

Referência: Pasquali, L. (2010). Testes referentes a construto: teoria e modelo de construção. In: L. Pasquali, *Instrumentação psicológica: fundamentos e práticas* (pp. 165-198). Porto Alegre: Artmed.

# 8

## TESTES REFERENTES A CONSTRUTO: TEORIA E MODELO DE CONSTRUÇÃO

*Luiz Pasquali*

### INTRODUÇÃO

Inicialmente, é importante alertar o leitor de que a tecnologia aqui apresentada de elaboração de instrumentos psicológicos exige o conhecimento de algumas disciplinas ensinadas nas universidades, bagagem sem a qual dificilmente o pesquisador poderá se considerar apto a construir instrumentos psicológicos. Entre essas disciplinas salientam-se particularmente as seguintes, às quais este livro remete sem poder substituí-las, apenas indicando o momento no processo de elaboração do instrumento em que elas têm seu espaço de aplicação:

- psicometria: fundamental para a teoria da medida em psicologia, particularmente o conhecimento da teoria da resposta ao item (TRI);
- disciplinas de teoria psicológica, tais como história e sistemas, teorias da personalidade, psicopatologia, psicologia social, etc.; essas disciplinas são básicas para os procedimentos teóricos;
- disciplinas de delineamento de pesquisa científica; esse conhecimento é fundamental para os procedimentos experimentais;
- disciplinas de estatística – estatística básica, análise de hipótese, análise fa-

torial; esses conhecimentos são decisivos nos procedimentos analíticos.

A teoria e o modelo de elaboração de instrumental psicológico apresentados neste capítulo são aplicáveis à construção de testes psicológicos de aptidão, de inventários de personalidade, de escalas psicométricas de atitude e do diferencial semântico. O modelo, que é detalhado na Figura 8.1, se baseia em três grandes polos, que chamaremos de procedimentos teóricos, procedimentos empíricos (experimentais) e procedimentos analíticos (estatísticos).

O *polo teórico* enfoca a questão da teoria que deve fundamentar qualquer empreendimento científico, no caso a explicitação da teoria sobre o construto ou objeto psicológico para o qual se quer desenvolver um instrumento de medida, bem como a operacionalização do construto em itens. Este polo expõe a teoria do traço latente, bem como a explicitação dos tipos e categorias de comportamentos que constituem uma representação adequada desse traço.

O *polo empírico* ou experimental define as etapas e técnicas da aplicação do instrumento piloto e da coleta válida da informação para proceder à avaliação da qualidade psicométrica do instrumento.

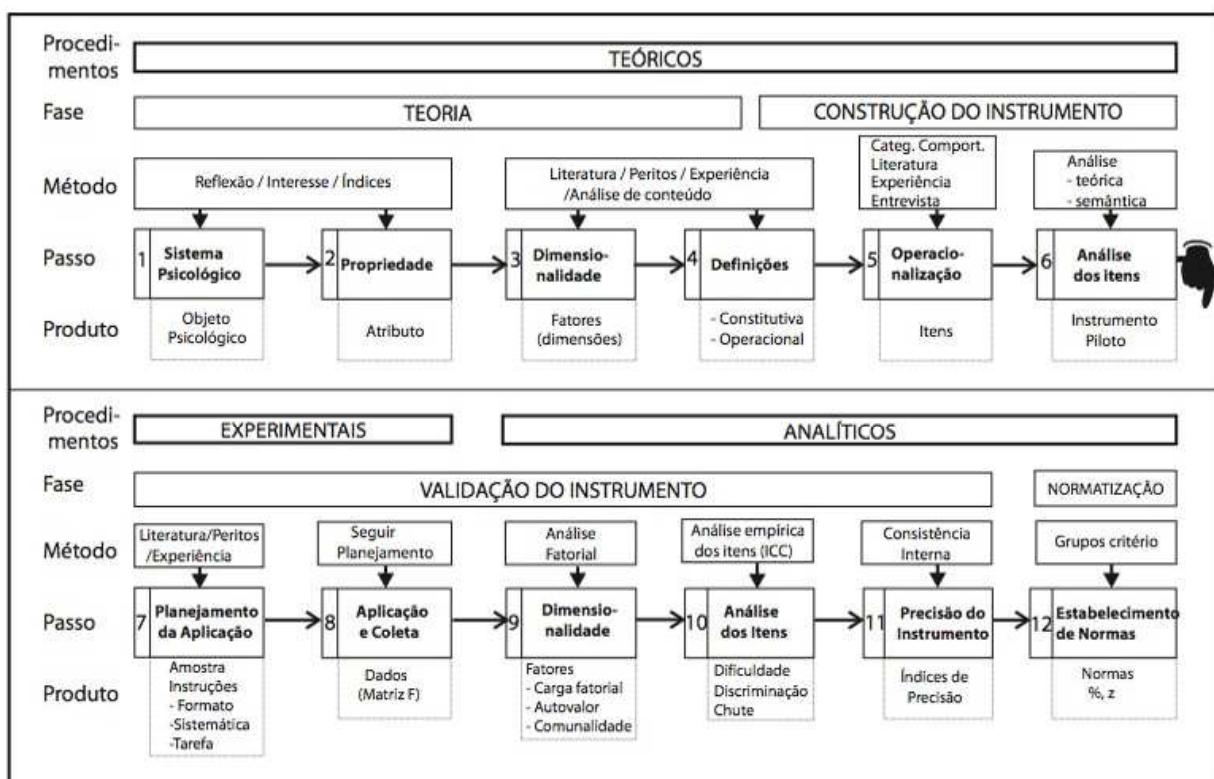
O polo analítico estabelece os procedimentos de análises estatísticas a serem efetuadas sobre os dados para levar a um instrumento válido, preciso e, se for o caso, normatizado.

A Figura 8.1 apresenta e detalha, para cada um desses três procedimentos, as etapas ou passos pelos quais se deve passar para se poder progredir sistematicamente na elaboração de um instrumento de medida psicológica baseado em construtos. Além disso, define, para cada passo, o método ou métodos a serem utilizados para superar o problema específico que constitui a tarefa a ser resolvida em cada passo, bem como o produto que decorre como resultado da solução do problema de cada passo. Além desses detalhes técnicos, a figura apresenta, para os três procedimentos, uma metanálise na qual se procura enquadrar e delimitar o evento ou eventos psicométricos que es-

tão ocorrendo; tal fenômeno vem identificado sob a égide do rótulo “fase”.

## PROCEDIMENTOS TEÓRICOS

Os procedimentos teóricos devem ser elaborados para cada instrumento, dependendo, portanto, da literatura existente sobre o construto psicológico que o instrumento pretende medir. A teoria ainda é, infelizmente, a parte mais fraca da pesquisa e do conhecimento psicológicos, o que tem como consequência a precariedade dos atuais instrumentos psicométricos de medida nesta área. Na verdade, os instrumentos baseados em uma teoria psicológica prévia mais elaborada (por exemplo, “Edwards personal preference schedule”) não são dos melhores no mercado. Tal ocorrência explica, em parte, por que os psicometristas sistematicamente fogem



**Figura 8.1**  
Organograma para elaboração de medida psicológica.

da explicitação de uma teoria preliminar e iniciam a construção do instrumento pela coleta intuitiva e mais ou menos aleatória de uma amostra de itens, que dizem possuir *face validity*, isto é, parecem cobrir o traço para o qual eles querem elaborar o instrumento de medida. Embora isso não pareça muito científico, infelizmente é o que ocorre mais frequentemente na construção de instrumental psicológico.

A inexistência de teorias sólidas sobre um construto não deve ser desculpa para o psicométrista fugir de toda a especulação teórica sobre ele. É obrigação dele levantar, pelo menos, toda a evidência empírica sobre o construto e procurar sistematizá-la e, assim, chegar a uma miniteoria sobre ele, que possa guiá-lo na

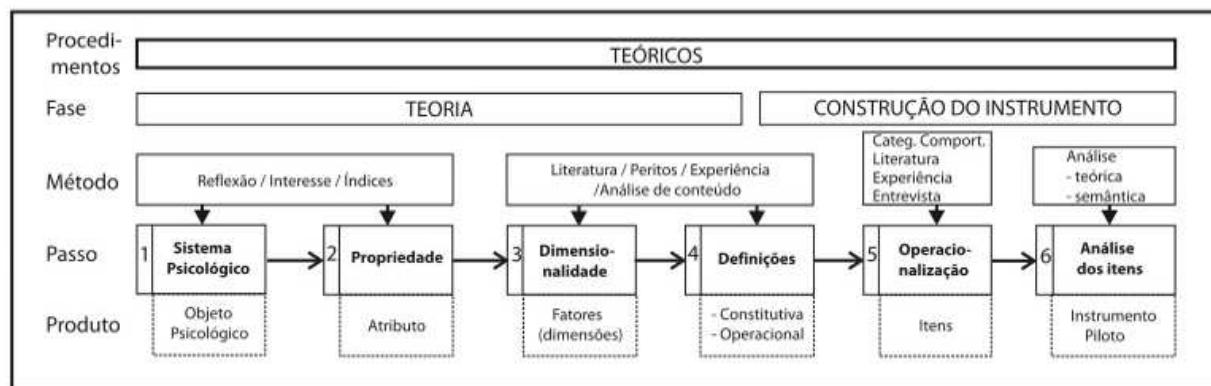
elaboração de um instrumento de medida para o tal construto. Apesar do avanço e da sofisticação estatísticos na psicometria, parece ser essa fraqueza da base teórica que vem maculando a imagem dos procedimentos psicométricos na observação dos fenômenos psicológicos. Na verdade, com uma base teórica coerente e, quando possível, completa, torna-se viável uma definição dos tipos e características dos comportamentos que irão constituir a representação empírica dos traços latentes e, assim, facilitar a tarefa do psicométrista em operacionalizá-los adequadamente (isto é, a construção dos itens se torna coerente e adequada).

De qualquer forma, a Figura 8.2 detalha esses procedimentos teóricos.

### NOTA EXPLICATIVA

A terminologia em ciência e, diria, particularmente em psicologia não é uniforme infelizmente. Por isso, é útil conceituar preliminarmente certas expressões aqui utilizadas, como segue:

- *Sistema*: sinônimo de objeto, coisa, ser, entidade que possui propriedades ou atributos. O sistema é definido não necessariamente pela natureza, mas pelo interesse do discurso, e existente neste mundo do discurso.
- *Atributo*: propriedade, qualidade, aspecto, componente do objeto. Ele é caracterizado por ser mensurável em um contínuo de pontos de magnitude.
- *Magnitude*: qualidade de um sistema que pode assumir diferentes valores de quantidade, isto é, ela pode ser mais ou maior que (>) ou menos ou menor que (<).
- *Isomorfismo*: afirmação de correspondência entre propriedades do número (matemática) e quantidades das propriedades dos sistemas da natureza (física ou não).
- *Definição*: delimitação de um conceito em termos de suas propriedades específicas. Ela é constitutiva ou formal se o conceito ou construto for definido em termos de outros construtos. Ela é operacional ou epistêmica se o conceito ou construto for definido em termos de fatos empíricos, da experiência ou observação.



**Figura 8.2**

Procedimentos teóricos na elaboração da medida psicológica.

### O sistema psicológico

Qualquer sistema ou objeto que possa eventualmente ser expresso em termos observáveis é suscetível de se tornar um objeto para fins de mensuração. Acontece, porém, que um objeto em si não pode ser medido. Os objetos podem apenas ser enumerados. O que pode ser medido são as propriedades ou atributos de um objeto, desde que apresentem magnitudes, isto é, diferenças individuais, tais como intensidade, peso, altura, distância, etc. Por isso esses atributos são geralmente chamados de variáveis, dado que não são invariantes entre sistemas individuais diferentes ou eles mesmos em diferentes ocasiões ou situações.

Se o sistema ou objeto representa o universo de interesse, o atributo dele constitui uma delimitação desse universo. O sistema realmente é definido pelo interesse do investigador. Como a ciência procura o conhecimento e não o poder ou a afirmação pessoal ou a política, então para o cientista não existe sistema privilegiado; todo e qualquer sistema é digno e válido para ser conhecido. Obviamente, interesses políticos, sociais, pedagógicos, financeiros podem ditar a escolha de um objeto de estudo. Assim, a relevância de um sistema de estudo não é ditada pelo saber em si, mas por fatores extrínsecos a ele; nem por isso esses fatores extrínsecos são negligenciáveis no contexto geral do universo da natureza e do ser humano, dado que o ser humano (pesquisador) está situado em um contexto e tem suas prioridades em parte ditadas por esse contexto. Assim, não há maior sabedoria em se estudar um grão de areia do que a sobrevivência do ser humano, embora para nós, seres humanos, esta última pareça bem mais relevante.

Enfim, o sistema representa o objeto de interesse, chamado também de objeto psicológico. A psicometria enfoca como

seu objeto específico as estruturas latentes, os traços psíquicos ou processos mentais, se quiser, que assim se constituem no seu objeto ou sistema direto de interesse. O sistema pode ser considerado de vários níveis, dependendo do interesse do pesquisador. Pode-se falar de um sistema universal e de sistemas locais, o universal sendo a estrutura psicológica total do ser humano e os locais, os vários subsistemas de interesse. Assim, a inteligência pode ser considerada um subsistema dos processos cognitivos e estes da estrutura latente geral, ou mesmo a inteligência, digamos, verbal pode ser considerada um sistema quando ela for o interesse imediato e na qual vários aspectos podem ser considerados, como a compreensão verbal e a fluência verbal. Sistema, portanto, constitui-se em sistema como o objeto imediato de interesse dentro de um delineamento de estudo, e não é uma entidade ontológica monolítica e unívoca.

Esses vários níveis de sistemas ocorrem mesmo nas coisas físicas. Assim, por exemplo, para o biólogo podem ser sistemas o organismo em sua totalidade ou parte dele, como é o sistema neurológico para o neurólogo, o sistema vascular para o cardiólogo, etc. O químico se interessa pelos elementos da tabela periódica, em que os seus sistemas naturais (água, ar, etc.) se reduzem a esses elementos de interesse desse profissional. O físico nuclear estuda seus sistemas reduzindo-os finalmente às partículas *quark* (*top*, *bottom*, *strange*, etc.), às forças glúons (força forte, fraca, gravitacional) e aos processos léptons (elétrons, pósitrons). Em psicologia, também encontramos tais níveis de sistemas. Considere, por exemplo, os processos cognitivos: Piaget e Spearman consideram a inteligência como uma grande estrutura (um sistema) que evolui geneticamente; os fatoristas consideram a inteligência no nível de estruturas menores, quando falam de raciocínio verbal,

numérico, abstrato, etc.; Sternberg vai ainda mais longe nessa elementarização dos sistemas (processos) cognitivos, buscando seus elementos no que ele chama de componentes cognitivos; e, finalmente, Newell e Simon levam ao extremo esse elementarismo quando defendem os processos elementares de informação (*elementary information process – EIP*) como os elementos últimos dos processos cognitivos. A qualquer desses níveis, o pesquisador pode se colocar e definir este nível como o nível do sistema de seu interesse. Não é preciso ver oposições teóricas antagônicas nesses vários autores quanto aos processos cognitivos. Eles simplesmente se põem em horizontes diferentes e, por isso, veem níveis diferentes de realidade, aliás, da mesma realidade. É apenas o exclusivismo, na verdade desnecessário, desses autores em afirmar que seu horizonte é o único ou o melhor para ver a realidade dos processos cognitivos. Isso vale, aliás, para qualquer outro processo psicológico, como a personalidade, por exemplo.

Enfim, o problema a ser resolvido neste passo praticamente se reduz a que o pesquisador, que pretende construir um instrumento, deve ter uma ideia, por mais vaga que seja, sobre o que é que ele quer trabalhar, para que tema da psicologia ele está interessado em construir um instrumento de medida e pesquisa. Este problema é, evidentemente, mais aparente em aluno de pós-graduação, ao qual se apresenta a necessidade de apresentar uma dissertação no final do curso e ainda não tem ideia sobre que assunto em psicologia ele quer desenvolver sua tese. Na falta de qualquer outra indicação ou interesse específico, tal indivíduo pode se dirigir aos livros índices onde estão elencados os principais trabalhos que se vêm fazendo em psicologia. Para o psicólogo há uma série de tais livros, sendo o mais útil o *Psychological Abstracts*, que é publicado mensalmente e onde aparece a quase to-

talidade dos artigos e trabalhos feitos em psicologia em nível mundial. Além deste, há o *Educational Index*, para os interessados na psicologia aplicada à educação, o *Index Medicus*, para os interessados em psicologia clínica, e o *Sociological Index*, onde aparecem temas referentes à psicologia social. Na falta de tais fontes ou se, ainda assim, o tema não surgiu na percepção do pesquisador (aluno), ele pode recorrer a peritos, que, no caso do aluno, é normalmente seu orientador. Enfim, o problema a ser resolvido nesse passo consiste em se ter uma ideia, um tema, um assunto para pesquisar. A esse tema chamamos de objeto psicológico, que representa o produto esperado desse passo na elaboração do instrumento. Agora, o sistema escolhido pode ser mais amplo ou mais restrito, como vimos anteriormente ao falarmos dos diferentes níveis de sistemas. Obviamente, quanto mais restrito ou elementar for o sistema, mais fácil se torna a construção de um instrumento de medida. Por isso, é relevante definir como sistema psicológico um processo ou traço latente o mais próximo do interesse direto do pesquisador. Tipicamente isso significa definir um sistema o mais elementar possível, dentro do interesse. Sistemas vagos e gerais dificultam depois sua operacionalização para fins de pesquisa empírica, como é a construção de um instrumento de medida.

### A propriedade do sistema psicológico

O sistema, já dissemos, não constitui objeto direto de mensuração, mas sim suas propriedades ou atributos, que são os vários aspectos que o caracterizam. Por exemplo, o sistema físico se apresenta com os atributos de massa, comprimento, etc. Similarmente, a psicometria concebe os seus sistemas como possuidores de propriedades/atributos que os definem,

sendo esses atributos o foco imediato de observação/medida. Assim, a estrutura psicológica apresenta atributos do tipo processos cognitivos, processos emotivos, processos motores, etc. A inteligência, como subsistema, pode apresentar atributos de tipo raciocínio verbal, raciocínio numérico, etc. O sistema se constitui como objeto hipotético que é abordado (conhecido) por meio da pesquisa de seus atributos.

O problema específico deste passo consiste em passar de um objeto psicológico, normalmente amplo demais para pesquisar, para a delimitação dos aspectos específicos dele que se deseja estudar e para os quais se quer construir um instrumento de medida. De fato, qualquer sistema apresenta ilimitado número de propriedades. A rosa, por exemplo, tem perfume, cor, peso, tamanho, beleza, ritmo de crescimento, etc. É relevante, para se poder escolher ou construir um instrumento de medida, definir qual ou quais propriedades do sistema serão objeto de estudo. Por exemplo, se meu interesse se focaliza sobre a criança, não é possível estudar, de uma só vez, tudo sobre a criança. Então, tenho que me decidir por um aspecto mais restrito referente à criança, que enfim vou pesquisar. Assim, da criança posso estudar o seu desenvolvimento psicomotor, o desenvolvimento cognitivo, o desenvolvimento da linguagem, a enurese, a timidez, a agressividade, etc. Em qual desses ou outros aspectos estou presentemente e diretamente mais interessado? Pois esse ou esses aspectos constituem a propriedade do objeto criança que presentemente quero abordar. A esses aspectos escolhidos chamamos de atributo. Ou o seu interesse pode se focalizar sobre a inteligência. Esta de fato já é em si mesma uma propriedade do sistema ser humano. Mas ela pode ser igualmente considerada um subsistema complexo, apresentando várias propriedades específicas dela, tais

como raciocínio verbal, raciocínio numérico, raciocínio abstrato, memória, percepção espacial, etc.

Para se definir um instrumento de medida, é preciso decidir qual ou quais dessas propriedades serão o objeto imediato de interesse. Ademais, à medida que o conhecimento sobre o sistema cresce, cresce também o número de novas propriedades descobertas, que, por sua vez, podem se tornar novos subsistemas de interesse, uma vez que se vão descobrindo propriedades diversas dentro desses novos subsistemas. Enfim, é de importância para se prosseguir sem transtornos e desvios de rumo que se defina claramente e preliminarmente as propriedades do sistema de interesse que se quer estudar. Tal definição evita que se misture, no prosseguimento do processo, "alhos e bugalhos", como, por exemplo, utilizar uma amostra de itens que mais medem aspectos de conhecimento de vocabulário, quando, de fato, se queria atingir o raciocínio verbal. Com isso, também não se está afirmado que entre tais propriedades de um sistema não haja correlações. Antes, pelo contrário, relações e interações entre as propriedades de um mesmo sistema são uma suposição não somente legítima, mas provável. Contudo, o que se está afirmado é que é preciso partir com conceituações e definições claras e precisas, bem como delimitadas, uma vez que a capacidade de conhecimento humano não é abrangente.

Como se decidir por este ou aquele aspecto? Novamente, recorro ao meu interesse, à ajuda dos livros índices e aos peritos (concretamente, ao meu orientador, se eu for aluno).

### **Dimensionalidade do atributo**

Se os dois primeiros passos anteriormente descritos possam parecer, para muitos, um mero exercício acadêmico, o

terceiro passo e os demais a seguir já não são tão simples, pois os problemas que eles apresentam são bem mais complexos.

A dimensionalidade do atributo diz respeito à sua estrutura interna, semântica. O atributo constitui uma unidade semântica única ou é ele uma síntese de componentes distintos ou até independentes? Deve ele ser concebido como uma dimensão homogênea ou devem-se nele distinguir aspectos diferenciados? A resposta a esse problema obviamente deve vir da teoria sobre o construto e/ou dos dados empíricos disponíveis sobre ele, sobretudo dados de pesquisas que utilizaram a análise fatorial na análise dos dados, pois o que está em jogo aqui é a questão de decidir se o construto é uni ou multifatorial. Os fatores que compõem o construto (o atributo) são o produto desse passo. Por exemplo: Tenho como objeto psicológico os processos cognitivos; a propriedade desse objeto psicológico que estou interessado em estudar é a inteligência verbal. Pergunta-se: é esta inteligência verbal um construto único ou devo distinguir nele componentes diferentes? Os dados empíricos disponíveis me mostram que a inteligência verbal é composta por, pelo menos, dois fatores bem distintos e praticamente independentes, a saber, compreensão verbal e fluência verbal. Consequentemente, se quiser pesquisar a inteligência verbal e construir para tal um instrumento de medida, não poderei prescindir de conhecer e levar em conta o fato de que a inteligência apresenta dois fatores distintos, cuja medida exige instrumentos diferentes. Claro, posso me decidir por estudar somente a inteligência verbal compreendida sob seu aspecto de compreensão verbal e prescindir de me preocupar com a fluência verbal. Porém, nesse caso o meu atributo de interesse de estudo não é mais a inteligência verbal, e sim a compreensão verbal. Mesmo tomando essa decisão de somente querer estu-

dar a compreensão verbal, não fico escusado de expor a teoria sobre a inteligência verbal em sua totalidade e, em seguida, justificar minha decisão pelo estudo de apenas um aspecto dela. Evidentemente, nessa justificativa pode ser e será suficiente o meu interesse específico por tal aspecto da inteligência. Isto é, eu devo saber o que estou fazendo, e demonstro isso na exposição da teoria que faço sobre o construto inteligência verbal.

Nesses dois passos, propriedade e dimensionalidade, entramos no ponto mais crítico na caminhada para a elaboração dos instrumentos psicológicos, porque toda esta parte resulta essencialmente da teoria psicológica, que concebe, define e estrutura os construtos psicológicos. A tarefa da construção da teoria psicológica não é específica do psicométrista, mas do psicólogo teórico. O psicométrista deveria poder contar com essa teoria e, com base nela, fundamentar a construção dos instrumentos de medida. A existência de teorias ou fantasias as mais variadas sobre praticamente qualquer construto em psicologia torna a tarefa do psicométrista quase uma tragédia quando quer construir instrumentos para medir construtos sobre os quais os psicólogos não se entendem. Dessa forma, o psicométrista acaba se decidindo em construir um instrumento para medir um construto concebido segundo algum psicólogo. E ali você tem uma fauna enorme de psicólogos teóricos, desde os behavioristas até os dialéticos, que falam linguagens quase totalmente estranhas um em relação ao outro. Infelizmente esta é a situação da teoria psicológica atual. Para caricaturar, imagine o seguinte: um físico vai construir um instrumento para medir o comprimento de objetos físicos. E se para poder efetuar tal empreendimento, ele tivesse que decidir sobre “bem, comprimento entendido segundo quem?” Tal pergunta careceria de sentido e seria ridícula fosse ela feita sobre comprimento

ou outras propriedades da matéria (pelo menos, na sua grande maioria). Mas, no caso do psicometrista, tal pergunta infelizmente é corriqueira, qualquer que seja o construto que ele queira estudar e medir, o que vem a mostrar o estado primitivo em que vive a teoria psicológica. Essa precariedade da teoria psicológica é a principal responsável pela fuga, por parte dos psicometristas, de basear a construção dos instrumentos psicológicos em uma teoria prévia e testá-los em seguida por meio da metodologia científica. Essa fuga acarreta que o psicometrista parte de uma coleção atabalhoadas de itens para em seguida ver o que eles estão medindo, se alguma coisa psicológica relevante.

Este estado de coisas deveria e deve obrigar o psicometrista a expor ou elaborar uma miniteoria sobre o que ele entende pelo construto que pretende medir. Felizmente, já existe razoável abundância de dados empíricos sobre muitos construtos psicológicos, com base nos quais o psicometrista poderá desenvolver uma tal miniteoria do construto, que irá guiar a construção do seu instrumento de medida. Os dados empíricos que serão coletados mediante o instrumento assim construído irá decidir se sua miniteoria tem ou não alguma consistência. Isso não é uma tragédia, é a própria lógica da pesquisa empírica, isto é, a testagem empírica que pode ou não confirmar a validade de uma teoria: a verdade científica é sempre relativa, nunca será um dogma, e, portanto, sempre reformável.

**Uma pausa.** Tipicamente, é nesses passos que o pesquisador (pós-graduando) se sente perdido, porque não consegue ver direito o que é exatamente o seu tema de pesquisa e como é que ele se enquadra no campo teórico do tema. Por isso vamos fazer um exercício. Vamos supor que eu quero construir um instrumento para medir o raciocínio verbal. Sendo isso verda-

de, então já tenho um pé no chão e posso fazer as seguintes considerações:

1. Se quero medir o raciocínio verbal, então sei que ele não é o objeto psicológico, porque a ciência não mede objetos, mas sim propriedades dele. Assim, raciocínio verbal é um atributo.
2. Se raciocínio verbal é um atributo, consequentemente ele é atributo de algum objeto. Dessa forma, devo descobrir qual é esse objeto do qual o raciocínio verbal é propriedade.
3. Se raciocínio verbal é um atributo de algum objeto, é de se supor que esse objeto tenha mais do que um atributo. Isto é, além de raciocínio verbal, o objeto tem outros atributos. Assim, devo descobrir quais são esses outros atributos, de tal forma que, depois, possa diferenciar (definir diferencialmente) o meu atributo de interesse – o raciocínio verbal – dos demais atributos do objeto em questão.

Com base na literatura, em peritos e na minha reflexão, resolvo essa equação da seguinte forma: Raciocínio verbal é atributo do processo cognitivo chamado raciocínio. Por sua vez, raciocínio tem como atributos, além de raciocínio verbal, também raciocínio numérico, raciocínio abstrato, raciocínio espacial, raciocínio mecânico e, talvez, outros. Mas, a literatura na área certamente fala desses cinco atributos de raciocínio. Dessa forma, posso ilustrar essa minha descoberta da seguinte forma:

Raciocínio	<ul style="list-style-type: none"> <li>• Raciocínio verbal</li> <li>• Raciocínio numérico</li> <li>• Raciocínio abstrato</li> <li>• Raciocínio espacial</li> <li>• Raciocínio mecânico</li> </ul>
------------	---

Pesquisando um pouco mais, verifico que raciocínio verbal não é unidimen-

sional. De fato, a literatura distingue dois tipos de raciocínio verbal, a saber, compreensão verbal e fluência verbal. Assim, a ilustração apresentada pode ser completada como segue:

- |                   |  |
|-------------------|--|
| <b>Raciocínio</b> | <ul style="list-style-type: none"> <li>• Raciocínio verbal</li> <li>• Compreensão verbal</li> <li>• Fluência verbal</li> </ul>                                       |
|                   | <ul style="list-style-type: none"> <li>• Raciocínio numérico</li> <li>• Raciocínio abstrato</li> <li>• Raciocínio espacial</li> <li>• Raciocínio mecânico</li> </ul> |

Com isso, tenho o meio de campo destrinchado. Agora falta definir diferencialmente todos esses atributos e estou bem encaminhado para a elaboração do meu instrumento de medida. Posso fazer tudo isso com qualquer tema para o qual queira construir um instrumento de mensuração. Boa sorte!

### **Definição dos construtos**

Definida a propriedade e suas dimensões, é preciso conceituar detalhadamente esses construtos, novamente baseando-se na literatura pertinente, nos peritos da área e na própria experiência. O problema deste passo é, portanto, a conceituação clara e precisa dos fatores para os quais se quer construir o instrumento de medida. A tarefa aqui é dupla, tendo como resultado dois produtos: as definições constitutivas e as definições operacionais dos construtos.

#### **Definição constitutiva**

Um construto definido por meio de outros construtos representa uma definição constitutiva. Nesse caso, o construto é concebido em termos de conceitos próprios da teoria em que ele se insere.

Definição constitutiva é a que tipicamente aparece como definição de termos em dicionários e enciclopédias: os conceitos são ali definidos em termos de outros conceitos; isto é, os conceitos, que são realidades abstratas, são definidos em termos de realidades abstratas. Por exemplo, se defino inteligência verbal como a capacidade de compreender a linguagem, estou diante de uma definição constitutiva, porque capacidade de compreender constitui uma realidade abstrata, um construto, um conceito.

As definições constitutivas são de extrema importância no contexto da construção dos instrumentos de medida, porque elas situam o construto exata e precisamente dentro da teoria desse construto, dando, portanto, as balizas e os limites que ele possui. Enfim, essas definições caracterizam o construto, dando as dimensões que ele deve assumir no espaço semântico da teoria em que está incluído. Assim, se defino assertividade como a capacidade de dizer não, a capacidade de expressar livremente sentimentos positivos e negativos, a capacidade de expor ideias sem receio, etc., estou dando os limites semânticos que esse conceito deve respeitar dentro da minha teoria de assertividade. Definições dessa natureza põem limitações definidas sobre o que devo explorar quando for medir o construto, limitações não somente em termos de fronteiras que não podem ser ultrapassadas, mas mais ainda em termos de fronteiras que devem ser atingidas. De fato, normalmente um instrumento que mede um construto não chega a cobrir toda a amplitude semântica de um conceito. Assim, boas definições constitutivas vão me permitir em seguida avaliar a qualidade do instrumento que mede o construto em termos de quanto dessa extensão semântica dele é coberta pelo instrumento, surgindo daí instrumentos melhores e piores uma vez que medem mais ou medem me-

nos da extensão conceptual do construto, extensão esta delimitada pela definição constitutiva desse mesmo construto.

### **Definição operacional**

Com as definições constitutivas nós estamos ainda no terreno da teoria, do abstrato. Um instrumento de medida já é uma operação concreta, empírica. A passagem do terreno abstrato para o concreto é precisamente viabilizada pelas definições operacionais dos construtos. Este é, talvez, o momento mais crítico na construção de medidas psicológicas, pois é aqui que se fundamenta a validade desses instrumentos; é aqui que se baseia a legitimidade da representação empírica (comportamental) dos traços latentes (os construtos). Duas preocupações são relevantes e decisivas neste momento:

1. as definições operacionais dos construtos devem ser realmente operacionais e
2. elas devem ser os mais abrangentes possíveis dos construtos.

Primeiramente, as definições operacionais devem ser realmente *operacionais*. Esta tautologia é proposital, porque se peca demais neste particular. Uma definição de um construto é operacional quando o mesmo construto é definido, não mais em termos de outros construtos, mas em termos de operações concretas, isto é, de comportamentos físicos por meio dos quais o tal construto se expressa. Assim, se defino inteligência verbal como a capacidade de compreender uma frase ou, mesmo, compreender uma frase, estou diante de uma definição constitutiva e não operacional. Isto porque compreender não é um comportamento, mas um construto.

Seria uma definição operacional de compreensão da frase *reproduzir a frase com outras palavras*. Mager (1981) dá

uma fórmula simples e perfeita para decidir se a definição é ou não operacional. Ela é operacional se você puder dizer ao sujeito “vá e faça...”. Assim, se defino inteligência verbal como compreender uma frase, o que devo pedir ao sujeito para fazer? Pois “vá e compreenda...” não diz ao sujeito nada que ele possa fazer. Ao passo que dizer “vá e reproduza a frase” indica claramente o que o sujeito deve fazer, como deve se comportar, e, portanto, esta última é uma definição operacional, pois ela define comportamentos que devem ocorrer, enquanto compreender a frase não indica nenhum comportamento concreto específico a ser exibido por parte do sujeito.

Em segundo lugar, a definição operacional deve ser a mais *abrangente* possível do construto. Nenhuma definição operacional esgota a amplitude semântica de um construto; assim, pode haver definições operacionais mais ou menos abrangentes do mesmo construto, e essa grandeza de abrangência, evidentemente, fala da boa, má ou pior qualidade da definição operacional, o que vai obviamente repercutir sobre o instrumento de medida do construto que será baseado nessa definição operacional do mesmo construto. Aliás, uma definição operacional pode ser perfeitamente operacional e também perfeitamente equivocada ou errada, quando não cobrir nada do espaço semântico próprio do construto. Assim, definir inteligência verbal como desenhar círculos na areia constitui uma definição perfeitamente operacional, pois todo mundo entende quando se manda desenhar círculos na areia; contudo, apesar de operacional, ela é uma definição perfeitamente equivocada de inteligência verbal, pois o comportamento de desenhar círculos na areia não tem nada a ver com o construto em questão. Disso segue que as definições operacionais podem representar um construto em uma escala que expressa uma

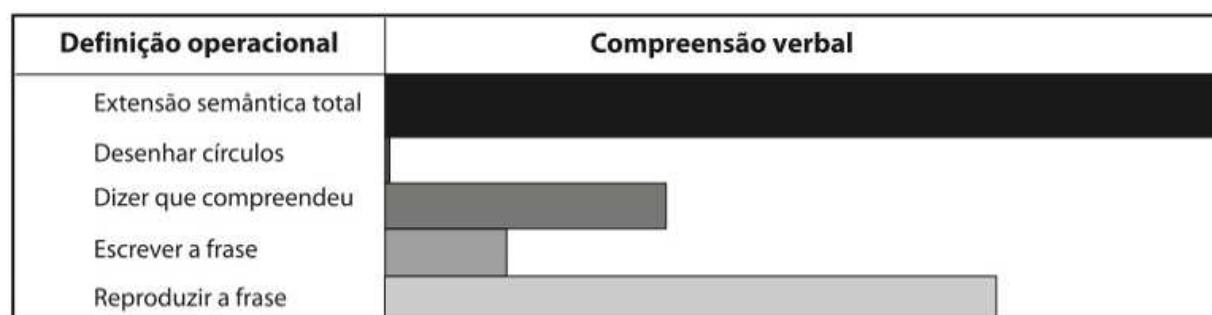
proporção de coincidência entre construto e definição operacional que vai de 0 a 1, sendo 0 quando a definição não cobre nada do construto e 1 quando ela cobre 100% do espaço semântico do construto. Como já dissemos, nenhuma definição operacional será capaz de cobrir 100% do construto, mas quanto maior covariância existir entre construto e definição operacional, maior qualidade se deve atribuir a essa definição do construto e, por consequência, maior chance terá o instrumento, que de tal definição resulta, de ser superior em qualidade. Dizemos maior chance porque a qualidade do instrumento não depende unicamente de boas definições operacionais, embora sem a boa qualidade destas o instrumento já começa de saída a ser inferior. A Figura 8.3 deixa visualizar a problemática da qualidade de representação comportamental de diferentes definições operacionais do construto compreensão verbal.

Para garantir melhor cobertura do construto, as definições operacionais deverão especificar e elencar aquelas categorias de comportamentos que seriam a representação comportamental do construto. Quanto melhor e mais completa for essa especificação, melhor será a garantia de que o instrumento que resultar para a medida do construto será válido e útil. Por exemplo, quais seriam as categorias de comportamentos que expressariam com-

portamentalmente a compreensão verbal? Seriam tais como reproduzir o texto, dar sinônimos e antônimos, explicar o texto, sublinhar alternativas, etc. Quanto mais completa essa listagem de categorias comportamentais, mais próximo estou da construção do instrumento, porque o passo seguinte será simplesmente expressar essas categorias em tarefas unitárias e específicas (os itens), e o instrumento piloto está construído. Por isso, nunca é demais gastar tempo na implementação detalhada das definições operacionais do construto.

Onde vou me inspirar para realizar adequadamente esta tarefa? Novamente, os métodos a serem utilizados para resolver o problema deste passo de construção de medidas psicológicas são a literatura pertinente sobre o construto, a opinião de peritos na área, a experiência do próprio pesquisador, bem como a análise de conteúdo do construto. Torna-se aqui, como se vê, indispensável o conhecimento aprofundado da literatura sobre o construto e sobre as técnicas de análise de conteúdo.

É bom lembrar nesse contexto que os instrumentos de medida psicológica visam medir traços latentes. Mas como medir traços latentes que são impérvios à observação empírica que é o método da ciência? Estamos aqui nos defrontando com o problema da representação: qual é a maneira adequada de se



**Figura 8.3**

Extensão semântica de definições operacionais de compreensão verbal.

representar esses atributos latentes para que possam ser cientificamente abordados? Embora o problema pareça, e é na verdade, grave, ele não é específico da psicometria; ele ocorre na própria física com a teoria quântica, por exemplo. Como o comportamento representa esses traços latentes? É precisamente o problema que as definições operacionais precisam resolver.

### **Operacionalização do construto**

Este é o passo da construção dos itens, que são a expressão da representação comportamental do construto, a saber, as tarefas que os sujeitos terão de executar para que se possa avaliar a magnitude de presença do construto (atributo).

#### **Fontes dos itens**

Se os passos até aqui discutidos foram adequadamente resolvidos, nós estamos agora diante das categorias comportamentais que expressam o construto de interesse, que dão praticamente a resposta à construção dos itens. Além disso, podemos apelar para outras duas fontes de itens: a entrevista e outros testes que medem o mesmo construto. A entrevista consiste em pedir a sujeitos representantes da população para a qual se deseja construir o instrumento para opinar em que tipo de comportamentos tal construto se manifesta. Por exemplo, se meu desejo é construir um instrumento sobre assertividade, posso me dirigir a representantes da população e perguntar “como é para você uma pessoa assertiva?”. De uma pesquisa dessa natureza pode surgir uma grande riqueza de comportamentos que expressam assertividade e que podem ser aproveitados como itens do instrumento. Ademais, posso me inspirar em itens que

compõem outros instrumentos disponíveis no mercado e que medem o mesmo construto no qual estou interessado. Assim, temos três fontes preciosas para a construção dos itens:

- literatura: outros testes que medem o construto;
- entrevista: levantamento junto à população meta;
- categorias comportamentais: definidas no passo das definições operacionais.

É importante notar que, no processo de elaboração do instrumento como o temos exposto, os itens não são mais coletados a esmo ou chutados; eles são elaborados ou, pelo menos, selecionados em função das definições operacionais de um construto que foi exaustivamente analisado em seus fundamentos teóricos e nas evidências (dados) empíricas disponíveis. Então, não é qualquer item que pareça medir o construto que é aceito, mas somente aquele que corresponde às suas definições teóricas (constitutivas) e às suas definições operacionais. Não é mais a malfadada *face validity* que impõe na seleção dos itens, e sim a sua pertinência (a esta altura, obviamente, ainda teórica) ao contexto teórico do construto. Aliás, os itens não são inventados ou pescados, eles são construídos para representar comportamentalmente o construto de interesse.

#### **Regras para construção de itens**

Dadas as fontes que baseiam a construção dos itens, é preciso dar agora algumas regras ou critérios fundamentais para a elaboração adequada dos próprios itens. Essas regras se aplicam, em parte, à construção de cada item individualmente, e em parte ao conjunto dos itens que medem um mesmo construto. Ademais,

dependendo do tipo de traço a ser medido (se é de aptidão ou de personalidade), algumas das regras se aplicam e outras não.

a) *Critérios para a construção dos itens:*

1. *Critério comportamental:* o item deve expressar um comportamento, não uma abstração ou construto. Segundo Mager (1981), o item deve poder permitir ao sujeito uma ação clara e precisa, de sorte que se possa dizer a ele “vá e faça”. Assim “reproduzir um texto” é um item comportamental (“vá e reproduza...”), ao passo que “compreender um texto” não o é, pois o sujeito não sabe o que fazer com “vá e comprehenda...”.
2. *Critério de objetividade ou de desejabilidade:* para o caso de escalas de aptidão, os itens devem cobrir comportamentos de fato, permitindo uma resposta certa ou errada. O sujeito respondente deve poder mostrar se conhece a resposta ou se é capaz de executar a tarefa proposta. Assim, por exemplo, se você quer saber se o sujeito entende o que seja “abstêmio”, faz mais sentido pedir a ele que dê um sinônimo do que pedir que diga se entendeu ou não. Já para o caso das atitudes e de personalidade em geral, os itens devem cobrir comportamentos desejáveis (atitude) ou característicos (personalidade). O respondente, nesse caso, deve poder concordar ou discordar ou opinar sobre se tal comportamento convém ou não para ele, isto é, os itens devem expressar desejabilidade ou preferência. Não existem neste caso respostas certas ou erradas; existem sim diferentes gostos, preferências, sentimentos e modos de ser.
3. *Critério da simplicidade:* um item deve expressar uma única ideia. Itens que

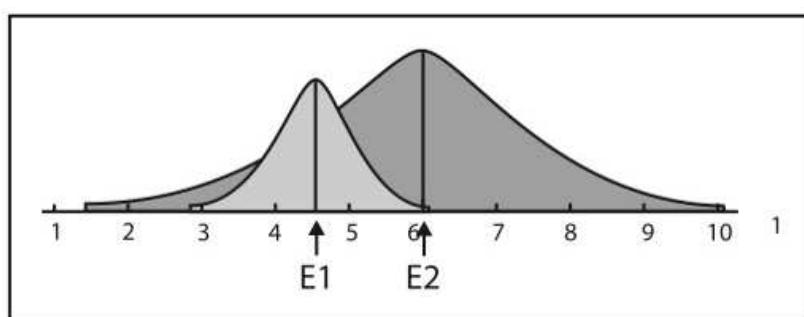
introduzem explicações de termos ou oferecem razões ou justificativas são normalmente confusos porque introduzem ideias variadas e confundem o respondente. Por exemplo: “Gosto de feijão porque é saudável”. O sujeito pode de fato gostar de feijão, mas não porque seja saudável; assim, ele não saberia como reagir a tal item: se porque o feijão é gostoso ou porque é saudável. O item exprime duas ideias. O mesmo vale para “a maçã é gostosa e saudável”.

4. *Critério da clareza:* o item deve ser inteligível até para o estrato mais baixo da população meta; daí utilizar frases curtas, com expressões simples e inequívocas. Frases longas e negativas incorrem facilmente na falta de clareza. Com referência às frases negativas: normalmente elas são mais confusas que as positivas; consequentemente, é melhor afirmar a negatividade do que negar uma afirmação. Por exemplo: fica mais inteligível dizer “detesto ser interrompido” do que “não gosto de ser interrompido”, ou em vez de “não me sinto feliz” é melhor dizer “sinto-me infeliz”. Neste contexto, é preciso igualmente ter atenção em não utilizar gírias, porque elas não são normalmente inteligíveis para todos os membros de uma população meta do instrumento, além de tipicamente ofender o estrato mais sofisticado da população, o que pecaria contra o critério número 10. Contudo, o linguajar típico da população meta deve ser utilizado na formulação dos itens. Assim, são admissíveis e são mais apropriadas expressões conhecidas por tal população, ainda que elas possam parecer linguisticamente menos castiças. A preocupação aqui é a compreensão das frases (que representam tarefas a serem entendidas e se possível resolvidas), não sua elegância artística.

5. *Critério da relevância* (pertinência, saturação, unidimensionalidade, correspondência): a expressão (frase) deve ser consistente com o traço (atributo, fator, propriedade psicológica) definido e com as outras frases que cobrem o mesmo atributo. Isto é, o item não deve insinuar atributo diferente do definido. O critério diz respeito à saturação que o item tem com o construto, representada pela carga factorial na análise factorial e que constitui a covariância (correlação) entre o item e o fator (traço). Veja o seguinte exemplo: seja o construto “compreensão verbal” definido como compreender o significado de palavras e frases. Dos três itens a seguir, um é pertinente, outro é mais ou menos e um é impertinente:
- Reproduzir a frase com as próprias palavras → pertinente.
  - Decorar uma sentença → pouco pertinente.
  - Falar em voz alta → impertinente.
6. *Critério da precisão*: o item deve possuir uma posição definida no contínuo do atributo e ser distinto dos demais itens que cobrem o mesmo contínuo. Este critério supõe que o item possa ser localizado em uma escala de estímulos; em termos de Thurstone, diríamos que o item deve ter uma posição escalar modal definida e um desvio pa-

drão reduzido. Em termos da TRI, este critério representa os parâmetros  $b$  (dificuldade) e  $a$  (discriminação) e pode realmente ser avaliado definitivamente somente após coleta de dados empíricos sobre os itens. Por exemplo, na escala de Thurstone (ver Figura 8.4), o item E1 é muito preciso, enquanto o E2 é impreciso.

7. *Critério da variedade*: dois aspectos especificam este critério.
- Deve-se variar a linguagem pois o uso dos mesmos termos em todos os itens confunde as frases e dificulta diferenciá-las, além de provocar monotonia, cansaço e aborrecimento. Exemplo: o EPPS (Edwards Personal Preference Schedule) em inglês começa quase todas as suas 500 frases com a expressão “I like...”. Depois de tantos “I like”, qualquer sujeito deve se sentir saturado!
  - No caso de escalas de preferências, deve-se formular a metade dos itens em termos favoráveis e metade em termos desfavoráveis, para evitar erro da resposta estereotipada à esquerda ou à direita da escala de resposta. É a recomendação que Likert já dava em 1932.
8. *Critério da modalidade*: formular frases com expressões de reação modal, isto é, não utilizar expressões extremadas,



**Figura 8.4**

Ilustração da precisão dos itens na escala de Thurstone.

como “excelente”, “miserável”, etc. Assim, ninguém é *infinitamente* inteligente, mas a maioria é *bastante* inteligente. A intensidade da reação do sujeito é dada na escala de resposta. Se o próprio item já vem apresentado em forma extremada, a resposta na escala de respostas já está viciada. Assim, se pergunto ao sujeito se está pouco ou muito de acordo (em uma escala, por exemplo, de sete pontos que vai de desacordo total a acordo total), um item formulado extremado tal como “meus pais são a melhor coisa do mundo” dificilmente receberia resposta 7 (totalmente de acordo) por parte da maioria dos sujeitos da população meta, simplesmente porque a formulação é exagerada. Se em lugar dela eu usasse uma expressão mais modal, tal como “eu gosto dos meus pais”, as chances de respostas mais variadas e inclusive extremadas (resposta 7) seriam de se esperar.

9. *Critério da tipicidade:* formar frases com expressões condizentes (típicas, próprias, inerentes) com o atributo. Assim, a beleza não é pesada, nem grossa, nem nojenta.

10. *Critério da credibilidade (face validity):* O item deve ser formulado de modo que não apareça sendo ridículo, despropositado ou infantil. Itens com esta última caracterização fazem o adulto se sentir ofendido, irritado ou coisa similar. Enfim, a formulação do item pode contribuir e contribui (Nevo, 1985; Nevo e Sfez, 1985) para uma atitude desfavorável para com o teste e assim para o aumento dos erros (vieses) de resposta. Este tema, às vezes, é discutido sob o que se chama de validade aparente (*face validity*), que não tem nada a ver com a validade objetiva do teste, mas pode afetar negativamente a resposta ao teste, ao afetar o indivíduo respondente.

b) *Critérios referentes ao conjunto dos itens (o instrumento todo):*

11. *Critério da amplitude:* este critério afirma que o conjunto dos itens referentes ao mesmo atributo deve cobrir toda a extensão de magnitude do contínuo desse atributo. Este critério é novamente satisfeito pela análise da distribuição dos parâmetros b da TRI. A razão disto é que um instrumento deve poder discriminar entre sujeitos de diferentes níveis de magnitude do traço latente, inclusive entre os que possuem um traço alto quanto entre os que possuem um traço pequeno, e não somente entre os de traço alto e traço baixo.
12. *Critério do equilíbrio:* os itens do mesmo contínuo devem cobrir igualmente ou proporcionalmente todos os segmentos (setores) do contínuo, devendo haver, portanto, itens fáceis, difíceis e médios (para aptidões) ou fracos, moderados e extremos (no caso das atitudes). De fato, os itens devem se distribuir sobre o contínuo em uma distribuição que se assemelha à da curva normal: maior parte dos itens de dificuldade mediana e diminuindo progressivamente em direção às caudas (itens fáceis e itens difíceis em número menor). A razão deste critério se encontra no fato de que a grande maioria dos traços latentes se distribui entre a população mais ou menos dentro da curva normal, isto é, a maioria dos sujeitos possui magnitudes medianas dos traços latentes, sendo que uns poucos possuem magnitudes grandes e outros, magnitudes pequenas. Assim, a distribuição dos itens em um instrumento deve ser mais ou menos segundo a curva normal, como mostrado na Figura 8.5 a seguir, onde se diz que 10% dos itens

devem ter dificuldade mínima ou máxima, 40% dificuldade mediana, etc.

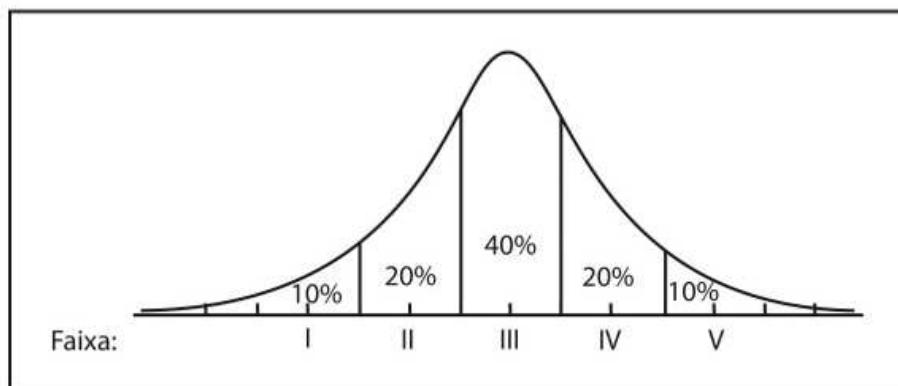
### **Quantidade de itens**

Para se cobrir a totalidade ou a maior parte ou, pelo menos, grande parte da extensão semântica do construto, explicitada nas definições constitutivas, normalmente se exige, no instrumento final, um número razoável de itens. O que é um número razoável? O bom senso de quem trabalha nesta área sugere que um construto, para ser bem representado, necessita de cerca de 20 itens. Há, evidentemente, construtos muito simples que difficilmente necessitam de tal número, sendo suficientes apenas uma meia dúzia ou menos deles. Por exemplo, em relação à satisfação com o salário. Quantas maneiras há de se verificar tal satisfação? Parece exagerado perguntar 20 vezes ao sujeito se está satisfeito com o seu salário. Posso, sim, perguntar se ele está contente com a quantia, com o poder de compra, com a pontualidade de entrega, e mais alguns aspectos. Mas parece difícil descobrir umas 20 maneiras de estar satisfeito com o salário. Entretanto, a grande maioria dos traços latentes normalmente possui uma gama bem maior de aspectos e, por

isso, exige maior número de itens para serem adequadamente representados.

Se o número final de itens, isto é, depois que o instrumento passou por todas as fases de construção e validação, deve ser em torno de 20, pergunta-se com quantos itens é preciso começar para que no final possamos salvar 20. A resposta dada no contexto da psicometria tradicional positivista é a de que se deve começar com, pelo menos, o triplo de itens para se poder assegurar, no final, um terço deles. Esta resposta se deve ao modo positivista ou ateórico de construir instrumentos psicológicos. Nesse enfoque, os itens não são construídos a partir de uma teoria; eles são coletados ou selecionados de uma tal *pool of items* que parece medir um dado construto e, em seguida, analisados estatisticamente para ver quais deles se salvam. Quer dizer, os itens são aqui simplesmente chutados; eles são selecionados simplesmente porque parecem medir o que quero medir.

Dentro da técnica de construção de instrumentos baseada na teoria dos traços latentes que estamos expondo, para se salvar 20 itens no final de toda a elaboração e validação do instrumento, não é necessário iniciar com mais do que 10% de itens além dos 20 requeridos no instrumento final. Isso porque os itens incluídos



**Figura 8.5**

Distribuição percentual dos itens em cinco faixas de dificuldade.

no instrumento piloto são itens que possuem validade teórica real, e não simplesmente que *parecem* ter validade.

### Análise teórica dos itens

Operacionalizado o construto por intermédio dos itens, estou diante da hipótese de que eles representam adequadamente o tal construto. Essa é a minha versão da hipótese a ser testada. Contudo, é importante avaliar a minha hipótese contra a opinião de outros para me assegurar de que ela apresenta garantias de validade. Essa avaliação ou análise da hipótese (análise dos itens) é obviamente ainda teórica porque consiste simplesmente em pedir outras opiniões sobre minha hipótese, sendo que os outros que a vão avaliar ainda não são uma amostra representativa da população para a qual construí o instrumento. Essa análise teórica é feita por juízes e comporta dois tipos distintos deles, segundo a análise incida sobre a compreensão dos itens (análise semântica) ou sobre a pertinência dos itens ao construto que representam (propriamente chamada de análise dos juízes). Assim, antes de partir para a validação final do instrumento piloto, este é submetido a uma análise teórica dos itens por meio da análise semântica e da análise dos juízes.

### Análise semântica dos itens

A análise semântica tem como objetivo precípua verificar se todos os itens são compreensíveis para todos os membros da população a que o instrumento se destina. Nela duas preocupações são relevantes:

1. verificar se os itens são inteligíveis para o estrato mais baixo (de habilidade) da população meta e, por isso, a amostra

para essa análise deve ser feita com esse estrato;

2. para evitar deselegância na formulação dos itens, a análise semântica deverá ser feita também com uma amostra mais sofisticada (de maior habilidade) da população meta (para garantir a chamada “validade aparente” do teste).

Entende-se por estrato mais baixo aquele segmento da população meta que apresenta menor nível de habilidades. Assim, por exemplo, se meu teste se destina a uma população que congrega sujeitos do ensino fundamental até universitários, obviamente o estrato mais baixo nesse contexto são os sujeitos do ensino fundamental, e o mais sofisticado será representado pelos sujeitos de nível universitário. De qualquer forma, a dificuldade na compreensão dos itens não deve se constituir em fator complicador na resposta dos indivíduos, uma vez que não se quer medir a compreensão deles (a não ser, obviamente, que o teste queira medir precisamente isso), mas sim a magnitude do atributo a que os itens se referem. Que técnica se deve utilizar para fazer essa análise? Há várias maneiras eficientes para tal tarefa, como, por exemplo, aplicar o instrumento a uma amostra de uns 30 sujeitos da população meta e em seguida discutir com eles as dúvidas que os itens suscitarem. Entretanto, uma técnica que se tem mostrado das mais eficazes na avaliação da compreensão dos itens consiste em checá-los com pequenos grupos de sujeitos (três ou quatro) em uma situação de *brainstorm*. Essa técnica funciona da seguinte forma: constitui-se um grupo de até quatro sujeitos, iniciando com sujeitos do estrato baixo da população meta, porque se supõe que, se tal estrato compreende os itens, *a fortiori* o estrato mais sofisticado também os compreenderá. A esse grupo apresenta-se item por item, pedindo que ele seja reproduzido pelos

membros do grupo. Se a reprodução do item não deixar nenhuma dúvida, o item é corretamente compreendido. Se surgirem divergências na reprodução do item ou se o pesquisador perceber que está sendo entendido diferentemente do que ele julga que deveria ser entendido, este item tem problemas. Dada essa situação, o pesquisador então explica ao grupo o que ele pretendia dizer com tal item. Normalmente, neste caso, os próprios sujeitos do grupo irão sugerir como se deveria formular o item para expressar o que o pesquisador queria dizer com ele; e aí está o item reformulado como deve ser. Quantos grupos são necessários para proceder a essa análise semântica? Bem, itens que não ofereceram qualquer dificuldade de compreensão em uma ou no máximo duas sessões não necessitam de checagem ulterior. Itens que continuam apresentando dificuldades após, digamos, no máximo de cinco sessões merecem ser simplesmente descartados. Em seguida a essas sessões, é importante pelo menos uma sessão de checagem dos itens com um grupo de sujeitos mais sofisticados. O objetivo desta última verificação consiste em evitar que os itens se apresentem demasiadamente primitivos para estes sujeitos e que assim percam a validade aparente. É que os itens devem também dar a impressão de seriedade – como diz o ditado de que a mulher de César não somente deve ser honesta,

mas deve também parecer honesta! (Ver regra número 10 dos critérios de construção de itens.)

### Análise dos juízes

Esta análise é, às vezes, chamada de análise de conteúdo, mas propriamente deve ser chamada de análise de construto, uma vez que precisamente procura verificar a adequação da representação comportamental do(s) atributo(s) latente(s).

Na análise de conteúdo, os juízes devem ser peritos na área do construto, pois sua tarefa consiste em ajuizar se os itens estão se referindo ou não ao traço em questão. Uma tabela de dupla entrada, com os itens arrolados na margem esquerda e os traços no cabeçalho, serve para coletar a informação. Uma concordância de, pelo menos, 80% entre os juízes pode servir de critério de decisão sobre a pertinência do item ao traço a que teoricamente se refere.

A técnica exige que sejam dadas aos juízes duas tabelas: uma com as definições constitutivas dos construtos/fatores para os quais se criaram os itens e outra tabela de dupla entrada com os fatores e os itens, como no Quadro 8.1, em que são avaliados os itens que medem os dois fatores (compreensão verbal e fluência verbal) de raciocínio verbal. Normalmente é

**QUADRO 8.1**  
Tabelas para a análise dos itens pelos juízes

Fatores	Definição	Itens	Compreensão verbal	Fluência verbal
Compreensão verbal	É a capacidade de...	1	X	
Fluência verbal	É a capacidade de...	2		X
		3		X
		...		
		N	X	

necessária uma terceira tabela que elenca os itens, uma vez que a tabela de dupla entrada geralmente não comporta a expressão completa do conteúdo dos itens.

Com base nessas tabelas, a função dos juízes consiste em colocar um X para o item debaixo do fator ao qual o juiz julga o item se referir. Uma meia dúzia de juízes será suficiente para realizar a tarefa. Itens que não atingirem uma concordância de aplicação aos fatores (cerca de 80%) obviamente apresentam problemas, e seria o caso de descartá-los do instrumento piloto. Isso vale, contudo, se o construto para o qual se está construindo o teste apresentar fatores (particularmente quando forem em maior quantidade) que se supõem ou se sabe que não são correlacionados. Quando os fatores se supõem que sejam correlacionados, acontece que uma mesma tarefa (item) pode se referir, certamente com níveis de saturação diferente, mas de fato se referir simultaneamente a mais de um fator, o que implicaria que os juízes iriam mostrar alguma discordância quanto à aplicação do item a este ou àquele fator. Nesse caso, a discordância deve ser considerada como concordância. Uma outra solução seria instruir os juízes a marcar, para cada item, não o fator, mas aqueles fatores aos quais o item se refere. Entretanto, com tal dica, você abre campo para muita divagação por parte dos juízes, e você perde a utilidade prática dessa análise. Seria melhor instruir os juízes para colocar, se possível, cada item sob um fator somente.

Com o trabalho dos juízes ficam completados os procedimentos teóricos na construção do instrumento de medida, que comportaram a explicitação da teoria do(s) construto(s) envolvido(s), bem como a elaboração do instrumento piloto que constitui a representação comportamental destes mesmos construtos e que se põe como a hipótese a ser empiricamente testada (validação do instrumento). Esta

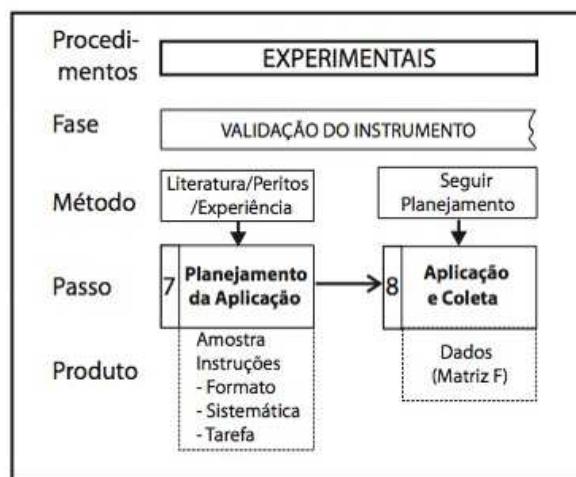
tarefa será iniciada com os procedimentos que seguirão, que consistem em coletar a informação empírica válida e submetê-la às análises estatísticas pertinentes em psicométrica, como veremos.

## PROCEDIMENTOS EXPERIMENTAIS

Os procedimentos envolvidos nesta etapa fazem apelo direto ao conteúdo da disciplina ensinada nas instituições universitárias sob o nome de delineamento ou planejamento de pesquisa, cujo conhecimento é absolutamente necessário, uma vez que garante a tecnologia da coleta válida da informação empírica. Aqui serão, por isso, explicitados apenas alguns pontos dessa tecnologia que têm mais a ver diretamente com o problema de elaboração de instrumentos psicológicos, mas o conhecimento aprofundado da citada disciplina é imprescindível.

Dois passos são salientados nestes procedimentos empíricos na validação do instrumento piloto: o planejamento da aplicação e a própria coleta da informação empírica, conforme detalha a Figura 8.6.

Com referência ao planejamento da aplicação do instrumento piloto, dois



**Figura 8.6**

Procedimentos empíricos na elaboração de medida psicológica.

pontos são particularmente relevantes: a definição da amostra e das instruções de como aplicar o instrumento.

Quanto à *amostra*: um instrumento é tipicamente construído para um certo tipo de população. Esta, consequentemente, deve ser claramente definida e delimitada em termos de suas características específicas. Assim, é necessário se determinar para que faixa etária o instrumento foi construído, para que nível socioeconômico, para que nível de escolaridade, etc. Enfim, é preciso dizer qual é o tipo de indivíduo, em termos de características biossociodemográficas, que constitui a população meta do instrumento. E é dessa população que sairá a amostra de sujeitos para a testagem da qualidade psicométrica do instrumento de medida. Obviamente, aqui se deve recorrer à teoria e às técnicas de amostragem, ensinadas na disciplina de planejamento de pesquisa ou similar.

Salientamos aqui apenas alguns aspectos relevantes da amostra para o caso específico de validação de instrumentos psicológicos. Como estamos elaborando um instrumento referente a construto, tipicamente a análise estatística a seguir utilizada para a análise dos dados será a análise fatorial e as análises multivariadas da TRI. Essas técnicas estatísticas fazem algumas exigências importantes dos dados, especificamente que eles produzam suficiente variância para que a análise seja consistente. Essa afirmação normalmente implica, pelo menos, que as amostras utilizadas sejam grandes. Quanto grandes?

Há duas dicas úteis para responder a essa pergunta. Primeiro, se eu estiver seguro de quantos fatores o meu instrumento mede (o que foi teoricamente definido quando se discutiu o passo da dimensionalidade do objeto psicológico que o instrumento iria medir), então a dica é de que a amostra deve conter um mínimo de 100 sujeitos por fator medido. Assim, se

meu instrumento mede dois fatores, necessito de 200 sujeitos na minha amostra. Estamos supondo aqui que a população meta seja homogênea em relação ao traço latente que o instrumento mede. Se o traço varia dentro da população, não somente em termos de magnitude, o que é de se esperar, mas em termos de estrutura, isto é, ele se torna de fato um traço psicologicamente diferente para diferentes estratos da mesma população, então estamos falando não mais de um traço latente, mas de dois ou mais. Nesse caso, estamos assumindo que instrumentos diferentes são necessários para avaliar traços diferentes. Mas se o traço se mantém qualitativamente (em termos de sua estrutura conceitual, de sistema) o mesmo na população, então essa população é homogênea. Um exemplo: um teste de inteligência para adultos não inclui crianças na sua população, pois a inteligência da criança é qualitativamente diferente da dos adultos, segundo teorias (Piaget, Spearman, etc.) e dados empíricos. Assim, a amostra para validação de um teste de inteligência para adultos deve ser selecionado de uma população de adultos exclusivamente, que, nesse sentido, se torna uma população homogênea.

Segunda dica: se houver dúvidas sérias quanto ao número de dimensões ou fatores que o instrumento mede, costuma-se dizer que são necessários para a amostra 10 sujeitos para cada item do instrumento. Assim, um instrumento com 100 itens demandaria 1.000 sujeitos. Isso equivaleria a supor que o instrumento estivesse medindo cerca de 10 fatores. Esse modo de pensar está mais ligado ao sistema positivista de construir instrumentos, em que os itens não são construídos via teoria e sim “pescados” aleatoriamente e em seguida analisados via análise fatorial para ver quantos fatores está medindo. De qualquer forma, é uma dica ainda útil, quando há dúvidas com respeito ao

número de fatores. Geralmente, entre 5 e 10 sujeitos por item do instrumento serão suficientes para responder à questão do tamanho da amostra, com a ressalva de que qualquer análise fatorial e da TRI com menos de 200 sujeitos dificilmente pode ser considerada adequada.

Quanto às *instruções*: estas se referem aos contornos da tarefa do sujeito que vai responder ao instrumento. Aqui são definidas a sistemática de aplicação do instrumento, o formato em que ele se apresenta e o que o sujeito tem que fazer ao respondê-lo. No tocante à *sistemática*, serão definidas as condições de aplicação: se será coletiva ou individual; se será preciso ou não aviso prévio aos testandos; são necessários contatos prévios com diretores, chefes dos sujeitos, etc. Enfim, devo saber em que estou “me metendo” e quais são as dificuldades que vou encontrar ao querer aplicar o instrumento em uma amostra definida de sujeitos, pois eles normalmente não estão gratuitamente disponíveis às minhas necessidades de pesquisador. Por isso, tenho que elaborar uma estratégia de convencimento, para os responsáveis dos sujeitos que entrarão na amostra, e uma estratégia operacional para poder viabilizar a aplicação do instrumento.

No referente ao formato do instrumento, deve-se decidir como a resposta do sujeito será dada para cada item. Aqui existe uma infinidade de formatos possíveis, como, por exemplo, o da *escolha forçada*, em que dois itens são apresentados simultaneamente, sendo a tarefa do sujeito escolher um deles como mais apropriado, mais típico, ou mais o que seja, bastante comuns em testes de personalidade e mais ainda em testes de interesse; o das *múltiplas alternativas*, mais comuns em testes de aptidão, em que o sujeito deve escolher a alternativa correta; o das *escalas tipo Likert*, em que a cada item segue uma escala de pontos (de 2 a mais de 10) que exprimem a intensidade de acordo do

sujeito com o que o item está afirmando. Este último formato é o mais utilizado no caso de testes de personalidade e escalas de atitudes. Todos esses e outros formatos apresentam vantagens e desvantagens. Por exemplo, o caso da escolha forçada, em testes de atitudes e personalidade, parece ser a maneira mais fácil de responder, pois o sujeito tem melhores condições de escolher entre duas alternativas do que dar uma resposta absoluta, como é o caso nas escalas de Likert. Contudo, dois problemas graves existem com este formato de escolha forçada: primeiro, se você vai comparar os itens do instrumento dois a dois, o instrumento se torna muito rapidamente de um comprimento incontrolável. Por exemplo, um teste com apenas 10 itens terá  $n(n-1)/n$  questões, isto é, 45 questões, e um de 100 itens terá 4.950! Além dessa dificuldade, existe o problema da chamada deseabilidade social, a saber, os dois itens que estão sendo comparados devem possuir mais ou menos o mesmo nível de atratividade, do contrário a própria questão já está dando a resposta ao sujeito se um dos itens da questão é socialmente desejável e o outro indesejável, como, por exemplo, escolher entre “A – sou uma pessoa simpática” e “B – sou uma pessoa fraca”. Nesse caso, a maioria das pessoas iria escolher a alternativa A. Certo?

No caso do formato de múltipla escolha, existem os problemas do número de alternativas e da qualidade das alternativas. Primeiramente, como se trata de respostas certas e erradas, apenas uma das alternativas será a correta. Mas, quando o sujeito não sabe a resposta correta, ele tem a chance de “chutar” e acertar por acaso; e isso é um problema, que é tanto mais grave quanto menor for o número de alternativas. Por exemplo, em um item com duas alternativas, o sujeito tem a chance de acertar por acaso em 50% das vezes, ao passo que em um item com cinco alternativas essa chance cai para 20%,

mas ainda não é zero. Então, deve-se ter maior do que menor número de alternativas para diminuir o acerto aleatório. Mas, fazendo isso, você torna o teste cada vez mais difícil de construir, porque não é tarefa fácil inventar alternativas, uma vez que elas devem de fato se apresentar como alternativas plausíveis e atrativas (e este é o segundo problema), isto é, elas devem ter alguma aparência de serem respostas corretas, do contrário não são alternativas. Assim se você constrói o seguinte item:

A camada mais externa da pele se chama:

- a) epiderme
  - b) paquiderme
  - c) dermatologia
  - d) epidemia

é claro que b, c, d não constituem alternativas plausíveis ou sérias.

Quanto às escalas tipo Likert, pergunta-se frequentemente qual é o número ideal de pontos que a escala de resposta deve ter e qual o formato ideal da escala.

Com respeito ao formato das escalas: Existem os mais variados modos de apresentar essas escalas, mas que finalmente se reduzem a escalas verbais, numéricas ou gráficas, sendo estas últimas normalmente ancoradas, ou combinação das três.

Vejamos:

Esses e outros tipos de formatos não parecem ter maior impacto sobre a resposta do sujeito, de sorte que o formato da escala depende mais do gosto pessoal do pesquisador do que qualquer outra razão técnica. Pessoalmente acho que quanto mais leve a escala, melhor, e a escala numérica e gráfica me parece muito pesada; mas sua opinião é tão boa quanto a minha.

Quanto ao número de pontos: normalmente as afirmações ou itens são respondidos em uma escala de 3 ou mais pontos, isto é, o sujeito tem que dizer se concorda, está em dúvida ou discorda do que a frase afirma sobre o objeto psicológico. O número de pontos na escala de resposta varia de 3 a mais de 10, sendo as mais utilizadas as escalas de 5 e 7 pontos. O número de pontos utilizados nas escalas Likert parece, novamente, ser algo irrelevante. Na pesquisa de Matell e Jacoby (1972), foram utilizadas escalas com 2 até 19 pontos. Com exceção das escalas de 2 e 3 pontos (por oferecerem poucos graus de liberdade), em todas as outras a porcentagem de uso dos pontos e o tempo de resposta não foram afetados de modo significativo. Outros estudos já haviam descoberto que o número de pontos da escala e a existência ou não de um ponto neutro não afetam a consistência interna da escala Likert (Bendig, 1954; Komorita, 1963; Matell e Jacoby, 1971), nem a estabilidade teste-reteste (Jones, 1968; Van der Veer, Howard e Austria, 1970; Goldsam,

1971; Matell e Jacoby, 1971) e nem a validade concorrente e preditiva (Matell e Jacoby, 1971; 1972).

As *instruções* que acompanham o instrumento têm a função única de tornar a tarefa do respondente inambígua. Consequentemente, elas devem poder deixar absolutamente claro o que o sujeito tem que fazer para responder corretamente o teste e, por isso, elas devem ser avaliadas na análise semântica. Algumas precauções: as instruções devem informar em termos gerais sobre que é o teste; devem ser as mais curtas possíveis, sem sacrificar a compreensão da tarefa por parte de todos os sujeitos da população meta; devem, tipicamente, conter um ou mais exemplos de como os itens devem ser respondidos; devem pôr o sujeito em um estado psicológico livre de tensão e ansiedade.

Finalmente, no que se refere à própria coleta da informação (passo 8), devem-se seguir todas as precauções exigidas em qualquer aplicação de instrumentos psicológicos, a saber, os sujeitos devem ser postos em um ambiente condizente e livre de distrações e de tensão,

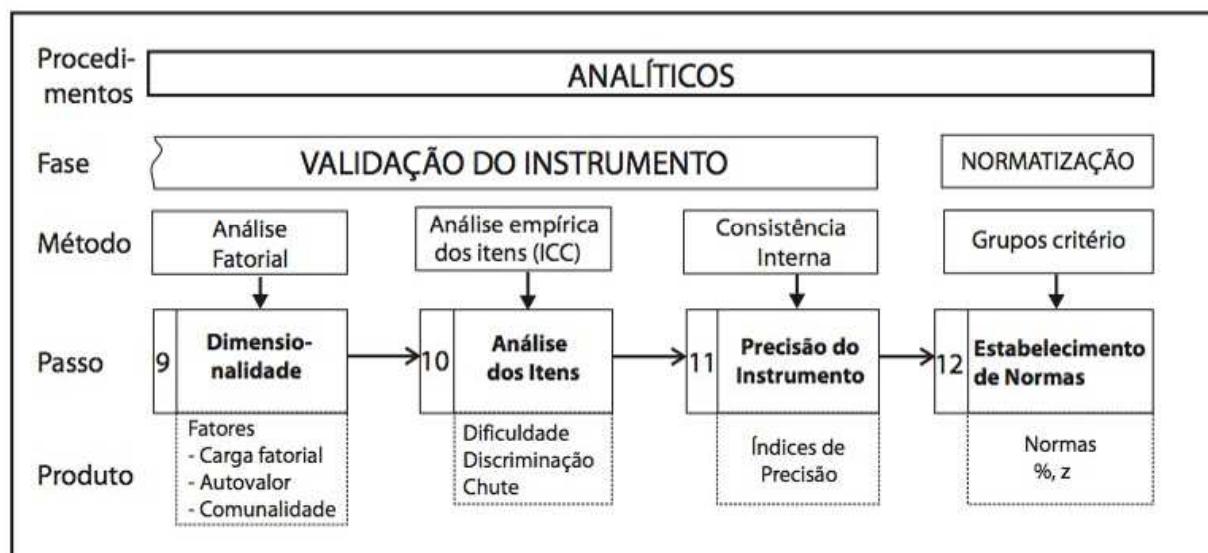
o aplicador deve ser competente para a tarefa, etc.

## PROCEDIMENTOS ANALÍTICOS

Esta parte da elaboração de instrumentos psicológicos (ver Figura 8.7) é aquela que mais atemoriza os psicólogos, dada a sua sofisticação estatística. Ela comporta igualmente a parte mais volumosa de qualquer livro sobre psicométrica. Entretanto, o conhecimento da estatística e da psicométrica não são aqui substituíveis. Felizmente, o psicólogo pode apelar neste particular para a ajuda de estatísticos ou de psicométristas. A sofisticação nesta área é tão grande que não é possível ser exposta neste capítulo. Para tanto, são recomendadas as obras que em seguida serão citadas, sendo a exposição de conteúdo neste capítulo apenas exemplificativa.

### Algumas obras básicas de análise psicométrica

Anastasi, A. (1988). *Psychological testing* (6th ed). New York: Macmillan.



**Figura 8.7**

Procedimentos analíticos na elaboração da medida psicológica.

- Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
- Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Norwell, MA: Kluwer Nijhoff.
- Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: Sage.
- Hambleton, R.K., & Zaal, J.N. (Eds.). (1991). *Advances in educational and psychological testing: Theory and applications*. Boston, MA: Kluwer Academic.
- Harman, H.H. (1967). *Modern factor analysis*. Chicago, IL: University of Chicago Press.
- Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Muñiz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Pirámide.
- Muñiz, J. (1992). *Teoría clásica de los tests*. Madrid: Pirámide.
- Muñiz, J. (1996). *Psicometría*. Madrid: Univérsitas.
- Nunnally, J.C., Jr. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Pasquali, L. (2009). *Psicometria: Teoria dos testes na psicologia e na educação* (3. ed). Rio de Janeiro, RJ: Vozes.
- Pasquali, L. (2007). *TRI – Teoria de resposta ao item: Teoria, procedimentos e aplicações*. Brasília, DF: LabPAM.
- Santisteban, C. (1990). *Psicometría*. Madrid: Norma.
- Thorndike, R.L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.
- Yela, M. (1987). *Introducción a la teoría de los tests*. Madrid: Facultad de Psicología, Universidad Complutense.

### **Algumas anotações sobre os procedimentos analíticos**

#### **A dimensionalidade do instrumento (validade)**

As análises estatísticas que se fazem de um instrumento psicológico, no seu todo e em cada item individual, fazem a suposição de que o instrumento seja unidimensional. Isso implica que todos os itens do instrumento estejam medindo um e o mesmo construto. Dessa forma, estando o instrumento medindo mais de um

fator, as análises estatísticas devem ser feitas independentemente para cada fator. Como inicialmente ainda não se sabe se o instrumento que acaba de ser construído e aplicado é ou não unidimensional, a primeira análise que se impõe sobre os dados empíricos coletados é a verificação da unidimensionalidade. Tipicamente, necessita-se proceder a uma análise factorial para definir a *dimensionalidade* do instrumento. Essa análise vai determinar quantos fatores o instrumento está de fato medindo. Essa exigência pode parecer um tanto frustrante, uma vez que, se o instrumento foi construído para medir um fator somente, por exemplo, então não se pode supor que esteja medindo somente este fator para o qual foi construído em primeiro lugar? É bom lembrar aqui que o instrumento constitui uma hipótese; mas, agora estamos verificando essa hipótese empiricamente, então é necessário se demonstrar, e não somente supor, que o instrumento de fato mede um único fator ou quantos e quais fatores ele está medindo. Aliás, essa análise factorial constitui a demonstração da própria validade do instrumento e representa igualmente a análise preliminar dos próprios itens.

A *análise factorial* (ver Pasquali, 2009) produz resultados importantes com os quais se pode tomar decisões sobre a qualidade dos itens, bem como do instrumento no seu todo. Na verdade, ela mostra o que o instrumento está medindo, isto é, os fatores, bem como os itens que compõem cada fator. Ela produz, para cada item, a carga factorial (saturação) deste no fator, e esta carga factorial indica a covariância entre o fator e o item. Isso quer dizer que a carga factorial mostra a porcentagem que existe de parentesco (covariância) entre o item e o fator, de forma que quanto mais próximo de 100% de covariância item-fator, melhor será o item, pois ele assim se constitui em um excelente representante comportamental do fator (do traço la-

tente). Qual é o montante de covariância entre o item e o fator necessário para se dizer que o item é um bom representante deste? As cargas fatoriais são expressas similarmente aos índices de correlação e, portanto, podem ir de -1,00 a +1,00. Uma carga de 0,00 significa que não há relação alguma entre o item e o fator; nesse caso, o item seria uma representação comportamental totalmente equivocada do fator. Então, que nível de magnitude de carga o item deve apresentar para ser um bom representante do fator? Costuma-se apontar o valor 0,30 (positivo ou negativo) como sendo uma carga mínima necessária para o item ser um representante útil do fator. Obviamente, quanto maior de 0,30 for a carga, melhor o item. Uma carga de 0,30 indica que há uma covariância de cerca de 10% ( $0,30^2 = 0,09$ ) entre o item e o fator, o que já pode ser considerado não negligível, embora não seja lá grande coisa. Obviamente, se todos os itens de um fator apresentam cargas fatoriais em torno de 0,30, este fator está muito mal representado, porque se esperam cargas bem maiores (acima de 0,50) para se dizer que o fator foi bem representado comportamentalmente. Você vê, então, que as cargas fatoriais falam tanto da qualidade de cada item como do conjunto deles, isto é, do próprio fator. Assim, se você construiu 25 itens para representar o traço latente e, destes 25 itens, 20 apresentam cargas acima de 0,50 e 5 apresentam cargas em torno de 0,30, você irá eliminar estes últimos 5 itens e trabalhar somente com os 20 que apresentaram cargas fatoriais respeitáveis. Veja o exemplo (fictício) da Tabela 8.1.

A Tabela 8.1 exemplifica uma típica matriz fatorial com as informações essenciais sobre os itens e os fatores. Nela se vê que, dos 20 itens, 9 (com cargas fatoriais em negrito) representam o fator 1, pois possuem cargas fatoriais altas neste fator e praticamente cargas nulas no fator 2; ao

contrário, os 10 últimos itens possuem cargas fortes no fator 2 e quase nada no fator 1. O item 10 não possui carga expressiva em nenhum dos dois fatores e será, por isso, descartado do teste. Observe que as cargas fatoriais podem ser tanto positivas quanto negativas e, assim mesmo, pertencerem ao mesmo fator, contanto que elas sejam altas. É que o fato de elas serem positivas e negativas no mesmo fator apenas indicam que um item está expressando o pólo positivo e o outro o pólo negativo do fator. Por exemplo, os itens “gosto de meus pais” e “detesto meus pais”, ambos se referem à questão da filiação, apenas o primeiro item expressa o pólo positivo da filiação e o segundo, o pólo negativo.

Assim, o teste mede dois fatores, um com 9 itens e o outro com 10, mostrando-se um item (o número 10) uma representação equivocada tanto do fator 1 quanto do fator 2. Os dois fatores explicam 47,73%

**TABELA 8.1 Matriz fatorial de 20 itens em dois fatores**

Item	Fator 1	Fator 2	$h^2$
1	<b>0,80</b>	0,10	0,65
2	<b>0,78</b>	-0,05	0,61
3	<b>0,78</b>	0,20	0,65
4	<b>0,70</b>	0,15	0,51
5	<b>0,65</b>	0,08	0,43
6	<b>0,64</b>	0,12	0,42
7	<b>-0,64</b>	-0,10	0,42
8	<b>0,60</b>	0,03	0,36
9	<b>-0,60</b>	-0,23	0,41
10	-0,25	0,19	0,10
11	0,30	<b>-0,83</b>	0,78
12	0,21	<b>-0,83</b>	0,68
13	0,04	<b>-0,78</b>	0,61
14	0,16	<b>-0,70</b>	0,52
15	-0,12	<b>0,70</b>	0,50
16	0,09	<b>0,66</b>	0,44
17	-0,00	<b>-0,65</b>	0,42
18	0,12	<b>-0,63</b>	0,41
19	-0,03	<b>0,56</b>	0,31
20	0,21	<b>0,50</b>	0,29
Autovalor	4,614	4,932	
% Var. total	23,07	24,66	
% Var. comum	48,33	51,67	

(= 23,07 + 24,66) da variância total do teste, sendo o restante da variância irrelevante ao conteúdo que o teste mede (como erros de medida e peculiaridades específicas dos itens). O  $h^2$  representa a comunabilidade que cada item possui com os dois fatores e mostra a covariância de item com os fatores e, por conseguinte, o tanto que o item tem a ver com os fatores. Assim, para o item 1, o  $h^2$  é 0,65, isto é, este item possui 65% de covariância (parentesco) com os dois fatores, sendo 0,64% ( $0,80^2$ ) com o fator 1 e apenas 1% ( $0,10^2$ ) com o fator 2; disso se deduz que o item 1 é uma excelente representação comportamental do fator 1 e nada do fator 2.

Nesta questão da validade do instrumento, outras técnicas são utilizadas além da análise factorial, tais como a técnica da *validação convergente-discriminante* (Campbell e Fiske, 1967); a utilização da *idade* como critério para a validação de construto de um teste quando este mede traços que são intrinsecamente dependentes de mudanças no desenvolvimento cognitivo/afetivo dos indivíduos, como é o caso, por exemplo, na teoria piagetiana do desenvolvimento dos processos cognitivos e da teoria de Spearman sobre a inteligência; a *correlação com outros testes* que meçam o mesmo traço do meu novo instrumento e o uso da *intervenção experimental* (ver Pasquali, 2009).

### **A análise empírica dos itens**

Os itens que se mostraram ser representantes satisfatórios do traço latente que o instrumento mede (no caso da Tabela 3.2 seriam os 9 itens para o fator 1 e os 10 do fator 2) devem ser submetidos a análises individuais ulteriores, com o objetivo de verificar outras características que eles devem apresentar dentro de um mesmo instrumento, além de serem legítimos representantes do traço latente. Essas características dos itens devem ser analisadas

dentro de cada fator (os 9 itens no fator 1 do nosso exemplo, e os 10 no fator 2) e normalmente se reduzem a duas: a dificuldade e a discriminação. A dificuldade do item diz respeito à magnitude do traço latente que o sujeito deve possuir para poder acertar (testes de aptidão) ou aceitar (testes de personalidade) o item. Assim, quanto maior for a magnitude do traço latente exigida para acertar ou aceitar o item, mais difícil este é dito ser. A discriminação do item diz respeito ao fato de ele poder diferenciar sujeitos que possuem magnitudes diferentes do mesmo traço latente. Assim, quanto mais próximas forem as magnitudes do traço que o item puder diferenciar, mais discriminativo ele será.

A psicometria tradicional faz análises estatísticas para determinar esses dois parâmetros psicométricos dos itens de uma forma que pode ser hoje considerada inferior diante dos avanços da psicometria mais moderna da TRI. A TRI introduziu técnicas nesta área da análise dos itens que, embora complicadas, devem ser as utilizadas neste passo da elaboração de qualquer instrumento psicológico (ver Hambleton, Swaminathan e Rogers, 1991; Muñiz, 1990; Pasquali, 2007). Um exemplo ajudará a entender esses procedimentos (ver Figura 8.6).

Primeiramente, deve-se atentar a que existem vários modelos matemáticos envolvidos na TRI. Na verdade, há três deles principais, dependendo do número de parâmetros que pretendem avaliar dos itens. Os parâmetros em questão são a dificuldade, a discriminação e a resposta aleatória (ou melhor, a resposta correta dada ao acaso). Assim, temos os modelos logísticos de um, dois ou três parâmetros.

Todos os modelos trabalham com traços latentes, isto é, teorizam sobre as estruturas latentes. Entendem os sistemas psicológicos latentes como possuindo dimensões, isto é, propriedades de diferentes magnitudes ou mensuráveis. Por isso,

esta teoria também é conhecida como a teoria do traço latente ou a teoria da curva característica do item ou *item characteristic curve* – ICC, pelo fato de produzir para cada item uma ogiva característica dele. A teoria supõe que o sujeito possui um certo nível de magnitude do traço latente, designado por teta ( $\theta$ ), que é determinado mediante a análise das respostas dos sujeitos, fazendo uso de diversas funções matemáticas. A função do modelo completo de três parâmetros é:

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{D_a(\theta-b_i)}}{1+e^{D_a(\theta-b_i)}}$$

A probabilidade de resposta correta, que define a posição ( $\theta$ ) do indivíduo no traço medido, é função de três parâmetros: “a” corresponde ao índice de discriminação do item e é determinado pela inclinação da curva no ponto de inflexão; “b” é o parâmetro da dificuldade/preferência e é expresso pelo valor no eixo dos X no ponto de inflexão da curva; “c” é o parâmetro que determina as respostas acertadas por acaso, sendo o D uma constante usualmente com valor 1.7.

Os três modelos de TRI mais conhecidos são os seguintes:

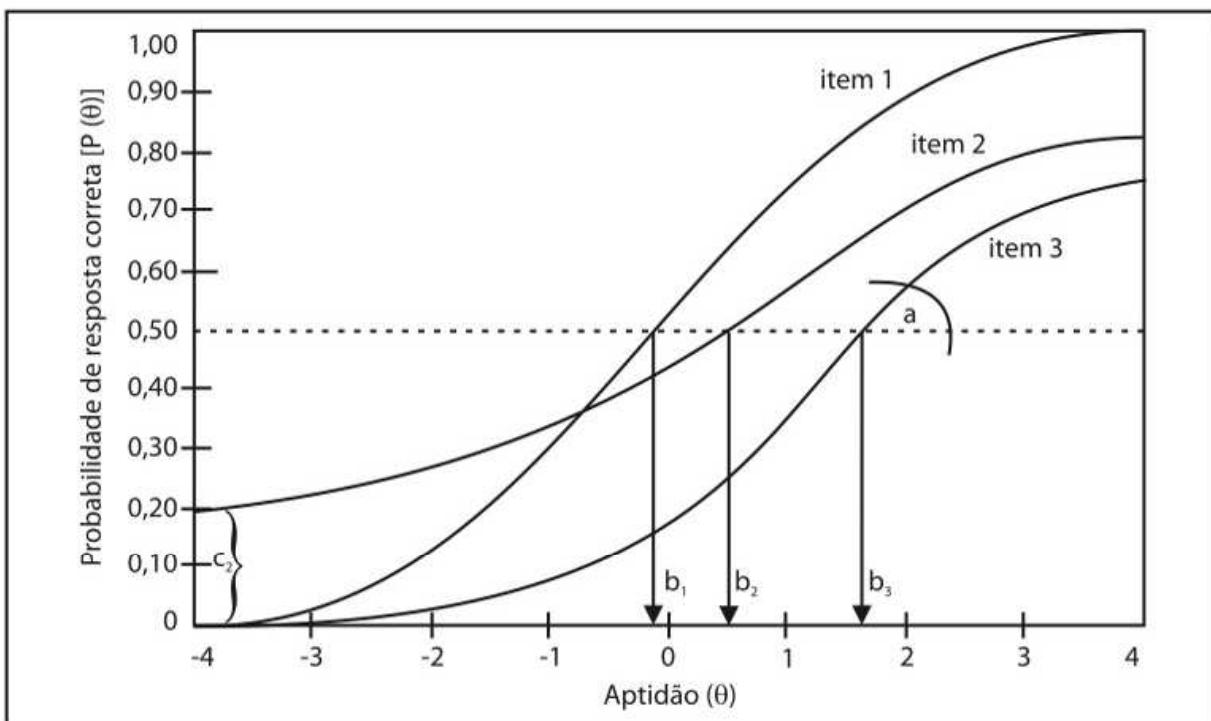
1. O modelo logístico de um parâmetro ou o modelo Rasch (1966). Rasch faz a suposição de que os itens possuem o mesmo nível de discriminação e que não há respostas dadas ao acaso, ficando como parâmetro a ser avaliado somente a dificuldade dos itens.
2. O modelo logístico de dois parâmetros (Birnbaum, 1968), que avalia a dificuldade e a discriminação dos itens, assumindo que não haja respostas dadas ao acaso.
3. O modelo de três parâmetros de Lord (1980) em que os três parâmetros dos itens são avaliados.

Exemplificando com o modelo de Lord: Os valores  $\theta$  são expressos em coordenadas cartesianas, tendo na ordenada a probabilidade de resposta correta, isto é, o  $P_i(\theta)$ , e na abscissa o traço latente, o próprio  $\theta$ . Este procedimento produz, para cada item, uma ogiva, chamada de CCI como na Figura 8.8.

Na ilustração da Figura 8.8, os três parâmetros aparecem nas seguintes posições: o “a” é representado pela inclinação da curva na altura do ponto de inflexão, isto é, onde a curva corta a linha que representa a probabilidade 0,50 de resposta correta (50%); quanto mais íngreme essa curvatura, isto é, quanto mais próxima de um ângulo de incidência de 90°, mais discriminativo é o item. O “b” é representado pela distância na linha dos X (abscissa), que corresponde ao ponto determinado pela perpendicular que vem do ponto de inflexão da curva. O “c” é definido pela assíntota inferior da curva; quando essa assíntota não atinge a abscissa há respostas dadas ao acaso, e o tamanho dessas respostas é definido pela distância que vai do ponto 0 na abscissa até o ponto onde a ogiva corta a ordenada; por exemplo, o item 2 tem cerca de 20% de resposta ao acaso. Vê-se também nesta Figura 8.8 que o item 3 é mais discriminativo do que os outros itens; igualmente, que os três itens possuem diferentes níveis de dificuldade, sendo o item 3 o mais difícil deles. Os dados oferecidos pela TRI são algebraicamente expressos em uma tabela como a que segue (onde aparecem os dados dos três itens do exemplo da Figura 8.8):

Item	a	b	c
1	1,00	-0,10	0,00
2	0,90	0,50	0,20
3	0,75	1,70	0,00
...	...	...	...

Nível ideal de dificuldade dos itens. Pode-se perguntar, ainda, se existe um



**Figura 8.8**  
CCI para três itens.

nível ideal de dificuldade para os itens de uma escala ou teste. Essa pergunta está relacionada com os critérios 11 e 12 (amplitude e equilíbrio dos itens no instrumento) das regras de construção dos itens. A resposta a essa indagação depende da finalidade do teste. Caso se deseje um teste para selecionar os melhores ou para determinar se um determinado patamar de conhecimento foi atingido (como nos testes educacionais de referência a critério), então os itens devem todos apresentar o nível de dificuldade do patamar que se quer como critério de seleção ou acima dele. Assim, se a intenção for de selecionar somente os 30% melhores candidatos, os índices de dificuldade dos itens devem ser em torno de 30% ( $p = 0,30$ ) ou menos, isto é, somente 30% dos sujeitos devem ter a probabilidade de acertar os itens. De fato, nesse caso, existe o interesse em apenas discriminar entre sujeitos de alta aptidão, sendo sem interesse itens que apenas discriminam sujeitos de menor aptidão.

Se, entretanto, o interesse consiste em avaliar a magnitude diferencial dos traços nos sujeitos de uma população, como geralmente é o caso em testes referentes a construto, então uma distribuição mais equilibrada dos itens em termos de dificuldade é requerida. Em casos como esse, o interesse se centra sobre o poder de um teste em discriminar diferentes níveis de habilidades nos sujeitos e, por conseguinte, os itens devem poder avaliar tanto os que possuem pouca quanto muita habilidade. Entretanto, é bom saber que itens que todos os sujeitos acertam ou aceitam e itens que ninguém acerta ou não aceitam são itens inúteis para fins de diferenciar indivíduos; de fato, tais itens não trazem nenhuma informação. Os itens que trazem maior informação são aqueles cujo índice de dificuldade se situa em torno de 50%, isto é, no valor 0 da escala dos sigmas, pois neste caso 50% dos sujeitos acertam e 50% erram, resultando  $50 \times 50 = 2.500$  comparações possíveis, ao passo que um item com dificuldade de 30% teria 70% de erros

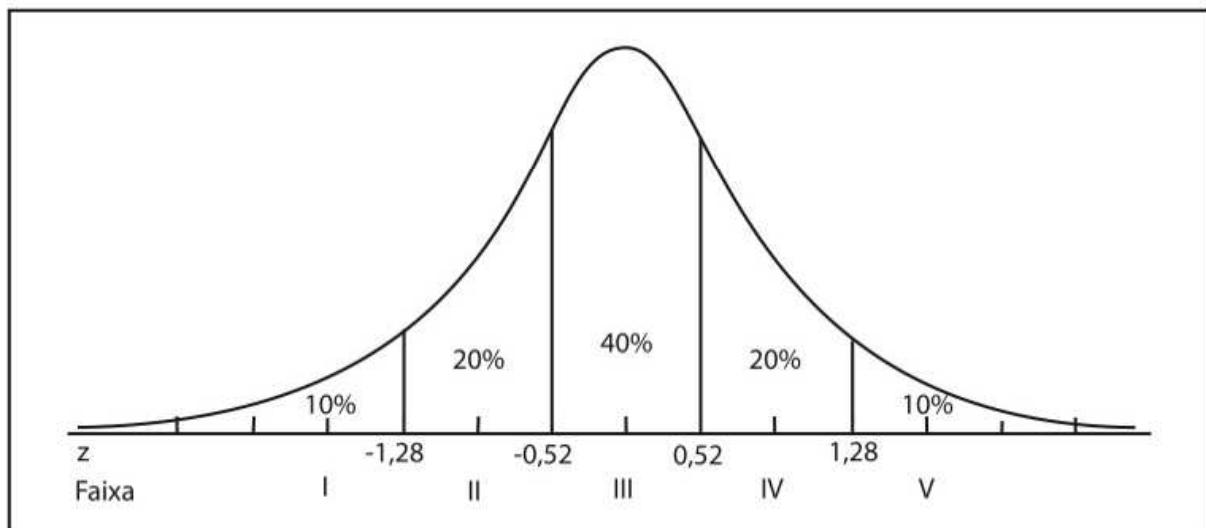
e 30% de acertos, resultando em um nível de  $30 \times 70 = 2.100$  bits de informação. Obviamente, um item com dificuldade 100% ou 0% produzirá zero informação. Deve-se concluir daí que todos os itens de um teste devem ter dificuldade de 50%? Embora grande parte dos itens deva apresentar tal índice de dificuldade, nem todos o deverão, pois que assim se poderia discriminar apenas dois níveis da magnitude do traço medido, dado que itens com o mesmo nível de dificuldade terão altas intercorrelações, determinadas pela circunstância de que serão os mesmos sujeitos que sempre acertam ou sempre erram os itens todos. Isto vale dizer que a dificuldade média dos itens do teste deve ser em torno de  $p = 0,50$ . Haveria, então, uma distribuição mais adequada dos itens de um teste em termos de dificuldade? Considerando que eles devem cobrir toda a extensão de magnitude do traço e que os itens de dificuldade 50% são os que produzem maior informação, pode-se sugerir que uma distribuição mais ou menos dentro de uma curva normal seria o ideal. Assim, se considerarmos a amplitude de um atributo ou traço em uma escala de 100 pontos, podemos dividí-la em cinco níveis de magnitudes: 0 a

20 ( $\sigma \leq -1,28$ ), 20 a 40 ( $\sigma$  entre  $-1,28$  e  $-0,52$ ), 40 a 60 ( $\sigma$  entre  $-0,52$  e  $+0,52$ ), 60 a 80 ( $\sigma$  entre  $0,52$  e  $1,28$ ) e 80 a 100 ( $\sigma \geq 1,28$ ), distribuindo os itens assim: 10% deles em cada uma das duas faixas extremas, 20% em cada uma das duas faixas seguintes e 40% na faixa média (ver Figura 8.9).

Essa discussão sobre a dificuldade ideal dos itens faz mais sentido dentro da teoria clássica dos testes. A TRI tem maneiras bem mais condizentes e apropriadas para fazer essa análise por meio do uso do índice de informação do item e do teste. Trabalhar com este último índice é bem mais complexo, mas existem softwares apropriados em abundância no mercado para auxiliar nessa tarefa. Ademais, para poder fazer uso inteligente de tal procedimento é necessário um conhecimento razoável da TRI. Por isso, o leitor deve se aprofundar no estudo de algum dos livros citados sobre a TRI.

### **Fidedignidade do instrumento**

O problema que se enquadra sob o conceito de fidedignidade vem relatado



**Figura 8.9**

Distribuição percentual dos itens em cinco faixas de dificuldade.

sob uma série de outras expressões, como precisão, fidedignidade, constância, consistência interna, confiabilidade, estabilidade, confiança, homogeneidade. As mais genéricas e, por isso, mais utilizadas são as expressões precisão e fidedignidade.

Essas diferentes expressões mostram a variabilidade de conceitos que precisão assume, dependendo do aspecto que esse parâmetro quer salientar do teste. Na verdade, fidedignidade cobre aspectos diferentes de um teste, mas todos eles se referem a quanto os escores de um sujeito se mantêm idênticos em ocasiões diferentes; por exemplo, os escores obtidos num tempo 1 e num tempo 2 para os mesmos sujeitos. Essa ocorrência (identidade dos escores) evidentemente supõe que o traço que o teste mede se mantenha constante sobre essas diferentes ocasiões, como é suposto ser o caso, por exemplo, na maioria dos traços de personalidade e de aptidão. Não seria o caso em um teste de humor, porque ele traço por natureza varia de um momento para outro, e um teste válido de humor produziria escores necessariamente diferentes em ocasiões diferentes. Assim, o conceito de fidedignidade, na verdade, se refere ao quanto o escore obtido no teste se aproxima do escore verdadeiro do sujeito em um traço qualquer; isto é, a fidedignidade de um teste está intimamente ligada ao conceito da variância erro, sendo este definido como a variabilidade nos escores produzida por fatores estranhos ao construto que o teste mede. Aparece, assim, claro que a fidedignidade de um teste depende da questão do erro da medida, especificamente do erro produzido pelo próprio instrumento: quanto o escore produzido pelo teste se distancia do escore verdadeiro do sujeito no traço em questão, isto é, a valor  $\theta$  individual na TRI.

Para melhor conceber esta problemática, é preciso se referir à variância verdadeira e à variância erro. Um procedi-

mento de medida qualquer, por exemplo, os escores em um teste, produz uma variabilidade nos resultados que, em parte, é provocada pelas diferenças no próprio traço medido entre diferentes sujeitos, e em parte pela imprecisão do próprio instrumento e em parte, ainda, por uma série de outros fatores aleatórios. A fidedignidade da medida depende do tamanho da variância erro, que é precisamente a variabilidade nos resultados provocada por esses fatores aleatórios e pela imprecisão do instrumento. Expressa mais positivamente, a fidedignidade de um instrumento diz respeito ao montante de variância verdadeira que ele produz *vis-à-vis* a variância erro, isto é, quanto maior a variância verdadeira e menor a variância erro, mais fidedigno o instrumento: um escore preciso é um escore que se aproxima do valor verdadeiro, expresso estatisticamente pelo erro padrão da medida (tratado mais adiante).

A definição estatística da fidedignidade é feita mediante a correlação entre escores de duas situações produzidos pelo mesmo teste. Se o teste é preciso, essa correlação deve ser não somente significativa, mas se aproximar da unidade (cerca de 0,90). De fato, uma correlação de 0,70, por exemplo, expressaria uma comunalidade de apenas 49% entre as duas situações provocadas pelo mesmo teste nos mesmos sujeitos. Nesse caso, a variância comum, digamos a variância verdadeira, seria menor que a variância erro, demonstrando que o teste não produz resultados fidedignos, isto é, o teste não possui precisão. Essa correlação, no caso do parâmetro de fidedignidade ou precisão, é referida como o coeficiente de precisão ou de fidedignidade.

Dependendo da técnica utilizada para cômputo da precisão de um teste, surgem vários tipos de precisão: teste-reteste, formas paralelas, consistência interna.

1. A precisão *teste-reteste* consiste em calcular a correlação entre as distribuições de escores obtidos em um mesmo teste pelos mesmos sujeitos em duas ocasiões diferentes de tempo. A correlação de 1,00 seria obtida se não houvesse variância erro provocada pelo teste ou outros fatores aleatórios, como fatores não controlados nos sujeitos ou na situação de testagem. Quanto mais longo o período de tempo entre a primeira e a segunda testagem, mais chances haverá de fatores aleatórios ocorrerem, diminuindo o coeficiente de precisão. Esse intervalo de tempo permite a ação dos fatores mencionados por Campbell e Stanley (1963) sob o tema de fontes de erro devido à história, à maturação, à retestagem e às interações entre esses fatores, bem como ao próprio instrumento. Por isso, veem-se as graves dificuldades que apresenta esse tipo de análise da fidedignidade de um teste; particularmente grave aparece aqui a questão da maturação, isto é, se o próprio traço matura (se desenvolve, modifica), essa análise da precisão torna-se errônea, dada sobretudo a eventualidade de que a maturação do traço se processasse diferencialmente para os diversos sujeitos testados. Além disso, e particularmente em testes de aptidão, a testagem constitui um treinamento, e provavelmente diferencial, para os sujeitos, o que provocará diferenças na retestagem entre eles, reduzindo novamente o coeficiente de precisão do teste. Para contornar essas dificuldades, outros tipos de análises foram elaborados, como a das formas alternativas ou análise da consistência interna.
2. Na precisão de *formas alternativas*, os sujeitos respondem a duas formas paralelas do mesmo teste, e a correlação entre as duas distribuições de escores constitui o coeficiente de precisão do

teste. A condição necessária para que essa análise seja válida se situa na demonstração de que as amostras de conteúdo (de itens) em ambas as formas sejam equivalentes, isto é, que os itens possuam níveis equivalentes de dificuldade e de discriminação em ambas. Esses parâmetros podem ser facilmente verificados por meio da TRI. Há, contudo, algumas dificuldades neste tipo de análise da precisão: as duas formas são aplicadas em sucessão imediata, não eliminando assim totalmente o efeito do intervalo de tempo, resultando na possível introdução de efeitos da história e do treinamento (prática) obtido ao responder à primeira das formas alternativas; aparece facilmente um efeito repetitório, uma vez que os itens de ambas as formas são similares, produzindo efeitos motivacionais negativos no respondente. Além disso, não é tarefa fácil construir formas alternativas, quando a construção de um só teste já é uma tarefa dispendiosa, razão pela qual poucos testes aparecem no mercado com formas alternativas.

3. A precisão da *consistência interna* é viabilizada por intermédio de várias técnicas estatísticas que visam verificar a homogeneidade da amostra de itens do teste, ou seja, a consistência interna do teste. As técnicas mais utilizadas são duas metades, Kuder-Richardson e alfa de Cronbach. Todas elas exigem aplicação do teste em uma única ocasião, evitando totalmente a questão da constância temporal.

No caso da precisão das *duas metades*, os sujeitos respondem a um único teste em uma única ocasião. O teste é dividido em duas partes equivalentes e a correlação é calculada entre os escores obtidos nas duas metades. Não é importante como o teste é dividido em duas metades, desde que estas sejam equivalentes.

Na prática, contudo, as duas formas mais normalmente utilizadas são a divisão do teste em primeira metade e segunda metade ou em itens pares e itens ímpares. Para efetuar essa análise da precisão, de fato o teste não precisa ser homogêneo, isto é, no qual todos os itens medem o mesmo traço (por exemplo, itens somente verbais ou numéricos); o que é fundamental é que as duas metades emparelhem itens homogêneos: verbal com verbal, numérico com numérico, etc.

Neste tipo de precisão, é preciso notar que o cálculo da correlação se baseia somente na metade do teste. Assim, em um teste de 100 itens, a correlação se basearia somente em 50 itens. Como o número de itens afeta o tamanho do coeficiente de correlação, é preciso corrigir esse coeficiente para que leve em consideração a extensão total do teste e, assim, produzir um coeficiente de precisão mais justo para o teste. Essa correção é feita pela fórmula de Spearman-Brown:

$$r_{tt} = \frac{nr_{12}}{1 + r_{12}}$$

onde,  $r_{tt}$  é o coeficiente de precisão calculado,  $r_{12}$  é o coeficiente de correlação entre as duas metades do teste e  $n$  é o número de vezes em que o teste foi dividido. Assim, em um teste dividido em duas metades, o  $n$  será 2, porque ele deve ser aumentado 2 vezes para se obter a forma total do teste.

A técnica de Kuder-Richardson (Kuder e Richardson, 1937) para verificar a fidedignidade de um teste se baseia na análise de cada item individual do teste. Os autores desenvolveram várias fórmulas, sendo a mais utilizada a fórmula 20, que segue:

$$r_{tt} = \left( \frac{n}{n - 1} \right) \frac{DP_t^2 - \Sigma pq}{DP_t^2}$$

onde,

$r_{tt}$  é o coeficiente de precisão do teste,  $n$ , o número de itens do teste,  $DP_t^2$ , o desvio padrão dos escores totais do teste e  $\Sigma pq$  é o somatório do produto da proporção de sujeitos que passaram ( $p$ ) e dos que não passaram ( $q$ ) cada item.

Cronbach (1951) mostrou que essa técnica produz um coeficiente de precisão do teste que corresponde à média dos coeficientes de todas as metades em que o teste possa ser dividido, mas somente quando se utiliza a fórmula de Rulon (1939), que trabalha com as variâncias das diferenças entre as duas metades, e não a simples correlação com a correção de Spearman-Brown, segundo observaram Novick e Lewis (1967). Esta equivalência de coeficientes, contudo, ocorre em testes homogêneos, porque nos testes heterogêneos os coeficientes de Kuder-Richardson são normalmente menores, dado que esta técnica não trabalha com diferenças entre pares de itens e sim com a variância de todos os itens.

O próprio Cronbach (1951) desenvolveu uma técnica geral para estabelecer a fidedignidade dos testes, o Alfa de Cronbach. Esta técnica constitui uma extensão da de Kuder-Richardson. Esta última é aplicável somente quando a resposta ao item é dicotômica: certo e errado, por exemplo. Entretanto, quando a resposta ao item pode assumir mais de duas alternativas, o valor  $\Sigma pq$  é substituído por  $\Sigma s_i^2$ , a soma das variâncias de cada item. Esta fórmula genérica é a seguinte:

$$r_{tt} = \left( \frac{n}{n - 1} \right) \frac{S_t^2 - \Sigma S_i^2}{S_t^2}$$

em que,  $S_t^2$  é a variância de todo o teste e  $\Sigma S_i^2$ , o somatório das variâncias de cada item do teste.

Um instrumento submetido à série de análises anteriormente mencionadas pode

ser considerado um instrumento válido e fidedigno e pronto para uso na pesquisa. No caso de o instrumento ser orientado para uso clínico (casos individuais), ele deve ser submetido à normatização para se poder interpretar os resultados que ele produz. Contudo, para fins de pesquisa, que tipicamente trabalha com comparações de grupos de sujeitos, a normatização não é necessária. Aliás, ela não acrescenta nada de novo e útil para a qualidade psicométrica do instrumento; apenas ela é útil para a interpretação dos resultados, pois constitui uma simples transformação dos resultados brutos do instrumento em resultados de alguma maneira padronizados e comparáveis.

## REFERÊNCIAS

- Anastasi, A. (1988). *Psychological testing* (6th ed). New York: Macmillan.
- Bendig, A.W. (1954). Reliability and the number of rating scale categories. *Journal of Applied Psychology*, 38, 38-40.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring a examinee's ability. In F.M. Ford & M. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin*, 56, 81-105.
- Campbell, D.T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.
- Campbell, D.T., & Stanley, J. (1973). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
- Embretson, S.E. (1984). A general latent trait model for response processes. *Psychometrika*, 49, 175-186.
- Embretson, S.E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics*. New York: Academic Press.
- Goldsamt, M.R. (1971). Effects of scoring method and rating scale length in extreme response style measurement. Tese de doutorado não publicada, University of Maryland.
- Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Norwell, MA: Kluwer Nijhoff.
- Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Beverly Hills, CA: Sage.
- Hambleton, R.K., & Zaal, J.N. (Eds.). (1991). *Advances in educational and psychological testing: Theory and applications*. Boston, MA: Kluwer Academic Publishers.
- Harman, H.H. (1967). *Modern factor analysis*. Chicago, IL: University of Chicago Press.
- Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- Jackson, D.N., & Messick, S. (1967). *Problems in human assessment* (Cap. 6). New York: McGraw-Hill.
- Jones, R.R. (1968). Differences in response consistency and subject's preferences for three personality inventory response formats. *Proceedings of the 67th Annual Convention of the American Psychological Association*, 3, 247-248.
- Komorita, S.S. (1963). Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology*, 61, 327-334.
- Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability *Psychometrika*, 2, 151-160.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 1-55.
- Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Mager, R.F. (1981). *Medindo os objetivos de ensino ou "conseguiu um par adequado"*. Porto Alegre: Globo.
- Matell, M.S., & Jacoby, J. (1971). Is there an optimal number of Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, 31, 657-674.
- Matell, M.S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? *Journal of Applied Psychology*, 56(6), 506-509.
- Muñiz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Pirámide.
- Muñiz, J. (1992). *Teoría clásica de los tests*. Madrid: Pirámide, S.A.
- Muñiz, J. (1996). *Psicometría*. Madrid: Univeristas.
- Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, 22, 287-293.
- Nevo, B., & Sfez, J. (1985). Examinees' feedback questionnaires. *Assessment and Evaluation in Higher Education*, 10, 236-249.
- Novick, M.R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1-13.
- Nunnally, J.C., Jr. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Pasquali, L. (Org.). (1996). *Teoria e métodos de medida em ciências do comportamento*. Brasília: INEP.

- Pasquali, L. (2004). *Psicometria: Teoria dos testes na psicologia e na educação* (2. ed). Rio de Janeiro, RJ: Vozes.
- Pasquali, L. (2006). *Análise fatorial para pesquisadores*. Brasília, DF: LabPAM.
- Pasquali, L. (2007). *TRI: Teoria de resposta ao item: Teoria, procedimentos e aplicações*. Brasília, DF: LabPAM.
- Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99-103.
- Santisteban, C. (1990). *Psicometría*. Madrid: Norma.
- Thorndike, R.L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.
- Van der Veer, F., Howard, K.I., & Austria, A.M. (1970). Stability and equivalence scores based on three different response formats. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 5, 99-100.
- Yela, M. (1987). *Introducción a la teoría de los tests*. Madrid: Facultad de Psicología, Universidad Complutense.