuitivamente, o parâmetro de discriminação não distingue se o indivíduo apresenta ou não o traço latente daquele nível, mas permite diferenciar pessoas que apresentam níveis semelhantes do traço latente medido por aquele determinado item. Assim, quanto maior for o valor do parâmetro de discriminação, mais o item é capaz de detectar diferenças entre indivíduos.

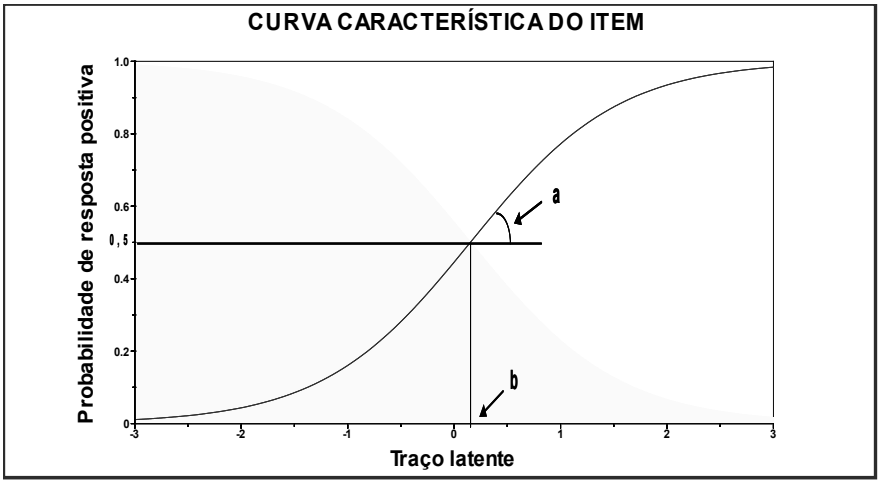

Figura 1. Exemplo de Curva Característica do Item (CCI) para itens dicotômicos, do modelo logístico de dois parâmetros. A CCI indica a probabilidade de resposta positiva a um dado item (eixo y) em função da intensidade (ou nível) do traço latente do indivíduo (valores de *b* no eixo x). A inclinação da curva indica a capacidade de discriminação do item e a projeção do valor de *a* no eixo x indica o valor de *b* para o item.

Para os itens politômicos, ou seja, aqueles que apresentam várias categorias como opção de resposta, deve ser utilizado um modelo adequado, como o modelo de resposta gradual, proposto por Samejima (1969). Para utilização deste modelo, as categorias de resposta devem ser ordinais e crescentes. Segundo Andrade et al. (2000), este modelo permite obter mais informação a partir das respostas dos indivíduos do que modelos dicotômicos, nos quais só existem dois tipos de respostas: afirmativas ou negativas.

No modelo, a probabilidade de um indivíduo *j* escolher uma categoria *k* (*k*=0, 1,...) ou outra mais alta do item *i* (na métrica normal) é dada por

$$P_{i,k}^{+}(\theta) = \frac{1}{1 + e^{-1,7.a_i(\theta_j - b_{i,k})}}$$

Neste caso, a fórmula deve ser interpretada como no modelo anterior. A letra *k* é a categoria do item *i*, $P_{i,K}(\theta)$ é a probabilidade do indivíduo com gravidade θ estar em uma particular categoria *k* ou outra mais alta. $b_{i,K}$ é o parâmetro de dificuldade da *k*-ésima categoria do item *i*.

Uma das principais suposições deste modelo é que as categorias de resposta de um item podem ser reagrupadas caso o parâmetro *a* indique pouca ou excessiva discriminação, ou seja, apresente valores menores do que 0,70 ou maiores do que 4,0. No modelo de resposta gradual, a CCI apresenta simultaneamente todas as curvas correspondentes à probabilidade de resposta afirmativa a cada uma das categorias. Este modelo fornece um único valor para o parâmetro *a* (englobando todas as categorias) e diferentes valores do parâmetro *b* (um para cada categoria), sendo mais comumente apresentada a média dos valores de *b*. Neste caso, o valor de *b* de cada categoria representa a localização das categorias no traço latente e a probabilidade de 50% de um indivíduo responder a essa categoria ou a categorias mais altas.

Para exemplificar, a Figura 2 mostra a CCI de um item com cinco opções de resposta, sendo possível visualizar que as probabilidades de um indivíduo responder afirmativamente nas categorias 1 e 5, são maiores do que nas categorias 2, 3 e 4. As curvas 2 e 3 estão praticamente sobrepostas, indicando que poderiam ser agrupadas, caso o valor de *a* indicasse baixa discriminação.

Embora os modelos possam ser usados separadamente, caso o instrumento possua itens de diferentes naturezas, os parâmetros dos itens politômicos e dicotômicos podem ser estimados conjuntamente de modo padronizado (escala com média=0 e *DP*=1). Inicialmente, devem ser estimados os parâmetros *a* e *b* incluindo todos os itens. A partir desta análise inicial, alguns itens dicotômicos podem ser retirados e alguns itens politômicos podem ter suas categorias de resposta reagrupadas, seguindo os seguintes critérios: (a) parâmetro *a*, de discriminação, menor que 0,70 ou maior que 4,0 e/ou; (b) parâmetro *b*, de dificuldade, maior que ± 4,0 e/ou; (c) estimativas com erros padrões muito grandes.

A seleção dos itens a serem incluídos no modelo deve ser feita de acordo com o objetivo do estudo (Embretson & Reise, 2000). Se o objetivo for, por exemplo, criar uma versão reduzida e representativa do instrumento, outra análise idêntica à primeira pode ser feita incluindo-se somente os itens mais discriminativos, adotando como ponto de corte $a \geq 0,70$.

Escala do Traço Latente (Níveis do Traço Latente)

Os modelos logísticos da TRI apresentados anteriormente geram uma escala final denominada "escala de níveis do traço latente". Podendo gerar uma interpretação

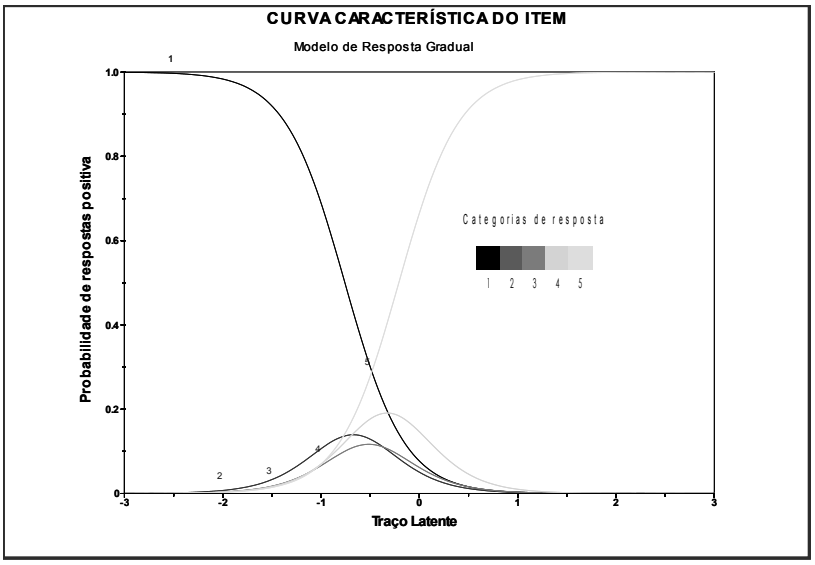

Figura 2. Exemplo de Curva Característica do Item (CCI) de um item politômico.

interessante das características de cada nível do traço latente, uma explicação sobre esta escala vale ser realizada. A escala gerada é padronizada (média = 0 e *DP*=1) e, assim como comentado em relação à métrica do parâmero *b*, em teoria, esta escala poderia variar de -∞ a +∞, mas na maioria dos softwares utilizados para análise com a TRI, é representada por oito níveis do traço latente variando entre -4 e +4 (Figura 3). Por exemplo, o nível 0 abrange os pontos da escala entre -0,6 e +0,5; o nível 1 abrange os pontos entre +0,6 e +1,5, enquanto o nível -1 abrange os pontos entre -0,6 e -1,5. Assim como o valor do parâmetro de dificuldade *b* indica a posição do item na escala do traço latente, na TRI posiciona-se também o indivíduo na mesma escala, de acordo com sua pontuação no traço latente avaliado. Os escores são estimados pela TRI utilizando o método da máxima verossimilhança, sendo atribuído a cada pessoa o escore que melhor identifica seu posicionamento na escala do traço latente.

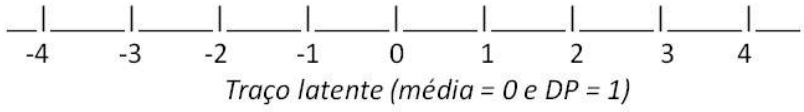

Figura 3. Representação da escala do traço latente.

Após a estimação dos parâmetros *a* e *b,* são realizados os cálculos das probabilidades de um indivíduo responder positivamente aos itens, a partir das quais os itens dicotômicos são posicionados nos níveis da escala do traço latente. Considerando que o modelo para itens politômicos calcula a probabilidade de resposta para cada categoria separadamente, o posicionamento dos itens politômicos deve ser feito a partir da média dos valores de *b* de cada item.

Portanto, a partir do posicionamento dos itens na escala, é possível identificar onde os itens se concentram caracterizando, assim, cada nível. Permite também avaliar qual a probabilidade de indivíduos que se encontram naquele nível do traço têm de responder positivamente àqueles itens. Alguns itens são mais característicos de determinados níveis e são denominados itens âncora. De acordo com Andrade et al. (2000), para um item ser considerado "âncora", mais do que 65% (0,65) das pessoas que se encontram no mesmo nível do traço latente do item devem tê-lo respondido afirmativamente e não mais do que 50% (0,50) das pessoas que se encontram no nível imediatamente inferior devem tê-lo feito. Além disso, a diferença de probabilidade de respostas afirmativas àquele item entre dois níveis subseqüentes deve ser de no mínimo 30% (0,30). Mesmo que um dado item não tenha sido considerado âncora, seu posicionamento na escala do traço latente pode ser feito baseado nesses critérios, ainda que nem todos tenham sido preenchidos.

Assim, quanto mais itens âncora são encontrados, melhor podemos caracterizar cada nível, sendo este um dos últimos passos da análise com a TRI. A interpretação de cada nível do traço latente, a partir dos itens que o compõe, além de facilitar sua compreensão, salienta-se como um interessante processo para facilitar aos clínicos a determinação do perfil dos indivíduos quanto ao traço latente que está sendo medido.

Discussão

Recentemente, o uso da TRI tem se expandido, levando diversos psicometristas e pesquisadores de diferentes áreas a utilizá-la para avaliação dos mais diversos tipos de instrumentos psiquiátricos e psicológicos. A TRI tem se mostrado como uma técnica atraente levando os pesquisadores a ansiar por adotá-la em suas análises, sem considerar que esta modalidade apresenta suposições muito mais fortes do que a TCT, além de ser mais complexa do ponto de vista matemático. De qualquer maneira, ao se observar as vantagens da TRI sobre a TCT, pode-se questionar por que a TCT continua ainda sendo mais utilizada. Alguns fatores podem explicar este fato.

Primeiro, a maioria dos cursos de graduação e até mesmo de pós-graduação continuam focando seu conhecimento sobre a TCT e pouco sobre a TRI. Segundo, porque a TCT é de mais simples entendimento do que a TRI. Um terceiro ponto refere-se ao tamanho da amostra, que necessita ser bem maior nas análises com a TRI do que com a TCT. Principalmente nos modelos mais sofisticados, em que são avaliados mais de um parâmetro dos itens ou itens politômicos, o uso da TRI pode se tornar inviável em alguns estudos. Um quarto aspecto, não menos importante, diz respeito ao fato de que os softwares mais comumente utilizados apresentam somente as análises clássicas, enquanto que os programas que permitem cálculos da TRI (como PARSCALE, MULTILOG, BILOG) são mais dispendiosos e menos disponíveis.

Em resumo, a TRI é um método mais complexo que exige do pesquisador disposição para sua compreensão, principalmente quando se trata de profissionais com formação humanística. Ambos os métodos (TRI e TCT) apresentam visões diferentes a respeito do problema, já que são baseados em padrões distintos do ponto de vista estatístico. Sendo assim, inicialmente, pode haver certa dificuldade na interpretação dos resultados da TRI por parte dos pesquisadores até então acostumados com o modelo da teoria clássica. No entanto, embora a TCT tenha sido capaz de desenvolver instrumentos bons o suficiente até então, o esforço para utilizar adequadamente a TRI pode ser compensado, considerando que possibilita uma melhor descrição das propriedades psicométricas dos instrumentos.

Não se pode deixar de ressaltar que embora a TCT apresente diversas limitações, o emprego deste método pode e deve continuar a ser utilizado na avaliação das propriedades psicométricas dos instrumentos. A TRI não deve ser entendida como um método que visa substituir a teoria clássica, mas sim, complementar suas análises. Desta maneira, o primeiro passo do pesquisador ao escolher o método mais adequado para avaliação de seu instrumento é definir seu objetivo, qual o tamanho de sua amostra, se seu construto pode ser unidimensional, qual método gostaria de usar, qual a natureza dos itens (dicotômicos, politômicos), enfim, se seu estudo apresenta as condições necessárias para aplicação de um ou de outro método. Nos casos em que o usuário deseja simplesmente classificar seu grupo de entrevistados e não visa generalizar seus dados para além da amostra utilizada, ou quando apresenta uma amostra pequena, a TCT pode ser bem empregada. Entretanto, caso o pesquisador perceba que a análise com a TCT não responde a todas as perguntas do estudo, pode-se pensar na aplicação da TRI, que poderá fornecer uma série de outras respostas e informações sobre o instrumento. No caso da opção pela TRI, o pesquisador deve ter conhecimento dos vários modelos propostos, e identificar aquele que melhor se adeque à sua proposta de estudo.

Por outro lado, pode ser muito útil que ambas as técnicas sejam aplicadas conjuntamente, o que possibilita uma avaliação estrutural mais completa do instrumento, além de minimizar as limitações de cada um dos métodos. Cabe salientar, nesse sentido, que a TCT e a TRI devem ser vistas como abordagens complementares, podendo fornecer informações úteis em diversas fases da análise.

A construção de instrumentos breves ou redução de instrumentos é um bom exemplo no qual se justifica a realização da análise conjunta. Considerando o pouco tempo disponível pelos profissionais para aplicação de instrumentos longos em ambientes (*settings*) clínicos, a construção de instrumentos breves ou redução de instrumentos já padronizados é bastante desejável. Sendo assim, ao visar a seleção de itens e avaliação dos construtos de um instrumento, a técnica clássica da análise de itens pode ser aplicada previamente para avaliar os parâmetros de dificuldade e discriminação e identificar os melhores itens por esta análise. Em outro contexto, a análise fatorial pode ser empregada para se obter o número de fatores do construto, ou avaliar se o instrumento apresenta unidimensionalidade, suposição exigida para aplicação da maioria dos modelos da TRI. Análises de descrição dos dados também são fundamentais para identificar possíveis erros de processo que podem ocorrer nas fases iniciais de tratamento dos dados. Posteriormente, a TRI pode ser aplicada para refinar as análises dos itens, avaliar os parâmetros de dificuldade, discriminação ou acerto ao acaso (de acordo com os objetivos do estudo), os escores dos indivíduos, descrever as características do diferentes níveis do traço latente, enfim, para complementar qualquer outra informação necessária ao estudo. Por fim, acreditamos que à medida que o uso da TRI se expandir e os programas de análise se tornarem mais acessíveis, baratos e amigáveis, novas escalas serão desenvolvidas usando esta família de técnicas.

Referências

Alterman, A. I., Cacciola, J. S., Habing, B., & Lynch, K. G. (2007). Addiction severity index recent and lifetime summary indexes based on Nonparametric Item Response Theory Methods. *Psychological Assessment*, *19*(1), 119-132.

Andrade, D. F. (2001). Comparando desempenhos de grupos de alunos por intermédio da Teoria de Resposta ao Item. *Estudos em Avaliação Educacional*, *23*, 31-69.

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria de Resposta ao Item: Conceitos e aplicações*. São Paulo, SP: Associação Brasileira de Estatística.

Andrade, D. F., & Valle, R. C. (1998). Introdução à Teoria de Resposta ao Item: Conceitos e aplicações. *Estudos em Avaliação Educacional*, *18*, 13-32.

Andriola, W. B. (1998). Utilização da Teoria da Resposta ao Item (TRI) para a organização de um banco de itens destinados a avaliação do raciocínio verbal. *Psicologia: Reflexão e Crítica, 11*(2), 295-308.

Andriola, W. B. (2001). Descrição dos principais métodos para detectar o Funcionamento Diferencial dos Itens (DIF). *Psicologia: Reflexão e Crítica*, *14*(3), 643-652.

Ark, L. A. (2007). Mokken Scale Analysis. *Journal of Statistical Software*, *20*(11), 1-19.

Armstrong, B. K., White, E., & Saracci, R. (1994). *Principles of exposure measurement in epidemiology*. Oxford, UK: Oxford University Press.

Bruscato, W. L. (1998). *Tradução, validade e confiabilidade de um inventário de avaliação de relações objetais (BORRTI--FORMA O)* (Tese de doutorado, Universidade Federal de São Paulo, SP, Brasil).

Bueno, J. M. H., Miguel, F. K., Primi, R., Muniz, M., Couto, G., & Noronha, A. P. P. (2009). Comparação entre dois sistemas de pontuação para o teste informatizado de percepções de emoção em fotos. *Estudos de Psicologia* (Campinas)*, 26*(1), 35-44.

Chachamovich, E., Fleck, M. P., Trentini, C. M., Laidlaw, K., & Power, M. J. (2008). Development and validation of the Brazilian version of the Attitudes to Aging Questionnaire (AAQ): An example of merging classical psychometric theory and the Rasch measurement model. *Health and Quality of Life Outcomes*, *21*(6), 5.

Cook, K. F., Choi, S. W., Crane, P. K., Deyo, R. A., Johnson, K. L., & Amtmann, D. (2008). Letting the CAT out of the bag: comparing computer adaptive tests and an 11-item short form of the Roland-Morris Disability Questionnaire. *Spine*, *33*(12), 1378-1383.

Cúri, M., Singer, J. M., & Andrade, D. F. (2011). A model for psychiatric questionnaires with embarrassing items. *Statistical Methods in Medical Research*, *20*(5), 451-470.

De Champlain, A. F. (2010). A primer on Classical Test Theory and Item Response Theory for Assessments in Medical Educacion. *Medical Education*, *44*, 109-117.

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care*, *44*(11), 50-59.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fernandes, D. C., Prieto, G., & Delgado, A. R. (2010). Construção de um teste transmodal de memória de reconhecimento. *Estudos de Psicologia* (Campinas), *27*(3), 283-291.

Fletcher, P. (1994). A teoria da resposta ao item: Medidas invariantes do desempenho escolar. *Ensaio: Avaliação e Políticas Públicas em Educação*, *1*(2), 21-28.

Gabriel, C. G., Vasconcelos, F. A., Andrade, D. F., & Schmitz, B. A. (2009). First law regulating school canteens in Brazil: Evaluation after seven years of implementation. *Archivos Latinoamericanos de Nutrición*, *59*(2), 128-138.

Grégoire, J., & Laveault, D. (2002). *Introdução às Teorias dos Testes em Ciências Humanas*. Porto, Portugal: Porto.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hambleton, R. K., & Slater, S. C. (1997). Item Response Theory Models and Testing Practices: Current international status and future directions. *European Journal of Psychological Assessment, 13*(1), 21-28.

Hambleton, R. K., & Swaminthan, H. (1985). *Item Response Theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Jiang, Y., & Hesser, J. E. (2009). Using Item Response Theory to analyze the relationship between health-related quality of life and health risk factors. *Preventing Chronic Disease*, *6*(1), A30.

Kopec, J. A., Badii, M., McKenna, M., Lima, V. D., Sayre, E. C., & Dvorak, M. (2008). Computerized adaptive testing in back pain: Validation of the CAT-5D-QOL. *Spine*, *33*(12), 1384-1390.

Lima, M. S., Soares, B. G. O., & Bacaltchu, J. (2000). Psiquiatria baseada em evidências. *Revista Brasileira de Psiquiatria*, *22*(3), 142-146.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Maroco, J., & Garcia-Marques, T. (2006). Qual a fiabilidade do *alpha* de Cronbach? Questões antigas e soluções modernas? *Laboratório de Psicologia*, *4*(1), 65-90.

Martin, C. S., & Chung, T. (2008). How should we revise diagnostic criteria for substance use disorders in the DSM-V? *Journal of Abnormal Psychology*, *117*(3), 561-575.

Menezes, P. R., & Nascimento, A. F. (2000). Validade e confiabilidade das escalas de avaliação em Psiquiatria. In C. Gorenstein, L. H. S. G. Andrade, & A. W. Zuardi (Eds.), *Escalas de Avaliação Clínica em Psiquiatria e Psicofarmacologia* (Vol. 1, pp. 23-28). São Paulo, SP: Lemos.

Nunes, M. F. O., Muniz, M., Nunes, C. H. S. S., Primi, R., & Miguel, F. K. (2010). Escala Fatorial de Socialização: Versão Reduzida: Seleção de itens e propriedades psicométricas. *Psicologia: Reflexão e Crítica, 23*(2), 345-353.

Pasquali, L. (1997). *Psicometria: Teoria e aplicações*. Brasília, DF: Editora da Universidade de Brasília.

Pasquali, L. (2009). *Psicometria: Teoria dos Testes na Psicologia e Educação*. Porto Alegre, RS: Vozes.

Pasquali, L., & Primi, R. (2003). Fundamentos da Teoria de Resposta ao Item – TRI. *Avaliação Psicológica*, *2*(2), 99-110.

Polanczyk, G., Laranjeira, R., Zaleski, M., Pinsky, I., Caetano, R., & Rohde, L. A. (2010). ADHD in a representative sample of the Brazilian population: Estimated prevalence and comparative adequacy of criteria between adolescents and adults according to the item response theory. *International Journals of Methods in Psychiatric Researchs, 19*(3), 177-184.

Portney, L. G., & Watkins, M. P. (1993). *Power analysis and determination of sample size. Foundations of Clinical Research: Applications to practice*. Norwalk, CT: Applenton & Lange.

Primi, R. (1998). *Desenvolvimento de um instrumento informatizado para avaliação do raciocínio analítico* (Tese de doutorado, Universidade de São Paulo, SP, Brasil).

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 321-334). Berkeley, CA: University of Chicago Press.

Reeve, B. B., & Fayers, P. (2005). Applying Item Response Theory modeling for evaluating questionnaire item and scale properties. In P. M. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods and practice* (Vol. 1, pp. 55-73. Oxford, UK: Oxford University Press.

Reis, E. (2001). *Estatística multivariada aplicada* (2. ed). Lisboa, Portugal: Edições Sílabo.

Reise, S. P., & Waller, N. G. (2009). Item Response Theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27-48.

Samejima, F. A. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, *17*.

Santos, A. A. A., Primi, R., Taxa, F. O. S., & Vendramini, C. M. M. (2002). O teste de Cloze na avaliação da compreensão em leitura. *Psicologia: Reflexão e Crítica*, *15*(3), 549-560.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350-353.

Schuckit, M. A., Smith, T. L, Danko, G. P., Kramer, J., Godinez, J, Bucholz, K. K., …Hesselbrock, V. (2005). Prospective evaluation of the four DSM-IV criteria for alcohol abuse in a large population. *The American Journal of Psychiatry, 162*, 350-360.

Sisto, F. F., Santo, A. A. A., & Noronha, A. P. P. (2007). Estudo sobre a dimensionalidade do teste R1 – Forma B. *Psicologia em Estudo*, *12*(1), 185-193.

Soares, T. M. (2005). Utilização da Teoria da Resposta ao Item na produção de indicadores sócio-econômicos. *Pesquisa Operacional*, *25*(1), 83-112.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1-25.

Streiner, D. L. (2010). Measure for measure: New developments in measurement and Item Response Theory. *La Revue Canadienne de Psychiatrie*, *55*(3), 180-187.

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research*, *16*(1), 33-42.

Vendramini, C. M. M., Silva, M. C., & Canale, M. (2004). Análise de itens de uma prova de raciocínio estatístico. *Psicologia em Estudo, 9*(3), 487-498.

Wainer, H. (2000). *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum.

Recebido: 30/05/2011
1ª revisão: 19/03/2012
Aceite final: 30/03/2012