2
Mensuração de variáveis latentes

Este capítulo trata das principais técnicas estatísticas empregadas na literatura para a medição de variáveis latentes.

2.1 Introdução

O ato de quantificar um construto latente, de forma indireta, através de um conjunto homogêneo de itens dicotômicos ou politômicos, com boas propriedades psicométricas, teve origem na Psicometria. Nesta área, desde o ano de 1904, várias técnicas estatísticas foram criadas com este objetivo. Por exemplo, a Análise Fatorial foi desenvolvida com o propósito de mensurar o traço latente associado a características psicológicas dos indivíduos, como pode ser visto em diversos estudos (Sperman, 1904a, 1904b, 1907, 1913; Fioravanti, 2006; Pasquali, 2009; Thurstone, *1935,* 1947; Burt, 1941; Viana, 2009). Essas técnicas ganharam aceitação em diversas áreas do conhecimento, entre elas, Enfermagem (Andrade *et al*, 2011; Araujo, Andrade e Bortolotti, 2009; Costa e Polak, 2009; Silva e Lacerda, 2010), Educação (Ribeiro e Soares, 2008; Ortigão, 2009), Psiquiatria (Artes, 1998; Pasquali, 1998), Cardiologia (Pinho *et al*, 2009), entre outras.

A teoria das variáveis latentes (Loehlin, 2004) apresenta vários modelos matemáticos para mensuração de um traço latente (construção de uma escala). Deste modo, ao invés do pesquisador ser obrigado a comparar inúmeras correlações - parciais e múltiplas - para descobrir padrões de homogeneidade nos itens de um instrumento de medida (Babbie, 1999), é possível empregar algumas dessas técnicas estatísticas multivariadas (Figura 2.1) que estão presentes nos programas estatísticos especializados como SPSS, R, MSP, dentre outros.

Estas técnicas podem ser classificadas em paramétricas e não paramétricas. Na primeira classificação estão incluídos os modelos nas versões exploratória e confirmatória da Análise Fatorial (AFE e AFC) (Gutierrez, 2005; Castillo, 2007, Vianna, 2009, Escocard, 2007), da Teoria de Resposta ao Item (TRI) (Andrade,

Tavares e Valle, 2000; Pasquali e Primi, 2003; Nunes, Sancineto e Primi, 2005; Soares, 2005; Guewhr, 2007; Justino e Andrade, 2007; Andriola, 2009; Tezza e Bornia, 2009) e os modelos da Teoria Clássica dos Testes (TCT) (Vendramini, Silva e Canale, 2004; Moura e Pasquali, 2006). No segundo grupo, os modelos da Teoria de Resposta ao Item Não Paramétrica (TRIN) (Andrade *et al*, 2011; Van Schuur, 2003; Gutierrez, 2005; Ferreira, 2008; Van der Ark, Croon e Sijtsma, 2008).

Ao utilizar estes modelos para construção de escalas é comum o uso de itens (questões) que são dicotômicos ou politômicos. O item politômico que compõem um instrumento de mensuração, geralmente, é medido em nível ordinal como uma escala de *Likert*. Desta forma, muitos pesquisadores assumem uma decisão prática: optam por considerar esta variável ordinal como intervalar, quando é razoável supor que esta tenha intervalos aproximadamente iguais. Diante desta conduta, os pesquisadores devem estar conscientes desta violação no uso das técnicas e confrontar a consistência dos resultados observados com as teorias que embasam o fenômeno que se pretende mensurar (Levin e Jack, 2004; Viana, 2009).

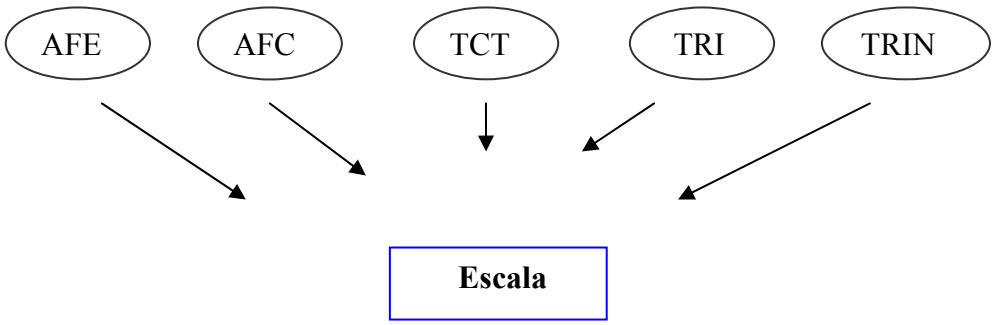

Figura 2.1: Técnicas estatísticas para a construção de escalas

2.2 Hipóteses dos modelos da TCT, TRI e TRIN

2.2.1 Unidimensionalidade

De acordo com a teoria psicológica, qualquer desempenho humano é sempre multideterminado ou multimotivado (Andriola, 2009; Pasquali, 2009). Neste sentido, no momento da resolução de questões de um teste de aptidão, por exemplo, várias dimensões do mesmo traço latente podem estar interferindo na sua mensuração. Por outro lado, mensurar aspectos multidimensionais de uma

variável latente, apesar de desejável do ponto de vista prático, é ainda um problema não solucionado de modo satisfatório na Psicometria moderna. Para contornar este detalhe, é postulado o enfoque unidimensional do teste.

Para definir o conceito de unidimensionalidade (homogeneidade) de um teste é assumida a presença de um fator dominante (dimensão única ou principal) em detrimento aos outros fatores secundários (possíveis dimensões do traço latente) que estejam em vigor no momento da mensuração e que possam ser considerados suficientemente pequenos para serem negligenciados da medição (Pasquali, 2009; Guewehr, 2007).

Este aspecto constitui uma das hipóteses relevantes que permeia os modelos matemáticos presentes nas técnicas de construção de escalas (ou testes) como TCT, TRI e TRIN e que deve ser examinado inicialmente no conjunto de itens de interesse. Nesta investigação preliminar podem ser empregados os modelos da análise fatorial exploratória, da análise fatorial *full information* (Vendramini, Silva e Canale, 2004) ou usar o modo exploratório para construção de escalas no enfoque da TRIN.

Além disso, a avaliação empírica da dimensionalidade do teste deve ser baseada num diálogo entre a teoria subjacente ao objeto de medida, a verificação da direção única dos itens no instrumento (itens positivos ou estritamente negativos) e os resultados das análises estatísticas. De tal modo, é possível investigar até que ponto as outras dimensões do traço latente (fatores secundários) possam ser consideradas desprezíveis (Junker e Stout, 1994; Mc Donald, 1994; Gessaroli, 1994; Guewehr, 2007; Ribeiro e Soares, 2008).

2.2.2 Independência local

Segundo Pasquali (2009), o postulado da independência local estabelece que "as respostas dos indivíduos a quaisquer dos itens do teste são estatisticamente independentes, já que são mantidas constantes (controladas) as aptidões que afetam o teste". De acordo com este autor, isto significa que "os itens são respondidos em função do traço latente dominante ou da "única" dimensão do traço latente dos sujeitos". Em outras palavras, este postulado nos diz que a probabilidade do indivíduo responder positivamente ao item *i*, não é afetada pelas respostas aos itens anteriores, nem tampouco a resposta dada ao item *i* afeta a

probabilidade de resposta aos itens subseqüentes. Nota-se que a unidimensionalidade e a independência local estão relacionadas. Desta forma, basta o pesquisador elaborar itens que satisfaçam à hipótese de unidimensionalidade (Guewehr, 2007).

Considere um conjunto de J itens dicotômicos tal que $i = 1, 2,..., J$ e seja β o traço latente dominante do indivíduo.

A independência local das respostas de dado indivíduo ao teste unidimensional com J itens dicotômicos é matematicamente definida por:

$$\Pr(U_1, U_2, ..., U_J | \beta) = \Pr(U_1 | \beta) \; P(U_2 | \beta) \ldots \Pr(U_J | \beta) = \prod_{i=1}^{J} \Pr(U_i | \beta) \qquad (2.1)$$

onde: U_i é a resposta positiva do indivíduo a um item dicotômico i e $\Pr(U_i | \beta)$ é a probabilidade da resposta positiva deste indivíduo ao item i dado seu traço latente β.

Na seqüência são descritas, de forma resumida, as técnicas mais utilizadas na literatura para a mensuração de variáveis latentes: Teoria Clássica dos Testes, Teoria de Resposta ao Item, Análise Fatorial com Informação Completa, Análise Fatorial Exploratória e Análise Fatorial Confirmatória.

2.3 Teoria Clássica dos Testes

A Teoria Clássica dos Testes também denominada de Psicometria Clássica ou Teoria Clássica das Medidas, abreviada como TCT - surgiu de forma axiomatizada nos trabalhos de Guilford (1936, 1950) e Gulliksen (1950) - com o intuito de produzir testes psicológicos e consequentemente uma medida de avaliação do sujeito pelo escore total observado no teste (a soma das alternativas de respostas de cada item respondido por cada participante do teste) (Pasquali, 2009).

Diante da abordagem da TCT, torna-se possível construir uma escala unidimensional e com boas propriedades psicométricas (estatísticas), mesmo de forma exploratória, a partir de um subconjunto de itens dicotômicos, politômicos nominais (questões de múltipla escolha pertencentes a um teste de habilidade) ou medidos na escala de *Likert* (politômicos ordinais). Para isto, a partir deste subconjunto de itens devem ser inspecionadas, além da hipótese de unidimensionalidade, as seguintes características: (1) dificuldade, (2)

discriminação, (3) fidedignidade (4) validade e (5) viés; sendo que as duas primeiras são aspectos individuais dos itens e as demais são características gerais do teste.

Na tradição positivista da Psicometria clássica para a construção de testes psicológicos, o teste (conjunto de itens) é construído através de um grupo de itens coletados de um universo (supostamente indefinido) de itens que parecem medir um dado construto. Diante deste ponto de vista, a hipótese da unidimensionalidade dos itens é assumida e em seguida é calculado o escore total observado do teste (Pasquali, 2009).

Hoje em dia, considerando esta abordagem para a construção de escalas (mensuração do traço latente), o pressuposto da unidimensionalidade deve ser avaliado previamente, antes do prosseguimento das análises das características dos itens e do teste. Em caso contrário, os resultados desta análise podem gerar conclusões errôneas (Vendramini, Silva e Canale, 2004; Quadros e Camey, 2010); já que todos os aspectos dos itens e do teste são calculados a partir do escore total observado no teste.

2.3.1 Parâmetros do item

2.3.1.1 Dificuldade

O parâmetro que avalia a dificuldade do item dicotômico, também conhecido como índice de facilidade ou índice de dificuldade, é dado pela proporção de acertos[1] do item (i.e. proporção de indivíduos que responderam corretamente ao item). Assim, à medida que a proporção de acertos se aproxima do valor 1, o item vai ficando cada vez mais fácil (Pasquali, 2009).

Cabe destacar que itens com nível de dificuldade nula ou igual a 1 devem ser removidos da análise, pois não colaboram com qualquer informação para a escala. No entanto, Pasquali (2009) sugere a seguinte distribuição em percentual para um conjunto de itens num teste, considerando 5 níveis de dificuldade: (I) 10%, (II) 20%, (III) 40%, (IV) 20% e (V) 10%. Em cada classe, a amplitude do

[1] Em testes de personalidade, atitudes, com itens dicotômicos, por exemplo, o acerto pode ser definido pelo pesquisador como a resposta à categoria de interesse no item.

intervalo é mantida no valor de 20. Dessa forma, o primeiro nível está representado pelos itens que possuem proporções de acertos no intervalo (0; 20]; enquanto que no último nível de dificuldade são incluídos os itens na faixa (80; 100]. Assim, os itens no teste apresentam uma distribuição mais equilibrada em termos de dificuldade.

2.3.1.2 Discriminação

A discriminação do item avalia a capacidade de diferenciar indivíduos com escore total observado alto no teste dos respondentes com escores baixos no mesmo teste. Para investigar estatisticamente este atributo num determinado item, a literatura (Pasquali, 2009; Soares, 2005) aponta duas direções complementares: (a) a criação de grupos-critério e (b) a correlação do item com o escore total observado dos itens.

A opção mais popular para criação de grupos-critérios é denominada de "regra 27". Neste procedimento, primeiramente são ordenados os indivíduos em termos do escore total observado. Em seguida, são escolhidos os participantes com os 27% escores mais altos (Grupo A) e os respondentes com os 27% escores mais baixos (Grupo B). Portanto, são formados os dois grupos. Em seguida, é computada em cada grupo a proporção de acertos e posteriormente, são empregados o *índice D* ou o *teste t de Student* (unilateral).

- O índice *D*

Calculado a partir da diferença entre as proporções de acertos do item *i* nos grupos A e B. Para que o item *i* seja discriminativo, o *índice D* deve ser positivo e elevado. Se D for nulo ou negativo, o item *i* é classificado como não-discriminativo. Cabe destacar que o uso deste índice foi utilizado no estudo de Vendramini, Silva e Canale (2004) para avaliar o poder de discriminação de itens de uma prova de estatística.

- O teste *t*

Baseado na distribuição *t* de *Student*, este teste unilateral pode ser empregado para avaliar se a média do grupo A é maior que a média do grupo B (Pasquali, 2009).

Diante da aplicação do teste *t*, um item é considerado não discriminativo ou inadequado, ao nível de significância α, quando o valor da estatística de teste ficar fora da região de rejeição do teste.

Convém destacar que as hipóteses de normalidade e variâncias iguais do escore total observado em ambos os grupos devem ser verificadas. Caso um dos pressupostos não seja satisfeito, é recomendável a escolha de um teste não paramétrico denominado como o teste de Wilcoxon (ou Mann-Whitney).

- Correlação ponto bisserial (r_{pb})

Quando o pesquisador escolhe avaliar o poder discriminativo de um item, sob a abordagem da TCT, através da correlação do item com o escore total observado dos itens (correlação item total), existem outras estatísticas mais populares que possuem resultados similares: *correlação ponto-bisserial*, *correlação bisserial*, correlação *phi* (Φ) ou *correlação tetracórica*. Para maiores detalhes consultar Pasquali (2009) e Soares (2005).

A correlação ponto bisserial (r_{pb}) do item *i* com o escore total observado do teste é definida pela seguinte expressão:

$$r_{pb} = \frac{\overline{M}_i - \overline{M}_{teste}}{d_{teste}} \sqrt{\frac{P_i}{1 - P_i}} \tag{2.2}$$

onde:

$\overline{M}_i$: média dos escores total dos respondentes que acertaram o item *i*; $\overline{M}_{teste}$: média do escore total observado de todos os respondentes; d_{teste} é o desvio-padrão do escore total observado do teste; P_i é a proporção de indivíduos que acertaram o item *i*.

De modo geral, itens com valores de correlação item-total inferiores a 0.30 são possíveis candidatos à exclusão (Paes, 2009; Castillo, 2007; Escocard, 2007; Vianna, 2009).

Na literatura ainda existem algumas controvérsias em escolher a TCT para uma análise sólida da qualidade dos itens e consequentemente a construção de uma escala. Dentre várias limitações que são discutidas a seguir, é importante destacar uma avaliação crítica do cálculo do índice de discriminação do item por esta teoria.

Segundo Pasquali (2009), na avaliação da discriminação de um item *i* existe a colaboração de outros itens (sem uma qualidade atestada previamente) que formam o escore total observado parcial. Mesmo que a hipótese de unidimensionalidade seja satisfeita no conjunto de itens e no cálculo do escore total observado parcial já esteja descontada a participação do item *i*, a presença de itens com baixa qualidade (em termos de dificuldade ou discriminação) pode mascarar o poder discriminativo do item *i* em questão. Na verdade, sob o enfoque da TCT, cada item depende dos outros itens do teste para sua caracterização individual.

Além disso, quando os itens são muito fáceis ou quando são muito difíceis, a construção dos grupos-critério para estabelecer o poder discriminativo do item, não é confiável, já que o índice de discriminação se aproxima de zero provocando talvez a exclusão errônea do item (Pasquali, 2009).

2.3.2 Parâmetros do teste

2.3.2.1 Fidedignidade

Com o intuito de avaliar o quanto um conjunto de itens *mede sem erros* um determinado construto de interesse, a Teoria Clássica dos Testes estabeleceu o conceito de fidedignidade (confiabilidade, consistência interna, precisão, etc.) que ainda é muito empregado atualmente (Dias e Vendramini, 2008; Pinho, Custódio *et al*, 2009).

Segundo Pasquali (2009), "medir sem erros significa que o mesmo teste, aplicado aos mesmos sujeitos em ocasiões diferentes; ou testes equivalentes medindo os mesmos sujeitos na mesma ocasião, produzem resultados idênticos. Assim, a correlação linear entre as duas medidas provenientes dos dois testes deve ser 1".

De modo geral, os coeficientes de confiabilidade existentes na literatura, que são casos específicos da fórmula do *alpha de Cronbach* (1951), a saber: Rulon (1939), Guttman (1945), Flanagan (1937), Kuder-Richardson (1937), etc. Além de produzirem conclusões semelhantes, tais coeficientes informam o grau de distanciamento da correlação linear máxima. Dessa forma, quanto mais o nível de precisão do teste se aproximar do valor unitário, mais fortes indícios que o mesmo esteja mensurando o construto de interesse com baixo nível de erro.

O coeficiente de Kuder-Richardson é recomendado quando os itens produzem respostas dicotômicas (Dias e Vendramini, 2008; Pinho, Custódio *et al*, 2009)

O índice de fidedignidade pode ser estimado através de duas medidas estatísticas: a correlação linear e o *alfa de Cronbach*. De acordo com o objetivo da pesquisa, este coeficiente será estimado de acordo com um dos três tipos de coleta de informações empíricas (delineamentos).

Os tipos de coleta de informações empíricas podem ser classificados, em termos de popularidade, como: (a) uma amostra (representativa) de sujeitos, um único teste e uma única ocasião, (b) uma amostra de sujeitos, dois testes e um único período; (c) uma amostra de indivíduos, uma única aplicação do teste e duas ocasiões (Yoshida e Colugnati, 2002).

Para os delineamentos (b) e (c), o índice de confiabilidade é estimado pela correlação linear. No entanto, a medida de correlação adequada para o caso (c) deve incorporar o efeito longitudinal.

Por ser o delineamento (a) o mais popular no contexto das mensurações de variáveis latentes presentes nas ciências humanas, torna-se necessário um maior aprofundamento no coeficiente mais conhecido de consistência interna denominado de *alfa de Cronbach* (1951). Este coeficiente é uma função que depende do tamanho do teste, da variância de cada item e da variância do escore total observado. Sua expressão é dada por:

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum \dot{s}_i^2}{\dot{s}_T^2}\right) \qquad (2.3)$$

onde: $\Sigma \dot{s}_i^2$: total das variâncias dos J itens e $\dot{s}_T^2$: variância total dos escores do teste.

Além disso, esta estatística que varia no intervalo fechado [0; 1][2] procura verificar o grau de covariância (homogeneidade) dos itens entre si, ou ainda, a congruência que cada item do teste tem com o restante dos itens do mesmo teste. Neste sentido, quanto menos variabilidade um mesmo item produz numa amostra de indivíduos, menos erros ele provoca e consequentemente mais preciso é o item.

[2] O limite inferior aceitável para o *alfa de Cronbach* é 0.70.

De modo geral, se os itens se comportarem com pouca variabilidade, o teste torna-se mais consistente e preciso (Pasquali, 2009; Thurstone, 1927, Erthal, 2009).

Na análise inicial para avaliar a qualidade dos dados pode ser calculado o *alfa de Cronbach*. Quando o resultado for inferior a 0.60, pode ser um indício de dados mal coletados (Castillo, 2007) ou a violação na hipótese de unidimensionalidade do teste. No segundo caso, é recomendável usar uma das técnicas da análise fatorial exploratória, por exemplo, para avaliar esta possível violação.

Este coeficiente de consistência interna empregado conjuntamente com a medida de correlação item-total, também denominado de "*alfa de Cronbach se o item é excluído*", permite avaliar até que ponto a consistência interna do teste é afetada pela remoção do item com baixo poder de discriminação (Castillo, 2007; Paes, 2009; Vianna, 2009, Dias e Vendramini, 2008). Isto constitui uma maneira exploratória de construção de escalas pela abordagem da TCT que é considerada por Van Schuur (2003) como uma construção de cima para baixo.

Além das questões ambientais, psicológicas e o tempo para realização do teste, a quantidade e a qualidade dos itens envolvidos, e o tamanho da amostra de participantes são fatores que interferem diretamente no índice de fidedignidade do teste (Castillo, 2007; Paes, 2009; Pasquali, 2009; Erthal, 2009).

Segundo Pasquali (2009), à medida que aumenta o número de itens em uma escala com boas propriedades psicométricas, maior será sua precisão. Por outro lado, um instrumento de medida com elevado número de questões pode desencadear reações desagradáveis (como fadiga, desinteresse, etc.) no grupo de respondentes e consequentemente diminuição na qualidade da medição. Diante deste dilema, não há um consenso na literatura sobre o tamanho ideal de um teste.

Além destes fatores descritos que influenciam diretamente na fidedignidade de um teste, convém destacar que itens muito fáceis ou muito difíceis também colaboram para diminuição da precisão. Erthal (2009) sugere que os itens de dificuldade média são os mais adequados, já que favorecem a variabilidade dos escores. De modo geral, cabe ao pesquisador construir um teste equilibrado em termos das características dos itens (dificuldade e discriminação), além de minimizar os outros fatores.

Green e Jang (2009), Paes (2009), Sijtsma (2009) e Bentler (2009) não recomendam o uso do índice de fidedignidade *alfa de Cronbach*, pelo fato desta medida não avaliar adequadamente a consistência interna de um teste.

2.3.2.2 Validade

Nesta etapa considera-se que o conjunto de itens (escala) já possui boas propriedades psicométricas tanto a nível individual quanto global (*unidimensionalidade* e consistência interna satisfatória). Dessa forma, para investigar até que ponto este teste mede o que supostamente pretende medir, torna-se indispensável à análise de validade do teste (Apolinário, Silva e Cardoso, 2011; Castillo, 2007; Paes, 2009; Pasquali, 2009; Vianna, 2009; Erthal, 2009; Anastasi e Urbina, 2000). Esta investigação pode ser conduzida por três aspectos complementares, a saber: validade de conteúdo, validade de construto (ou conceito) e validade de critério. Para o esclarecimento deste último aspecto consultar Pasquali (2009), Anastasi e Urbina (2000), Erthal (2009), dentre outros.

Na validade de conteúdo os itens que compõem a escala são avaliados por uma comissão de especialistas (juízes) que verificam se as questões selecionadas representam o comportamento que se quer medir (Paes, 2009).

Na validade de construto, segundo Vianna (p.80, 2009), avalia-se "se as variáveis que formam o teste apresentam o verdadeiro significado teórico do conceito que o instrumento deseja abarcar. A avaliação da validade de construto pode ser feita através da comparação com outros instrumentos de medida para aquele mesmo conceito. O que se espera é que o instrumento testado apresente uma correlação alta com o outro (validade convergente). Uma outra maneira de assegurar que uma escala é adequada para avaliar um determinado conceito é compará-la com um outro instrumento que apresenta uma finalidade diferente. Neste caso, não é esperada uma correlação alta entre os dois instrumentos. Este procedimento é denominado de validade divergente ou discriminante. Para ambos os propósitos da investigação da validade de construto pode ser utilizado o teste de correlação de Spearman".

No estudo desenvolvido por Vianna (2009) os instrumentos MASC-VB (*escala multidimensional de ansiedade para crianças*) e RCMAS (*Revised Manifest Anxiery Scale for Children*) foram utilizados para medir a ansiedade

infantil, enquanto o *Inventário de Depressão Infantil* (CDI) foi empregado para mensurar a depressão em crianças. Os dois primeiros instrumentos evidenciaram ao nível de 5% uma correlação alta, ou seja, validade convergente satisfatória. Quando comparados individualmente com o inventário CDI, há evidências ao nível de 5% de boa validade divergente para cada escala.

De acordo com Pasquali (2009), é recomendável o uso de mais de uma técnica para investigar a validade de construto de um teste, dado que a convergência de resultados das várias técnicas constitui uma garantia para a validade de um instrumento.

Segundo Erthal (2009), "a fidedignidade e a validade de um teste são dois conceitos intimamente relacionados que denotam a eficiência de um teste" e assim estes aspectos devem considerados na construção de um bom instrumento de mensuração.

2.3.3 Limitações da TCT

De acordo com Pasquali (2009) e Pasquali e Primi (2003) a Teoria Clássica dos Testes apresenta algumas limitações, a saber:

- De modo geral, os testes elaborados sob este enfoque são dependentes dos itens que os compõem e consequentemente a validade destes instrumentos de medida é limitada.
- Coletadas duas amostras[3] probabilísticas em duas ocasiões t_1 e t_2, não é possível fazer comparações entre os grupos, considerando o mesmo teste aplicado.
- Suposição da variância constante dos *erros de medida* para todos os respondentes do teste. A partir desta hipótese é definido conceito de testes com formas paralelas e posteriormente a definição de fidedignidade do teste. Como esta hipótese é pouco provável na prática, ambos os conceitos descritos acima ficam comprometidos.
- Não é possível posicionar o item nem o respondente na escala obtida pelo escore total observado.

[3] Homogêneos em termos das características de interesse do fenômeno psicológico subjacente ao teste.

- Quando dois itens apresentam a mesma proporção de acertos, isto dificulta o ordenamento dos itens pelo parâmetro de dificuldade. Esta limitação pode ser contornada ao empregar o *modelo de Rasch* para a estimação desse parâmetro, como foi apontado no estudo de Quadros e Camey (2010).
- Os parâmetros descritivos dos itens (dificuldade e discriminação) e do teste (escore total observado) são afetados pelo grupo de sujeitos examinados pelo teste.

Convém ressaltar que esta última limitação pode ser corrigida a partir do uso de amostras probabilísticas representativas da população de indivíduos a que se destina o teste.

Apesar da Teoria Clássica dos Testes apresentar algumas limitações, os parâmetros descritivos dos itens, os coeficientes de consistência interna *alfa de Cronbach* e Kuder- Richardson; são aspectos desta teoria que são empregados na prática para a construção de uma escala (Yoshida e Colugnati, 2002; Dias e Vendramini, 2008; Ribeiro e Soares, 2008; Pinho, Custódio *et al* 2009; Filgueiras, 2011; Borges e Pasquali, 2011, Vendramini, Dias e Canale, 2004).

2.4 Teoria de Resposta ao Item

A Teoria de Resposta ao Item (TRI), também denominada de Psicometria Moderna, surgiu com estudos de Richardson (1936), Lawley (1943, 1944), Tucker (1946), Lazerfeld (1950, 1959), Lord (1952, 1953), Rasch (1960), Samejima (1969, 1972) e foi axiomatizada por Birnbaum (1957, 1968) e Lord (1980) com o intuito de solucionar algumas limitações da Teoria Clássica dos Testes (TCT) (Hambleton *et al*., 1991; Pasquali e Primi, 2003; Pasquali, 2009), a saber:

- O conjunto de itens bem como os parâmetros dos itens (dificuldade e discriminação) e a estimativa do traço latente são dependentes da amostra de respondentes e do instrumento de medida.
- Não posicionamento dos itens e das estimativas do traço latente na escala produzida.

2.4.1 TRI no Brasil

Diferente da Teoria de Resposta ao Item não Paramétrica, desde a década de 80, a TRI tem se tornado a técnica predominante no contexto da construção de testes e escalas, apesar de fazer algumas suposições para a estimação dos parâmetros dos modelos matemáticos, como será visto mais adiante. Além disso, esta predominância pode ser observada tanto no número crescente de artigos científicos nos periódicos especializados, no desenvolvimento de novos modelos matemáticos[4] e da sua utilização dominante na área de avaliação educacional.

No contexto da avaliação educacional brasileira a participação desta teoria vem crescendo desde 1995, com a construção de testes e banco de itens calibrados para avaliar o desempenho dos alunos em diversas disciplinas e séries. Várias avaliações em larga escala utilizam esta metodologia: o Sistema de Avaliação da Educação Básica (SAEB), o Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (SARESP), o Programa de Avaliação Educacional do Estado de Minas Gerais (SIMAVE) e a Prova Brasil (MEC, 2007), dentre outros.

2.4.2 Modelos da TRI

Inicialmente, esta teoria considerava como ponto de partida um conjunto de itens dicotômicos de um teste de habilidade (desempenho, aptidão). Através de modelos probabilísticos (normal e logístico) relacionava as variáveis observáveis (respostas aos itens) com traços hipotéticos não observáveis (traço latente, aptidão, etc.). Hoje em dia, de uma forma mais abrangente, são considerados modelos matemáticos (acumulativos ou não-acumulativos) e também para dados oriundos de amostras complexas (Thomas, 2001; Thomas e Cyr, 2001; Carle, 2007). Maiores detalhes sobre modelos não acumulativos veja Andrade *et al* (2000).

A Teoria de Resposta ao Item assume algumas suposições básicas para utilização dos seus modelos acumulativos, de acordo com Andrade *et al* (2000), Pasquali e Primi (2003) e Pasquali (2009):

a) Unidimensionalidade.

[4] Modelos matemáticos que incorporam as observações que são oriundas de amostras complexas.

b) Independência local.

c) O comportamento (resposta ao item do teste) é uma função do traço latente β. Esta relação pode ser descrita por uma função logística apropriada na escala de aptidão β, como mostra a Figura 9.2. Desta forma, a probabilidade de acerto ao item *i*, é modelada por esta função que incorpora os parâmetros do item *i* e o traço latente do indivíduo (estimado pelo teste). Assim, à medida que o valor do traço latente cresce a probabilidade de acerto ao item *i* aumenta.

2.4.2.1 Escala do traço latente

A escala de habilidade (traço latente) é uma reta real na qual são posicionadas as estimativas do traço latente, os parâmetros dos itens e posteriormente os itens. Na prática, admite-se que a origem desta escala (ou métrica) é o valor zero (valor médio do traço latente na população em estudo) e o desvio-padrão, o valor unitário. Desta forma, muitos autores definem esta escala como métrica usual ou escala (0; 1).

2.4.2.2 Alguns modelos unidimensionais

Para a medição de traços latentes como inteligência (Quadros e Camey, 2010), raciocínio estatístico (Vendramini, Silva e Canale, 2004), proficiência em Matemática (Andrade, 2003; Rodrigues, 2006; Conde, 2007) e condição socioeconômica (Soares, 2005); podem ser empregados modelos estatísticos específicos que levam em consideração o traço latente β e os parâmetros dos itens. Nesse contexto, estes modelos acumulativos necessitam de um teste unidimensional com itens dicotômicos e são apresentados na literatura como *modelo de Rasch* (Rasch, 1960), *modelo logístico de dois parâmetros* (2PL) (Birnbaum, 1968) e *modelo logístico de três parâmetros* (Lord, 1980). Apenas o modelo 2PL será abordado neste capítulo.

No *modelo logístico de dois parâmetros*, além do aspecto de dificuldade do item, é acrescentado o parâmetro *a* (discriminação) do item.

Para $i = 1, 2, \ldots, J$ e $k = 1, 2, \ldots, N$ sua expressão é dada por:

$$\Pr(U_{ik} = 1 | \beta_k) = \frac{e^{Da_i(\beta_k - b_i)}}{1 + e^{Da_i(\beta_k - b_i)}}, \qquad (2.4)$$

na qual:

a_i: parâmetro de discriminação do item *i*;

$\Pr(U_{ik} = 1|\beta_k)$: probabilidade de acerto do item *i*, dado o traço latente β_k do *k-ésimo* indivíduo, a dificuldade e a discriminação do item *i*;

J é o número de itens no teste;

b_i : parâmetro de dificuldade do item *i;*

D: uma constante que assume os valores 1.7 ou 1.

Cabe enfatizar que o ajuste do modelo 2PL permite, além dos aspectos de cada item (dificuldade e discriminação) representados graficamente na curva *CCI*, avaliar a contribuição do item (curva de informação do item) e a informação total dos itens (curva de informação do teste).

A curva de informação do item *i* apresenta o intervalo no eixo do traço latente β no qual existe mais informação (contribuição) deste item para a medida estimada do traço latente. Esta contribuição se apresenta em torno do valor do parâmetro de dificuldade (b) do item *i*.

A curva de informação do teste (informação total) mostra o intervalo, também no eixo do traço latente, onde existe mais contribuição do conjunto de itens para a escala. De acordo com Andrade et al (2000), a informação total fornecida para a escala é simplesmente a soma das informações produzidas por cada item.

2.4.2.2.1 Parâmetros do item

Nessa breve exposição, maior ênfase será dada para o modelo acumulativo logístico unidimensional de dois parâmetros. Desta forma, os parâmetros deste modelo: a dificuldade e a discriminação do item são abordados nesta subseção e podem ser visualizados através da Curva Característica do Item (CCI), como ilustra a Figura 9.2:

1. A dificuldade do item *i* (parâmetro b_i) é o ponto na escala de habilidade na qual a probabilidade de acerto é 0.5. Considerando a métrica usual, os valores (típicos) de b_i variam no intervalo [-3; 3]. Para os itens fáceis, os valores de b_i estão próximos de -3 enquanto que os itens muito difíceis possuem valores de b_i perto de 3.

2. A discriminação do item *i* (parâmetro a_i) é estabelecida pela inclinação de uma reta tangente ao ponto de inflexão[5] da curva CCI. Na prática, na métrica usual, Baker (2001) apresenta a seguinte classificação do parâmetro de discriminação por faixa de valores: 0 (*nenhuma discriminação*); de 0.01 até 0.34 (*discriminação muito baixa*); de 0.35 até 0.64 (*discriminação baixa*); de 0.65 até 1.34 (*discriminação moderada*); de 1.35 até 1.69 (*discriminação alta*); maior que 1.70 (*discriminação muito alta*). Vale ressaltar que os valores a_i >1 são os mais apropriados.
3. O parâmetro de discriminação (*a*) do item pode ser interpretado da seguinte forma de acordo com Andrade et al. (2000):
   - Itens com parâmetro $a < 0$ não são esperados, uma vez que indicariam que a probabilidade de responder corretamente ao item diminui com o aumento do traço latente.
   - Baixos valores de *a* indicam que o item tem pouco poder de discriminação, ou seja, tanto indivíduos com baixo traço latente quanto indivíduos com alto traço latente têm praticamente a mesma probabilidade de responder corretamente ao item.
   - Valores muito altos do parâmetro *a*, por sua vez, indicam itens com curvas características (CCI) muito "íngremes", que discriminam os indivíduos basicamente em dois grupos: os que possuem habilidade abaixo do valor do parâmetro *b* e os que possuem habilidades acima do parâmetro *b*.

2.4.2.2.2 Estimação dos parâmetros dos modelos unidimensionais

Na prática, a estimação dos modelos unidimensionais logísticos, a partir das respostas de uma amostra aleatória simples de indivíduos que participaram de um teste de habilidade, pode ocorrer de 3 formas diferentes: a estimação das habilidades quando já se conhecem os parâmetros dos itens; a estimação dos parâmetros dos itens quando já são conhecidas as estimativas das habilidades; a estimação conjunta dos parâmetros dos itens e das habilidades.

[5] Onde a probabilidade atinge o valor de 0.5.

Os métodos empregados para a estimação podem ser bayesianos ou baseados no método de *máxima verossimilhança* (Lord, 1986). Para o processo de estimação conjunta dos parâmetros dos itens e das habilidades pode ser empregado o *método de máxima verossimilhança marginal* (Bock e Lieberman,1970) ou o método de *máxima verossimilhança conjunta* ou métodos bayesianos. Indiferente do método escolhido, este processo de estimação ocorre em duas etapas. Na primeira, são estimados os parâmetros dos itens. Em seguida, é efetuada a estimação dos traços latente dos indivíduos, considerando as estimativas dos parâmetros dos itens. Para tal propósito, os métodos de estimação descritos acima estão disponíveis nos seguintes pacotes estatísticos: BICAL (Wright *et al*, 1979), LOGIST (Wingersky, Barton e Lord, 1982), BILOG (Mislevy e Bock, 1984), RASCAL (1995), XCALIBRE (1995) e R (versão 2.18, com as bibliotecas *ltm* e *irtoys*).

Convém destacar que a etapa preliminar da estimação dos parâmetros dos modelos unidimensionais logísticos, em particular do modelo 2PL, é baseada na proporção de acertos de cada item e na correlação bisserial dos itens que são os aspectos da análise da Teoria Clássica dos Testes (Vendramini, Silva e Canale, 2004) e também considerados como valores iniciais na estimação.

Segundo Pasquali (2009), é importante fazer a avaliação estatística da adequação dos ajustes dos modelos aos dados empíricos utilizados na mensuração do traço latente β. Por exemplo, um modelo logístico de um parâmetro é adequado (pelo aspecto do ajuste), quando os valores estimados da probabilidade $\Pr(U_{ik}=1|\beta,b_i)$ não diferem significativamente dos valores da proporção de acertos do item *i* que foram obtidos empiricamente.

Na literatura, apesar de existirem muitos procedimentos estatísticos para uma avaliação completa dos pré-requisitos do modelo (unidimensionalidade, independência local, invariância das estimativas do traço latente, invariância das estimativas dos parâmetros dos itens, ajuste do modelo, etc.), ainda não existe um consenso. As mais usuais para a verificação de ajuste do modelo são: análise dos resíduos e uma estatística de teste semelhante ao *Qui-quadrado* (Muñiz, 1990; Pasquali, 2009).

2.4.2.3 Invariância das estimativas dos parâmetros

Diferente das outras técnicas de mensuração de variáveis latentes, a TRI possui uma característica muito marcante: a *invariância das estimativas dos parâmetros do* construto latente e dos parâmetros dos itens. Isto significa que, independente da amostra de respondentes[6] ou de questões (que meçam o mesmo traço latente), tanto as estimativas dos escores dos indivíduos quanto as estimativas dos parâmetros dos itens (*a*, *b* e *c*) permanecem inalteradas[7] .

Convém lembrar que o princípio da invariância das estimativas dos parâmetros se aplica adequadamente quando o modelo escolhido apresenta boa adequação aos dados empíricos (Hambleton, Swaminathan e Rogers, 1991) e, além disso, a amostra de itens ou de indivíduos seja apropriada do ponto de vista estatístico. De tal modo, mediante este precedente, torna-se possível ultrapassar certas limitações da TCT, assim como a construção de banco de itens calibrados e de testes adaptativos (*computerized adaptive testing*).

2.4.3 As relações entre TCT e TRI

Cada parâmetro do item *i* - a_i e b_i - pode ser estimado, de modo apropriado, mediante relações matemáticas entre aspectos individuais item *i* definidas pela TCT como: índice de dificuldade, correlação bisserial (r_{bi}) e escore total observado.

Neste caso, mantendo-se as condições de normalidade de β, o parâmetro de dificuldade do item *i* (b_i) é definido pela razão entre o índice de dificuldade e a correlação bisserial. Além disso, o parâmetro a_i (discriminação) é dado pela seguinte expressão (Andriola, 2009):

$$a_i \cong \frac{r_{bi}}{\sqrt{1-(r_{bi})^2}} \tag{2.8}$$

Os estudos de Vendramini *et al* (2004) e Quadros & Camey (2010) mostram as relações dos parâmetros dos itens da TRI (dificuldade e discriminação) com os aspectos individuais dos itens sob a TCT (dificuldade e correlação bisserial).

---

[6] A amostra de indivíduos, além de representativa da população de interesse, deve ser semelhante em termos das características da amostra inicial do estudo.

[7] Quando as diferenças entre as estimativas, oriundas de cada amostra, de cada parâmetro do item forem aproximadamente nulas.

2.4.4 Limitações da TRI

Ainda não existe consenso sobre o tamanho mínimo da amostra de respondentes e do conjunto de itens necessários para a estimação dos parâmetros dos modelos unidimensionais logísticos. Alguns autores (Soares, 2005; Nunes e Primi, 2005) afirmam que neste cálculo de dimensionamento devem ser levados em consideração: a distribuição da variável latente de interesse, o tipo do modelo, o número de itens e os métodos de estimação.

2.4.5 Análise Fatorial com Informação Completa

Com o intuito de investigar o princípio da *unidimensionalidade* nos itens sem empregar os modelos tradicionais da análise fatorial (baseados em correlações lineares), Bock e Aitkin (1981) desenvolveram um novo método, fundamentado na Teoria de Resposta ao Item, adequado tanto para itens dicotômicos quanto politômicos, denominado de análise fatorial com informação completa (*full information*), abreviada como FIFA (Bartholomew, 1980).

Este procedimento utiliza o modelo fatorial de Thurstone (1947) baseado nas estimativas de máxima verossimilhança marginal (*marginal maximum likelihood*) e no algoritmo EM (Dempster, Laird e Rubin, 1977). Este modelo é empregado para a análise dos padrões diferentes de respostas[8] dadas ao conjunto de itens. Para aprofundamento desta técnica veja Pasquali (2009).

A aplicação deste novo método aparece em alguns estudos nacionais (Dias e Vendramini, 2008; Soares, 2008) com o uso do pacote estatístico TESTFACT (Wilson, Wood e Gibbons, 1991).

2.5 Análise Fatorial Exploratória

O principal propósito da análise fatorial exploratória (AFE) também denominada como análise fatorial tradicional é identificar, se possível, dentro de um instrumento de mensuração, as estruturas de correlação linear (interdependências) presentes entre os itens (variáveis observáveis). Assim, os

[8]Para itens dicotômicos (tipo certo ou errado), é uma seqüência composta de 0 ou 1 atribuído à resposta a cada item pelo indivíduo.

itens com maiores cargas fatoriais[9] (ou correlações) podem ser agrupados em variáveis independentes não observáveis denominadas de fatores (ou variáveis hipotéticas) (Johnson e Wichern, 1992).

Os fatores explicam parte da variabilidade total dos dados, expressa através da soma das variâncias das variáveis originais (itens). Os itens com uma maior variabilidade (variância) podem vir a ser predominantes na construção dos fatores, mascarando, eventualmente, a presença de itens com menor variabilidade. Nesses casos, sugere-se trabalhar com as variáveis padronizadas[10] e considerar a matriz de correlação de Pearson na análise (Artes, 1998; Johnson e Wichern, 1992; Yoshida e Colugnati, 2002; Gouveia *et al*, 2008; Gomes e Borges, 2009).

Além disso, quando a análise fatorial é satisfatória, o pesquisador pode considerar um número menor de itens sem perda de informações e ainda usar os fatores (possíveis escalas unidimensionais) como representantes da dimensionalidade do traço latente, caso estes estejam em conformidade com a teoria subjacente.

Cabe ao pesquisador considerar os fatores como possíveis escalas unidimensionais (dimensões) do traço latente ou escolher o fator com maior variância explicada para ser considerado a dimensão predominante (única) da variável latente.

O ajuste de modelos da AFE está disponível em diversos pacotes estatísticos como: R, SPSS, SAS, STATA, etc. Para tais modelos, os métodos mais populares para a estimação dos parâmetros são: componentes principais e máxima verossimilhança. Geralmente são combinados com as técnicas de rotação dos eixos ou fatores tais como *varimax* (rotação ortogonal), *rotação oblíqua* (rotação não ortogonal) ou *promax*, etc. empregadas para melhorar a interpretação dos fatores. Aliás, é possível avaliar, através do *teste KMO* (*Kaiser-Meyer-Olkin*) (Castillo, 2007; Dias e Vendramini, 2008; Paes, 2009) a adequação dos modelos da análise fatorial exploratória para detectar a estrutura latente no conjunto de itens, antes de prosseguir na extração e rotação de fatores.

[9] Carga fatorial alta significa um valor de pelo menos 0.30.

[10] Cujas variâncias são iguais a um e cujas covariâncias correspondem às correlações entre as variáveis originais.

A partir da escolha do modelo da análise fatorial tradicional, a determinação da quantidade de fatores a serem retidos influencia a caracterização dos fatores extraídos, principalmente quando é escolhida a *rotação oblíqua*, conforme assinala Andriola (2009). Além do mais, apesar de existir enorme variedade de métodos na literatura, para a determinação da quantidade de fatores a serem retidos, há sugestões do método de *Kaiser-Guttman* e o *Scree-Plot* (Andriola, 2009).

Ao escolher o método de componentes principais é possível conduzir a análise da interdependência dos itens pela matriz de correlação ou pela matriz de variância-covariância. No caso de variáveis dicotômicas, Yoshida e Colugnati (2002), Andriola (2009) e Silva e Lacerda (2010) sugerem como forma apropriada o uso da matriz tetracórica que está disponível no pacote estatístico MicroFACT (Waller, 1995).

Para a aplicação adequada dos métodos de extração da análise fatorial exploratória, as respostas dadas a cada item do instrumento devem ser independentes, medidas em nível intervalar ou razão e modeladas por uma distribuição normal. Este último pressuposto deve ser satisfeito, caso seja escolhido o método de máxima verossimilhança. Dessa forma, com a avaliação de cada pressuposto exigido pelos modelos da análise fatorial exploratória os resultados estatísticos tornam-se mais confiáveis.

Além disso, o tamanho da amostra de respondentes e do conjunto de itens devem ser adequados. Assim, de acordo com Hair *et al* (1995) e Reis (1997, 1998), o dimensionamento da amostra de respondentes deve variar entre 5 a 20 vezes o número de itens, sendo que na estimação dos parâmetros são exigidas pelo menos 100 observações. Para amostras inferiores a 50 observações não é aconselhável aplicar esta técnica (Artes, 1998).

Quando há suspeita de uma relação não linear entre as respostas dadas aos itens, a literatura apresenta uma versão não-linear da análise fatorial (Hattie, 1985), mas ainda não existe um consenso sobre a eficiência deste método (Pasquali, 2009).

Em resumo, os modelos tradicionais da análise fatorial são recomendados para avaliar empiricamente a hipótese de unidimensionalidade de um dado grupo de itens e, ao mesmo tempo, em algumas situações, sugerir uma escala.

2.6 Análise Fatorial Confirmatória

É crescente o número de estudos nacionais que empregam a análise fatorial confirmatória (AFC) como mostram alguns trabalhos de Gouveia et. al (2001), Adanez e Velasco (2002), Pereira, Camino e Costa (2004), Lopes (2005), Pilati e Abbad (2005), Gomes e Borges (2009), Borrego, Chicau *et al* (2010) e Fernandes e Vasconcelos (2010).

A análise fatorial confirmatória é uma técnica adequada quando o pesquisador já possui uma escala previamente construída e pretende testar (confirmar) a partir da estrutura empírica observada neste conjunto de itens se há evidências significativas da mensuração do construto teórico de interesse. Para avaliar esta hipótese será necessário o uso de modelos estatísticos específicos pertencentes à família dos modelos estruturais (Hair *et al*, 1995) que estão disponíveis nos programas estatísticos como *EQS 6.0* (Bentler, 1995; Bentler & Wu, 1993), AMOS 4.0 e LISREL 8 (Joreskög & Sörbom, 1989), dentre outros.

Para uma modelagem adequada, via AFC, há exigência de alguns pressupostos nos itens que compõem a escala prévia, a saber: medidos como variáveis intervalares[11] ou razão, normalidade da distribuição, linearidade e homogeneidade de variâncias, ausência de multicolinearidade ou de singularidade (Tabachnick; Fidell, 2001; Malhotra, 2001; Lopes, 2005). Além disso, há uma recomendação para que se tenham entre 10 e 15 respondentes para cada item incluído na pesquisa (Hair JR. *et al*, 1998; Lopes, 2005).

O processo de estimação do modelo consiste de quatro etapas: a) especificação, b) estimação, c) avaliação e, às vezes, uma quarta, d) modificação do modelo.

No próximo capítulo será abordada, com mais detalhes, a Teoria de Resposta ao Item não Paramétrica (TRIN).

[11] Os itens medidos como variáveis dicotômicas apresentam propriedades matemáticas limitadas que violam as hipóteses exigidas na modelagem pela AFC. Dessa forma, não é aconselhável incluir escalas formadas por tais itens e sim, recorrer a outras técnicas como a TRIN, por exemplo.