[capítulo 4]

FUNDAMENTOS EM FIDEDIGNIDADE

O termo *fidedignidade* sugere confiabilidade. Quando decisões de qualquer tipo devem ser tomadas, no todo ou em parte, com base em escores de testes, seus usuários precisam ter certeza de que estes escores são razoavelmente confiáveis. Quando usada no contexto dos testes e medidas, a fidedignidade se baseia na consistência e precisão dos resultados do processo de mensuração. Para terem um certo grau de confiança nos escores, os usuários de testes exigem evidências de que os escores obtidos seriam consistentes, se os testes fossem repetidos com os mesmos indivíduos ou grupos, e de que são razoavelmente precisos.

Enquanto a fidedignidade na mensuração implica consistência e precisão, a falta de fidedignidade implica inconsistência e imprecisão, ambas resultam em erros de mensuração. No contexto da testagem, um *erro de mensuração* pode ser definido como qualquer flutuação nos escores resultante de fatores relacionados aos processos de mensuração que são irrelevantes ao que está sendo medido. A *fidedignidade*, portanto, é a qualidade dos escores de teste que sugere que eles são suficientemente consistentes e livres de erros de mensuração para serem úteis.

Observe que, para serem úteis, os escores de teste não precisam ser totalmente consistentes e livres de erros. Como vimos no Capítulo 1, mesmo nas ciências físicas – algumas das quais se orgulham de uma instrumentação incrivelmente confiável – as mensurações sempre estão sujeitas a um certo grau de erro e flutuação. Nas ciências sociais e comportamentais, as mensurações são muito mais propensas a erro devido à natureza complexa dos constructos e ao fato de que os dados comportamentais, a partir dos quais eles são avaliados, podem ser afetados por muito mais fatores imprevisíveis do que outros tipos de dados (ver quadro Consulta Rápida 5.2, Desconstruindo Constructos, no Capítulo 5). Os escores de testes psicológicos, em particular, são especialmente suscetíveis a influências de diversas fontes – incluindo o testando, o examinador e o contexto no qual a testagem ocorre –, todas podem resultar em uma variabilidade alheia aos objetivos do teste.

VERDADE E ERRO NA MENSURAÇÃO PSICOLÓGICA

Uma das abordagens mais tradicionais ao tópico da fidedignidade na teoria clássica dos testes é a noção do *escore verdadeiro* (Gulliksen, 1950). De certa forma, pode-se dizer que esta idéia representa o Santo Graal da psicometria. Embora os escores verdadeiros não existam na realidade, ainda assim é possível imaginar sua existência: são as entidades hipotéticas que resultariam de mensurações inteiramente livres de erros. Os métodos para estimar a fidedignidade dos escores oferecem um meio de estimar escores verdadeiros, ou pelo menos os limites dentro dos quais estes poderiam se localizar. Os conceitos de fidedignidade e erro nos escores de teste – que obviamente devem ser levados em consideração ao se lidar com qualquer escore – são aplicados de forma paralela, mas um tanto diferente, quando se trata de um ou mais escores de um mesmo indivíduo e quando se lida com escores de grupos.

O conceito do escore verdadeiro nos dados individuais

Na teoria clássica dos testes, o *escore verdadeiro* de um indivíduo é conceitualizado como o escore médio em uma distribuição hipotética que seria obtida se o indivíduo se submetesse ao mesmo teste um número infinito de vezes. Na prática, obviamente, é impossível obter tal escore até mesmo para um único indivíduo, o que dirá para muitos. Ao invés de escores verdadeiros, o que derivamos dos testes são os *escores observados* (isto é, os escores que os indivíduos efetivamente obtêm).

Em relação a um único escore, as idéias apresentadas até este ponto podem ser representadas sucintamente por meio da seguinte equação:

$$X_O = X_{verdadeiro} + X_{erro} \qquad (4.1)$$

que expressa o conceito de que qualquer escore observado (X_o) tem dois componentes: um componente de escore verdadeiro ($X_{verdadeiro}$) e um componente de erro (X_{erro}). De um ponto de vista realista, as magnitudes destes dois componentes sempre serão desconhecidas. Não obstante, em teoria, o componente do escore verdadeiro é entendido como aquela parte do escore observado que reflete a habilidade, traço ou característica avaliada pelo teste. Inversamente, o componente de erro, que é definido como a diferença entre o escore observado e o escore verdadeiro, representa quaisquer outros fatores que possam influenciar o escore observado como conseqüência do processo de mensuração.

Escores verdadeiros em dados de grupo

A importância singular da variabilidade entre os indivíduos já foi discutida no Capítulo 2, no qual foi destacado que a utilidade da testagem psicológica depende da obtenção de algum grau de variabilidade entre os testandos. Sem a variabilidade dos escores, os testes não serviriam como auxílio para a tomada de decisões com-

parativas entre pessoas. Também devemos recordar que, no mesmo capítulo, a variância de amostra (s^2) foi definida como a quantidade média de variabilidade em um grupo de escores. Com base nesta informação, a Fórmula (4.1) – que diz respeito a um único escore de teste – pode ser extrapolada e aplicada à distribuição dos escores obtidos em uma amostra, ou uma população, da seguinte maneira:

$$\text{Variância de amostra} = s^2 = s_v^2 + s_e^2 \qquad (4.2)$$

ou

$$\text{Variância populacional} = \sigma^2 = \sigma_v^2 + \sigma_e^2 \qquad (4.3)$$

Ambas as fórmulas expressam a mesma idéia, qual seja, de que a variância em um conjunto de escores observados de uma amostra (s^2) ou população (σ^2) consiste em uma porção de variância verdadeira (s_v^2 ou σ_v^2) e uma porção de variância de erro (s_e^2 ou σ_e^2). A variância verdadeira consiste nas diferenças entre os escores dos indivíduos de um grupo que refletem sua posição na característica avaliada pelo teste. A variância de erro é composta pelas diferenças entre os escores que refletem fatores irrelevantes ao que o teste avalia. As fórmulas (4.2) e (4.3) também implicam que a fidedignidade dos escores aumenta à medida que o componente de erro diminui. De fato, o *coeficiente de fidedignidade* (r_{xx}) – que será discutido mais adiante neste capítulo – pode ser definido como a razão entre a variância verdadeira (s_v^2) e a variância total (s^2), ou,

$$r_{xx} = \frac{s_v^2}{s^2} \qquad (4.4)$$

Em outras palavras, se toda a variância nos escores de teste fosse variância verdadeira, a fidedignidade dos escores seria perfeita (1,00). Um coeficiente de fidedignidade pode ser visto como um número que estima a proporção da variância em um grupo de escores que é explicada por erros oriundos de uma ou mais fontes. Nesta perspectiva, a avaliação da fidedignidade de um escore envolve um processo em dois tempos que consiste em (a) determinar quais as possíveis fontes de erro que podem interferir nos escores e (b) estimar a magnitude destes erros.

A RELATIVIDADE DA FIDEDIGNIDADE

Embora a prática de descrever os *testes* como fidedignos seja comum, o fato é que a qualidade da fidedignidade, caso exista, pertence não aos testes, mas aos *escores* deles obtidos. Esta distinção é enfatizada consistentemente pela maioria dos autores que contribuem para a literatura contemporânea sobre a fidedignidade (quadro Consulta Rápida 4.1). Embora possa parecer sutil à primeira vista, esta distinção é fundamental para a compreensão das implicações do conceito de fidedignidade em relação ao uso de testes e à interpretação de seus escores. Se um teste for descrito como fidedigno, sugere-se que sua confiabilidade foi estabelecida perma-

CONSULTA RÁPIDA 4.1

Uma excelente compilação de textos sobre fidedignidade, com explicações mais extensas e detalhadas de muitos tópicos abordados neste capítulo, pode ser encontrado em *Score Reliability: Contemporary Thinking on Reliability Issues*, organizada por Bruce Thompson (2003b).

nentemente, em todos os aspectos, para todos os usos e com todos os usuários. Isso é o mesmo que dizer que um piano bem-afinado sempre vai estar afinado e produzirá sons igualmente bons independentemente do tipo de música tocada ou de quem o tocar. Na verdade, a qualidade do som de um piano é uma função não apenas do instrumento em si como também de variáveis relacionadas à música, ao pianista e ao ambiente (p. ex., a acústica da sala) onde o piano é tocado. Da mesma forma, embora a fidedignidade na testagem dependa, em um grau significativo, das características do teste, a fidedignidade dos escores – que é o que resulta do uso do instrumento e, assim como a música produzida pelo piano, é o que realmente importa – também pode ser afetada por muitas outras variáveis.

Além disso, mesmo aplicada aos escores de teste, a qualidade da fidedignidade é relativa. O escore que uma pessoa obtém em um teste não é fidedigno num sentido absoluto, mas pode ser mais ou menos confiável devido a fatores pertinentes unicamente ao testando (p. ex., fadiga, falta de motivação, influência de drogas, etc.) ou condições da situação de testagem (p. ex., presença de ruídos que causam distração, a personalidade do examinador, a rigidez com que o limite de tempo é observado, etc.). Todos estes fatores podem afetar individual ou conjuntamente o escore obtido em maior ou menor grau, até mesmo ao ponto dos escores se tornarem tão pouco confiáveis que precisem ser descartados. Embora não tenham relação com o teste em si, todas estas questões precisam ser levadas em consideração no processo de avaliação.

Em contraposição, quando a fidedignidade (r_{xx}) é considerada do ponto de vista dos dados de escores obtidos *a partir de uma amostra grande em condições padronizadas*, presume-se que os erros de mensuração que podem afetar os escores individuais de membros da amostra, embora ainda presentes, estejam distribuídos aleatoriamente. Uma vez que os erros aleatórios têm igual probabilidade de influen-

Não Esqueça

- A fidedignidade é uma característica dos escores de teste, e não dos testes em si.
- A fidedignidade de qualquer mensuração, e dos escores de testes psicológicos em particular, não é absoluta nem imutável. Por isso, as possíveis fontes de erro na mensuração e o grau em que estas influenciam qualquer uso específico de um teste devem ser levados em consideração, estimados e relatados sempre que escores de teste forem empregados (AERA, APA, NCME, 1999).

ciar os escores em direção positiva ou negativa, também se pode presumir que se anulem mutuamente. Mesmo neste caso, no entanto, as estimativas de fidedignidade vão variar de amostra para amostra dependendo de sua composição e das circunstâncias em que a testagem ocorre. Por exemplo, se um teste tem como alvo adultos de 18 a 90 anos, a fidedignidade estimada de seus escores será suscetível à influência de diferentes fatores, dependendo da estimativa ser baseada em dados obtidos das faixas etárias mais velhas, mais jovens ou de um grupo que seja representativo da faixa etária completa para a qual o teste se destina.

Uma observação a respeito de verdade e erro

A necessidade de identificar e investigar os componentes de verdade e erro dos escores é analisada mais detalhadamente na seção a seguir, mas é importante enfatizarmos que estes julgamentos sempre devem ser feitos em relação ao que cada teste pretende avaliar e às circunstâncias nas quais ele é administrado. Por exemplo, os escores de um teste que denotam a velocidade com que os indivíduos são capazes de encaixar 100 peças em um tabuleiro, se aplicados em condições padronizadas, de modo geral iriam refletir os níveis típicos de destreza manual dos testandos de forma bastante fidedigna. Se o mesmo teste for administrado (a) em condições planejadas de modo a distrair os testandos ou (b) em condições que distraíssem sem querer os testandos, ambos os conjuntos de escores iriam refletir níveis de destreza manual sob condições de distração. No entanto, a influência das distrações somente seria vista como uma fonte de variância de erro, ou como redutora da fidedignidade daquilo que os escores pretendem indicar, no segundo caso.

FONTES DE ERRO NA TESTAGEM PSICOLÓGICA

Como vimos, o erro pode influenciar os escores dos testes psicológicos devido a um número enorme de razões, muitas das quais estão fora do âmbito das estimativas psicométricas de fidedignidade. De modo geral, no entanto, os erros que influenciam os escores de teste podem ser categorizados como oriundos de uma ou mais das seguintes três fontes: (a) o contexto no qual a testagem ocorre (incluindo fatores relacionados ao administrador do teste, ao avaliador e ao ambiente, bem como aos motivos da aplicação do teste), (b) o testando e (c) o teste em si. Alguns erros oriundos destas fontes podem ser minimizados ou eliminados desde que práticas apropriadas de testagem sejam observadas pelas partes envolvidas no processo de desenvolvimento, seleção, administração e pontuação dos instrumentos. Outros, como a negligência do testando ou tentativas de manipular a impressão gerada por suas respostas, não podem ser eliminados, mas podem ser detectados por vários mecanismos de checagem embutidos nos testes. As práticas relacionadas ao uso apropriado dos testes – a maioria voltada para a redução do erro nos escores – são discutidas mais detalhadamente no Capítulo 7, que trata de questões relevantes à seleção, administração e pontuação, entre outras.

Para os fins da discussão sobre fidedignidade, podemos pressupor que os usuários, administradores e avaliadores de testes selecionam cuidadosamente os instrumentos mais apropriados, preparam ambientes adequados, estabelecem um bom *rapport* com os testandos e administram e pontuam os escores de acordo com procedimentos padronizados bem-estabelecidos. Além disso, também vamos pressupor que os testandos são motivados e preparados adequadamente para realizar os testes. Quer estas premissas se apliquem ou não a casos específicos, permanece o fato de que os comportamentos que elas acarretam estão sujeitos ao controle de um ou mais indivíduos envolvidos no processo de testagem e não são pertinentes ao teste de forma direta. Se estas premissas estiverem corretas, os erros nos escores que se originam de fontes não relacionadas aos testes podem obviamente ser eliminados ou pelo menos minimizados.

Ao considerarmos a fidedignidade no restante deste capítulo, as fontes de erro discutidas dizem respeito basicamente a fatores que estão fora do controle consciente das partes envolvidas no processo de testagem, ou seja, fatores aleatórios ou casuais. Antes de prosseguirmos, no entanto, devemos observar que o erro de mensuração pode ser sistemático e consistente, bem como aleatório. Assim como uma balança pode acrescentar ou diminuir alguns quilos, um teste pode ter características intrínsecas que afetem todos os testandos. As estimativas tradicionais de fidedignidade podem não detectar este tipo de erro consistente, dependendo de sua fonte, porque se baseiam em métodos criados para detectar inconsistências nos resultados. Erros consistentes e sistemáticos de mensuração afetam não apenas a fidedignidade, mas também a validade dos resultados. Para detectá-los, é preciso comparar os resultados de um instrumento com os de outras ferramentas que avaliem o mesmo constructo, mas não compartilhem o fator que causa o erro consistente. Para detectar o erro no caso de uma balança desregulada, por exemplo, seria necessário pesar a mesma pessoa ou objeto em uma ou mais balanças bem-calibradas.

O quadro Consulta Rápida 4.2 lista algumas possíveis fontes de erro que podem tornar os testes inconsistentes. Esta lista categoriza as fontes de erro avaliadas pelas estimativas tradicionais de fidedignidade juntamente com os tipos de teste aos quais dizem respeito mais diretamente e os coeficientes de fidedignidade tipicamente usados para estimá-las. Uma explicação conceitual de cada fonte de erro e estimativas de fidedignidade é apresentada a seguir, na mesma ordem em que são listadas no quadro Consulta Rápida 4.2. Considerações a respeito de quando e como estes conceitos e procedimentos são aplicados no processo de uso dos testes são discutidas em uma seção subseqüente deste capítulo.

Diferenças entre avaliadores

Diferenças entre avaliadores (ou *entre pontuadores*) é o nome dado aos erros que podem influenciar escores sempre que o elemento da subjetividade desempenha um papel na avaliação de um teste. Presume-se que juízes diferentes nem sempre vão designar exatamente os mesmos escores ou notas ao desempenho em um mesmo teste, mesmo se (a) as instruções de pontuação especificadas no manual do teste forem explícitas e detalhadas e (b) os avaliadores forem conscienciosos ao

CONSULTA RÁPIDA 4.2

Fontes de erro de mensuração com os coeficientes de fidedignidade típicos usados para estimá-las

Fonte de erro	Tipo de teste propenso a cada fonte de erro	Medidas apropriadas para estimar erros
Diferenças entre avaliadores	Testes avaliados com algum grau de subjetividade	Fidedignidade do avaliador
Erro de amostragem de tempo	Testes de traços ou comportamentos relativamente estáveis	Fidedignidade de teste-reteste (r_{tt}), ou coeficiente de estabilidade
Erro de amostragem de conteúdo	Testes para os quais a consistência de resultados é desejada como um todo	Fidedignidade de forma alternativa (r_{tt}) ou fidedignidade pelo método das metades (*split-half*)
Inconsistência entre itens	Testes que requerem consistência entre os itens	Fidedignidade pelo método das metades ou medidas mais rígidas de consistência interna, como a fidedignidade de Kuder-Richardson 20 (K-R 20) ou o coeficiente alfa (α)
Inconsistência entre itens e heterogeneidade de conteúdo combinadas	Testes que requerem consistência e homogeneidade entre os itens	Medidas de consistência interna e evidências adicionais de homogeneidade
Erros de amostragem de tempo e conteúdo combinados	Testes que requerem estabilidade e consistência dos resultados como um todo	Fidedignidade de forma alternativa com intervalo

aplicar estas instruções. Em outras palavras, a variabilidade dos escores que se deve a diferenças entre avaliadores não implica negligência nem na preparação das instruções para a pontuação, nem na avaliação do teste, mas sim se refere a variações que têm origem em diferenças no julgamento subjetivo dos avaliadores.

Fidedignidade do avaliador

O método básico para estimar erros devidos a diferenças entre avaliadores consiste em fazer com que pelo menos dois indivíduos diferentes avaliem o mesmo conjunto de testes, para que o desempenho de cada testando gere dois ou mais escores independentes. As correlações entre os conjuntos de escores gerados desta maneira são índices de *fidedignidade do avaliador*. Correlações muito altas e positivas, da

ordem de 0,90 ou mais, sugerem que a proporção de erro devida às diferenças entre avaliadores é de 10% ou menos, uma vez que $1 - (\geq 0{,}90) = \leq 0{,}10$.

Erro de amostragem de tempo

O *erro de amostragem de tempo* se refere à variabilidade inerente aos escores de teste como função do fato de serem obtidos em um determinado momento do tempo e não em outro. Este conceito se baseia em duas noções relacionadas, quais sejam: (a) que qualquer constructo ou comportamento avaliado por um teste é passível de flutuação com o tempo e (b) que alguns dos constructos e comportamentos avaliados estão muito menos sujeitos a mudanças ou se alteram em ritmo muito mais lento do que outros. Por exemplo, constructos psicológicos relacionados a habilidades, como compreensão verbal ou aptidão mecânica, geralmente são vistos como menos propensos a flutuações do que constructos relacionados à personalidade, como cordialidade ou empatia. No campo do funcionamento emocional e da personalidade, a diferença entre constructos mais e menos estáveis foi codificada na distinção tradicional entre *traços* – que são entendidos como características relativamente duradouras – e *estados*, que são por definição condições temporárias. Em certo grau, esta distinção também pode ser aplicada às características cognitivas. A habilidade verbal, por exemplo, é considerada muito mais estável em um indivíduo do que sua capacidade de atenção e memória, ambas são mais suscetíveis às influências de condições transitórias ou estados emocionais. De qualquer modo, está claro que, embora se pressuponha que uma certa quantidade de erro de amostragem de tempo está presente em todos os escores de teste, via de regra deve-se esperar que sua influência seja menor naqueles testes que avaliam traços relativamente estáveis.

Fidedignidade de Teste-reteste

Para gerar estimativas da quantidade de erro de amostragem de tempo que afeta os escores de um dado teste, costuma-se administrar o mesmo teste em duas ocasiões diferentes, separadas por um certo intervalo de tempo, a um ou mais grupos de indivíduos. A correlação entre os escores obtidos nas duas administrações é um *coeficiente de fidedignidade de teste-reteste* (ou *estabilidade*) (r_{tt}) e pode ser vista como um índice do grau em que os escores podem flutuar como resultado de erro de amostragem de tempo. Quando este procedimento é usado, o intervalo de tempo entre as duas administrações do teste sempre tem que ser especificado, pois obviamente vai afetar a estabilidade dos escores. Na realidade, no entanto, existem muitos fatores que podem afetar de modo diferente os escores de testes derivados de um grupo de pessoas em duas ocasiões distintas. Por isso, não existe um intervalo fixo que possa ser recomendado para todos os testes. Se o intervalo for muito curto, por exemplo, os testandos podem se lembrar das respostas que deram na primeira ocasião, o que pode afetar seus escores na repetição. Por outro lado, se o intervalo for longo demais, sempre existe a possibilidade de que experiências

intervenientes – incluindo passos que os testandos possam ter dado como reação à primeira administração – venham a afetar os escores da segunda ocasião. Além destas considerações, os usuários de testes também devem avaliar os coeficientes de estabilidade do ponto de vista das expectativas teóricas pertinentes aos traços e comportamentos avaliados pelo teste. Um exemplo seriam as diferenças na taxa de mudanças que podem ser esperadas como função da idade dos testandos. A compreensão da leitura, por exemplo, pode mudar muito rapidamente em crianças pequenas, mas deve se manter estável durante a vida adulta, a menos que seja afetada por alguma circunstância incomum – como um treinamento especial ou uma lesão cerebral.

Erro de amostragem de conteúdo

Erro de amostragem de conteúdo é o termo usado para indicar a variabilidade irrelevante aos traços que pode influenciar os escores de teste como resultado de fatores fortuitos relacionados ao conteúdo de itens específicos. Um exame simples de como o erro de amostragem de conteúdo pode influenciar escores é apresentado no quadro Consulta Rápida 4.3. Esta ilustração é um tanto limitada porque diz respeito a um erro resultante de falhas na construção do teste e que, portanto, poderia ser evitado com facilidade. No exemplo, a seleção dos itens feita pelo pro-

CONSULTA RÁPIDA 4.3

Uma ilustração simples de erro de amostragem de conteúdo resultante de falhas na construção de um teste

Tomemos o caso de um teste de sala de aula referenciado no domínio que busca avaliar o conhecimento de todo o material contido em cinco capítulos de um livro. Suponhamos que o professor que prepara o teste desenvolva a maioria dos itens a partir do conteúdo de apenas três capítulos, deixando de incluir itens dos dois capítulos restantes. Suponhamos ainda que vários alunos também tenham se concentrado em apenas três capítulos ao estudarem para o teste.

O erro de amostragem de conteúdo nos escores deste teste resultaria primordialmente da amostragem irregular do material que ele deveria cobrir. *Não havendo outras falhas*, as conseqüências da amostragem incorreta de conteúdo por parte do professor seriam as seguintes: (a) aqueles alunos que estudaram os mesmos três capítulos dos quais foi retirado o conteúdo do teste teriam escore próximo de 100%; (b) aqueles que se concentraram em dois destes capítulos e um dos outros teriam notas de cerca de 67%, e (c) aqueles que tiveram o azar de se concentrar em apenas um dos capítulos "certos" e nos dois capítulos que não foram incluídos no teste teriam escores de aproximadamente 33%.

Se assumimos que todos os alunos que estudaram apenas três dos cinco capítulos tinham dominado 60% do material que o teste deveria cobrir, seus escores verdadeiros deveriam ter aproximadamente esta percentagem. As discrepâncias entre os escores obtidos e seu verdadeiro nível de competência do material é o erro de amostragem de conteúdo. Neste caso em particular, o erro nos escores reduziria não apenas a confiabilidade dos escores como também sua validade.

fessor resulta em uma cobertura inadequada do conteúdo de conhecimento que o teste pretende avaliar. Conseqüentemente, uma boa parte da variabilidade dos escores não está relacionada, por parte dos alunos, ao nível de competência do material tornando seus escores não apenas menos confiáveis mas também menos válidos do que poderiam ser. Um exemplo mais típico são casos em que – por razões que fogem ao controle do desenvolvedor – o conteúdo específico de um teste favorece ou prejudica alguns testandos, devido a suas experiências de vida diferentes. Por exemplo, um teste que visa avaliar a compreensão da leitura pode acidentalmente incluir diversas passagens que são conhecidas por alguns testandos e não por outros. Obviamente, os testandos familiarizados com as passagens poderão responder às perguntas baseadas nelas com mais facilidade e rapidez do que o resto, devido à sua maior familiaridade com o material, mais do que a um nível mais alto de realização em compreensão de leitura.

Fidedignidade de forma alternativa

Os *procedimentos de fidedignidade de forma alternativa* procuram estimar a quantidade de erro nos escores de teste que pode ser atribuída ao erro de amostragem de conteúdo. Para investigar este tipo de fidedignidade, duas ou mais formas diferentes de um teste – com objetivos idênticos, mas conteúdo específico diferente – precisam ser preparadas e administradas ao mesmo grupo de sujeitos. Os escores dos testandos em cada versão são então correlacionados para se obter *coeficientes de fidedignidade de forma alternativa* (r_{11}). Já que é improvável que os mesmos fatores do acaso que favorecem alguns testandos, e não outros, venham a afetar as diferentes formas do teste, correlações altas e positivas (p. ex., 0,90 ou mais) entre escores nas várias formas podem ser tomadas como indicação de que o erro de amostragem de conteúdo não exerce grande influência nos escores (p. ex., 10% ou menos). A fidedignidade de forma alternativa com intervalo, uma variação deste procedimento usada para avaliar os efeitos combinados da amostragem de tempo e conteúdo, é discutida mais adiante nesta seção.

Não esqueça

A expressão *não havendo outras falhas*, que aparece no quadro Consulta Rápida 4.3 e em outras partes deste livro, é um artifício retórico para sugerir que todas as considerações pertinentes, além do conceito em discussão no momento, devem ser desconsideradas temporariamente para fins de isolamento e esclarecimento da questão em destaque.

É importante lembrarmos, no entanto, que a premissa de que "não há outras falhas" raramente é realista na testagem, assim como em outros aspectos da vida. A expressão tem por objetivo (a) alertar o leitor para a possibilidade de que várias outras coisas precisam ser levadas em conta, além do conceito específico em discussão, e (b) estimular o leitor a refletir sobre quais elas podem ser.

Fidedignidade pelo método das metades (split-half)

O desenvolvimento de formas alternativas de um teste, ou a administração do mesmo teste duas vezes, costuma envolver problemas teóricos e práticos que dificultam esses cursos de ação. Uma solução é simplesmente administrar um teste a um grupo de indivíduos e criar dois escores para cada pessoa, dividindo o teste pela metade.

Como dividir um teste pela metade. A melhor forma de dividir os testes para calcular coeficientes de fidedignidade pelo método das metades depende de seu delineamento. Em particular, é imperativo considerar duas possibilidades: (a) se alguns itens diferem sistematicamente de outros ao longo de todo o teste e (b) se a velocidade tem um papel significativo no desempenho dos testandos. Ambas as condições podem ter efeitos profundos na magnitude dos coeficientes de fidedignidade das metades.

1. *Diferenças sistemáticas entre itens* podem ocorrer devido a várias razões. Por exemplo, muitos testes de habilidades começam com os itens mais fáceis e se tornam progressivamente mais difíceis, ou são divididos em partes ou subtestes que cobrem conteúdos diferentes. Ainda outros, como o Teste Wonderlic de Pessoal, são estruturados em *formato de coletânea em espiral*, de modo que itens ligados a tarefas verbais, numéricas, espaciais e analíticas se alternem sistematicamente. Muitos inventários de personalidade também são arranjados de modo que itens de diferentes escalas reapareçam ao longo do teste.
2. *Quando o desempenho no teste depende primariamente da velocidade,* os itens geralmente são formulados num nível de dificuldade baixo o bastante para que todos os testandos possam completá-los corretamente, mas são fixados limites de tempo para que a maioria deles não consiga terminar o teste. Por exemplo, testes de aptidão administrativa muitas vezes incluem tarefas que exigem que os testandos examinem uma longa lista de pares de números, letras ou símbolos em um período de tempo breve e indiquem se cada par é ou não idêntico. Neste tipo de teste de alta velocidade, os escores dependem primariamente do número de itens completados, mais do que do número de respostas corretas. Como a maioria dos testandos vai apresentar um desempenho perfeito ou quase perfeito em todos os itens, qualquer divisão de tal teste em termos de itens – bem como qualquer medida de consistência interna – vai produzir coeficientes próximos da perfeição.

O quadro Consulta Rápida 4.4 apresenta algumas possíveis soluções para o problema de como dividir vários tipos de testes. Depois que isso é feito, a correlação entre os escores nas duas metades (r_{hh}) é usada para derivar um *coeficiente de fidedignidade pelo método das metades (split-half)*. Como na verdade estima a consistência dos escores nos dois meios-testes, a *fórmula de Spearman- Brown (S-B)* é

CONSULTA RÁPIDA 4.4

Algumas soluções para o problema de como dividir um teste pela metade

- Uma regra básica para dividir testes de vários tipos para o cálculo de coeficientes de fidedignidade pelo método das metades é *dividir o teste nas duas metades mais comparáveis*. Embora isso possa ser feito de muitas formas, costuma-se fazer uma divisão em pares e ímpares, com os itens pares (2, 4, 6, etc) e ímpares (1, 3, 5, etc) compondo as duas metades.
- Quando a velocidade tem um papel no desempenho no teste, qualquer estimativa de fidedignidade obtido com uma única administração – como o método das metades – vai produzir resultados espúrios altos. Isso ocorre porque, para testes significativamente acelerados, a fidedignidade dos escores é primariamente uma função da consistência da velocidade com que os testandos realizam o teste, em oposição à consistência do calibre de suas respostas. Assim, para testes acelerados, uma possível solução é usar *métodos de fidedignidade com dupla administração*, como o de teste-reteste ou de formas alternativas. Outra é dividir o teste em termos de metades, com tempos contados separadamente, e então calcular o coeficiente de fidedignidade do mesmo modo que no habitual método das metades.
- *Por que isso é importante para um potencial usuário de testes?* Se o método usado para calcular estimativas da consistência interna dos escores não for apropriado ao delineamento do teste, os coeficientes de fidedignidade resultantes serão enganosos. Os potenciais usuários de testes que estão considerando a questão da fidedignidade no processo de seleção de um teste precisam atentar para estas questões.

aplicada ao r_{hh} para se obter a estimativa para o teste completo. Esta fórmula se baseia na proposição da teoria clássica dos testes de que um número maior de observações vai produzir um resultado mais confiável do que um número menor. Em outras palavras, não havendo outras falhas, um escore baseado em um teste mais longo vai estar mais próximo do escore verdadeiro do que outro baseado em um teste mais curto. A versão geral da fórmula S-B é

$$r_{S\text{-}B} = \frac{nr_{xx}}{1 + (n - 1)r_{xx}} \qquad (4.5)$$

em que

$r_{S\text{-}B}$ = a estimativa de um coeficiente de fidedignidade de Spearman-Brown,
n = o multiplicador pelo qual a extensão do teste deve aumentar ou diminuir, e
r_{xx} = o coeficiente de fidedignidade obtido com a extensão original do teste.

Esta fórmula pode ser usada para estimar que aumentar um teste ou diminuí-lo a qualquer fração de seu tamanho original terá efeito no coeficiente obtido. Por exemplo, se o r_{xx} obtido para um teste de 30 itens for 0,80 e desejarmos estimar qual seria o coeficiente de fidedignidade se o teste fosse aumentado para 90 itens, acrescentando-se 60 itens *comparáveis*, descobriríamos que n seria 3 e $r_{S\text{-}B}$ seria 0,92. Se quiséssemos encurtar o mesmo teste para 15 itens, n seria ½ e $r_{S\text{-}B}$ diminuiria para

0,67. Quando aplicada a um coeficiente de fidedignidade pelo método das metades (r_{hh}), que envolve estimar a fidedignidade do teste como um todo baseado na correlação entre suas duas metades, a fórmula de S-B pode ser simplificada da seguinte maneira:

$$r_{S\text{-}B} = \frac{2r_{hh}}{1 + r_{hh}} \qquad (4.6)$$

Inconsistência entre itens

A *inconsistência entre itens* se refere a erros nos escores que resultam de flutuações nos *itens* ao longo do teste, em oposição ao erro de amostragem de conteúdo que emana da configuração particular dos itens incluídos no teste como um todo. Embora as inconsistências possam ficar aparentes, em um exame cuidadoso do conteúdo dos itens – e dos processos cognitivos que podem estar em jogo ao se responder aos diferentes itens de um teste –, do ponto de vista estatístico, elas se manifestam em correlações baixas entre eles. Estas inconsistências podem se dever a uma variedade de fatores (incluindo erro de amostragem de conteúdo), muitos dos quais são fortuitos e imprevisíveis, e também podem resultar da heterogeneidade do conteúdo.

Heterogeneidade de conteúdo

A *heterogeneidade de conteúdo* resulta da inclusão de itens, ou conjuntos de itens, que exploram o conhecimento de conteúdos ou funções psicológicas que diferem daquelas exploradas por outros itens no mesmo teste. Este fator está em grande parte sob o controle dos criadores de testes, que devem determinar o grau de heterogeneidade do conteúdo do teste com base em seus objetivos e no tipo de população para o qual ele se destina. Se um teste é delineado com a finalidade de obter amostras de conteúdo heterogêneo, esta heterogeneidade não pode ser considerada uma fonte de erro. A heterogeneidade no conteúdo ou nas funções cognitivas exploradas pelos diferentes itens é fonte de erro somente quando o teste pretende ser homogêneo em um ou mais aspectos ao longo de todos os seus itens. O quadro Consulta Rápida 4.5 mostra alguns conjuntos de itens que variam em termos de heterogeneidade.

Medidas de consistência interna

Medidas de consistência interna são procedimentos estatísticos que procuram avaliar a extensão da inconsistência entre os itens de um teste. Os coeficientes de fidedignidade pelo método das metades realizam esta tarefa em certa medida. No entanto, mesmo um teste muito curto pode ser dividido de várias formas diferentes – por exemplo, um teste de quatro itens pode ser dividido de três formas diferentes; um

Não Esqueça

Desconstruindo a heterogeneidade e a homogeneidade

Na testagem psicológica, os conceitos de homogeneidade e heterogeneidade são usados em referência à composição de: (a) amostras de comportamento ou itens que formam um teste e (b) grupos de testandos, como amostras de padronização ou populações. Uma vez que ambos estes aspectos podem afetar todas as funções estatísticas usadas para avaliar testes (ver, p. ex., a seção sobre restrição de amplitude e correlação, bem como a Figura 2.7 do Capítulo 2), é importante recordar o seguinte:

- *Heterogeneidade* e *homogeneidade sempre são termos relativos*. Qualquer entidade que seja composta de elementos separados é heterogênea se seus elementos são desiguais em algum aspecto. Portanto, qualquer grupo composto de constituintes multidimensionais, como pessoas ou itens de teste, é heterogêneo em algum aspecto. Seguindo a mesma lógica, nenhum destes grupos é homogêneo em todos os aspectos.
- *Para se caracterizar um grupo* como heterogêneo ou homogêneo é necessário decidir qual variável ou variáveis vão servir como base para avaliar a semelhança ou a diferença. Por exemplo:
- *Os itens de um teste* podem ser heterogêneos em relação a conteúdo e formato – se alguns consistem em palavras e outros consistem em números, ou se alguns são apresentados oralmente e outros o são por escrito –, mas homogêneos em relação à função cognitiva se todos eles envolvem a memória (p. ex., recordar palavras e números).
- *Um grupo de pessoas* pode ser heterogêneo em relação a idade e sexo, se inclui homens e mulheres entre 17 e 45 anos, mas homogêneo em relação ao nível de escolaridade, se incluir apenas calouros de universidade.

teste de seis itens, de 10 formas, etc. – e cada uma delas pode produzir uma correlação diferente entre as metades. Um modo de superar este problema logístico é fazer uma divisão entre pares e ímpares – com metade do teste consistindo em itens pares e metade em itens ímpares – ou qualquer outra divisão que resulte em duas metades mais comparáveis (ver o quadro Consulta Rápida 4.4).

Outra solução é oferecida por fórmulas que levam em conta a correlação entre itens (isto é, a correlação entre o desempenho em *todos os itens* de um teste). As duas fórmulas usadas com maior freqüência para calcular a consistência entre itens são a fórmula *Kuder-Richardson 20 (K-R 20)* e o *coeficiente alfa* (α), também conhecido como *alfa de Cronbach* (Cronbach, 1951), que é simplesmente um caso mais geral do postulado da K-R 20. Tanto a K-R 20 como o coeficiente alfa requerem uma única administração do teste a um grupo de indivíduos. A magnitude dos coeficientes K-R 20 e alfa é uma função de dois fatores: (a) o número de itens do teste e (b) a razão entre a variabilidade no desempenho dos testandos em todos os itens e a variância total nos escores do teste. Não havendo outras falhas, a magnitude da K-R 20 e do coeficiente alfa serão mais altas: (a) à medida que o número de itens aumenta e (b) à medida que a razão entre a variância nos itens do teste e a variância total deste diminui. Conceitualmente, tanto a K-R 20 como o coeficiente alfa produzem estimativas de fidedignidade equivalentes à média de todos os coeficientes das metades possíveis que resultariam de todas as possíveis formas de se dividir o teste. Assim sendo, representam uma estimativa combinada do erro de

CONSULTA RÁPIDA 4.5

Exemplos de conjuntos de itens do conteúdo mais heterogêneo ao menos heterogêneo

Conjunto (A)

Item 1: Qual é o próximo número na seguinte série?

3 6 12 24 _____

Item 2: Qual dos cinco itens listados difere mais dos outros quatro?

Porco Vaca Galinha Atum Vitela

Item 3: Um trem percorre 12m em meio segundo. Nesta mesma velocidade, que distância vai percorrer em quatro segundos?

Conjunto (B)

Item 1: 4 + 10 = _____

Item 2: Se uma dúzia de ovos custa $ 1,03, quanto vão custar três dúzias?

Item 3: O preço de um item é reduzido em 60% durante uma liquidação. Em que percentagem ele deve ser aumentado para voltar ao preço original?

60% 80% 100% 120% 150%

Conjunto (C)

Item 1: 4 x 5 = _____

Item 2: 7 x 11 = _____

Item 3: 15 x 15 = _____

- *O conjunto A é o mais heterogêneo:* Os itens diferem em termos de domínios de conteúdo, formatos e habilidades exigidas.
- *O conjunto B vem a seguir:* Os itens são do mesmo domínio de conteúdo (matemática), mas diferem em formato e habilidades exigidas (isto é, adição, multiplicação e frações em matemática mais habilidades básicas de leitura).
- *O conjunto C é o mais homogêneo:* Seus itens têm em comum o domínio de conteúdo, o formato e as habilidades exigidas (compreensão da operação de multiplicação e seus símbolos).

amostragem de conteúdo, bem como da sua heterogeneidade. Portanto, a menos que um teste seja altamente homogêneo, as fidedignidades da K-R 2 e do coeficiente alfa serão mais baixas do que qualquer um dos coeficientes calculados pelo método das metades. O quadro Consulta Rápida 4.6 contém a fórmula K-R 20 e uma versão da fórmula do coeficiente alfa, juntamente com uma explicação básica de seus componentes e de sua aplicabilidade.

Uma vez que a K-R 20 e o coeficiente alfa dependem muito da quantidade de variabilidade entre os itens de um teste, fica óbvio que qualquer falta de uniformidade, como a heterogeneidade de conteúdo, vai diminuir estes coeficientes. Por exemplo, suponhamos que fossem calculados coeficientes de consistência interna para três testes de igual extensão – compostos por itens como os apresentados nos conjuntos A, B e C do quadro Consulta Rápida 4.5. Se isso fosse feito, o teste semelhante à mistura de itens do Conjunto A em termos de heterogeneidade teria

CONSULTA RÁPIDA 4.6

Fórmulas para cálculo da consistência interna

Fórmula Kuder-Richardson 20 (K-R 20)

$$r_{K-R\,20} = \left(\frac{n}{n-1}\right)\frac{s_t^2 - \sum pq}{s_t^2}$$

em que

n = número de itens do teste
S^2T = variância dos escores totais
$\sum pq$ = soma de p vezes q para cada item do teste
P = proporção de pessoas que passam em cada item ou o respondem em uma direção específica
Q = proporção de pessoas que são reprovadas em cada item ou o respondem na direção oposta

A fórmula $r_{K\,R\,20}$ se aplica a testes cujos itens são avaliados como certos ou errados, ou de qualquer outra forma dicotômica, como verdadeiro ou falso, se todos os itens forem formulados de tal modo que o sentido de cada alternativa é uniforme ao longo de todo o teste.

Coeficiente alfa (α) ou alfa de Cronbach

$$\alpha = \left(\frac{n}{n-1}\right)\frac{s_t^2 - \sum (s_i^2)}{s_t^2}$$

em que

n = número de itens do teste
S_t^2 = variância dos escores totais
$\sum (S_i^2)$ = soma das variâncias dos escores de itens

Esta fórmula do coeficiente alfa é uma variação de item padronizado conhecida como *alfa*, que usa a correlação média entre itens em vez das variâncias de escores de itens e do escore total, são usadas para testes cujos itens têm múltiplas respostas possíveis (p. ex., *concordo totalmente, concordo, discordo* e *discordo totalmente*). Cortina (1993) faz uma extensa discussão do significado das fórmulas do coeficiente alfa e os vários fatores que podem afetar seus resultados.

a consistência interna mais baixa, porque as diferenças entre a maestria dos testandos nas várias habilidades e domínios de conteúdo explorados pelos itens estariam refletidas em seu desempenho. O teste com os itens mais homogêneos, ou seja, itens como os do Conjunto C, teria o coeficiente mais alto.

Por que isto é importante? Quando um teste é delineado intencionalmente, de modo a incluir itens diversos em termos de uma ou mais dimensões, a K-R 20 e o coeficiente alfa vão superestimar o erro de amostragem de conteúdo, sendo por isso inadequados. Dependendo do delineamento do teste, e com base no exame de seu conteúdo, itens homogêneos podem ser colocados em subtestes ou segregados de alguma outra forma para permitir o cálculo de medidas separadas de consistência entre itens e entre grupos de itens semelhantes. Por outro lado, quando a homogeneidade entre todos os itens é desejada, a magnitude da K-R 20 ou do coeficiente

alfa é um índice do grau em que seu objetivo foi atingido. Na verdade, a diferença entre a magnitude do coeficiente de fidedignidade calculado pelo método das metades mais apropriado e os coeficientes K-R 20 ou alfa pode ser tomada como uma indicação da quantidade de heterogeneidade dos itens de um teste. Quanto mais próximas as duas estimativas estiverem, mais homogêneo será o conteúdo.

Técnicas de análise fatorial também podem ser usadas para investigar a heterogeneidade e a possível multidimensionalidade de itens de teste. Estas técnicas, discutidas mais longamente no Capítulo 5, são usadas para detectar semelhanças entre um conjunto de variáveis – como respostas a itens – com base na interrelação de seus padrões de variabilidade entre um ou mais grupos de testandos.

Erros de amostragem de tempo e conteúdo combinados

Erros de amostragem de tempo e de conteúdo podem ser estimados de forma combinada para testes que requerem ao mesmo tempo estabilidade e consistência de resultados. Como veremos mais adiante, também é possível estimar os efeitos combinados de outras fontes de erro em escores de testes através de outros meios. No entanto, o delineamento da forma alternativa com intervalo oferece um bom método para se estimar erros de amostragem de tempo e conteúdo com um único coeficiente.

Fidedignidade de forma alternativa com intervalo

Coeficientes de fidedignidade de forma alternativa com intervalo podem ser calculados quando duas ou mais formas alternativas do mesmo teste são administradas em duas ocasiões diferentes, separados por um certo intervalo de tempo, a um ou mais grupos de indivíduos. Assim como na fidedignidade de teste-reteste, o intervalo entre as duas administrações precisa ser especificado claramente, juntamente com a composição das amostras e outras condições que podem afetar a magnitude dos coeficientes obtidos. Se as duas formas de um teste são administradas em sucessão próxima ou imediata, o coeficiente de forma alternativa resultante será basicamente uma função da fidedignidade entre as formas. Com intervalos mais longos entre as administrações, a variância de erro nos escores vai refletir não só as flutuações de tempo como o erro de amostragem de conteúdo do teste.

Uma observação sobre efeitos da prática. Uma conseqüência inevitável de se usar o mesmo teste ou formas alternativas de um teste repetidamente com os mesmos sujeitos é que isso introduz uma fonte adicional de variabilidade, indesejada nos escores devido a *efeitos da prática*. Naturalmente, a duração do intervalo entre as administrações afeta o grau no qual os escores da segunda administração ou das administrações subseqüentes vão estar sujeitos a esses efeitos. Com intervalos curtos, os efeitos da prática podem ser muito significativos, especialmente quando os itens do teste envolvem tarefas novas que exigem que os testandos aprendam certas estratégias de solução de problemas com probabilidade de serem relembradas. Os métodos de administração única para estimar a fidedignidade dos escores, como

a técnica das metades e o coeficiente alfa, não são vulneráveis aos efeitos da prática, enquanto procedimentos de administração dupla, como o teste-reteste e a fidedignidade de forma alternativa geralmente o são. Quando os indivíduos diferem na quantidade de melhora demonstrada na retestagem devido à prática, as correlações obtidas entre as duas administrações devem ser reduzidas. Mais importante, no entanto, é o fato de que, quando os testes são administrados repetidamente para fins de avaliação de mudança ao longo do tempo, como nos estudos longitudinais, os efeitos da prática podem ser uma variável complicadora significativa, que deve ser levada em consideração (Kaufman e Lichtenberger, 2002, p. 163-165).

FIDEDIGNIDADE NO USO DE TESTES

A fidedignidade dos escores é uma consideração pertne na testagem psicológica, devido à possibilidade sempre presente de que erros de várias fontes influenciem os resultados de testes. No entanto, o modo como a fidedignidade é considerada difere em vários pontos do processo do desenvolvimento de um teste, bem como em sua aplicação. Do ponto de vista de um usuário de testes, que é o mais pertinente para nossos objetivos, as estimativas de fidedignidade devem ser cuidadosamente consideradas e aplicadas nos estágios de (a) seleção do teste e (b) interpretação dos escores. O Capítulo 7 trata destas questões mais detalhadamente, mas como os usos das estimativas estatísticas de fidedignidade são apresentados neste capítulo, as diferentes maneiras como esta é considerada em cada estágio serão introduzidas agora para proporcionar um contexto para a discussão futura.

Considerações sobre a fidedignidade na seleção de testes

Quando precisam escolher que teste usar para um dado objetivo, os usuários devem se voltar para os dados que já foram coletados a respeito da fidedignidade dos escores de cada instrumento específico. Estes dados geralmente podem ser encontrados nos manuais, guias e artigos preparados pelos autores ou criadores dos testes, mas também podem aparecer na literatura psicológica como resultado do trabalho de investigadores independentes. Tipicamente, os dados sobre fidedignidade são apresentados na forma de coeficientes de correlação. Devido ao uso generalizado do coeficiente *r* de Pearson para se avaliar a fidedignidade e a validade dos escores de teste, os aspectos essenciais deste método correlacional, incluindo suas limitações (discutidas no Capítulo 2), devem ser plenamente compreendidos antes de nos aprofundarmos nesses tópicos. Um fato particularmente relevante a ser lembrado é que a magnitude dos coeficientes de correlação depende, em certo grau, da variabilidade das amostras com as quais eles foram calculados (ver a seção dobre Restrição da Amplitude e Correlação no Capítulo 2).

Os vários tipos de coeficientes que podem ser calculados para se estimar erros de mensuração, juntamente com as fontes mais pertinentes de erro, já foram descritos, ainda que de forma abstrata. No momento de selecionarem um teste, os

usuários em potencial precisam aplicar estas noções às situações particulares nas quais desejam empregar essa ferramenta. O quadro Consulta Rápida 4.7 lista os passos básicos envolvidos na seleção de um teste do ponto de vista da fidedignidade, e uma discussão mais extensa destas considerações é apresentada nos parágrafos a seguir. Como advertência preliminar, deve-se observar que, ao avaliarmos as características psicométricas de um teste – seja em relação à fidedignidade, validade, dados normativos ou qualquer outro aspecto técnico de um instrumento – não há regras fixas aplicáveis a todos os testes ou seus usos.

CONSULTA RÁPIDA 4.7

Considerações sobre fidedignidade na seleção de testes

Passo 1 Determinar as fontes potenciais de erro que podem influenciar os escores dos instrumentos sob revisão.

Passo 2 Examinar os dados de fidedignidade disponíveis para esses instrumentos, incluindo os tipos de amostras das quais foram obtidos.

Passo 3 Avaliar os dados de fidedignidade à luz de todos os outros atributos dos testes em questão, como dados normativos e válidos, limitações de custo e tempo, etc.

Passo 4 Não havendo outras falhas, selecionar o teste que prometa produzir os escores mais fidedignos para os fins e a população em questão.

Avaliando possíveis fontes de erro em escores de testes

A precaução primordial para minimizar o erro em escores de teste é aderir estritamente a procedimentos padronizados para sua administração e avaliação (ver Capítulo 7). Além disso, os usuários de testes precisam avaliar a possível relevância de cada uma das fontes de erro listadas no quadro Consulta Rápida 4.2, tendo em vista as escolhas disponíveis de instrumentos e os objetivos para os quais eles podem ser empregados. Por exemplo:

- Se a avaliação de um teste envolve julgamentos subjetivos, a fidedignidade do avaliador deve ser considerada.
- Se um teste vai ser usado para avaliar mudanças ao longo do tempo, como uma possível melhora resultante de uma intervenção terapêutica, a estimativa do erro de amostragem de tempo – bem como de possíveis efeitos da prática – nos escores dos instrumentos sob consideração é essencial.
- Se existe a possibilidade de que uma pessoa tenha que ser retestada em um momento posterior para confirmar ou ratificar achados prévios, a disponibilidade de uma forma alternativa do teste, com alto escore de fidedignidade de forma alternativa com intervalo, seria altamente desejável.
- Se são desejadas homogeneidade e consistência ao longo de todo o teste, deve-se buscar um coeficiente alfa ou K-R-20 alto.

Avaliando dados de fidedignidade

Os coeficientes de fidedignidade fornecem algumas informações aos usuários de testes a respeito da magnitude do erro que podem influenciar os escores a partir de várias fontes. No entanto, ao avaliarmos dados de fidedignidade, devemos ter em mente o fato de que estas estimativas são afetadas pelas características da amostra com a qual foram calculadas e podem ser ou não generalizáveis para outros grupos de testandos. Entre outras coisas, isso significa que pequenas oscilações na magnitude dos coeficientes de diferentes testes provavelmente não têm a mesma significância que outras considerações. Além disso, à luz da variedade de fatores que podem influenciar a fidedignidade dos escores de testes, existe um reconhecimento crescente de que os investigadores devem incluir rotineiramente dados sobre a fidedignidade dos escores de suas próprias amostras ao relatarem os resultados de seus estudos de pesquisa (Baugh, 2003; Onwuegbuzie e Daniel, 2002).

Quando um teste se destina a ser usado em avaliações individuais, e não em pesquisas que envolvem dados de grupos, a importância de se examinar criticamente as informações publicadas sobre a fidedignidade, antes de selecionar um instrumento, é ainda maior. Além de avaliar as amostras com as quais os dados foram obtidos em relação a tamanho, representatividade e variabilidade, os usuários de testes devem ponderar se os coeficientes disponíveis são os mais apropriados para o tipo de instrumento em questão e para os usos pretendidos do teste. Além disso, se um teste é composto de subtestes ou outras partes cujos escores devem ser interpretados isoladamente ou em combinação, as estimativas de fidedignidade para cada escore parcial devem estar disponíveis, além das estimativas para o escore total.

De certa forma, um coeficiente de fidedignidade pode ser descrito como a correlação do teste consigo mesmo. Embora não seja totalmente precisa, esta descrição é um lembrete de que os coeficientes de fidedignidade se baseiam em dados – como duas administrações do mesmo teste, duas versões do mesmo teste, correlações entre itens, etc. – que *precisam ser* altamente consistentes. Estimativas baixas de fidedignidade (abaixo de 0,70) sugerem que o escore derivado de um teste pode não ser muito confiável. Por isso, embora não haja um limiar mínimo para que um coeficiente de fidedignidade seja considerado adequado para todos os fins, entende-se que, não havendo outras falhas, quanto mais alto o coeficiente, melhor. A maioria dos usuários de testes buscam coeficientes pelo menos da faixa de 0,80 ou mais.

Avaliando dados de fidedignidade de escores à luz de outros atributos

As decisões relativas à seleção de testes devem ser feitas caso a caso, levando em consideração todas as características dos instrumentos disponíveis, bem como os requisitos da situação específica na qual os escores serão usados. Embora fundamental, a fidedignidade dos escores não é de forma alguma a única consideração na seleção de um teste. Além da questão da fidedignidade, os dados de validade (Capítulo 5) e a disponibilidade de informações normativas ou referenciadas em

critérios para a interpretação dos escores (Capítulo 3) são de suprema importância. Embora considerações práticas – como custos, facilidade de administração e avaliação, limitações de tempo, etc – necessariamente tenham um papel na seleção dos testes, quando o uso destes provavelmente tiver um impacto significativo nos testandos, estas considerações não devem ser os fatores determinantes na escolha do instrumento.

Avaliação de erros de múltiplas fontes

A maioria dos escores de teste são suscetíveis a erros de mensuração oriundos de mais de uma fonte. Na teoria clássica dos testes, esta possibilidade realista é acomodada por (a) métodos que estimam a influência combinada de duas fontes, como a fidedignidade de forma alternativa com intervalo, que estima tanto o erro de amostragem de tempo como o de conteúdo ou (b) pela soma das quantidades de variância de erro estimadas por todos os coeficientes de fidedignidade pertinentes para se chegar a uma estimativa de variância de erro total. Ambas as estratégias dependem do fato de que os coeficientes de fidedignidade podem ser interpretados como estimativas da proporção de variância do escore atribuíveis a erros de várias fontes (ver Fórmula [4.4]). Por exemplo, se o coeficiente de fidedignidade de forma alternativa com intervalo de um teste é de 0,75, 75% da variância do escore podem ser interpretados como variância verdadeira e 25% (1 – 0,75 = 0,25) podem ser atribuídos à influência combinada de erro de amostragem de tempo e de conteúdo. Se os escores podem ser afetados por várias fontes de erro, as estimativas de fidedignidade que avaliam o erro de diferentes fontes podem ser combinadas. O quadro Consulta Rápida 4.8 descreve esta análise de fontes de variância de erro para os escores do subteste de Vocabulário da WAIS-III (Psychological Corporation, 1997).

Teoria da generalizabilidade

Uma abordagem alternativa à fidedignidade que busca ser mais abrangente do que a que discutimos até agora passou a ser conhecida como *teoria da generalizabilidade*, ou simplesmente *teoria G* (Cronbach, Gleser, Nanda e Rajaratnam, 1972). A teoria da generalizabilidade é uma extensão da teoria clássica dos testes que usa métodos da análise de variância (ANOVA) para avaliar os efeitos combinados de múltiplas fontes de variância de erro em escores de teste simultaneamente.

Uma vantagem distinta da teoria G – comparada ao método para combinar estimativas de fidedignidade ilustrado no quadro Consulta Rápida 4.8 – é que ela também permite a avaliação dos efeitos de interação de diferentes tipos de fontes de erro. Por isso, é um procedimento mais completo para identificar o componente de variância de erro que pode influenciar os escores. Por outro lado, para se aplicarem os delineamentos experimentais requeridos pela teoria G, é necessário obter múltiplas observações do mesmo grupo de indivíduos em todas as variáveis independentes que podem contribuir para a variância de erro em um dado teste (p. ex.,

CONSULTA RÁPIDA 4.8

Análise de múltiplas fontes de variância de erro em escores de um único teste

O subteste de Vocabulário da Escala de Inteligência Wechsler para Adultos – Terceira Edição (WAIS-III) consiste em uma série de palavras de dificuldade crescente que são lidas para o testando pelo examinador e simultaneamente apresentadas visualmente em um livreto de estímulos. As definições do testando, fornecidas oralmente, são registradas literalmente e imediatamente pontuadas pelo examinador, usando uma escala de 2, 1 ou 0 pontos, dependendo da qualidade das respostas. Ao avaliarem as respostas, os examinadores são guiados por uma ampla familiaridade com as amostras de respostas fornecidas no manual para cada uma das palavras – nos três níveis de pontuação – bem como pelas definições dicionarizadas de cada palavra (Psychological Corporation, 1997).

O escore total para o subteste de Vocabulário é a soma dos pontos obtidos pelo examinando em todos os itens (palavras). Um escore deste tipo está sujeito a erros de amostragem de tempo e conteúdo, bem como à possibilidade de diferenças entre avaliadores. As estimativas médias de fidedignidade oferecidas no manual da WAIS-III (que pode ser consultado por aqueles que desejarem informações mais detalhadas) para o subteste de Vocabulário são as seguintes:

Fonte de erro/ Tipo de fidedignidade	Coeficiente médio	Proporção e percentagem(%) de variância de erro
Amostragem de tempo/ estabilidade (teste-reteste)	0,91	1 – 0,91 = 0,09 (9%)
Amostragem de conteúdo/ consistência interna	0,93	1 – 0,93 = 0,07 (7%)
Diferenças entre avaliadores/ pontuadores	0,95	1 – 0,95 = 0,05 (5%)
Variância de erro total medida		0,9 + 0,7 + 0,5 = 0,21 (21%)
Variância verdadeira estimada		1 – 0,21 = 0,79 (79%)

A partir dos cálculos acima, deve ficar evidente que utilizar uma estimativa de fidedignidade de *fonte única* para um teste do tipo exemplificado pelo subteste de Vocabulário da WAIS-III produziria uma impressão altamente enganosa da possível quantidade de erro em seus escores. Além disso, este exemplo aponta que, para que escores que estão sujeitos a múltiplas fontes de erro sejam suficientemente confiáveis, as estimativas de fidedignidade para cada fonte, de maneira isolada, precisam ser bastante altas, na faixa de 0,90 ou mais.

Ver Exemplo 1: Aplicando o EMP no texto para uma aplicação específica desta análise de múltiplas fontes de erro e seus efeitos na fidedignidade de um escore do subteste de Vocabulário da WAIS-III.

escores em todas as ocasiões, por todos os avaliadores, entre formas alternativas, etc). De modo geral, no entanto, quando isto é viável, os resultados fornecem uma estimativa melhor da fidedignidade dos escores do que as abordagens descritas anteriormente. Apesar da teoria G ter sido introduzida originalmente no início dos anos de 1960, poucos autores de testes a aplicaram ao desenvolverem novos ins-

trumentos. No entanto, à medida que a familiaridade com esta técnica se disseminar, ela deverá ganhar popularidade. Os leitores que quiserem conhecer os procedimentos básicos da teoria G devem consultar uma introdução breve apresentada por Thompson (2003a), que inclui um exemplo simples de cálculo. Um tratamento mais abrangente e detalhado do referencial conceitual e dos aspectos estatísticos da teoria da generalizabilidade pode ser encontrado na obra de Robert Brennan sobre este tópico (2001).

Abordagem da fidedignidade na teoria da resposta ao item

Métodos mais sofisticados para estimar a fidedignidade estão disponíveis na teoria da resposta ao item (TRI) (apresentada no Capítulo 3 e discutida mais extensamente no Capítulo 6). Uma explicação completa dos aspectos técnicos dos modelos da TRI está além do âmbito deste texto, mas as vantagens que estes modelos oferecem, especialmente para a testagem em larga escala e a testagem adaptativa computadorizada, têm estimulado seu desenvolvimento e aplicação nas últimas décadas. Com os métodos da TRI, a fidedignidade e o erro de mensuração são abordados desde o ponto de vista da função de informação de itens individuais do teste, em oposição ao teste como um todo. Como o nível de dificuldade e o poder discriminativo de itens individuais – em relação ao traço avaliado pelo teste – podem ser calibrados mais cuidadosamente pelos métodos da TRI, as informações oferecidas pela resposta de cada testando são mais precisas e, por isso, mais fidedignas. No tipo de testagem adaptativa computadorizada permitida por estes métodos, a seleção do item mais apropriado a ser apresentado aos testandos é determinada por suas respostas anteriores. Usando a metodologia TRI e a testagem adaptativa, a fidedignidade adequada pode ser obtida com erro de mensuração mínimo em testes mais curtos do que os tradicionais (que apresentam o mesmo conteúdo fixo a todos os testandos), desde que um banco de itens suficientemente extenso e inclusivo esteja disponível. Este é apenas um dos muitos aspectos fundamentais em que a versão de mensuração baseada no modelo conhecido como TRI difere das regras e premissas da teoria clássica dos testes (Embretson e Reise, 2000, p.13-39).

CONSIDERAÇÕES SOBRE A FIDEDIGNIDADE NA INTERPRETAÇÃO DE TESTES

Depois que um teste foi escolhido, administrado e avaliado, os dados de fidedignidade são aplicados no processo da interpretação do teste para duas finalidades distintas porém relacionadas. A primeira é reconhecer e quantificar a margem de erro nos escores obtidos. A segunda é avaliar a significância estatística da diferença entre os escores obtidos para ajudar a determinar a importância destas diferenças em termos do que os escores representam.

Quantificando o erro nos escores de teste: O erro de mensuração padrão (EMP)

Na interpretação de qualquer escore – ou média de escores – de um teste, os dados de fidedignidade são usados para derivar os limites inferiores e superiores da faixa dentro da qual os escores verdadeiros dos testandos provavelmente se encaixarão. Um intervalo de confiança é calculado para um escore obtido a partir da fidedignidade estimada dos escores dos testes em questão. O tamanho do intervalo depende do nível de probabilidade escolhido.

Exemplo 1: Aplicando o EMP

A fidedignidade estimada dos escores do subteste de Vocabulário da WAIS-III descrita no quadro Consulta Rápida 4.8 – após a subtração da variância estimada de erro de três fontes relevantes – é de 0,79. Como todos os subtestes da Wechsler, o subteste de Vocabulário têm escores escalonados que podem ir de 1 a 19, com M = 10 e DP = 3. Para ilustrar a aplicação mais básica dos dados de fidedignidade, vamos supor que uma testanda chamada Maria obtém um escore de 15 no subteste de Vocabulário da WAIS-III .

Passo 1. Para obtermos um intervalo de confiança para o escore de 15 obtido por Maria (Xa), precisamos do *erro de mensuração padrão (EMP) para o subteste* de Vocabulário. O EMP é uma função estatística que representa o desvio padrão da distribuição hipotética que teríamos se Maria realizasse este subteste um número infinito de vezes. Como já foi mencionado neste capítulo, a média desta distribuição hipotética seria o *escore verdadeiro* de Maria no subteste de Vocabulário. Um exame da fórmula no quadro Consulta Rápida 4.9 revela que o EMP é uma função do coeficiente de fidedignidade dos escores do teste em questão, que é expresso em termos da unidade de desvio padrão deste, e, por isso, seu tamanho não pode ser tomado por si só como um índice de fidedignidade. Testes com unidades grandes de desvio padrão, como o SAT (DP = 100), terão EMPs muito maiores do que testes com unidades pequenas de desvio padrão, o que é o caso dos subtestes da escala Wechsler (DP = 3), mesmo que seus coeficientes de fidedignidade sejam iguais em magnitude.

Passo 2. Como não podemos obter múltiplos escores no subteste de Vocabulário para Maria nem fazer sua média para encontrar uma estimativa de seu escore verdadeiro, devemos optar por um escore disponível que possa ser colocado no centro do intervalo a ser criado pelo EMP. Aqui surgem duas possibilidades: (a) podemos usar o escore obtido, X_o, como estimativa do escore verdadeiro de Maria, ou (b) podemos estimar seu escore verdadeiro (T') com a seguinte fórmula, baseada em Dudek (1979):

$$T' = r_{xx}(X_o - M) + M \qquad (4.7)$$

em que

T' = o escore verdadeiro estimado do indivíduo

CONSULTA RÁPIDA 4.9

Fórmula do erro de mensuração padrão (EMP)

$$EMP = DP_t\sqrt{1 - r_{xx}}$$

em que

DP_t = desvio padrão do teste
r_{xx} = coeficiente de fidedignidade

Fórmulas do erro padrão da diferença entre dois escores (EPdif)

Fórmula EPdif 1:

$$EP_{dif} = DP\sqrt{2 - r_{11} - r_{22}}$$

onde

DP = desvio padrão do Teste 1 e do Teste 2
r_{11} = estimativa de fidedignidade para os escores do Teste 1
r_{22} = estimativa de fidedignidade para os escores no Teste 2

Fórmula EPdif 2:

$$EP_{dif} = \sqrt{(EMP_1)^2 + (EMP_2)^2}$$

em que

EMP1 = erro de mensuração padrão do Teste 1
EMP2 = erro de mensuração padrão do Teste 2

A Fórmula EP_{dif} 1 é usada quando os dois escores comparados são expressos na mesma escala, e a Fórmula EP_{dif} 2 é usada quando as escalas são diferentes.

r_{xx} = a fidedignidade estimada dos escores do teste
X_o = o escore obtido pelo indivíduo
M = a média da distribuição dos escores do teste

No caso de Maria, como $X_o = 15$, $r_{xx} = 0{,}79$ e $M = 10$ para o subteste de Vocabulário, assim como para todos os escores de subtestes da Wechsler, seu escore verdadeiro estimado é 14 ($T' = (0{,}79)\ (15\text{-}10) + 10 + 13{,}95$, ou 14). Observe que, uma vez que seu escore está acima da média, *seu escore verdadeiro estimado é mais baixo do que o escore obtido*. Em contraste, um escore obtido de 5 no mesmo subteste – que se desvia tanto da média quanto o de Maria, mas na direção oposta – resultaria em um escore verdadeiro estimado de 6, que é *mais alto* do que X_o (se $X_o = 5$, $T' = (0{,}79)\ (5\text{-}10) + 10 = 6{,}05$, ou 6). O motivo para essa diferença nas estimativas de escore verdadeiro para escores obtidos acima ou abaixo da média é que o procedimento de estimativa leva em conta o efeito da regressão em direção à média. Pelo mesmo raciocínio, se $X_o = M$, a melhor estimativa de escore verdadeiro seria a própria média.

Passo 3. A necessidade ou não de se calcular *T'* para criar um intervalo de confiança depende de quanto um escore obtido se desvia da média. Se o escore obtido está próximo da média, o escore verdadeiro estimado não vai diferir muito; por outro lado, se os escores obtidos se aproximam dos extremos, calcular escores verdadeiros estimados mais próximos da média se torna mais aconselhável. De qualquer forma, o Passo 3 envolve o cálculo do *EMP*. Usando a fórmula do quadro Consulta Rápida 4.9, constatamos que $EMP = 3\sqrt{1 - 0,79} = 1,37$. Já que este *EMP*, como os outros erros padrões descritos no Capítulo 2, representa o desvio padrão de uma distribuição hipotética de escores considerada normal, podemos interpretá-lo em termos das freqüências da curva normal. Devemos recordar do Capítulo 3 em que aproximadamente 68% da área sob a curva normal estão incluídos dentro de ± 1 *DP* da média, 95% estão dentro de ±1,96 *DP*s, etc. Aplicando estas percentagens ao escore verdadeiro estimado (*T'*) de 14 de Maria, e aplicando o *EMP* obtido de 1,37, podemos dizer que (a) há uma chance de 68/100, ou $p = 0,32$, de que o escore verdadeiro de Maria se localize no intervalo de 14 ± 1,37, ou seja, entre 13 e 15; e (b) há uma chance de 95/100, ou $p = 0,05$, de que seu escore verdadeiro esteja dentro de 14 ± (1,37) (1,96), ou seja, entre 11 e 17.

Interpretando a significância das diferenças entre escores

Com freqüência os objetivos da avaliação acarretam comparações (a) entre dois ou mais escores obtidos pelo mesmo indivíduo em diferentes partes de uma bateria de testes, como quando se comparam níveis de desempenho em diferentes domínios, ou (b) entre os escores de duas ou mais pessoas no mesmo teste, para fins de avaliar seus méritos ou características relativas. Em ambos os casos, os dados de fidedignidade podem ser usados para derivar afirmações sobre a probabilidade de que as diferenças obtidas entre escores – e o que estes representam – possam se dever ao acaso. A função estatística usada para esta finalidade é o *erro padrão da diferença entre escores*, ou EP_{dif}, que pode ser calculado usando-se uma das duas fórmulas listadas no quadro Consulta Rápida 4.9, dependendo se os escores a serem comparados estão expressos na mesma escala (Fórmula 1) ou não (Fórmula 2). Independentemente da fórmula usada, o EP_{dif} vai ser maior do que o *EMP* dos dois escores envolvidos na comparação, porque a avaliação das diferenças entre os escores tem que levar em conta o erro presente em ambos.

Exemplo 2: Aplicando o EP_{dif}

Para ilustrar o uso do erro padrão da diferença entre escores, vamos supor que desejamos estimar a significância estatística da diferença entre os escores obtidos por Maria em dois subtestes da WAIS-III: seu escore de 15 no subteste de Vocabulário e seu escore de 10 no subteste de Informação. O subteste de Vocabulário é descrito no quadro Consulta Rápida 4.8; o subteste de Informação avalia o conhecimento sobre eventos comuns, objetos, lugares e pessoas.

Passo 1. Já que queremos estimar a significância de uma diferença entre dois escores obtidos, o primeiro passo é calcular esta diferença. Neste caso, 15 – 10 = 5. Existe uma diferença de 5 pontos entre o escore de Maria no subteste de Vocabulário e seu escore no subteste de Informação. De posse desta informação, podemos a seguir avaliar se a diferença obtida é estatisticamente significativa (isto é, se não aconteceu por acaso).

Passo 2. Precisamos calcular o erro padrão da diferença entre escores nos subtestes de Vocabulário e Informação da WAIS-III. Como os dois subtestes são expressos na mesma escala de escores (M = 10 e DP = 3), podemos usar a Fórmula 1 do quadro Consulta Rápida 4.9. Isso requer que conheçamos os coeficientes de fidedignidade para os escores dos subtestes. O coeficiente combinado para Vocabulário é 0,79, conforme o estimado no quadro Consulta Rápida 4.8. Para o subteste de Informação, o coeficiente é estimado em 0,85, baseado na combinação dos coeficientes de consistência interna e estabilidade de 0,91 e 0,94 respectivamente, disponíveis no *Manual Técnico da WAIS-III/WMS-III* (Psychological Corporation, 1997). Portanto, o erro padrão da diferença entre os escores nos subtestes de Vocabulário e Informação é

$$EP_{dif} = 3\sqrt{2 - 0{,}79 - 0{,}85} = 1{,}80$$

*Passo 3. Para d*eterminar a significância estatística da diferença de cinco pontos entre os escores obtidos, dividimos esta diferença pelo EP_{dif} e obtemos um valor crítico de 5/1,80 = 2,78.

*Passo 4. Consul*tando a Tabela de Áreas da Curva Normal no Apêndice C, para um valor z de 2,78, constatamos que a área na porção menor que é cortada por este valor z é de 0,0027. Como não havia motivos para pressupormos que algum dos escores nos subtestes (Vocabulário ou Informação) seria mais alto do que o outro, um teste bilateral de significância para a hipótese nula de nenhuma diferença entre os escores é apropriado. Assim, multiplicamos 0,0027 por 2 e obtemos 0,0054, que indica que a probabilidade de que os escores de Maria nos dois subtestes difiram em cinco pontos devido ao acaso é de 5,4 em 1000. Dado este alto nível de significância para a diferença, não havendo outras falhas, podemos inferir com segurança que realmente existe uma diferença: o conhecimento vocabular de Maria, medido pelo subteste de Vocabulário da WAIS-III, provavelmente excede seu conhecimento de informações gerais a respeito de eventos comuns, lugares, objetos e pessoas, conforme medido pelo subteste de Informação.

Por que é importante criar intervalos de confiança para escores obtidos e para diferenças entre eles? Duas razões básicas podem ser citadas em resposta a esta pergunta. A primeira é que os intervalos de confiança para escores obtidos nos lembram que os escores de teste não são tão precisos quanto sua natureza numérica pode sugerir. Por isso, sempre que decisões importantes devem ser tomadas com a ajuda de escores de teste, especialmente quando são usados escores de corte, é preciso considerar seriamente o erro de mensuração quantificado pelo *EMP*. A segunda razão, que está relacionada à primeira, é que os intervalos de confiança evitam que designemos um sentido indevido a diferenças de escore que podem ser

insignificantes à luz do erro de mensuração. Reconhecendo a importância desses fatos, muitos manuais de testes incluem tabelas listando os erros padrões de mensuração para seus escores, bem como as faixas numéricas para cada escore possível que podem ser derivadas de um teste, juntamente com os níveis de confiança para cada faixa de escore. Por exemplo, para um QI Total obtido de 110 na WAIS-III, o intervalo de nível de confiança de 90% fica entre 96 e 113. (Wechsler, 1997, p.198). A disponibilidade desta informação nos manuais de teste encoraja os usuários a aplicar intervalos de confiança na interpretação dos escores sem ter que calculá-los. No entanto, cabe ao usuário determinar se os números publicados são aplicáveis e significativos em cada caso do uso do teste.

A Tabela 4.1 e a Figura 4.1 ilustram como os dados de fidedignidade de escores e o *EMP* podem ser usados na análise de 4 – de um total de 14 – escores de subtestes que se pode obter com a WAIS-III. Além dos escores de Maria nos subtestes de Vocabulário e Informação, já utilizados nos exemplos anteriores, dois outros escores fictícios em subtestes da WAIS-III, Aritmética e Dígitos, foram acrescentados ao seu perfil. Estes subtestes foram selecionados porque as habilidades primárias que avaliam – habilidade quantitativa e memória auditiva de curto prazo, respectivamente – são suficientemente singulares para tornar uma comparação entre eles e os outros dois escores interessante e plausível. As faixas de erro calculadas no intervalo de confiança de 90%, baseado nos *EMP*s para os respectivos subtestes, são apresentadas na Tabela 4.1. Os *EMP*s, por sua vez, foram calculados com base em combinações de *todos* os números pertinentes de fidedignidade apresentados no *Manual Técnico da WAIS-III/WMS-III* (Psychological Corporation, 1997) para cada um dos quatro subtestes. Esta prática mais rigorosa se contrapõe ao uso dos *EMP*s (menores), baseados em apenas uma estimativa de fidedignidade, que são fornecidos nas tabelas do manual do teste para os escores-índice e de QI da WAIS-III (ver Wechsler, 1997, p.195-202). A Figura 4.1 mostra os dados da Tabela 4.1 em forma gráfica. O exame desta figura rapidamente revela que, quando o *EMP* é levado em consideração, as faixas prováveis dos escores de Maria se sobrepõem consideravelmente, o que significa que parte da diferença entre os escores obtida pode ser devida a erro de mensuração. Ao mesmo tempo, este tipo de análise de perfil permite ao usuário do teste explorar *hipóteses* a respeito do possível sentido das diferenças no desempenho de Maria nas habilidades exploradas pelos subtestes cujas faixas de erro não se sobrepõem (isto é, Vocabulário e Informação, Aritmética e Dígitos e Dígitos e Informação). Por exemplo, Maria aparentemente tem um relativo potencial em capacidade memória de curto prazo e vocabulário, enquanto que seu estoque de informações gerais pode ser um ponto relativamente fraco. Naturalmente, quaisquer conclusões baseadas nestas diferenças estão sujeitas à confirmação ou revisão à luz de dados adicionais. Não obstante, quando uma avaliação psicológica exige a avaliação dos potenciais e pontos fracos de um indivíduo, seja na área intelectual ou em outro aspecto – interesses vocacionais, por exemplo –, este tipo de análise exploratória, embora não seja definitiva, pode ser bastante útil.

O erro padrão da diferença entre escores (EP_{dif}) discutido antes serve a um propósito semelhante ao das análises de perfil mostradas na Figura 4.1. Ele fornece dados a respeito de discrepâncias de escore que podem ter significância prática ou psicológica. Tipicamente, os valores do EP_{dif} – também encontrados em muitos

Tabela 4.1 Perfil dos escores obtidos por Maria em quatro subtestes da WAIS-III com *EMP*s e faixas de erro no nível de confiança de 90%

Subteste WAIS-III	Escore obtido por Maria (Xo)	Coeficiente de fidedignidade estimado[a]	*EMP*[b] (1.64)[c] = faixa de erro	Faixa de erro ± do Xo de Maria no nível de confiança de 90%
Vocabulário	15	0,79	1,37(1,64) = 2,25	15 ± 2,25 = 12,75 a 15,25
Aritmética	12	0,74	1,53(1,64) = 2,51	12 ± 2,51 = 9,49 a 14,51
Dígitos	17	0,83	1,24(1,64) = 2,03	17 ± 2,03 = 14,97 a 19,03
Informação	10	0,85	1,16(1,64) = 1,90	10 ± 1,90 = 8,10 a 11,90

Nota: WAIS-III = Escala Wechsler de Inteligência para Adultos-Terceira Edição; *EMP* = erro de mensuração padrão; X_o = escore observado; *DP* = desvio padrão.

[a]Fidedignidade estimada após subtração de todas as estimativas pertinentes de variância de erro, como no exemplo do quadro Consulta Rápida 4.8.

[b]$EMP = DP\sqrt{1 - rxx}$

[c]1,64 é o valor z para $p = 0,10$ (90% de nível de confiança).

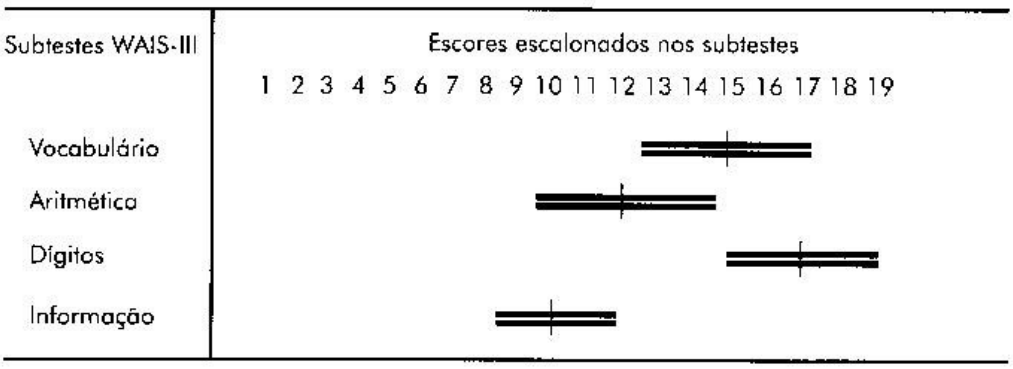

Figura 4.1 Perfil gráfico dos escores de Maria nos subtestes WAIS-III, ilustrando o uso de faixas de erro na Tabela 4.1.

manuais de testes – são usados para avaliar a significância estatística das diferenças obtidas entre escores que são de interesse especial. Os manuais das edições recentes das escalas Wechsler, por exemplo, rotineiramente incluem tabelas com as diferenças de pontos nos escores índice e QIs (QI Verbal e QI de Desempenho) necessários para significância estatística nos níveis de confiança de 90 e 95%. Reconhecendo que uma diferença *estatisticamente* significativa pode não o ser no aspecto psicológico, os autores destes manuais – e também os de outros testes – também fornecem tabelas que mostram as freqüências com que diferenças de escore de várias magnitudes foram encontradas entre as amostras de padronização dos testes em questão. Esta informação contempla a questão de como podem ser raras ou comuns as diferenças entre a amostra normativa. Sua importância deriva da premissa (nem sempre justificada, mas merecedora de reflexão) de que, se diferenças de uma certa magnitude ocorrem freqüentemente, elas provavelmente têm menos significância interpretativa do que aquelas que ocorrem raramente.

O EP_{dif} também pode, é claro, ser usado para calcular a probabilidade de que uma diferença obtida entre os escores de dois indivíduos no mesmo teste possa ser devida a erro de mensuração. Por exemplo, em decisões de seleção educacional ou profissional tomadas com a ajuda de testes, as diferenças entre os escores dos candidatos podem ser avaliadas em termos de significância à luz do EP_{dif}, além de outros fatores relevantes. Mais uma vez, assim como na análise de perfil, este pro-

Não esqueça

O uso de erros de mensuração padrão (*EMPs*) para escores de teste e erros padrões para diferenças entre escores (EP_{dif}), ambos derivados de estimativas de fidedignidade dos escores, são informações essenciais porque

1. os *EMPs* fornecem intervalos de confiança para escores obtidos que alertam os usuários dos testes para o fato de que os escores estão sujeitos a flutuações devidas a erros de mensuração, e
2. os intervalos de confiança obtidos com o uso de estatísticas de EP_{dif} evitam a supervalorização das diferenças de escore, que podem ser insignificantes à luz do erro de mensuração.

cesso chama atenção para o fato de que as diferenças de escore não podem ser tomadas por seu valor nominal.

A relação entre fidedignidade e validade

Na perspectiva psicométrica, as evidências de fidedignidade de escores são consideradas uma condição necessária, porém não suficiente, para a validade (Sawilowsky, 2003). Na verdade, como veremos no próximo capítulo, os dois conceitos estão intrinsecamente relacionados se a fidedignidade do escore pode ser entendida como uma evidência mínima para obtenção de uma medida válida de amostra de comportamento.

Os profissionais da avaliação geralmente concordam que as evidências de fidedignidade de escores não são base suficiente para se fazer inferências válidas quanto ao sentido destes. No entanto, existem algumas discordâncias quanto à extensão em que as evidências de fidedignidade são consideradas *essenciais* para uma avaliação válida de todos os tipos de amostra de comportamento que podem ser coletadas por meio de testes. Por exemplo, quando os escores são derivados de amostras de comportamento singulares ou idiossincráticas, elas podem não ser repetíveis ou consistentes. Testes que revelam o nível ótimo de desempenho do indivíduo, como amostras de trabalho ou portifólios, podem produzir resultados válidos e fidedignos em termos de precisão, mas não em termos de consistência ou estabilidade (Moss, 1994). Da mesma forma, instrumentos administrados individualmente, como muitas escalas de inteligência ou técnicas projetivas, são altamente suscetíveis a influências oriundas da qualidade do *rapport* entre o examinador e o testando, bem como outros fatores motivacionais e situacionais. No contexto da avaliação individual, estes instrumentos podem fornecer noções válidas de aspectos da constituição psicológica de uma pessoa que poderiam não ser reproduzidas com um examinador diferente ou em circunstâncias diferentes, mesmo se procedimentos de padronização forem rigidamente observados (Masling, 1960; McClelland, 1958; Smith, 1992).

CONCLUSÃO

O uso de testes psicológicos seria grandemente simplificado se os coeficientes de fidedignidade e os *EMPs* pudessem ser tomados por seu valor nominal ao se avaliar escores. Como este capítulo atesta, porém, a fidedignidade dos escores é um julgamento relativo baseado tanto nos dados psicométricos quanto no contexto em que os testes são administrados. Veremos no Capítulo 5 que o mesmo se aplica à validade dos dados de escores. Portanto, embora a disponibilidade de dados psicométricos apropriados sobre fidedignidade seja um pré-requisito básico para qualquer uso de escores de teste, o contexto no qual a testagem psicológica acontece também é uma consideração fundamental na interpretação dos escores obtidos por indivíduos ou grupos. À medida que o potencial de impacto das decisões a serem tomadas com o auxílio dos escores de teste aumenta, ambos os fatores assumem maior importância.

Teste a si mesmo

1. Um escore verdadeiro é
   (a) uma entidade hipotética
   (b) uma entidade real
   (c) igual ao escore observado
   (d) igual ao escore observado mais o erro

2. Se a fidedignidade de um teste é bem estabelecida, os usuários podem pressupor que os escores obtidos com aquele teste serão confiáveis. Verdadeiro ou Falso?

3. Quais dos seguintes fontes de erro em escores de teste não é avaliada pelas estimativas tradicionais de fidedignidade?
   (a) diferenças entre avaliadores
   (b) amostragem de tempo
   (c) amostragem de conteúdo
   (d) desvios dos procedimentos padronizados

4. Coeficientes de fidedignidade _____ são usados para estimar erro de amostragem de tempo em escores de teste.
   (a) de teste-reteste
   (b) de forma alternativa
   (c) de avaliador
   (d) calculados pelo método das metades

5. Qual dos seguintes tipos de coeficiente de fidedignidade resulta em uma estimativa combinada de erros oriundos de duas fontes diferentes?
   (a) de avaliador
   (b) de teste-reteste
   (c) de forma alternativa
   (d) de forma alternativa com intervalo

6. Não havendo outras falhas, os escores obtidos em testes mais longos são _____ os obtidos em testes comparáveis porém mais curtos.
   (a) menos confiáveis do que
   (b) mais confiáveis do que
   (c) tão confiáveis quanto

7. A magnitude de um coeficiente de fidedignidade tem maior probabilidade de ser afetada por _____ do que por _____ da amostra com a qual é calculada.
   (a) tamanho/heterogeneidade
   (b) heterogeneidade/tamanho

8. Uma das vantagens distintas da teoria da generalizabilidade em relação às abordagens tradicionais da fidedignidade dos escores é que ela
   (a) requer um número menor de observações
   (b) resulta em componentes de erro menores
   (c) permite a avaliação de efeitos de interação
   (d) usa métodos estatísticos menos complicados

9. Suponha que um aluno obtém um escore de 110 em um teste com $M = 100$, $DP = 20$ e fidedignidade estimada em 0,96. Existe uma chance de 68 em 100 de que o escore verdadeiro do aluno fique em algum ponto entre
   (a) 100 e 110
   (b) 102 e 112
   (c) 106 e 114
   (d) 110 e 120

10. O erro de mensuração padrão do Teste A é 5, e o do Teste B é 8. O erro padrão da diferença para a comparação de escores dos dois teste será
    (a) menor que 8
    (b) menor que 5
    (c) entre 5 e 8
    (d) maior do que 8

Respostas: 1. a; 2. b; 3. d; 4. a; 5. d; 6. b; 7. b; 8. c; 9. c; 10. d.

[capítulo 5]

FUNDAMENTOS EM VALIDADE

Os testes psicológicos existem para nos ajudar a fazer inferências a respeito de pessoas e seu comportamento. A validade – que é, sem dúvida alguma, a questão mais fundamental relativa aos escores de testes e seus usos – depende das evidências que podemos reunir para corroborar qualquer inferência feita a partir de resultados de testes. A primazia das considerações sobre a validade é reconhecida nos *Padrões de Testagem* atuais pela colocação deste tópico no primeiro capítulo, que define *validade* como "o grau em que todas as evidências acumuladas corroboram a interpretação pretendida dos escores de um teste para os fins propostos" (AERA, APA, NCME, 1999, p.11). Nesta definição estão implícitas três idéias relacionadas que refletem a visão atual dos profissionais da testagem a respeito deste conceito central e multifacetado:

1. A validade dos escores de teste resulta das evidências acumuladas que corroboram sua interpretação e seu uso. Portanto, a validade sempre é uma questão de grau, e não uma determinação do estilo tudo-ou-nada. A *validação* – o processo por meio do qual as evidências de validade são coletadas – começa com uma afirmação explícita do referencial conceitual e dos fundamentos teóricos de um teste feita por seu criador, mas é, por natureza, aberta porque inclui todas as informações que se somam à nossa compreensão dos resultados do teste.
2. À medida que a compreensão teórica e as evidências empíricas para interpretações dos escores de um teste se acumulam, a validade das inferências (isto é, hipóteses) feitas a partir delas para vários objetivos pode aumentar ou diminuir. Um corolário para esta noção, incluído nos *Padrões de Testagem* (AERA, APA, NCME, 1999), é "a validação é de responsabilidade conjunta do desenvolvedor do teste [que fornece as evidências e a fundamentação teórica para seu uso pretendido] e do usuário [que avalia as evidências disponíveis no contexto em que o teste vai ser usado]" (p.11).

3. Devido às muitas diferentes finalidades possíveis dos escores de teste, as bases confirmatórias para sua interpretação podem ser derivadas de uma variedade de métodos. As contribuições para evidências de validade de escores podem ser feitas por qualquer pesquisa sistemática que corrobore ou acrescente algo ao seu sentido, independentemente de quem a conduz ou quando ela ocorre. Desde que existam evidências científicas sólidas para um uso proposto dos escores de um teste, usuários qualificados são livres para empregá-los para seus fins, independentemente destes terem sido previstos pelos desenvolvedores do teste. Esta proposição ajuda a explicar a natureza multifacetada da pesquisa de validação, bem como seus achados muitas vezes redundantes e, às vezes, conflitantes. Também explica a longevidade de alguns instrumentos, como o MMPI e as escalas Wechsler, sobre os quais foi acumulada uma vasta literatura – que engloba numerosas aplicações em uma variedade de contextos – ao longo de décadas de pesquisa básica e aplicada.

O leitor atento pode já ter concluído que a validade, assim como a fidedignidade, não é uma qualidade que caracteriza abstratamente os testes ou qualquer teste específico ou seus dados. Mais do que isso, a validade é uma questão de *julgamento* que diz respeito aos escores de teste, como são empregados para um determinado objetivo em um dado contexto. Por isso, o processo de validação é semelhante à testagem de hipóteses, incluindo as noções de sentido e fidedignidade dos escores, discutidas nos dois capítulos anteriores, bem como os modos como as aplicações dos dados de teste à pesquisa e à prática psicológica podem ser justificados, que será o tópico abordado no presente capítulo. O quadro Consulta Rápida 5.1 lista algumas das contribuições mais significativas ao tópico da validade dos anos de 1950 aos 1990.

CONSULTA RÁPIDA 5.1

Consultas básicas em validade

Samuel Messick articulou suas idéias sobre a validade mais explicitamente em um capítulo de *Educational Measurement* (3.ed., p.13-103), uma obra notável organizada por Robert L. Linn e publicada conjuntamente pelo Conselho Americano de Educação e a Mcmillan em 1989. O capítulo de Messick sobre a validade e seus outros trabalhos sobre este tópico (Messick, 1988, 1995) influenciaram diretamente seu tratamento na versão atual dos *Padrões de Testagem* (AERA, APA, NCME, 1999). Outras contribuições-chave que são amplamente reconhecidas como formadoras na evolução dos conceitos teóricos da validade incluem as seguintes:

- Cronbach, L.J., e Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph Supplement]. *Psychological Reports, 3,* 635-694.
- Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.
- Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer e H.I. Braun (Orgs), *Test Validity* (pp. 3-17). Hillsdale, NJ; Erlbaum.

Não esqueça

Talvez nenhum teórico tenha influenciado mais a reformulação do conceito de validade do que Samuel Messick. Segundo Messick (1989, p.13), "a validade é um julgamento avaliativo integrado do grau em que as evidências empíricas e a fundamentação teórica corroboram a *adequação* e a *propriedade* de *inferências* e ações baseadas em escores de teste ou outros modos de avaliação".

PERSPECTIVAS HISTÓRICAS SOBRE A VALIDADE

O advento da testagem psicológica moderna aconteceu mais ou menos ao mesmo tempo em que a psicologia estava se tornando uma disciplina científica estabelecida. Ambos os campos datam seu início do final do século XIX e primeiras décadas do século XX. Como resultado desta coincidência histórica, nossa compreensão da natureza, funções e metodologia dos testes e mensurações psicológicas evoluiu ao longo do século passado juntamente com o desenvolvimento e a sofisticação crescente da ciência psicológica.

Em seu início, a psicologia científica era primariamente dedicada ao estabelecimento de leis psicofísicas, utilizando a investigação experimental da relação funcional entre estímulos físicos e as respostas sensoriais e perceptivas que eles despertavam nos humanos. A psicologia teórica consistia basicamente em especulações de natureza filosófica, até o primeiro quarto do século XX. Nenhuma destas afirmações sugere que as contribuições dos pioneiros da psicologia não foram valiosas (Boring, 1950; James, 1890). Não obstante, contra este pano de fundo, os primeiros testes psicológicos passaram a ser vistos, um tanto ingenuamente, como ferramentas científicas que mediam um catálogo cada vez maior de habilidades mentais e traços de personalidade, da mesma forma que os psicofísicistas estavam medindo as respostas auditivas e visuais e outras reações sensoriais e perceptivas a estímulos como sons, luzes e cores de vários tipos e intensidades. Além disso, como vimos no Capítulo 1, o sucesso da Stanford-Binet e da Army Alpha em auxiliar na tomada de decisões práticas a respeito de indivíduos no contexto da educação e do emprego levou a uma rápida proliferação dos testes nas primeiras duas décadas do século XX. A ampla gama de aplicações para as quais estes instrumentos foram usados logo superou os fundamentos científicos e teóricos disponíveis para eles na época. Em suma, muitos dos primeiros testes psicológicos foram desenvolvidos e usados sem o benefício da teoria psicométrica, princípios éticos e diretrizes práticas que começariam a se acumular em décadas posteriores (von Mayrhauser, 1992).

A definição clássica da validade

O reconhecimento desse estado de coisas na profissão resultou nas primeiras tentativas de delinear as características que iriam distinguir um bom teste de um teste ruim. Assim, a primeira definição de *validade* como "o grau em que um teste mede o que pretende medir" foi formulada em 1921 pela Associação Nacional

dos Diretores de Pesquisa Educacional (T.B. Rogers, 1995, p.25) e foi ratificada por muitos especialistas em testagem – incluindo Anne Anastasi em todas as edições de sua influente obra *Psychological Testing* (1954-1988), bem como Anastasi e Urbina (1997, p.8). A visão de que "a validade de um teste diz respeito a *o que* o teste mede e *com que eficácia* ele o faz" (Anastasi e Urbina, p. 113) ainda é considerada por muitos como a essência da questão da validade. Apesar de sua aparente simplicidade, esta visão traz uma série de problemas, especialmente quando vista da perspectiva dos *Padrões de Testagem* atuais (AERA, APA, NCME, 1999) e do esforço que ainda existe para definir alguns dos constructos mais básicos do campo da psicologia.

Aspectos problemáticos da visão tradicional da validade

As questões levantadas pela definição clássica da validade giram em torno de suas premissas implícitas, porém claras, de que

1. a validade é uma propriedade dos testes, e não das interpretações de seus escores;
2. para serem válidos, os escores de teste devem medir algum suposto constructo diretamente;
3. a validade de um escore é, pelo menos em certo grau, uma função da compreensão do autor ou desenvolvedor do teste a respeito do constructo que ele pretende medir.

Embora essas premissas possam ser justificáveis em alguns casos, definitivamente não o são em todos. A primeira premissa, por exemplo, se sustenta somente se os dados de validação corroborarem a finalidade pretendida do teste e este for usado especificamente para esta finalidade e com o tipo de população para a qual os dados de validade tiverem sido coletados. A segunda e a terceira premissas se justificam apenas para testes que medem comportamentos que podem ser ligados a constructos psicológicos de maneira bastante inequívoca, como certas funções de memória, velocidade e precisão no desempenho de várias tarefas de processamento cognitivo ou extensão de conhecimentos sobre um universo de conteúdo bem definido. Elas não são necessariamente defensáveis para (a) testes delineados para avaliar constructos teóricos complexos ou multidimensionais sobre os quais ainda há muita discussão, como a inteligência ou o autoconceito; (b) testes desenvolvidos com base em relações estritamente empíricas – em oposição a teóricas ou lógicas – entre escores e critérios externos, como o MMPI original; (c) técnicas cujo objetivo é revelar aspectos encobertos ou inconscientes da personalidade, como os instrumentos projetivos. Para instrumentos desta natureza, o que está sendo medido é o comportamento que pode ser ligado mais ou menos diretamente aos constructos que são de real interesse, primariamente por uma rede de evidências correlacionais. O quadro Consulta Rápida 5.2 define os vários sentidos da palavra *constructo* e pode ajudar a esclarecer as distinções feitas acima, bem como as que surgirão mais adiante neste capítulo.

CONSULTA RÁPIDA 5.2

Desconstruindo constructos

Como o termo *constructo* é usado com tanta freqüência neste capítulo, um esclarecimento do seu sentido é necessário. De modo geral, um *constructo* é qualquer coisa criada pela mente humana que não seja diretamente observável. Os constructos são abstrações que podem se referir a conceitos, idéias, entidades teóricas, hipóteses ou invenções de muitos tipos.

Na psicologia, o termo *constructo* é aplicado a conceitos como traços, e às relações teóricas entre conceitos que são inferidas de observações empíricas consistentes de dados comportamentais. Os constructos psicológicos diferem amplamente em termos de

- sua amplitude e complexidade,
- sua aplicabilidade potencial e
- grau de abstração necessário para inferi-los a partir dos dados disponíveis.

Como regra, constructos de definição estrita requerem menos abstração, mas têm uma gama menor de aplicações. Além disso, como é mais fácil obter consenso a respeito de constructos estritos, simples e menos abstratos, estes também são avaliados com mais facilidade do que constructos mais amplos e multifacetados que podem ter adquirido sentidos diferentes em vários contextos, culturas e períodos históricos.

Exemplos:

- Enquanto a *destreza manual* é um constructo que pode ser relacionado prontamente a dados comportamentais específicos, a *criatividade* é muito mais abstrata. Por isso, quando é necessário avaliar esses traços, determinar quem tem mais destreza manual é muito mais fácil do que determinar quem é mais criativo.
- A *introversão* é um constructo mais simples e de definição mais estrita do que a *conscienciosidade*. Embora esta seja potencialmente útil na predição de uma gama mais ampla de comportamentos, ela também é mais difícil de avaliar.

Sinônimos: os termos *constructo* e *variável latente* muitas vezes são usados de forma equivalente. Uma *variável latente* é uma característica que presumivelmente subjaz a um fenômeno observado, mas não é diretamente mensurável ou observável. Todos os traços psicológicos são variáveis latentes, ou constructos, assim como as denominações dadas a fatores que emergem de pesquisas de análise fatorial, como compreensão verbal ou neuroticismo.

A idéia de que a validade dos escores é uma função do grau em que os testes medem o que pretendem medir também leva a uma certa confusão entre a consistência ou precisão das mensurações (isto é, sua fidedignidade) e sua validade. Como vimos no Capítulo 4, se um teste mede *bem* o que pretende medir, seus escores podem ser considerados fidedignos (consistentes, precisos ou confiáveis), mas não serão necessariamente válidos no sentido contemporâneo mais amplo do termo. Em outras palavras, os escores de teste podem ser relativamente livres de erros de mensuração, e ainda assim não ser muito úteis como base para as inferências que precisamos fazer.

Além disso, a implicação de que um escore reflete o que o autor do teste pretende que ele reflita tem sido uma fonte de mal-entendidos. Um deles diz respeito aos títulos dos testes, que jamais deveriam ser – mas muitas vezes são – aceitos sem questionamento. Os títulos variam dos muito precisos e empiricamente defensáveis àqueles que meramente refletem as intenções (não-realizadas) dos

autores ou as preocupações de comercialização das editoras. Um segundo problema ainda mais importante ligado à noção de que escores válidos refletem a finalidade expressa dos testes é que eles podem levar a definições empíricas superficiais, ou fáceis, de constructos psicológicos. Possivelmente o exemplo mais famoso disto seja a definição de E. G. Boring, de 1923, da *inteligência* como "o que quer que seja que os testes de inteligência medem" (citado por Sternberg, 1968, p.2).

Como resultado desses mal-entendidos, o campo da testagem psicológica está sobrecarregado de instrumentos – que pretendem medir constructos maldefinidos ou efêmeros – cujas promessas superam muito o que eles podem realmente fazer, cujo uso na pesquisa psicológica impede ou retarda o progresso da disciplina e cuja existência, por associação, diminui a imagem do campo como um todo. Medidas antigas de masculinidade-feminilidade são um bom exemplo deste tipo de problema (Constantinople, 1973; Lenney, 1991; Spence, 1993), embora haja muitos outros.

Talvez a conseqüência mais significativa da definição tradicional de validade é que esta passou a ser associada aos testes e ao que eles pretendem medir, e não aos *escores* de teste e às interpretações que se podem basear neles. Por implicação, então, qualquer evidência rotulada como *validade de teste* passou a ser vista como prova de que o teste em questão era válido e digno de uso, independentemente da natureza da ligação entre os dados dos escores e as inferências que se pretendessem fazer a partir deles. Conseqüentemente, numerosos estudos na literatura psicológica usaram escores de um único instrumento para classificar participantes de pesquisas em grupos experimentais, muitos clínicos se valeram exclusivamente de escores de teste para diagnósticos e planejamento de tratamentos e um número desconhecido de decisões no contexto educacional e de emprego foram baseadas em escores de corte de um único teste. Com muita freqüência, escolhas como estas são feitas sem consideração à sua adequação ao contexto específico, ou sem referência a fontes adicionais de dados, e justificadas simplesmente com o argumento de que o teste em questão é considerado "uma medida válida de ..." qualquer coisa que seu manual afirme.

Um marco importante na evolução do conceito da validade foi a publicação de *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (Recomendações técnicas para testes psicológicos e técnicas diagnósticas; APA, 1954), o primeiro da série de padrões de testagem que foram rebatizados, revisados e atualizados em 1955, 1966, 1974, 1985 e, mais recentemente, em 1999 (AERA, APA, NCME). Com cada revisão subseqüente, os *Padrões de Testagem* – discutidos no Capítulo 1 – buscaram promover práticas sólidas para a construção e o uso de testes e esclarecer as bases para a avaliação da qualidade dos testes e práticas de testagem.

O *Technical Recommendations*, publicado em 1954, introduziu uma classificação da validade em quatro categorias que serão discutidas mais adiante neste capítulo: validade de conteúdo, validade preditiva, validade concorrente e validade de constructo. Subseqüentemente, os *Padrões* de 1974 reduziram estas categorias a três, englobando a validade preditiva e a concorrente na rubrica da validade relacionada ao critério, e especificaram ainda que a validade de conteúdo, a validade relacionada ao critério e a validade de constructo são *aspectos* e não *tipos* de validade. No mesmo ano, os *Padrões* também introduziram a noção de que a validade "se

refere à adequação das inferências feitas a partir de escores de teste ou outras formas de avaliação" (AERA, APA, NCME, 1974, p.25).

Apesar das especificações propostas pelos *Padrões* de 1974 há mais de um quarto de século, a divisão da validade em três tipos (que passaram a ser conhecidos como a *visão tripartite* da validade) tornou-se muito arraigada, sobrevivendo até hoje em muitos manuais e revisões de testes, bem como em grande parte das pesquisas conduzidas sobre instrumentos psicométricos. Não obstante, revisões sucessivas dos *Padrões* – especialmente a atual – acrescentaram estipulações que esclarecem cada vez melhor que qualquer classificação usada para os conceitos de validade deve estar ligada ao tipo de evidências citadas para a interpretação dos escores de teste, e não aos testes em si. Com isso em mente, nos voltamos agora para a consideração da visão prevalente da validade como um conceito unitário e para as várias fontes de evidências que podem ser usadas na avaliação de possíveis interpretações de escores para finalidades específicas. Mais informações sobre a evolução da validade e conceitos relacionados, ver Anastasi (1986), Angoff (1988) e Landy (1986).

PERSPECTIVAS ATUAIS SOBRE A VALIDADE

Desde os anos de 1970 até o presente, tem havido um esforço coordenado entre os profissionais da testagem para refinar e revisar a noção de validade e fornecer uma teoria unificadora que englobe as muitas linhas de evidências das quais os escores de teste derivam sua significância e sentido. Um tema consistente desse esforço tem sido a integração de quase todas as formas de evidência de validade como aspectos da validade de constructo (Guion, 1991; Messick 1980, 1988, 1989; Tenopyr, 1986). Isso, por sua vez, estimulou um reexame do sentido de *constructo* – definido em termos gerais no quadro Consulta Rápida 5.2 – aplicado especificamente no contexto da validade na testagem e avaliação psicológica (Braun, Jackson e Wiley, 2002; Embretson, 1983).

A função integrativa dos constructos na validação de testes

Na testagem psicológica, o termo *constructo* é usado, muitas vezes indistintamente, de duas formas:

1. Para designar os *traços, processos, reservas de conhecimento* ou *características* cuja presença e extensão desejamos determinar por meio das amostras específicas de comportamento coletadas pelos testes. Nesse sentido da palavra, um constructo é simplesmente aquilo que o autor do teste pretende medir – isto é, qualquer entidade hipotética derivada da teoria psicológica, da pesquisa ou observação do comportamento, como ansiedade, assertividade, capacidade de raciocínio lógico, flexibilidade, etc.
2. Para designar as inferências que podem ser feitas a partir dos escores de teste. Quando usado desta forma, o termo *constructo* se refere a uma interpretação específica de dados de teste, ou qualquer outro dado

comportamental – como a presença de depressão clínica ou uma alta probabilidade de sucesso em alguma iniciativa –, que pode ser feita a partir de uma rede de relações teóricas e empíricas preestabelecidas entre os escores e outras variáveis.

Diversos teóricos tentaram explicar como esses dois sentidos se relacionam à noção de validade de escore de teste. Uma das primeiras formulações foi a classificação da validade de Cronbach (1949) da validade em dois tipos, quais sejam, lógica e empírica. Subseqüentemente, em um trabalho influente em co-autoria com Meehl, em 1955, Cronbach sugeriu o uso do termo *validade de constructo* para designar a *rede nomológica*, ou rede de inter-relações entre os elementos teóricos e observáveis que dão suporte a um constructo. Em uma tentativa de esclarecer como esses dois sentidos podiam ser distinguidos no processo de desenvolvimento, construção e avaliação dos testes, Embretson (1983) propôs uma separação entre dois aspectos da pesquisa de validação de constructos, quais sejam, a *representação de constructo* e o *intervalo nomotético*. Segundo Embretson (p.180), a pesquisa de *representação de constructo* "busca identificar os mecanismos teóricos que embasam o desempenho em uma tarefa". Na perspectiva do processamento de informação, a meta da representação de constructo é a *decomposição da tarefa*. O processo de decomposição da tarefa pode ser aplicado a uma variedade de tarefas cognitivas, incluindo inferências interpessoais e julgamentos sociais, e acarretar um exame das respostas do teste do ponto de vista dos processos, estratégias e reservas de conhecimento envolvidas em seu desempenho. O *intervalo nomotético*, por outro lado, "diz respeito à rede de relações de um teste com outras medidas" (Embretson, p.180), isto é, refere-se à força, freqüência e padrão de relações significativas entre escores de teste e outras medidas dos mesmos traços – ou traços diferentes –, entre escores de testes e medidas de critério, etc.

Embretson (1983) descreveu outras características dos conceitos de representação de constructo e intervalo nomotético que ajudam a esclarecer as diferenças entre esses dois aspectos da pesquisa de validação de constructo. Duas questões destacadas por ela pertinentes à distinção entre as funções dos dois tipos de pesquisa são particularmente úteis quando se considera o papel das fontes de evidência de validade:

1. *A pesquisa de representação de constructo busca primariamente identificar diferenças nas tarefas de um teste, enquanto que a pesquisa de intervalo nomotético tem como foco as diferenças entre os testandos.* Na pesquisa de representação de constructo, um processo, estratégia ou reserva de conhecimento identificado pela decomposição de tarefa (p. ex., codificação fonética, raciocínio seqüencial ou habilidade de compreender textos de nível elementar) podem ser considerados essenciais ao desempenho em uma tarefa do teste, mas não produzir diferenças sistemáticas em uma população de testandos composta de leitores. Por outro lado, para investigar o intervalo nomotético dos escores (isto é, a rede de relações entre eles e outras medidas), é necessário ter dados sobre diferenças individuais e variabilidade entre os testandos. Isso reforça a importância crucial

da variabilidade de escores para a derivação de informações que possam ser usadas para se fazer determinações ou tomar decisões a respeito de pessoas, o que foi discutido nos Capítulos 2 e 3. Se os escores de um grupo de pessoas, por exemplo, em um teste delineado para avaliar a habilidade de compreender textos de nível elementar devem ser correlacionados com alguma outra coisa – ou usados para determinar qualquer coisa além do fato de essas pessoas, como grupo, possuírem ou não esta habilidade – é preciso haver alguma variabilidade nos escores.

2. *A validação do aspecto de representação de constructo das tarefas de um teste é independente das evidências confirmatórias que podem ser coletadas em termos do intervalo nomotético dos escores, e vice-versa.* Em outras palavras, embora possamos saber precisamente quais processos estão envolvidos no desempenho das tarefas de um teste, na ausência de correlações significativas com comportamentos ou medidas extra-teste relevantes, seus escores podem ter uso limitado. Seguindo o mesmo raciocínio, é possível obter uma forte rede de relações entre escores de teste e outras medidas sem ter uma noção clara do constructo que esses escores representam. O exemplo usado por Embretson (1983) é o dos escores em testes de inteligência, que têm um forte intervalo nomotético (na medida que se correlacionam com consistência mais ou menos forte com uma variedade de outras medidas), mas ainda assim têm bases teóricas relativamente pouco claras.

Não esqueça

As pessoas fazem inferências a partir de observações e amostras de comportamento o tempo todo. Por exemplo, se escutamos alguém falar com muitos erros gramaticais, podemos inferir que esta pessoa tem baixo nível de escolaridade. Se uma pessoa chega invariavelmente na hora marcada, podemos inferir que ela é pontual. Algumas de nossas inferências são corretas, e algumas não. Algumas são importantes, e outras não.

Se as inferências que fazemos são importantes o bastante para desejarmos determinar sua correção, ou seja, validá-las, precisamos

1. definir nossos termos inequivocamente (p. ex., o que queremos dizer com "escolaridade"? "Chegar sempre na hora marcada" representa plenamente o conceito de pontualidade?);
2. investigar a fidedignidade de nossas observações (p. ex., a pessoa sempre comete erros gramaticais, ou apenas em algumas circunstâncias? Nosso amigo chega na hora marcada em todos os seus compromissos, ou apenas naqueles que tivemos oportunidade de observar?);
3. decidir se existem evidências suficientes para justificar as inferências que queremos fazer com base em nossas definições e nos dados disponíveis (p. ex., chegar na hora marcada em todos os compromissos é base suficiente para se julgar a pontualidade de uma pessoa), ou se precisamos corroborar nossas inferências com mais dados (p. ex., a pessoa demonstra outros indicadores daquilo que queremos dizer com "baixo nível de escolaridade"?).

Os testes psicológicos são ferramentas criadas para ajudar a refinar e quantificar observações comportamentais para fins de interferências a respeito de indivíduos, grupos ou constructos psicológicos. Fundamentalmente, os escores de testes psicológicos são válidos se podem nos ajudar a fazer inferências precisas.

O esquema conceitual exposto por Embretson (1983) mantém a noção de validação de constructo como uma forma unitária e abrangente de expressar a abordagem científica da integração de qualquer evidência relacionada com o sentido ou interpretação dos escores de teste. Ao mesmo tempo, fornece uma base para a distinção entre (a) as fontes de evidência da validade de escores de teste ligadas primariamente à identificação do que estamos medindo (isto é, representação do constructo) e (b) aquelas que lidam principalmente com as inferências que podemos fazer a partir do que estamos medindo (isto é, intervalo nomotético). Deve-se observar que estas fontes de evidências podem ser, e muitas vezes são, inter-relacionadas e que ambas envolvem elementos teóricos e observáveis, bem como modelos ou postulados a respeito das inter-relações entre seus elementos.

FONTES DE EVIDÊNCIAS DE VALIDADE

Em geral, a essência dos julgamentos a respeito da validade dos escores de teste está centrada no relacionamento entre aquilo que os escores representam e as perguntas que os usuários de testes querem responder com seu uso. As perguntas que fazemos determinam o tipo de evidência de que precisamos, bem como as relações lógicas – indutivas e dedutivas – que devem ser estabelecidas para contemplar as questões de (a) o que estamos medindo com os testes e (b) que inferências podemos fazer a partir de seus escores. Nesse ponto também deve estar claro que quanto maior a significância ou o impacto potencial das respostas que queremos, mais convincentes precisam ser as evidências. No restante deste capítulo, vamos discutir os tipos de evidências necessárias para a validação de inferências feitas a partir de escores de teste, com o entendimento de que a interpretação proposta de um escore determina o referencial conceitual para sua validação. A Tabela 5.1 apresenta uma lista das principais categorias em que os aspectos da validade podem ser classificados, juntamente com as principais fontes de evidências para cada uma delas, que serão discutidas no restante deste capítulo. É importante reconhecer desde o início que nem os aspectos da validade, nem as fontes ou os tipos de evidências associados a eles são mutuamente exclusivos. As estratégias de validação devem, na verdade, incorporar tantas fontes de evidências quantas forem possíveis ou apropriadas à finalidade de um teste.

Evidências de validade baseadas no conteúdo do teste e processos de resposta

Alguns testes psicológicos são delineados para coletar amostras de comportamento que podem ser mais ou menos relacionadas diretamente às inferências que desejamos fazer a partir de seus escores. De modo geral, esses instrumentos se encaixam na categoria de testes referenciados no critério ou conteúdo, já discutidos mais aprofundadamente no Capítulo 3. A maioria desses testes é usada no contexto educacional e ocupacional, embora também possa ser aplicada em campos (p. ex.,

Tabela 5.1 Aspectos da validade de constructo e fontes de evidências relacionadas

Aspecto da validade do constructo	Fontes de evidências[a]
Relacionada ao conteúdo	Relevância e representatividade do conteúdo do teste e dos processos de resposta às tarefas Validade de face (isto é, aparência superficial)
Padrões de convergência e divergência	Consistência interna de resultados do teste e outras medidas de fidedignidade Correlações entre testes e subtestes Matriz multitraço-multimétodo Diferenciação de escores de acordo com diferenças esperadas com base na idade e outras variáveis de *status* Resultados experimentais (isto é, correspondência entre escores de teste e os efeitos preditos de intervenções experimentais ou hipóteses baseadas em teorias) Análise fatorial exploratória Técnicas de modelagem de equação estrutural
Relacionada ao critério	Precisão das decisões baseadas na validação concorrente (isto é, correlações entre escores de teste e critérios existentes) Precisão de decisões ou predições baseadas na validação preditiva (isto é, correlações entre escores de testes e critérios preditos)

[a] Ver Capítulo 5 para explicações dos termos

avaliação neuropsicológica) em que é necessário determinar se uma pessoa é capaz ou incapaz de realizar tarefas de significância diagnóstica. Esses testes ou são compostos de itens que colhem amostras de conhecimento de um domínio de conteúdo definido ou requerem que os testandos demonstrem que possuem uma determinada habilidade ou competência. Os procedimentos de validação para testes deste tipo são o aspecto mais simples e de maior consenso no desenvolvimento de testes, porque as evidências a partir das quais as inferências serão feitas podem ser defendidas com argumentos lógicos e relações demonstráveis entre o conteúdo do teste e o constructo que este pretende representar.

As evidências de validade de escores derivadas do conteúdo de um teste podem ser embutidas em um novo instrumento, desde o início, pela escolha de seus itens ou tarefas. O requisito primário para o desenvolvimento de testes deste tipo é uma especificação cuidadosa dos domínios de conteúdo, processos cognitivos, habilidade ou tipos de desempenho dos quais serão coletadas amostras e de sua importância ou peso relativo.

- No contexto educacional, exemplos destas especificações podem ser encontrados nos currículos escolares, ementas de cursos, bibliografias e qualquer outro material que delineie, defina ou priorize os objetivos das experiências educacionais ou de treinamento tanto em termos do conhecimento de conteúdos quanto das capacidades de desempenho. O processo de delimitar o domínio de conhecimento e determinar os resultados deseja-

dos da instrução está dentro da esfera de ação dos professores, instrutores e outros especialistas que determinam currículos ou escrevem os livros que servem como texto básico em várias disciplinas.

- No contexto ocupacional, as especificações da habilidade ou domínio de conteúdo que um teste vai avaliar se baseiam em análises de função (*job analyses*). *Análise de função* se refere a qualquer um de vários métodos que buscam descobrir a natureza de um dado emprego pela descrição dos elementos, atividades, tarefas e deveres a ele relacionados (Brannick e Levine, 2002). A metodologia da análise de função tem uma variedade de aplicações no manejo de recursos humanos, incluindo avaliações de desempenho e determinação de necessidades de treinamento, entre outras. No contexto da seleção e classificação ocupacional, as análises de função – baseadas em informações de empregadores, supervisores e/ou colegas – são usadas para delinear as habilidades e reservas de conhecimento necessárias para o desempenho no trabalho.
- Na avaliação neuropsicológica, as especificações dos processos e capacidades cognitivas a serem avaliadas derivam do conhecimento teórico e empírico sobre as ligações entre o sistema nervoso central e as funções comportamentais. A natureza do conteúdo das ferramentas de avaliação neuropsicológica se baseia em evidências clínicas e científicas acumuladas a respeito das relações entre mente e comportamento.

O quadro Consulta Rápida 3.7, no Capítulo 3, lista alguns exemplos simples de objetivos e itens típicos de testes referenciados no domínio. Exemplos mais extensos, incluindo tabelas de especificações para testes referenciados no conteúdo e orientação a respeito de sua preparação, estão disponíveis em Gronlund (2003) e Linn e Gronlund (1995, p.119-125). Independentemente do contexto em que um teste referenciado no conteúdo é aplicado, após a especificação dos conhecimentos, habilidades ou processos a serem medidos através dele os procedimentos de validação do conteúdo envolvem a revisão crítica e o exame do conteúdo do teste a partir de duas perspectivas. A primeira é a *relevância* do conteúdo apresentado pelo teste para o domínio específico, e a segunda é a sua *representatividade* em relação às especificações do domínio que ele pretende cobrir. Embora dependa de um consenso de especialistas no assunto em questão, a questão da relevância também pode ser corroborada por achados empíricos, como diferenças nos escores de estudantes em séries sucessivas ou de indivíduos em vários estágios do processo de treinamento. Os autores e criadores de testes devem respaldar suas afirmações de validade relacionada ao conteúdo dos escores em manuais, livros técnicos e outras fontes de documentação confirmatória para testes. Quando a base primária das evidências de validação de um instrumento está centrada no conteúdo, habilidade ou processos cognitivos específicos que ele avalia, é necessária uma descrição dos procedimentos sistemáticos usados para garantir a relevância e a representatividade destes para os domínios-alvo do teste. O quadro Consulta Rápida 4.3 apresenta um exemplo de como a representatividade inadequada da cobertura de conteúdo tanto pode minar a fidedignidade quanto a validade dos escores de testes referenciados no conteúdo.

Testagem educacional

Escores que derivam sua validade de uma conexão direta e demonstrável entre o conteúdo do teste e as especificações usadas em seu desenvolvimento abundam em todos níveis da educação e do treinamento nos quais os resultados da instrução podem ser definidos sem ambigüidade. Quase todos os testes de sala de aula criados por professores se encaixam nessa categoria, assim como muitos testes padronizados publicados pela ETS, a ACT e organizações semelhantes. O primeiro objetivo destes instrumentos é medir a realização educacional – isto é, o que os estudantes aprenderam através da escolarização. Os escores destes testes podem responder mais diretamente a perguntas como "quanto de um domínio específico o aluno registrou?" ou "que grau de competência ou proficiência o testando atingiu na habilidade em questão?". Instrumentos referenciados no conteúdo ou domínio podem ser aplicados a uma variedade de decisões, incluindo atribuição de notas em uma disciplina, fornecimento de créditos através de exames, colação de grau ou distribuição de diplomas após um programa de estudos, certificação ou licenciamento de indivíduos para a prática profissional em um determinado campo, ou mesmo determinação da prontidão para passar a um nível mais avançado de treinamento. Tipicamente, as decisões baseadas nestes testes dependem dos níveis de maestria demonstrados pelos testandos, que, por sua vez, podem ser medidos em termos de notas de percentagem, ordens de percentil em comparação com grupos normativos apropriados ou determinações simples de aprovação ou reprovação com base em critérios preestabelecidos, como foi discutido no Capítulo 3. O quadro Consulta Rápida 5.3 lista alguns exemplos típicos de testes educacionais padronizados, juntamente com os objetivos para os quais eles foram desenvolvidos (isto é, o conhecimento e as habilidades que eles pretendem avaliar) e suas aplicações básicas. Amostras de questões e descrições mais elaboradas desses testes estão disponíveis nos *sites* da Internet listados na última coluna.

Testagem ocupacional

Muitos instrumentos usados para selecionar ou determinar o cargo mais adequado para candidatos a empregos consistem em amostras ou simulações de trabalho que exigem o desempenho de tarefas que compõem o emprego (p. ex., testes de digitação) ou amostras de comportamento que podem ser ligadas diretamente ao desempenho profissional por meio de análises de função. Alguns desses testes são desenvolvidos nas próprias empresas por seus funcionários e usam normas ou critérios de desempenho locais. Outros são instrumentos padronizados que oferecem escores normativos para indivíduos em várias ocupações e medem constructos de vários graus de amplitude. Numerosos exemplos destes testes podem ser encontrados na seção Vocações do Índice por Classificação de Assunto do *Tests in Print VI* (Murphy et al., 2002). Dois exemplos que podem ser usados para ilustrar a diversidade entre testes desse tipo são descritos no quadro Consulta Rápida 5.4.

CONSULTA RÁPIDA 5.3

Exemplos de testes educacionais padronizados que usam evidências baseadas no conteúdo como principal fonte de validação

Título do teste	Objetivo principal	Aplicações primárias	Site da Internet com descrição e amostras do teste
Test of English as a Foreign Language (TOEFL)	Avaliar a proficiência em inglês de pessoas cuja língua nativa não seja o inglês	Determinar se estudantes estrangeiros possuem conhecimento suficiente de inglês para serem admitidos em faculdades americanas	http://www.toefl.org
College-Level Examination Program (CLEP) Introductory Psychology Test	Medir o conhecimento dos materiais habitualmente ensinados em disciplina introdutória de psicologia em um semestre	Determinar se os estudantes têm conhecimento suficiente de psicologia introdutória para receber um crédito universitário através de exame	http://www.collegeboard.com/clep
ACT Assessment Science Reasoning Test	Medir as habilidades de interpretação, análise, avaliação, raciocínio e solução de problemas necessárias no campo das ciências naturais, incluindo biologia, química, física e ciências espaciais	Avaliar o conhecimento e as habilidades adquiridas por um estudante para determinar sua capacidade para assumir empregos de nível universitário	http://www.act.org
National Assessment of Educational Progress (NAEP)	Medir conhecimentos e habilidades em leitura, matemática, ciências, escrita, história dos EUA, geografia e artes	Fornecer informações a respeito do desempenho de populações e subgrupos de estudantes em todos os EUA e estados participantes	http://nces.ed.gov

Exemplos de testes ocupacionais padronizados que usam evidências baseadas no conteúdo como fonte de validação

Título do teste	Constructo avaliado	Descrição	Aplicação primária
Crawford Small Parts Dexterity Test (CSPDT)[a]	Coordenação visual-manual e destreza motora fina	O CSPDT consiste em duas tarefas: (a) trabalhar com pinças inserindo pequenos alfinetes nos orifícios de uma bandeja e depois colocar pequenos aros sobre as partes projetadas dos alfinetes; (b) inserir parafusos na bandeja e depois apertá-los com uma chave de fenda. A velocidade do desempenho é o principal fator na avaliação deste teste.	Usado para determinar se um indivíduo tem a destreza manual necessária para qualquer emprego que envolva trabalho de precisão com as mãos, como entalhes ou conserto de relógios.
Clerical Abilities Battery (CAB)[a]	Diversos componentes de uma ampla gama de ocupações administrativas identificadas pela análise de função de comportamentos administrativos gerais	O CAB tem sete subtestes auto-explicativos: Arquivamento, Comparação de Informações, Cópia de Informações, Uso de Tabelas, Revisão, Habilidades Básicas em Matemática e Raciocínio Numérico.	Usado para recrutamento e avaliação de funcionários administrativos. Em uma revisão do *Mental Measurements Yearbook*, Randhawa (1992) afirma que a amostragem e a amplitude das tarefas dos subtestes do CAB não são suficientemente representativos, e sugere que são necessários mais dados de padronização, fidedignidade e validade preditiva. No entanto, ele admite que o processo de desenvolvimento e o formato da bateria são adequados e fornecem as bases para uma ferramenta potencialmente excelente.

[a]Publicado por Psychological Corporation (http://www.PsychCorp.com).

Em parte devido ao custo e à dificuldade envolvida no desenvolvimento e na validação de instrumentos de avaliação de habilidades no nível local, a organização ACT (antigo *American College Testing Program*) (www.act.org) iniciou um programa conhecido como sistema *WorkKeys* que combina um número de componentes voltados para auxiliar companhias a recrutar, selecionar, contratar e treinar empregados. O componente de perfil de cargo permite aos funcionários ou seus supervisores, em consulta com especialistas da ACT, selecionar as tarefas mais importantes para um dado emprego e identificar as habilidades e níveis de habilidade necessários para o sucesso no seu desempenho. O aspecto de avaliação do WorkKeys fornece instrumentos padronizados para avaliar os níveis de habilidade dos candidatos ou funcionários em várias áreas críticas, como Tecnologia Aplicada, Redação Comercial, Localização de Informações, Capacidade de Escuta e Trabalho em Equipe, entre outras. A avaliação de habilidades na Localização de Informações, por exemplo, apresenta questões em quatro níveis sucessivos de complexidade e mede habilidades que vão de encontrar informações constantes em gráficos elementares – como formulários simples, gráficos de barras e plantas – a tirar conclusões a partir de informações apresentadas em tabelas, gráficos e plantas muito detalhadas, etc. Com base em comparações das informações oferecidas por estas ferramentas de avaliação de habilidades e os níveis mínimos de habilidade requeridos nos perfis de cargo, os empregadores podem avaliar as qualificações dos candidatos ou as necessidades de treinamento de seus funcionários.

Evidências de validade de conteúdo em outros contextos de avaliação

O grau em que os itens de testes são relevantes e representativos de um constructo pode ser uma fonte adicional de evidências de validade para instrumentos em praticamente qualquer campo. Por exemplo:

- *Na avaliação neuropsicológica,* como já foi mencionado, as especificações dos processos cognitivos e da capacidade de comportamento a serem avaliadas são derivadas do conhecimento teórico e empírico bem estabelecido das relações entre funções cognitivas ou comportamentais e as bases neurológicas presumíveis destas funções. Assim sendo, em grande parte, o processo da avaliação neuropsicológica se vale do conhecimento especializado das evidências científicas acumuladas a respeito das relações entre a mente e o comportamento. Uma bateria neuropsicológica adequada deve incluir itens que exploram uma gama de comportamentos suficientemente ampla e representativa para coletar evidências de capacidade ou comprometimento funcional nos vários sistemas que pretende avaliar (ver, p. ex., Franzen, 2000; Lezak, 1995). O *Boston Diagnostic Aphasia Examination* (Goodglass, Kaplan e Barresi, 2001), por exemplo, fornece uma amostragem sistemática de diversas funções de comunicação, tais como compreensão auditiva e expressão oral, para auxiliar no diagnóstico de síndromes afásicas e transtornos da linguagem.
- *Na avaliação da personalidade,* muitas ferramentas de auto-relato – como listas de itens, inventários e levantamentos de atitudes ou opiniões – se

valem em grande parte do conteúdo de seus itens para ajudar a gerar hipóteses ou fazer inferências a respeito dos constructos particulares que pretende avaliar. Procedimentos de observação estruturados, bem como vários inventários usados para coletar dados baseados nos relatos de pares, pais, cônjuges, professores e outros observadores também usam o conteúdo de itens como fonte básica de evidências de validade. Da mesma forma, testes psicológicos delineados para auxiliar no diagnóstico de transtornos psiquiátricos muitas vezes incluem itens, ou podem até ser compostos inteiramente por eles, que refletem aspectos sintomáticos críticos das síndromes que pretendem diagnosticar. Mais uma vez, a relevância e a representatividade dos itens destes instrumentos é de importância crucial para determinar sua utilidade para fins diagnósticos. Exemplo de testes deste tipo incluem o *Inventário de Depressão de Beck (BDI)*, o *State-Trait Anxiety Inventory (STAI)*, a *Symptom Checklist-90-Revised (SCL-90-R)* e o *Attitudes Toward Women Scale (AWS;* Spence e Helmreich,1972).

Evidências de validade do ponto de vista dos testandos

A relevância e a representatividade do conteúdo dos testes também é pertinente em relação a uma questão que é menos substantiva do que a validade dos escores, mas que mesmo assim é bastante importante. A *validade aparente* se refere à aparência superficial daquilo que o teste mede na perspectiva de um testando ou de qualquer outro observador leigo. Todos os instrumentos discutidos até agora têm alguma validade aparente quando usados nos contextos que têm sido discutidos. Para os testandos, eles parecem estar em consonância com as finalidades educacionais, ocupacionais, clínicas ou investigativas expressas das situações de avaliação nas quais são tipicamente aplicados. Embora a validade aparente não seja necessariamente uma indicação de validade na perspectiva psicométrica, ela não deixa de ser uma característica desejável dos testes, porque promove o *rapport* e a aceitação da testagem e de seus resultados por parte dos testandos. Se o conteúdo de um teste parece ser impróprio ou irrelevante, a disposição dos testandos em cooperar com o processo de testagem pode ser minada. Por isso, os criadores de testes precisam levar em conta a aparência de validade na perspectiva de todas as partes envolvidas – incluindo os testandos e outros leigos – e, sempre que possível, incorporar conteúdos que pareçam relevantes e apropriados às situações nas quais o teste deverá ser usado.

Evidências de validade baseadas na exploração de padrões de convergência e divergência

Quando vai além dos relacionamentos diretos e bastante claros entre o conteúdo do teste e as reservas de conhecimento, habilidades e processos funcionais que eles pretendem avaliar, a interpretação dos escores começa a depender de fontes cada vez mais indiretas de evidências de validade. Isso se aplica especialmente aos tes-

tes na área da personalidade, não apenas porque os constructos que eles avaliam geralmente são mais teóricos e abstratos do que os avaliados por testes cognitivos, mas também porque as respostas dos testandos às ferramentas de avaliação da personalidade são influenciadas por muito mais determinantes de situação e de estilo pessoal do que as respostas a testes cognitivos.

Existe um número grande e sempre crescente de métodos que podem ser usados para melhorar o sentido dos escores de teste para além da relevância e da representatividade de seu conteúdo. O denominador comum de todos estes procedimentos é sua produção de evidências na forma de padrões de convergência e divergência entre os escores de teste e outras variáveis (ver Tabela 5.1). Embora uma explicação detalhada desses métodos esteja muito além do âmbito deste livro, uma descrição básica dos procedimentos encontrados com maior freqüência é justificada.

A fidedignidade dos escores como fonte de evidência de validade

As investigações sobre a fidedignidade dos escores de teste do ponto de vista da estabilidade, diferenças entre avaliadores, erro de amostragem de conteúdo e sua heterogeneidade podem fornecer evidências a respeito da coesão, ou distintividade, do conteúdo de um teste. Como foi discutido no Capítulo 4, a fidedignidade de um escore pode por si só ser vista como evidência preliminar da obtenção de uma medida confiável de amostra de comportamento, podendo contribuir com evidências indiretas da validade de um escore de teste. Se, por exemplo, o teste for delineado para avaliar um constructo unidimensional, como a habilidade de soletrar, altos coeficientes de consistência interna iriam confirmar a alegação de unidimensionalidade. Da mesma forma, se for obtida uma consistência de escores entre diferentes avaliadores, pode-se supor que todos eles estão empregando os mesmos critérios e, assim, provavelmente avaliando as mesmas características. Se o constructo que está sendo avaliado for estável – por exemplo, um traço ou tipo de personalidade – uma alta fidedignidade de teste-reteste nos escores seria um pré-requisito essencial para evidências de validade.

Correlações entre testes e subtestes

Um modo simples e freqüentemente usado para coletar evidências de que um teste em particular mede o constructo que pretende medir é estabelecer altas correlações entre seus escores e os de outros instrumentos que também avaliam o mesmo constructo. Um dos exemplos mais básicos deste tipo de procedimento ocorre quando os testes são revisados e renormatizados. Nestes casos, os manuais quase invariavelmente citam altas correlações entre as edições anteriores e as novas como evidência de que ambas estão medindo os mesmos constructos. Isto se assemelha ao cálculo de correlações entre formas alternativas de um teste para estabelecer a fidedignidade ou consistência de escore entre as diferentes formas. Podemos recordar do Capítulo 3, no entanto, que mesmo se as correlações entre a versão antiga e a versão revisada de um teste forem muito altas, os escores normativos para as

versões repadronizadas tendem a flutuar em uma direção ou outra devido a mudanças na população em diferentes períodos de tempo.

De forma semelhante, os criadores de testes tipicamente apresentam correlações entre os escores de seus testes e os de instrumentos comparáveis como evidência de validade de escores. Por exemplo, todos os manuais das principais escalas de inteligência individual citam correlações entre seus escores e os de outros instrumentos bem estabelecidos do mesmo tipo. Examinando esses dados podemos constatar, por exemplo, que a correlação entre o QI Total da WAIS-III e o escore global composto da Stanford-Binet-IV (SB-IV), calculada para uma amostra de 26 indivíduos que se submeteram a ambos os testes, foi de 0,88 (Psychological Corporation, 1997, p.84), ou que a correlação obtida entre os escores compostos da SB-IV e do *Kaufman Adolescent e Adult Intelligence Scale* (KAIT) para uma amostra de 72 indivíduos testados com ambos os instrumentos foi de 0,87 (Kaufman e Kaufman, 1993, p.100). Coeficientes de correlação deste tamanho são típicos das principais escalas de inteligência e servem para corroborar o fato de que boa parte da variância de escores nestes testes é compartilhada.

Também podem ser obtidos coeficientes de correlação entre escores em subtestes de escalas diferentes. Exemplos típicos disso seriam as correlações entre os escores de várias escalas de depressão, como, por exemplo, da escala de Depressão do MMPI-2 e da escala de Distimia (r = 0,68) do *Millon Clinical Multiaxial Inventory-III* (MCMI-III), ou os escores do Inventário de Depressão de Beck e os da escala de Depressão Maior do MCMI-III (r = 0,71; Millon, Millon e Davis, 1994, p.126, 129). Como seria de se esperar, coeficientes de correlação calculados entre vários tipos de testes e subtestes são citados em profusão em manuais de teste e na literatura psicológica, embora esses índices com freqüência não sejam muito convincentes ou informativos.

As correlações entre testes são tão abundantes porque os dados para estudos correlacionais em pequenas amostras de conveniência são fáceis de coletar, especialmente para os testes de lápis e papel, que podem ser administrados facilmente a grupos. Correlações obtidas dessa forma podem, é claro, variar de zero a $\pm 1{,}00$, dependendo dos escores em questão e da natureza das amostras usadas (cf. Capítulo 2, especialmente a seção sobre Restrição da Amplitude e Correlação). Embora o sentido de qualquer coeficiente isolado esteja aberto a interpretações, se forem acumulados dados suficientes demonstrando correlações consistentemente altas ou baixas entre as medidas, alguns padrões de convergência e divergência poderão ser discernidos. Esses padrões informam aos usuários dos testes a quantidade aproximada de variância comum ou compartilhada entre conjuntos de escores e, indiretamente, o sentido dos escores em si. Correlações consistentemente altas entre medidas delineadas para avaliar um dado constructo – tais como as correlações citadas nos parágrafos anteriores entre escalas de depressão – podem ser tomadas como evidências de *validade convergente*, isto é, evidências da semelhança ou identidade dos constructos avaliados. Seguindo o mesmo raciocínio, as evidências de *validade discriminante*, baseadas em correlações consistentemente baixas entre medidas que devem diferir, também podem ser usadas para substanciar a identidade dos constructos que elas exploram. Um exemplo desse tipo de padrão divergente pode ser visto nas correlações entre escores na escala Bipolar, Maníaca do MCMI-III e a escala de De-

pressão do MMPI-2 (r = 0,06), bem como entre os escores da escala de Depressão Maior do MCMI-III e a escala de Hipomania do MMPI-2 (r = 0,08), ambas calculadas com uma amostra de 132 indivíduos (Millon, Millon e Davis, 1994, p.129-130).

A Matriz Multitraços-Multimétodos

Em um esforço para organizar a coleta e a apresentação de dados de validação convergentes e discriminantes, D.T. Campbell e Fiske (1959) propuseram um delineamento denominado *matriz multitraços-multimétodos* (MTMMM). Esta abordagem se refere a uma estratégia de validação que requer a coleta de dados sobre dois ou mais traços distintos (p. ex., ansiedade, afiliação e dominância) por dois ou mais métodos diferentes (p. ex., questionários de auto-relato, observações comportamentais e técnicas projetivas). Depois que esses dados são coletados e todas as suas intercorrelações são calculadas, eles podem ser apresentados na forma de uma matriz, como a da Tabela 5.2. As matrizes multitraços-multimétodos exibem (a) coeficientes de fidedignidade para cada medida, (b) correlações entre escores no mesmo traço avaliados por diferentes métodos (isto é, dados de validade convergente) e (c) correlações entre escores em traços diferentes medidos pelos mesmos métodos, bem como (d) entre escores em traços diferentes avaliados por métodos diferentes (ambos constituem dados de validade discriminante). A Tabela 5.2 é um MTMMM hipotético com um padrão de resultados que seria considerado exemplar para este tipo de delineamento de validação. A matriz desta tabela mostra:

- os coeficientes mais altos, que são indicativos de fidedignidade de escore adequada (entre parênteses), na diagonal principal;

Tabela 5.2 Uma matriz multitraços-multimétodos hipotética (MTMMM)

		Auto-Relato			Observação			Projetiva		
Método	Traço	Ans	Afi	Dom	Ans	Afi	Dom	Ans	Afi	Dom
Auto-Relato	Ans	(0,90)								
	Afi	*0,45*	(0,88)							
	Dom	*0,35*	*0,38*	(0,80)						
Observação	Ans	**0,60**	0,23	0,10	(0,92)					
	Afi	0,25	**0,58**	-0,08	*0,47*	(0,93)				
	Dom	0,12	-0,12	**0,55**	*0,30*	*0,32*	(0,86)			
Projetiva	Ans	**0,56**	0,22	0,11	**0,65**	0,40	0,31	(0,94)		
	Afi	0,23	**0,57**	0,05	0,38	**0,70**	0,29	*0,44*	(0,89)	
	Dom	0,13	-0,10	**0,53**	0,19	0,26	**0,68**	*0,40*	*0,44*	(0,86)

Nota: Ans = ansiedade; Afi = afiliação; Dom = dominância. Os coeficientes de fidedignidade estão entre parênteses, ao longo da diagonal principal. Os coeficientes de validade (mesmo traço avaliado por diferentes métodos) estão em negrito. Todos os outros coeficientes são índices da validade discriminante de escores de traços diferentes avaliados por um único método (representando a variância com o mesmo método em itálico) e traços diferentes avaliados por métodos diferentes (letra simples).

- os coeficientes mais altos seguintes – em negrito – entre medidas do mesmo traço feitas com métodos **diferentes**, indicando convergência entre seus escores;
- os coeficientes mais alto seguintes – em *itálico* – entre medidas de traços *diferentes* avaliados pelo *mesmo* método, indicando que uma boa quantidade da variância nos escores se deve aos métodos empregados;
- os menores coeficientes – em letra simples – entre medidas de traços diferentes avaliados por métodos diferentes, indicando que as medidas realmente discriminam bem entre traços distintos.

O delineamento MTMMM é ideal para investigar padrões de convergência e divergência entre escores de testes e dados coletados com outros tipos de instrumentos de avaliação. No entanto, ele constitui um padrão de validação um tanto rigoroso que costuma ser difícil de atingir, especialmente para instrumentos de avaliação da personalidade, cujos escores são propensos a exibir um alto índice de *variação devida ao método* (isto é, variabilidade relacionada a características inerentes a suas metodologias). Além disso, o delineamento MTMMM não é aplicado em sua forma completa com muita freqüência, porque coletar informações através de múltiplos métodos é bastante trabalhoso (Terrill, Friedman, Gottschalk e Haaga, 2002). Não obstante, variações mais simples do esquema MTMMM, baseadas em escores de testes que tanto medem constructos semelhantes quanto diferentes, ainda que por métodos semelhantes, estão sendo cada vez mais empregadas no processo de validação de testes. Além disso, alguns instrumentos têm características que facilitam a coleta de dados de diferentes fontes, que podem então ser usados para estudar padrões de convergência e discriminação. Por exemplo, o *Revised NEO Personality Inventory* (NEO PI-R) fornece versões paralelas dos mesmos conjuntos de itens – Forma S para auto-relatos e Forma R para avaliações feitas por observadores como pares ou cônjuges – que podem ser usadas para correlacionar e comparar cores derivados de ambas as fontes (Costa e McCrae, 1992, p.48-50).

Diferenciação de idade

Resultados de testes consistentes com tendências desenvolvimentais bem estabelecidas entre faixas etárias costumam ser vistos como evidências de validade dos escores. Na verdade, o critério da diferenciação de idade é uma das fontes mais antigas de evidências para a validação de testes de habilidade. Podemos recordar do Capítulo 1 que o sucesso das escalas Binet-Simon originais foi medido basicamente por estudos que provaram que suas amostragens das funções cognitivas produziam resultados que podiam ser usados para descrever quantitativamente os níveis de habilidade das crianças, em termos de faixas etárias às quais seu desempenho correspondia. Na maioria dos testes de habilidade, o desempenho das crianças e adolescentes das amostras normativas mostra tipicamente uma tendência ascendente em idades cronológicas sucessivas. No outro extremo do espectro etário, observa-se um declínio do desempenho entre amostras de adultos mais velhos em instrumentos que medem habilidades que tendem a diminuir com a idade, como

testes de memória e testes que avaliam a velocidade de desempenho. A diferenciação de idade também é evidente em estudos cuidadosamente delineados sobre tendências de longo prazo no desempenho de indivíduos de várias idades em testes de habilidade mental, como o *Seattle Longitudinal Study* (Schaie, 1994). Os aumentos ou declínios em escores consistentes com expectativas relacionadas à idade fornecem evidências de que é necessário, ainda que não suficiente, mostrar que um teste está medindo os constructos de habilidade que foi delineado para medir.

Resultados experimentais

Outra fonte indireta de evidências que pode ser útil na validação de escores é fornecida por investigações que usam escores de testes psicológicos como variável dependente para avaliar os efeitos de intervenções experimentais. Na área da testagem de habilidades, estas evidências derivam primariamente de diferenças entre os escores pré e pós-teste, após intervenções com o objetivo de remediar deficiências ou melhorar o desempenho em várias habilidades cognitivas e intelectuais. Por exemplo, se os escores em um teste do desenvolvimento conceitual básico em crianças pequenas (p. ex., o *Bracken Basic Concept Scale-Revised*) mostrassem um aumento significativo para um grupo exposto a um programa de reforço de curto prazo – comparado a nenhuma mudança para um grupo pareado que não participasse do programa –, a mudança nos escores poderia ser vista como evidência de sua validade, bem como da eficácia do programa. Contrastes pré e pós-teste semelhantes costumam ser usados para documentar a validade de escores derivados de ferramentas de avaliação da personalidade. Um exemplo desse tipo de estudo de validação pode ser encontrado no manual do *Quality of Life Inventory* (QOLI; Frisch, 1994, p.15-16), uma ferramenta para a avaliação de níveis de satisfação com a vida e bem-estar subjetivo que pode ser usada – entre outras coisas – para medir a eficácia do aconselhamento ou intervenções psicoterapêuticas.

Análise fatorial

Uma forma de lidar com o número imenso de constructos explorados pelos testes existentes – e com o número maciço de correlações que podem ser obtidas de seus escores globais, escores de subtestes e escores de itens – é através de uma série de procedimentos estatísticos conhecidos coletivamente como *análise fatorial* (AF). A principal meta da análise fatorial é reduzir o número de dimensões necessárias para se descrever dados derivados de um grande número de medidas. Ela é feita por meio de uma série de cálculos matemáticos, baseados na álgebra matricial, que buscam extrair padrões de intercorrelação entre um conjunto de variáveis.

Existem dois modos básicos de conduzir análises fatoriais. A abordagem original é de natureza exploratória, e por isso é conhecida como *análise fatorial exploratória*, ou AFE. Seu objetivo é descobrir quais fatores (isto é, variáveis latentes ou constructos) subjazem às variáveis em análise. Uma abordagem mais recente é denominada *análise fatorial confirmatória* (AFC) porque busca testar hipóteses ou confirmar teorias a respeito de fatores presumidamente existentes. Ambas as

abordagens podem ser usadas na análise de dados de testes psicológicos, bem como em muitos outros tipos de conjuntos de dados. As análises confirmatórias são mais sofisticadas do ponto de vista metodológico e serão discutidas mais adiante neste capítulo como um subconjunto das técnicas para análise de estruturas de covariância conhecidas como modelagem de equação estrutural.

Quais são os passos envolvidos na análise fatorial dos escores de testes psicológicos? As análises fatoriais exploratórias começam com uma *matriz de correlação*, uma tabela que exibe as intercorrelações entre os escores obtidos por uma amostra de indivíduos em uma ampla variedade de testes (ou subtestes ou itens). O fato de este ser o ponto de partida da AFE é importante para a compreensão dos resultados das pesquisas em análise fatorial, porque aponta duas características cruciais da AF que muitas vezes são esquecidas. Ambas as características dizem respeito a limitações da aplicabilidade dos resultados oriundos de qualquer AFE isolada, quais sejam, os resultados dependem em grande parte (a) da escolha das medidas incluídas na análise, e (b) da composição específica da amostra cujos escores fornecem dados para a análise.

A Tabela 5.3, Parte A, mostra um exemplo de uma matriz de correlação simples. Essa matriz foi derivada dos escores obtidos por 95 estudantes universitários

Tabela 5.3

A. Matriz de correlação: intercorrelações de escores em cinco subtestes do Beta III para 95 estudantes universitários

Subteste	Codificação	Completar desenhos	Checagem administrativa	Absurdos em desenhos	Raciocínio matricial
Codificação	1,00	0,13	0,62**	0,20	0,05
Completar desenhos		1,00	0,09	0,21*	0,11
Checagem administrativa			1,00	0,18	0,20
Absurdos em desenhos				1,00	0,31**
Raciocínio matricial					1,00

Nota: Os dados são de Urbina e Ringby (2001).
*p= .05 **p= .01

B. Matriz fatorial para os dois fatores extraídos da análise fatorial exploratória (AFE) de cinco subtestes do Beta III[a]

Subteste	Cargas no fator 1	Cargas no fator 2
Codificação	**0,90**	0,06
Completar desenhos	0,07	**0,54**
Checagem administrativa	**0,88**	0,14
Absurdos em desenhos	0,15	**0,75**
Raciocínio matricial	0,02	**0,73**

Nota: Os números em negrito indicam as cargas mais altas nos dois fatores transformados por rotação varimax.
[a]Os fatores 1 e 2 respondem por 61% da variância nos escores dos subtestes; os 39% restantes da variância são explicados por fatores específicos de cada subteste e pela variância de erro.

nos cinco subtestes do Beta III, um teste não-verbal de habilidade intelectual derivado do *Army Beta* (ver Capítulo 1). Os dados fazem parte de um estudo sobre diferenças de gênero em habilidades cognitivas (Urbina e Ringby, 2001).

Os passos seguintes na análise fatorial dependem da escolha específica das técnicas empregadas pelo investigador. Diversos procedimentos diferentes podem ser usados para conduzir essas análises e para extrair fatores (Bryant e Yarnold, 1995; Comrey e Lee, 1992). Uma discussão dos procedimentos de análise fatorial está além de nossos objetivos, pois iria envolver o aprofundamento em questões técnicas, como métodos para a extração e rotação de fatores, que são bastante complexas. Mesmo assim, o fato de existirem várias abordagens da análise fatorial deve ser observado e registrado porque podemos chegar a diferentes soluções, dependendo das premissas e métodos usados. As diferenças nas soluções geralmente dizem respeito ao número de fatores extraídos e à sua relativa independência mútua. Para mais informações sobre esta e outras questões relacionadas à metodologia da análise fatorial, ver Russel (2002).

O produto final das análises fatoriais é uma *matriz fatorial*, uma tabela que lista as cargas de cada uma das variáveis originais nos fatores extraídos com as análises. As *cargas fatoriais* são correlações entre as medidas originais na matriz de correlação e os fatores que foram extraídos. A Parte B da Tabela 5.3 mostra a matriz fatorial para os dois fatores extraídos de uma análise fatorial dos dados na matriz de correlação da Parte A da mesma tabela. Esta matriz fatorial indica que os dois subtestes (Codificação e Checagem Administrativa) têm cargas muito altas no Fator 1 e cargas insignificantes no Fator 2, enquanto que os outros três subtestes (Completar Figuras, Absurdos em Desenhos e Raciocínio Matricial) mostram o padrão inverso em suas cargas fatoriais.

Interpretando os resultados de análises fatoriais. Depois de obtidas, as matrizes fatoriais podem ser examinadas para se determinar a natureza dos fatores que explicam a maior parte da variância do conjunto original de dados. Os fatores em si são identificados a partir da lógica indutiva. Para identificar e nomear os fatores, devemos examinar as características distintivas das medidas com carga maior e menor em cada um dos fatores da matriz. Em nosso exemplo, a matriz fatorial da Tabela 5.3 sugere que o primeiro fator envolve a *velocidade de desempenho*, porque os dois subtestes que têm carga mais pesada naquele fator envolvem tarefas extremamente simples com limites de tempo muito breves. O segundo fator têm cargas altas nos três subtestes restantes, envolvendo problemas que requerem raciocínio baseado em estímulos *gráficos* ou *pictóricos*. Este padrão de matriz fatorial coincide com os derivados das análises fatoriais exploratória e confirmatória dos dados da amostra de padronização do Beta III, que são apresentados em seu manual. Os Fatores 1 e 2 são denominados "Velocidade de Processamento" e "Raciocínio Não-Verbal", respectivamente (Kellog e Morton, 1999).

A análise fatorial foi desenvolvida por psicólogos na tentativa de investigar as bases das inter-relações entre escores de teste, e entre escores de vários tipos de testes de habilidade em particular. No entanto, as técnicas de AF foram prontamente aplicadas a dados de testes de personalidade e descrições de traços de persona-

lidade. A história da AF, tanto na área das habilidades como na da personalidade, sempre esteve carregada de controvérsias a respeito da adequação de seus vários métodos e das extrapolações que podem ou não ser feitas a partir de seus resultados (Cowles, 2001, Capítulo 11).

No campo das habilidades cognitivas, parte da controvérsia se concentrou no fator geral da habilidade mental, ou *g* (originalmente postulado por Charles Spearman), especialmente em questões relacionadas à sua significância e hereditariedade (Jensen, 1998). Teorizações adicionais e pesquisas básicas neste campo trataram principalmente de questões pertinentes à natureza, número e organização dos traços intelectuais. Uma excelente compilação de boa parte da literatura sobre a análise fatorial das habilidades cognitivas humanas, juntamente com uma teoria hierárquica amplamente aceita da organização dos traços cognitivos, pode ser encontrada no livro de John Carroll (1993) sobre este tema.

No campo da avaliação da personalidade, a análise fatorial foi aplicada à tarefa de identificar e medir as principais dimensões necessárias para uma descrição abrangente da personalidade, uma questão sobre a qual também tem havido grandes discordâncias. Dentro desta área, duas tradições separadas de pesquisa em análise fatorial surgiram independentemente. Uma delas centrou-se desde o início no uso de dados de questionários de personalidade. A outra – conhecida como a tradição *léxica* – começou reduzindo a miríade de palavras usadas para descrever os atributos da personalidade a um número mais manejável por meio da combinação de sinônimos. Isso foi seguido por uma tentativa de identificar as dimensões primárias da personalidade pelas intercorrelações e análises fatoriais de classificações em vários traços atribuídos a grupos heterogêneos de indivíduos por seus associados, bem como por dados de questionários de auto-relato. Mais recentemente, as pesquisas em ambas as tradições se aproximaram e chegaram a um certo grau de consenso. O modelo prevalente se concentra no uso de um padrão hierárquico de análise para simplificar a coleta de dados de graus variáveis de generalidade pertinentes ao funcionamento da personalidade, e passou a ser conhecido como *modelo de cinco fatores* (*MCF*; Carroll, 2002; Costa e McCrae, 1992; Digman, 1990; Wiggins e Pincus, 1992).

Apesar de seus problemas e limitações, a tradição da análise fatorial tem sido extraordinariamente fértil para a testagem psicológica e, de modo mais geral, para a teorização psicológica. A longevidade destes métodos, bem como seu contínuo refinamento, criaram um rico acervo de ferramentas e dados a partir dos quais continuamos a ampliar nossa compreensão dos traços e testes psicológicos. O quadro Consulta Rápida 5.5 apresenta alguns dos benefícios que podem ser derivados da AF, bem como suas principais limitações.

Não esqueça

Fatores não são entidades "reais", embora muitas vezes sejam discutidos como se fossem. Eles são simplesmente constructos ou variáveis latentes que podem ser inferidos a partir dos padrões de covariância revelados por análises estatísticas.

CONSULTA RÁPIDA 5.5

Parte I: Benefícios da análise fatorial

Validação de constructos: Agrupando grande número de medidas e examinando os fatores que parecem ser responsáveis pela variância compartilhada por elas, podemos aprender mais a respeito da composição das tarefas que compõem os testes psicológicos e da organização dos traços, em termos de sua generalidade e especificidade.

Aplicação prática: Quando uma bateria é composta de um grande número de testes, os resultados da análise fatorial fornecem um meio de simplificar a interpretação e o relato dos escores dos subtestes. Isto é feito por meio de escores fatoriais, que essencialmente são índices que agregam os escores de subtestes a um número menor de categorias coesas de constructos derivadas da análise fatorial.

Parte II: Limitações da análise fatorial

A interpelação isolada dos resultados de qualquer estudo de análise fatorial não pode ir além dos dados usados na análise, seja em termos do que está sendo medido ou de sua generalizabilidade para outras populações.

O que está sendo medido? Tanto o manual do Beta III (um teste que pretende medir a habilidade intelectual não-verbal) como a análise mostrada na Tabela 5.3 sugerem que os cinco subtestes do Beta III podem ser configurados em dois grupos com uma boa quantidade de variância comum. O exame das tarefas envolvidas nos cinco subtestes confirma que os dois fatores são não-verbais. No entanto, esses dados não podem revelar se ou em que grau esses fatores capturam os aspectos essenciais da habilidade intelectual não-verbal.

Generalizabilidade dos resultados da análise fatorial: os dados correlacionais derivados de 95 estudantes universitários (Urbina e Ringby, 2001) apresentados na Tabela 5.3 produzem resultados semelhantes aos obtidos pelo grupo de padronização do Beta III, uma amostra muito maior (N = 1260) e bem mais representativa da população americana. Embora esta convergência de resultados corrobore a estrutura de fatores obtida em ambas as investigações, ela deixa em aberto a questão de se esta estrutura pode ser generalizada para outras populações, como pessoas de diferentes culturas, indivíduos com deficiências auditivas ou visuais não corrigidas ou qualquer outro grupo cujo histórico de experiências difere significativamente do das amostras usadas nas duas análises em questão.

Técnicas de modelagem de equação estrutural

A análise fatorial exploratória é apenas um dos diversos tipos de técnicas estatísticas multivariadas que permitem aos investigadores examinar as relações entre múltiplas medidas para tentar determinar os constructos subjacentes que explicam a variabilidade observada. A maior disponibilidade de computadores e *softwares* poderosos nas últimas décadas aumentou muito a facilidade com que as técnicas de análise fatorial – e outros métodos sofisticados de análise de dados correlacionais multivariados, tais como análises regressivas múltiplas – podem ser usadas para investigar variáveis latentes (isto é, constructos) e as possíveis ligações causais diretas e indiretas ou rotas de influência entre elas.

Um conjunto de procedimentos de rápida evolução que pode ser usado para testar a plausibilidade de hipóteses de inter-relações entre constructos, bem como as relações entre os constructos e as medidas usadas para avaliá-los, é conhecido como *modelagem de equação estrutural* (MEE). A idéia essencial da MEE é criar um

ou mais modelos – baseados em teorias, achados prévios ou análises exploratórias anteriores – das relações entre um conjunto de constructos ou variáveis latentes e comparar as estruturas ou matrizes de co-variância implicadas pelos modelos com as matrizes de co-variância efetivamente obtidas com um novo conjunto de dados. Em outras palavras, as relações obtidas com dados empíricos em variáveis que avaliam os vários constructos (fatores ou variáveis latentes) são comparadas às relações preditas pelos modelos. A correspondência entre os dados e os modelos é avaliada por estatísticas, apropriadamente denominadas estatísticas de *qualidade de ajuste*. A MEE oferece diversas vantagens em relação à análise regressiva tradicional, que derivam basicamente de duas características: (a) a MEE se baseia na análise de *estruturas de covariância* (isto é, padrões de comparação entre variáveis latentes ou constructos) que podem representar as influências diretas e indiretas de variáveis umas sobre as outras e (b) a MEE usa tipicamente múltiplos indicadores para as variáveis dependentes e independentes nos modelos, fornecendo assim um modo de explicar o erro de mensuração em todas as variáveis observadas. Os leitores que desejarem se aprofundar no tópico da MEE e em técnicas relacionadas, como análise de trajetória, podem consultar uma ou mais das fontes sugeridas no quadro Consulta Rápida 5.6.

No que diz respeito à validação de testes psicológicos, as técnicas de MEE são usadas para a exploração sistemática de constructos e teorias psicológicas por meio de pesquisas que empregam a testagem como um método de coleta de dados para um ou mais indicadores de um modelo. A MEE pode fornecer evidências confirmatórias da fidedignidade dos escores de teste, bem como de sua utilidade como medida de um ou mais constructos de um modelo. No entanto, presentemente, a aplicação mais extensa das técnicas de MEE na validação de escores de teste se dá através da análise fatorial confirmatória.

A *análise fatorial confirmatória* (AFC), mencionada rapidamente em uma seção anterior, envolve a especificação *a priori* de um ou mais modelos das relações

CONSULTA RÁPIDA 5.6 Fontes de informação sobre a modelagem de equação estrutural (MEE)

- Para uma boa introdução à MEEa que não pressupõe o conhecimento de métodos estatísticos além da análise regressiva, ver:
  Raykov, T., Marcoulides, G.A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Erlbaum.
- Uma apresentação mais avançada da MEE que inclui contribuições de muitas das principais autoridades nesta metodologia pode ser encontrada em:
  Bollen, K.A., Long, J.S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
- A Internet é uma excelente fonte de informações sobre muitos tópicos, incluindo técnicas de MEE que estão sendo usadas em vários campos. Um dos melhores pontos de partida é o site criado e mantido por Ed Rigdon, professor do Departamento de *Marketing* da Georgia State University:
  http://www.gsu.edu/~mkteer/sem.html

entre escores de teste e os fatores ou constructos que eles devem avaliar. Na AFC, assim como em todas as outras técnicas de MEE, a direção e a força das inter-relações estimadas por vários modelos são testadas em comparação com resultados obtidos de dados reais em termos de qualidade de ajuste. Estas análises foram facilitadas pelo desenvolvimento de programas de computador – como o LISREL (Jöreskog e Sörbom, 1993) – que geram valores para os modelos hipotéticos que podem então ser testados em comparação com os dados reais.

Os exemplos desse tipo de trabalho estão se tornando cada vez mais abundantes tanto na literatura psicológica quanto nas seções de validade dos manuais de teste. As análises fatoriais confirmatórias conduzidas com os dados da amostra de padronização da WAIS-III (Psychological Corporation, 1997, p.106-110) são típicas dos estudos que buscam fornecer evidências de validade para escores de testes psicológicos. Nessas análises, quatro modelos estruturais possíveis – um de dois, um de três, um de quatro e um de cinco fatores – foram avaliados sucessivamente e comparados a um modelo geral de um fator para determinar qual deles oferecia o melhor ajuste para os dados da amostra total e na maior parte das faixas etárias do grupo normativo da WAIS-III. Os resultados da AFC indicaram que o modelo de quatro fatores oferecia a melhor solução geral e confirmaram os padrões obtidos anteriormente com análises fatoriais exploratórias dos mesmos dados. Esses resultados, por sua vez, foram usados como base para determinar a composição dos quatro escores-índice (Compreensão Verbal, Organização Perceptiva, Memória de Trabalho e Velocidade de Processamento) que podem servir para organizar os resultados de 11 dos 14 subtestes da WAIS-III em domínios separados do funcionamento cognitivo.

Em contraste com esse tipo de AFC conduzida com os dados da WAIS-III, outros estudos de AFC envolvem um trabalho mais básico voltado para o esclarecimento da organização dos traços cognitivos e específicos de personalidade. Um exemplo desse tipo de estudo pode ser encontrado na descrição de Gustafsson (2002) de sua reanálise de dados coletados por Holzinger e Swineford nos anos 30 com um grupo de estudantes de 7ª e 8ª séries (N = 301) que foi testado com uma bateria de 24 testes delineados para explorar habilidades em cinco áreas amplas (expressão, espaço, memória, velocidade e dedução matemática). Usando dois modelos diferentes de AFC, Gustafsson encontrou suporte para a hipótese de sobreposição entre os fatores *G* (inteligência geral) e *Gf* (inteligência fluida ou capacidade de raciocínio), que havia sido sugerida pelas análises feitas nos anos de 1930. Gustafsson também usou contrastes nos padrões de resultados da análise fatorial original feita nos anos de 1930 e suas AFCs contemporâneas para ilustrar as implicações significativas de várias abordagens de mensuração, bem como da composição da amostra, para a validação de constructo dos escores de testes de habilidade.

A análise fatorial confirmatória e outras técnicas de modelagem estrutural ainda estão evoluindo e estão muito longe de oferecer qualquer conclusão definitiva. No entanto, a fusão de modelos teóricos, observações empíricas e análises estatísticas sofisticadas que caracteriza estas técnicas é muito promissora em termos do avanço de nossa compreensão das medidas que usamos e das relações entre os escores de teste e os constructos que eles devem avaliar.

Evidências de validade baseadas nas relações entre escores de teste e critérios

Se os objetivos da testagem psicológica se limitassem simplesmente à descrição do desempenho dos testandos em termos dos referenciais discutidos no Capítulo 3 ou ao aumento de nossa compreensão dos constructos psicológicos e suas inter-relações, as fontes de evidências já discutidas poderiam ser suficientes. No entanto, a interpretação válida dos escores de teste muitas vezes acarreta a aplicação do sentido inerente aos escores – seja ele baseado em normas, conteúdo do teste, processos de resposta, padrões estabelecidos de convergência e divergência ou qualquer combinação dessas fontes de evidências às inferências pragmáticas necessárias para a tomada de decisões a respeito de pessoas. Nesses casos, as evidências de validade precisam contemplar a significância que os escores podem ter em questões que vão além deles mesmos ou em campos que estão fora da esfera de ação direta do teste. Em outras palavras, é preciso demonstrar que os escores de teste se correlacionam com os vários critérios usados na tomada de decisões e predições.

Alguns fatos essenciais a respeito de critérios

O *Merriam-Webster's collegiate dictionary* (1995) define *critério* como "um padrão no qual um julgamento ou decisão pode ser baseado" ou "uma marca ou traço característico". Embora a forma plural, critérios, seja usada com freqüência como singular, no presente contexto é necessário aplicar as duas formas da palavra apropriadamente, porque ambas são centrais para nossos objetivos. Um critério também pode ser definido, de forma menos estrita, como aquilo que *realmente* queremos saber. Esta última definição, embora menos formal do que a do dicionário, enfatiza o contraste entre o que os escores de teste nos dizem e as razões práticas por que usamos os testes.

Para os testes psicológicos que são usados em julgamentos ou decisões a respeito de pessoas, a evidência de uma relação entre os escores e medidas de critério é uma base indispensável, porém não necessariamente suficiente, para a avaliação da validade. *Medidas de critério* são índices dos critérios, que os testes pretendem avaliar ou predizer, coletados independentemente do teste em questão. O quadro Consulta Rápida 5.7 fornece uma lista dos tipos de critérios tipicamente usados na validação de escores de teste. Uma vez que a natureza dos critérios depende das perguntas que se quer responder com a ajuda dos testes, segue-se que os procedimentos de validação baseados parcial ou inteiramente nas relações entre escores e medidas de critério devem produzir evidências de uma ligação entre os *preditores* (escores de teste) e os critérios.

As medidas ou estimativas de critério podem ser naturalmente *dicotômicas* (p. ex., colação de grau *versus* abandono de curso) ou artificialmente *dicotomizadas* (p. ex., sucesso *versus* fracasso); *politômicas* (p. ex., diagnósticos de transtorno de ansiedade *versus* transtorno de humor *versus* transtorno dissociativo, ou preferên-

CONSULTA RÁPIDA 5.7

Critérios típicos usados na validação de escores de teste

Embora haja um número quase infinito de medidas de critério que podem ser empregadas na validação de escores de teste, dependendo dos objetivos da testagem, as categorias mais freqüentes são as seguintes:

- *Índices de realização acadêmica ou desempenho em treinamento especializado*, tais como, notas escolares, históricos de graduação, menções honrosas, prêmios ou demonstrações de competência em áreas de treinamento por desempenho bem-sucedido (p. ex., em piano, trabalhos mecânicos, pilotagem, programação de computadores, exames de ordem ou provas de certificação).
- *Índices de desempenho no trabalho*, tais como históricos de vendas, históricos de produção, promoções, aumentos de salário, estabilidade em empregos que exigem competência, ausência de acidentes de trabalho ou avaliação por supervisores, pares, estudantes, empregados, clientes, etc.
- *Afiliação a grupos contrastados*, baseada em diagnósticos psiquiátricos, *status* ocupacional, realização educacional ou qualquer outra variável relevante.
- *Avaliações* de comportamento ou de traços de personalidade feitas por observadores independentes, parentes, pares ou quaisquer outros associados que tenham bases suficientes para fornecê-las.
- *Escores em outros testes relevantes.*

cia por ocupações artísticas *versus* científicas *versus* literárias); ou *contínuas* (p. ex., média de notas, número de unidades vendidas, escores em um inventário de depressão, etc). Enquanto a natureza dos critérios depende das decisões ou predições a serem feitas com a ajuda dos escores de teste, os métodos usados para estabelecer as relações entre escores e critérios variam dependendo tanto das características formais dos escores como das medidas de critério. Em geral, quando a medida de critério é expressa de forma dicotômica (p. ex., sucesso *versus* fracasso) ou em termos de um sistema categórico (p. ex., afiliação a grupos contrastados), a validade dos escores de teste é avaliada em termos de *taxas de acerto*. As taxas de acerto tipicamente indicam a percentagem de decisões ou classificações corretas feitas com o uso dos escores de teste, embora diferenças médias e índices de correlação adequados também possam ser usados. Quando as medidas de critério são contínuas (p. ex., escores em testes de realização, notas, avaliações, etc) as principais ferramentas usadas para indicar a extensão da relação entre os escores de teste e a medida de critério são coeficientes de correlação. No entanto, se um determinado valor em um critério contínuo, como uma nota média de 2,0, é usado como ponto de corte para determinar um resultado específico, como a colação de grau universitário, os escores no teste preditor também podem ser avaliados em termos de diferenciações ou não entre os testandos que satisfazem ou excedem o critério de corte e aqueles que não o fazem.

A história de testagem psicológica nas últimas décadas reflete não apenas uma evolução da compreensão da natureza e das limitações dos testes e seus escores, mas também a maior apreciação da significância da complexidade das medidas de critério (James, 1973; Tenopyr, 1986; Wallace, 1965). Como resultado, com

raras exceções, a noção de que existe algo como "um critério", em relação ao qual um teste pode ser validado, deixou de ser defensável tanto quanto a proposição de que a validade de um teste pode ser determinada em termos de tudo ou nada. Em vez disso, os seguintes fatos a respeito dos critérios agora são compreendidos de maneira geral:

1. Na maioria dos estudos de validação, existem muitos índices possíveis (quantitativos e qualitativos) que podem ser usados como medidas de critério, incluindo escores de testes que não aqueles submetidos à validação. Por isso, uma grande atenção deve ser dada à seleção dos critérios e medidas de critério.
2. Algumas medidas de critério são mais fidedignas e válidas do que outras. Por isso, a fidedignidade e a validade das medidas de critério precisam ser avaliadas assim como as dos escores de teste.
3. Alguns critérios são mais complexos do que outros. Como resultado, pode haver ou não uma correlação entre medidas de critério, especialmente quando os critérios são multifacetados.
4. Alguns critérios podem ser avaliados no momento da testagem; outros evoluem com o tempo. Isso significa que pode haver ou não correlações substanciais entre medidas de critério que estão disponíveis logo após a testagem e critérios mais distantes que podem ser avaliados somente ao longo de um período maior de tempo.
5. As relações entre escores de teste e medidas de critério podem ou não se generalizar para outros grupos, contextos ou períodos de tempo. Por isso, as evidências de validade relacionadas ao critério precisam ser demonstradas novamente para populações que diferem das amostras originais de validação em aspectos que podem afetar a relação entre escores e critérios, bem como entre vários contextos e momentos.
6. A força ou qualidade das evidências de validade em relação à avaliação ou predição de um critério é uma função das características do teste e das medidas de critério empregadas. Se as medidas de critério não são fidedignas ou são arbitrárias, os índices da validade dos escores serão enfraquecidos, independentemente da qualidade do teste usado para avaliar ou predizer os critérios.

Procedimentos de validação relacionados ao critério

As decisões relacionadas ao critério para as quais os escores de teste têm a possibilidade de ser úteis podem ser classificadas em dois tipos básicos: (a) aquelas que envolvem a determinação do *status* atual de uma pessoa e (b) aquelas que envolvem a predição de um desempenho ou comportamento futuro. Em certo sentido, esta dicotomia é artificial porque, quer precisemos saber algo a respeito do *status* atual de uma pessoa ou de seu desempenho futuro, a única informação que os escores de teste podem transmitir deriva de seu comportamento atual – isto é, do desempenho do testando no momento da testagem. Mesmo assim, os procedimen-

tos de validação relacionados ao critério freqüentemente são categorizados como concorrentes ou preditivos, dependendo das medidas empregadas bem como de seus objetivos primários.

Validação concorrente e preditiva

As evidências de *validação concorrente* são coletadas quando os índices dos critérios que os escores de teste pretendem avaliar estão disponíveis no momento em que os estudos de validação são conduzidos. Falando estritamente, a validação concorrente é apropriada para escores de testes que serão empregados para determinar o *status* atual de uma pessoa em relação a algum esquema classificatório, como categorias diagnósticas ou níveis de desempenho. As evidências de *validação preditiva*, por outro lado, são relevantes para escores de teste que serão usados na tomada de decisões baseadas na estimativa de níveis de desempenho ou resultados comportamentais futuros. Idealmente, os procedimentos de validação preditiva requerem que sejam coletados dados sobre a variável preditora (escores de teste) e que se espere que os dados de critério se tornem disponíveis para que os dois conjuntos de dados possam ser correlacionados. Este processo muitas vezes não é prático devido ao elemento temporal envolvido na espera para que os critérios amadureçam e também devido à dificuldade de encontrar amostras adequadas para tais estudos. Como resultado, a validação concorrente costuma ser usada como substituta da validação preditiva, mesmo para testes usados para estimar o desempenho futuro, como admissões em universidades ou seleções para emprego. Nestes casos, o teste em desenvolvimento é administrado a um grupo de pessoas, como estudantes universitários ou empregados, para o qual os dados de critérios já estão disponíveis.

Muitas vezes, a distinção entre dois tipos de procedimento de validação depende do modo como os usuários formulam as perguntas que querem responder com a ajuda do teste. O quadro Consulta Rápida 5.8 contém exemplos de algumas perguntas e situações de tomada de decisões típicas que podem demandar evidências de validação concorrente ou preditiva, dependendo de como a pergunta é formulada e do referencial de tempo escolhido. Para ilustrar a distinção entre estratégias de validação concorrente e preditiva, um exemplo relativamente simples de cada tipo de estudo será apresentado, seguido por uma discussão das principais questões pertinentes à validação relacionada ao critério.

Exemplo de validação concorrente: o Índice Whitaker de Pensamento Esquizofrênico

Testes que são usados para identificar transtornos psiquiátricos, como esquizofrenia ou depressão, geralmente passam por validação concorrente. Tipicamente, esses estudos empregam duas ou mais amostras de indivíduos que diferem em relação ao seu *status* diagnóstico estabelecido independentemente. Um dos muitos

CONSULTA RÁPIDA 5.8

Relações entre perguntas, decisões e predições que requerem validação relacionada ao critério

Perguntas sobre o *status* atual	Objetivo imediato: decisões	Perguntas implícitas: predições
John X está sofrendo de esquizofrenia, transtorno de pânico, depressão clínica, déficit de atenção ou algum outro transtorno mental?	John X deve receber o tratamento (medicação, psicoterapia, etc) recomendado para o transtorno em questão?	John X vai se beneficiar (um pouco, muito ou nada) do tratamento recomendado?
Mary Y está sujeita a impulsos suicidas (ou homicidas) que pode não ser capaz de controlar sozinha?	Mary Y deve continuar hospitalizada (ou presa)?	Mary Y vai tentar se matar (ou matar outra pessoa) se deixada em liberdade?
Joe Z é superdotado (ou sofre de retardo mental severo)?	Joe Z deve ser admitido em um programa de educação especial para indivíduos superdotados (ou com retardo mental)?	Joe Z vai se beneficiar (um pouco, muito ou nada) de um programa de educação especial?
Tom P é honesto e confiável (ou motivado para o trabalho com vendas)?	Tom P deve ser contratado para trabalhar como caixa (ou vendedor)?	Tom P vai ser um caixa consciencioso (ou um vendedor de sucesso), ou vai roubar dinheiro (ou fazer poucas vendas)?
Jane Q é capaz de realizar trabalhos de nível universitário (ou pilotar um avião)?	Jane Q deve ser admitida na universidade (ou receber uma licença para pilotar aviões)?	Jane Q será capaz de terminar a faculdade com média de notas altas o bastante para se formar (ou será capaz de pilotar um avião sem causar acidentes)?

instrumentos cujos escores são validados desta maneira é o Índice Whitaker de Pensamento Esquizofrênico [*Whitaker Index of Schizophrenic Thinking* (WIST; Whitaker, 1980)]. O WIST foi delineado para identificar o tipo de comprometimento do pensamento que costuma acompanhar as síndromes esquizofrênicas. Ambas as suas formas (A e B) consistem em 25 itens de múltipla escolha.

Na padronização do WIST, Whitaker usou amostras de pacientes esquizofrênicos agudos e crônicos (E), bem como três grupos de não-esquizofrênicos (NE), para derivar escores de corte que iriam diferenciar otimamente os indivíduos E dos NE. Os escores de corte estabelecidos com os grupos de padronização discriminavam entre grupos E e NE com 80% de eficiência para a Forma A e 76% de eficiência para a Forma B. Portanto, dependendo da forma, o índice de corte resultava em 20 a 24% de decisões incorretas. Com a Forma A, as decisões incorretas do tipo *falso-negativo* – aquelas nas quais sujeitos E eram classificados como NE pelo índice –

eram muito mais altas (33%) do que as decisões tipo *falso-positivo*, com as quais sujeitos NE eram classificados como E (10%). O mesmo padrão (38% de falsos-negativos *versus* 13% de falsos-positivos) foram obtidos com a Forma B. Ver o Capítulo 7 para mais explicações sobre a terminologia usada para designar decisões incorretas (p. ex., decisões tipo falso-positivo e falso-negativo).

Evidências adicionais de validação para o WIST incluem estudos feitos no México e Espanha com traduções do instrumento para o espanhol. Esses estudos mostram algum suporte para a capacidade do WIST de detectar o comprometimento do pensamento associado à esquizofrenia, ainda que com diferentes taxas de eficiência e diferentes escores de corte.

A taxa de acerto obtida quando os escores de corte do WIST são usados para discriminar sujeitos E de NE é bastante típica para testes desse tipo. Devido às taxas de erro relativamente altas que podem produzir, os escores de corte do WIST nunca devem ser usados como o único veículo para o estabelecimento de um diagnóstico de esquizofrenia, não mais do que qualquer outro indicador isolado. No entanto, dependendo do local e do contexto da testagem, o WIST pode se mostrar útil como parte de uma bateria de triagem ou como índice de mudança na sintomatologia de pacientes diagnosticados com esquizofrenia. Embora as discriminações feitas com o uso dos escores de corte do WIST estejam longe de ser perfeitas, elas ainda assim podem ser úteis para investigar a possibilidade de que uma pessoa sofra de comprometimento do pensamento. Para uma revisão do WIST, ver Flanagan (1992).

Advertência

- Os resultados de testes psicológicos, assim como os resultados de testes médicos e de muitos outros campos, não alcançam 100% de precisão.
- Devido à natureza menos do que perfeita de todos os indicadores diagnósticos, os psicólogos envolvidos em avaliações clínicas – bem como especialistas em diagnósticos de outros campos – jamais devem se valer de um único indicador.

Validação preditiva: Um exemplo hipotético

Testes que são usados para predizer o desempenho no contexto ocupacional e educacional requerem estratégias de validação voltadas para a predição. O delineamento ideal para um estudo de validação preditiva nesses campos deve envolver os seguintes passos: (a) testar um grupo não-selecionado de candidatos com um teste de habilidade ou bateria de testes; (b) contratá-los ou admiti-los sem considerar seus escores no teste; (c) esperar até que medidas de critério de desempenho no trabalho ou estudos estejam disponíveis; (d) obter correlações entre os escores no teste antes da contratação ou admissão e as medidas de critério e (e) usar os dados correlacionais para derivar uma equação regressiva que estime ou prediga o desempenho de futuros candidatos no critério. Por motivos óbvios, a maioria dos empregadores e administradores escolares não está disposta a contratar ou aceitar todos os candidatos – especialmente quando estes excedem o número de vagas

disponíveis – para conduzir um estudo de validação. Mesmo assim, para simplificar e ilustrar o modo como as predições podem ser feitas a partir dos escores de teste, um exemplo hipotético simples deste tipo de estudo será descrito.

Para fornecer um contexto para este exemplo – de modo a garantir um nível lucrativo de produção –, digamos que o proprietário de uma pequena fábrica queira contratar pessoas que sejam capazes de processar pelo menos 50 peças/hora na linha de montagem. Como não há outros requisitos para esse emprego, além daquele de comparecer ao trabalho na hora correta, o critério a ser predito é simplesmente o número de peças produzidas por hora na linha de montagem. A produção da linha de montagem é primariamente uma função da velocidade e da precisão dos trabalhadores em tarefas manuais. Por isso, um teste de destreza manual é selecionado como preditor. A Tabela 5.4 mostra os *dados bivariados* para 10 candidatos hipotéticos ao emprego na linha de produção que se submetem ao *teste de destreza manual*. São contratados e treinados e têm sua *produção por hora na linha de montagem* medida em repetidas ocasiões, com cálculo da média para produzir uma medida de critério viável.

Como foi discutido no Capítulo 2, quando duas variáveis exibem uma relação linear e uma forte correlação entre si, é possível predizer uma baseada no conhecimento da outra aplicando-se o modelo da regressão linear. A Figura 5.1 mostra o diagrama de dispersão para os dados bivariados da Tabela 5.4 e indica que a relação entre os escores no teste de destreza manual e a produção por hora na linha de montagem é forte, positiva e linear.

Nesse caso, é necessário resolver uma equação de regressão linear para predizer o critério para a produção na linha de montagem (a variável Y) a partir dos escores no teste de destreza manual (a variável X). Esta equação expressa a relação entre X e Y e contém os dois principais componentes necessários para se traçar a

Tabela 5.4 Dados para o exemplo de validação preditiva

Candidato	Escore no teste (X)	Produção (Y)	$X - M_x$ (x)	$Y - M_y$ (y)	x^2	y^2	xy
1	18	56	5	6	25	36	30
2	12	50	-1	0	1	0	0
3	8	47	-5	-3	25	9	15
4	20	52	7	2	49	4	14
5	14	52	1	2	1	4	2
6	5	42	-8	-8	64	64	64
7	10	48	-3	-2	9	4	6
8	12	49	-1	-1	1	1	1
9	16	50	3	0	9	0	0
10	15	54	2	4	4	16	8
Soma 130	500	0	0	188	138	140	
Média 13	50						

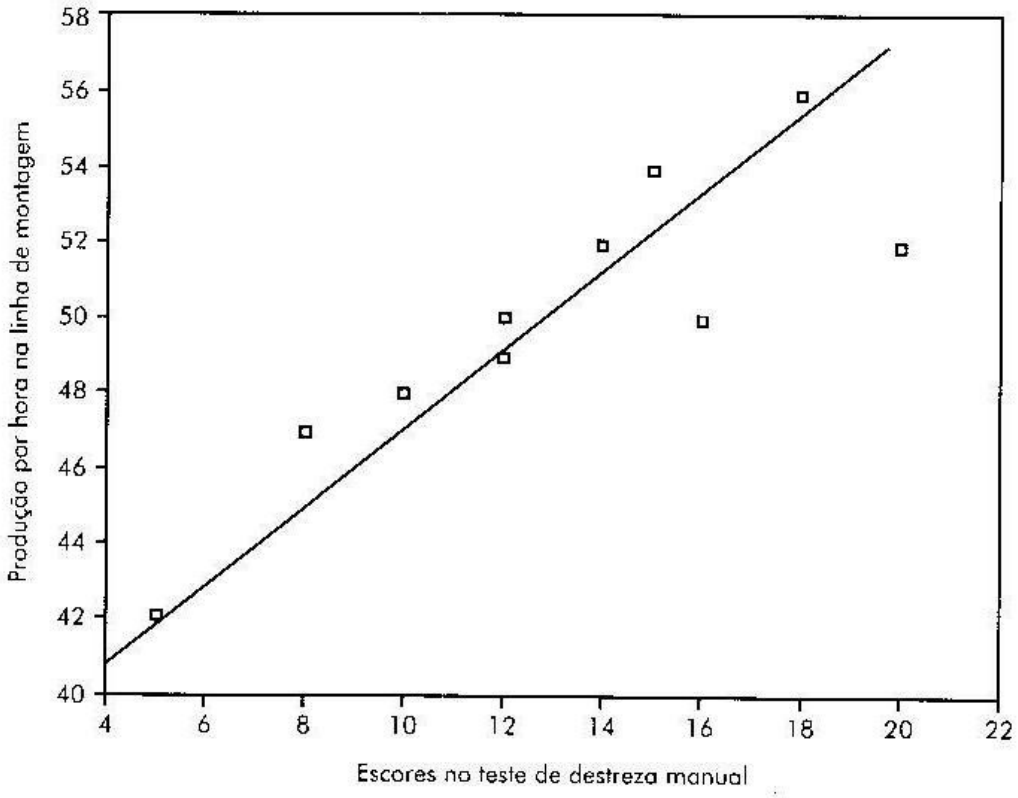

Figura 5.1 Diagrama de dispersão de dados e linha de regressão para o exemplo de validação preditiva.

linha de regressão mostrada na Figura 5.1. A linha de regressão é a linha que melhor se ajusta aos dados bivariados, uma vez que minimiza erros na predição de Y a partir de X. Os dois componentes cruciais da equação de regressão linear são (a) o intercepto Y, que é o ponto no qual a linha encontra o eixo vertical que representa a variável Y e (b) a declividade da linha, que é a razão de mudança na variável Y para cada unidade de mudança na variável X. Esses valores, bem como o coeficiente de correlação para X e Y, são calculados a partir dos dados bivariados. As análises necessárias são apresentadas na Tabela 5.5. Esta tabela mostra que o *r* de Pearson (r_{xy}) é 0,87 (significativo no nível 0,001). O *coeficiente de determinação*, obtido calculando-se r_{xy} é 0,755. Isto indica que 75,5% da variação em NY (produção na linha de montagem) estão associados com a variante em X (escores no teste de destreza). Ambos os coeficientes confirmam que existe uma forte relação entre X e Y, tornando possível predizer a medida de critério a partir dos escores de teste com precisão substancial. O proprietário da fábrica de nosso exemplo pode usar essas informações para estimar a produção na linha de montagem dos candidatos subseqüentes administrando o teste de destreza manual e inserindo seus escores na equação regressiva, como se segue.

Suponhamos que a 11ª pessoa a se candidatar ao emprego (depois dos 10 que foram usados no estudo hipotético de validação) obtenha um escore de 17 no teste

Tabela 5.5 Análise dos dados de validação preditiva

Estatística descritiva	Número de observações (*N*)	Média	Desvio padrão (*DP*)
X = preditor (escores de teste)	10	13	4,57
Y = critério (produção)	10	50	3,91

Pearson r $$r_{xy} = \frac{\sum xy}{(N-1)(DPx)(DPy)} = \frac{140}{(9)(4,57)(3,91)} = 0,87$$

Coeficiente de determinação $r_{xy}^2 = 0,755$

Equação de regressão linear $Y' = a_{yx} + b_{yx}(X)$

Dados do exemplo $a_{yx} = 40,32$
$b_{yx} = 0,745$

Equação regressiva para predizer a produção da linha de montagem com base nos escores do teste de destreza

$$\text{Produção predita} = Y' = 40,32 + 0,745(X)$$

"Onde Y' = escore predito no critério; $a_{yx} = My - b_{yx}(M_x)$ = intercepto da linha de regressão; $b_{yx} = (\sum_{xy})/(\sum x^2)$ = declividade da linha de regressão; e X = escore no preditor (escore no teste de destreza manual).

de destreza manual. Usando os coeficientes mostrados na Tabela 5.5 para resolver a equação regressiva Y' = 40,32 + 0,745 (17), o proprietário da fábrica ficaria sabendo que a produção por hora na linha de montagem estimada para o 11º candidato é de 53 peças, ou 3 acima do número mínimo desejado.

As correlações entre os escores de teste e as medidas de critério (r_{xy}) geralmente são denominadas *coeficientes de validade*. Se r_{xy} fosse 1,00, indicando uma correlação perfeita entre os escores no teste e o critério, uma predição perfeita seria possível. Embora o r_{xy} de 0,87 obtido com os dados de nosso exemplo seja extremamente alto – muito mais alto do que os coeficientes de validade preditiva típicos obtidos com dados reais –, ele não é perfeito. Por isso, é certo que haverá algum erro nas estimativas de critério feitas com o uso dos escores de teste. Este erro é medido pelo *erro padrão de estimativa* (EP_{est}), uma estatística que expressa, na escala usada para a medida de critério, o erro em predições que se baseiam em correlações imperfeitas.

Fórmulas para o EP_{est} e para um termo de correção aplicado ao EP_{est} devido a tamanhos pequenos de amostra são apresentadas no quadro Consulta Rápida 5.9, juntamente com os resultados para os dados de nosso exemplo. A interpretação do EP_{est} pressupõe (a) que o escore predito no critério, ou Y', é o valor médio em uma distribuição normal hipotética de todos os escores possíveis no critério para o candidato em questão e (b) que o EP_{est} é o desvio padrão desta distribuição. Essas premissas nos permitem atribuir um nível de probabilidade às predições feitas a partir de escores de teste baseados nas áreas da curva normal do Apêndice C. Para isso, calcula-se a amplitude contida dentro de Y' ± EP_{est} (z), em que z é o valor que corresponde ao nível de probabilidade desejado. Para o 11º candidato de nosso exemplo, o escore de 17 no teste de destreza resultou em um Y' predito de 53. O EP_{est} para os escores de teste, calculado no quadro Consulta Rápida 5.9, é de 2,05

Não esqueça

Avaliar a adequação de coeficientes de validade de várias magnitudes é uma questão relativa que depende dos objetivos para os quais os escores de teste serão usados. De modo geral, existem dois modos principais de abordar a questão:

1. Quando a validade é expressa na forma de um coeficiente de correlação (r_{xy}), a proporção de variância compartilhada pelo preditor (X) e o critério (Y) costumam ser estimados elevando-se r_{xy} ao quadrado e obtendo-se o *coeficiente de determinação*, o r^2_{xy}. Um coeficiente de determinação, discutido no Capítulo 2, expressa a proporção da variância em Y que está associada à variância em X. Por exemplo, o coeficiente de validade preditiva de .87 obtido no exemplo apresentado na Tabela 5.5 indica que os escores no teste de destreza manual usados como preditores podem explicar aproximadamente 76% dos variantes (0,87 x 0,87 = 0,76) no critério de produção na linha de montagem.
2. Os coeficientes de validade também podem ser avaliados em termos da *validade incremental* de um teste – isto é, o grau em que o uso de um teste (ou qualquer outro instrumento) opera um aumento na eficiência das decisões tomadas em uma dada situação comparado à validade de outros métodos. Este aspecto da validade diz respeito à utilidade dos testes e será mais discutido no Capítulo 7.

Coeficientes na casa de 0,20 e 0,30 não são incomuns nos estudos de validade preditiva. Como regra geral, coeficientes de 0,40 ou mais são considerados aceitáveis. Embora alguns desses números possam parecer baixos, eles devem ser vistos no contexto da natureza multideterminada do desempenho na maioria das atividades e devem ser pesados à luz da eficiência preditiva dos métodos alternativos de tomada de decisões de seleção.

CONSULTA RÁPIDA 5.9

Erro padrão de estimativa (EP_{est})

$$EP_{est} = DP_y \sqrt{1 - r^2_{xy}}$$

EP_{est} para o exemplo de validação preditiva:

$$EP_{est} = 3{,}91 \sqrt{1 - 0{,}755} = 3{,}91(0{,}495) = 1{,}935$$

EP_{est} corrigido para tamanho de amostra:

$$EP_{est}\text{ corrigido} = EP_{est} \sqrt{\frac{N-1}{N-2}} = 1{,}935 \sqrt{\frac{10-1}{10-2}} = 1{,}935(1{,}06) = 2{,}05$$

(ou 2). Com isso, podemos predizer que as chances de que este candidato vá produzir entre 51 e 55 peças por hora (isto é, 53 ± 2) na linha de montagem são de 68/100. Para fazer uma predição no nível de confiança de 95% (em vez de 68%), usaríamos um valor z de 1,96, que demarca o nível de significância de 0,05 para o teste bicaudal. Resolvendo a equação $Y' \pm EP_{est}$ (1,96), poderíamos predizer que as chances são de 95/100 de que o candidato em questão venha a produzir entre 49 e 57 peças por hora (53 ± 4).

Questões relativas aos estudos de validação relacionada ao critério

Algumas das dificuldades inerentes aos estudos de validação relacionada ao critério já foram mencionadas neste capítulo em conexão com a discussão da noção de critérios. Considerações adicionais que merecem um exame minucioso por parte dos usuários de instrumentos apoiados neste tipo de evidência de validade serão discutidas a seguir, usando algumas das características dos exemplos descritos anteriormente como ponto de partida. Embora nossa discussão não permita uma exploração extensa destas questões – ou dos vários métodos que podem ser usados para lidar com elas – algumas das mais salientes precisam ser descritas de forma breve.

Características das medidas de critério

Conforme mencionado anteriormente, os critérios com os quais os escores de testes são validados podem diferir muito em termos de sua própria fidedignidade e validade. No caso do WIST, o critério usado para estabelecer a validade dos escores era o *status* diagnóstico (E *versus* NE) dos indivíduos nas amostras de validação. Se a classificação inicial dos sujeitos desses estudos incluísse alguns diagnósticos incorretos (alguns E que eram NE, ou vice-versa), os dados de validade obviamente seriam enfraquecidos. Uma advertência semelhante se aplica a todos os procedimentos de validação que se valem de critérios subjetivos, tais como avaliações ou outros julgamentos qualitativos que são usados para categorizar pessoas em grupos de critério. Além disso, a validade das medidas de critério pode ser prejudicada quando os responsáveis por determinar a posição dos indivíduos das amostras de validação em relação ao critério têm acesso aos escores no teste que é usado como preditor. Este tipo de erro, conhecido como *contaminação de critério*, é facilmente evitado assegurando que professores, supervisores, responsáveis por diagnósticos e outros que atribuem classificações ou fazem julgamentos relacionados a medidas de critério não tenham acesso e não sejam influenciados pelo conhecimento de escores de teste. Em relação às avaliações em si, bem como a outras medidas de critério que dependem do julgamento subjetivo, os criadores de testes precisam fornecer evidências de que os instrumentos e métodos usados para avaliar ou classificar os grupos de critério empregados nos estudos de validação são fidedignos e válidos. Quando o critério consiste em afiliação a um grupo como uma determinada categoria diagnóstica, sua fidedignidade e validade podem ser melhoradas por meio de uma seleção cuidadosa dos sujeitos baseada em evidências de fontes múltiplas e preferencialmente independentes. No que diz respeito a critérios de avaliação, sua fidedignidade e validade também devem ser determinadas. Para esse fim, existe uma extensa literatura dedicada às formas de treinar avaliadores para minimizar vieses e explorar os melhores formatos e metodologias de avaliação (Guion, 1998, Capítulo 12).

Nosso estudo hipotético da validade dos escores do teste de destreza manual – usado para ilustrar os procedimentos de validação preditiva da maneira mais simples possível – contém diversas características pouco realistas. Por exemplo, nele o

critério de produção da linha de montagem podia ser avaliado de forma confiável e precisa contando-se o número de peças produzidas pelos trabalhadores. Nem todos os critérios são tão simples ou fáceis de avaliar. O sucesso em muitas atividades, tais como gerência, medicina, magistério, etc., pode ser julgado a partir de critérios que diferem em termos da fidedignidade com que podem ser avaliados, de quanto controle o trabalhador tem em relação a eles e do valor que a organização lhes confere. Por exemplo, o sucesso de um gerente pode ser medido em termos: (a) da produtividade dos funcionários e (b) da satisfação dos funcionários, entre outras coisas. As habilidades e características pessoais que tornam os gerentes bem-sucedidos em relação a (A) não são necessariamente as mesmas que levam ao sucesso em (B), e em algumas situações esses dois critérios podem até mesmo entrar em conflito um com outro. Além disso, cada critério pode ser avaliado por vários métodos. A produtividade pode ser medida pela quantidade ou qualidade da produção, ou ambas; a satisfação dos funcionários pode ser medida pelas avaliações dos supervisores, pela rotatividade de pessoal em uma unidade, etc. Escores de testes que podem predizer uma faceta do critério podem não ter correlação, ou até mesmo estar correlacionados negativamente, com os que predizem a outra.

Usando múltiplos preditores

A maneira tradicional de lidar com a predição de critérios complexos, como o desempenho no trabalho, tem sido usar uma bateria de testes. Neste contexto, o termo *bateria* se refere a uma combinação de preditores selecionados especialmente para predizer um ou mais critérios. Este sentido contrasta com o uso do termo no contexto clínico ou no de aconselhamento, em que uma bateria geralmente se refere a qualquer grupo de testes administrados a um indivíduo no processo da avaliação psicológica. Os escores nos preditores separados de uma bateria de testes podem ser combinados de várias formas, dependendo dos requisitos da seleção ou problema de classificação. As *técnicas de regressão múltipla*, por exemplo, combinam os escores de cada teste na bateria inserindo-os em uma equação de regressão linear que inclui um peso numérico para cada escore da bateria. As *equações de regressão múltipla* são extensões do método da regressão linear simples apresentado na Tabela 5.5, mas envolvem múltiplos preditores em vez de um único preditor. O peso de cada preditor é diretamente proporcional à sua correlação com o critério e inversamente à sua correlação com os outros preditores da bateria, de modo que os escores de testes com validade mais alta e menor quantidade de sobreposição com outros escores têm peso maior. Um coeficiente de correlação múltipla *(R)* pode então ser calculado para representar a correlação entre a combinação otimamente ponderada dos escores de teste e o critério. Um procedimento alternativo para combinar os escores de uma bateria de testes é por meio da *análise de perfil*. Este método envolve o estabelecimento de um escore de corte para cada preditor – baseado em sua relação com o critério – e resulta na rejeição de todos os candidatos cujos escores fiquem abaixo do mínimo em qualquer um dos testes ou, possivelmente, apenas naqueles que avaliam as habilidades que são consideradas críticas para um desempenho bem-sucedido no critério.

Os dois métodos descritos anteriormente têm algumas desvantagens. As equações de regressão múltipla permitem que deficiências em um ou mais preditores sejam compensadas por um desempenho superior em outros. O peso particular dos preditores nestas equações também deve ser verificado com amostras independentes daquelas usadas para derivar as equações, para ver se a correlação múltipla *(R)* se mantém. A replicação das relações preditor-critério em amostras separadas – um processo conhecido como *validação cruzada* – é necessária porque qualquer coeficiente de correlação, independentemente de sua magnitude, é, em certo grau, dependente de erros específicos da amostra. Alguma redução na magnitude do *R* original, ou *diminuição (shrinkage)*, é esperada na validação cruzada. Quando a diminuição é insignificante, os pesos originais podem ser considerados estáveis o bastante para serem aplicados sem necessidade de outras análises.

Uma das principais desvantagens de usar o método da análise de perfil, juntamente com os escores de corte, é que esse método deixa tipicamente de levar em conta a possível falta de fidedignidade dos escores (ver Quantificando o erro nos escores de teste: O erro de mensuração padrão, no Capítulo 4). Outra dificuldade tem origem no fato de que a existência de múltiplos escores de corte pode resultar na rejeição de um excesso de candidatos, especialmente aqueles com histórico de desvantagem que podem ter pontuação abaixo do corte em um ou mais dos testes de habilidade, mas que poderiam ser capazes de superar essas deficiências pelo treinamento ou por uma forte motivação. Em geral, o uso de escores de corte se justifica apenas em situações em que um déficit em uma habilidade específica teria conseqüências prejudiciais sérias para o desempenho no trabalho. Uma solução possível neste caso é selecionar, a partir dos escores de corte, apenas para testes que avaliam as habilidades que são críticas para o emprego e usar uma equação regressiva para os outros preditores da bateria.

O problema da amplitude restrita nas amostras de validação

Como já mencionado, a outra característica pouco realista do estudo de validação hipotético envolvendo escores em um teste de destreza manual foi a previsão de uma amostra heterogênea de 10 candidatos para os quais as medidas de critério ainda não existiam no momento da testagem. A maioria dos estudos de validade preditiva não procede desta maneira. Eles costumam usar amostras de indivíduos para os quais os dados de critério já estão disponíveis, como empregados ou alunos que já assumiram os empregos ou cursos para os quais o teste em validação será usado. Em outras palavras, a maioria desses estudos usa estratégias de validação concorrente para desenvolver evidências de validade preditiva. O tipo de indivíduos para os quais as medidas de critério já estão disponíveis difere daquele nos quais o teste eventualmente será usado já que eles foram selecionados para o emprego ou admissão acadêmica *e* permaneceram no emprego ou na escola sem ser demitidos ou abandonar o curso. Por isso, podemos quase sempre pressupor que seus escores nos testes preditores que estão sendo avaliados e também nas medidas de critério vão ter uma amplitude mais estreita do que a de uma amostra não-selecionada de candidatos. Podemos recordar do Capítulo 2 que o efeito de uma restrição na ampli-

tude de qualquer uma das variáveis é reduzir o tamanho dos coeficientes de correlação. Por isso, como conseqüência da restrição de amplitude, as correlações entre os escores de teste e os critérios (isto é, os coeficientes de validade) que resultam desses estudos de validação retrospectiva geralmente são menores do que seriam se as amostras fossem retiradas de uma população mais heterogênea, como, por exemplo, *todos* os candidatos aos empregos ou programas acadêmicos em questão.

Generalização da validade

A magnitude dos índices de validade preditiva obtidos para escores de teste depende de quatro elementos básicos: (a) a composição das amostras de validação em termos de tamanho e variabilidade; (b) a natureza e a complexidade do critério a ser predito; (c) as características do teste em si e (d) as interações entre os elementos anteriores. Como cada um desses quatro fatores pode alterar os resultados dos estudos de validação, os usuários de testes precisam considerá-los cuidadosamente antes de pressupor que as evidências publicadas da validade dos escores derivadas de um único estudo serão aplicáveis a seus objetivos e a suas populações de testandos.

Em relação à composição das amostras de validação, já discutimos os problemas pertinentes a medidas de critério e restrição de amplitude. O tamanho pequeno de amostra freqüentemente também é problemático. A maioria dos empregadores não tem grande número de funcionários na mesma categoria de emprego, e os achados de pesquisas de avaliação baseados em amostras pequenas são mais propensos a erros específicos de amostra do que aqueles baseados em amostras grandes. O conjunto de dados bivariados para 10 candidatos a emprego de nosso exemplo hipotético de validade preditiva produziu um coeficiente de validade alto demais de 0,87. Embora seja possível obter correlações de bom tamanho quando os indivíduos que participam de estudos de validação não são selecionados e os critérios são definidos de maneira estrita, como no exemplo, permanece o fato de que os estudos de validação locais conduzidos com amostras pequenas, com amplitude restrita de escores e critérios não fidedignos, produzem estimativas tipicamente baixas e instáveis da correlação entre o preditor e o critério.

Variáveis confundidoras

Uma questão adicional relacionada à composição das amostras em estudos de validação preditiva diz respeito ao possível papel das variáveis confundidoras. Uma *variável confundidora* é qualquer característica de um subgrupo de pessoas de uma amostra que influencia o grau de correlação de duas outras variáveis. Em teoria, praticamente qualquer característica demográfica (p. ex., sexo, etnia, nível de escolaridade, classe social, localização geográfica, etc) ou traço psicológico (interesses, motivação, nível de ansiedade, etc.) pode agir como variável confundidora em estudos de validade preditiva e produzir um efeito interativo que diminua ou aumente a correlação preditor-critério. Para verificar esta possibilidade é necessário conduzir estudos de validação separados ou dividir as amostras de validação em

subgrupos que difiram em relação à variável que se supõe ser confundidora dos coeficientes de validade.

Diferenças consideráveis em favor de brancos e asiáticos comparados a negros ou hispânicos têm sido encontradas consistentemente nos escores médios de testes de habilidades acadêmicas obtidos por pessoas de diferentes grupos raciais ou étnicos. Essas diferenças engendraram a suspeita de que a raça ou etnia pode confundir a validade preditiva dos testes de seleção. Como resultado, muitos estudos conduziram análises separadas da magnitude da correlação preditor-critério e de coeficientes de regressão para brancos, negros, hispânicos e membros de outras minorias raciais ou étnicas. A finalidade desses estudos é determinar se os escores de teste têm validade comparável para diferentes grupos e predizem igualmente bem para todos, ou se têm validade ou viés diferente para diferentes grupos. Neste contexto, o termo *viés* é usado para indicar qualquer diferença sistemática na relação entre preditores e critérios para pessoas que pertencem a diferentes grupos. Diferenças sistemáticas podem se manifestar de duas formas, quais sejam, validade diferencial e predição diferencial.

A *validade diferencial*, no contexto do viés de teste, se refere a diferenças no tamanho das correlações obtidas entre preditores e critérios para membros de diferentes grupos. Diferenças na magnitude dos coeficientes de validade sugerem que os escores de teste predizem mais precisamente para membros do grupo com coeficiente maior. Evidências gráficas da validade diferencial são vistas quando a declividade das linhas de regressão para os dois grupos em questão é diferente; a declividade da linha de regressão é mais pronunciada para o grupo com o coeficiente de validade mais alto. Devido a isso, o problema da validade diferencial também é denominado *viés de declividade* (ver Tabela 5.5 e Figura 5.1).

A *predição diferencial*, por outro lado, ocorre quando os escores de teste subpredizem ou superpredizem o desempenho de um grupo no critério comparado ao outro. Este problema é denominado *viés de intersecção*, porque quando um preditor subprediz ou superprediz o desempenho de um grupo em um critério, o intercepto Y, ou ponto de origem da linha de regressão daquele grupo no eixo Y, é diferente para os outros grupos. Com relação aos problemas de validade diferencial e predição diferencial para escores de teste, diversos resultados são possíveis, embora não igualmente prováveis. Os escores de teste podem demonstrar (a) nenhum viés com relação a diferentes grupos, (b) validade diferencial e predição diferencial ao mesmo tempo, (c) validade diferencial sem predição diferencial ou (d) predição diferencial sem validade diferencial. Em geral, a busca por evidências de que a raça age como variável confundidora que resulta em validade e predição diferenciais para membros de minorias raciais baseada em escores de testes de habilidade não tem sido muito fértil. Na verdade, estudos que investigaram diferenças entre grupos étnicos na precisão das predições de desempenho no critério indicam que os testes muitas vezes tendem a *superpredizer* o desempenho de negros e hispânicos comparados a brancos e asiáticos. Por outro lado, alguns testes – especialmente os usados em decisões de admissão educacional – às vezes subpredizem o desempenho de mulheres, ainda que em grau menor do que aquele com que superpredizem o desempenho de alguns grupos de minorias étnicas ou raciais (Young, 2001; Zwick, 2002, Capítulos 5 e 6).

Naturalmente, os motivos por que alguns escores de teste geralmente superpredizem o desempenho no critério – tipicamente na forma de médias de notas – para membros de certos grupos de minorias raciais ou étnicas e, muitas vezes, subpredizem o desempenho de mulheres têm sido objeto de muitas conjecturas e debates. A baixa fidedignidade e o possível viés do critério de notas freqüentemente são citados como possíveis explicações para a superpredição do desempenho acadêmico de minorias raciais ou étnicas, assim como as disparidades em sua criação ou na qualidade de suas experiências educacionais anteriores. Uma explicação mais recente gira em torno da noção de *ameaça de estereótipo*, que se refere aos efeitos que o medo de confirmar estereótipos raciais negativos parece ter no desempenho de alguns membros de grupos minoritários em testes (Steele, 1997). No que diz respeito à subpredição do desempenho universitário de mulheres, as conjeturas citadas com maior freqüência estão centradas no fato de que, ao contrário dos homens, as mulheres como grupo (a) tendem a escolher cursos que são avaliados com menos rigidez ou (b) são mais sérias quanto a seus estudos. Embora possa ser verdade que estas e outras variáveis relacionadas ao *status* de gênero e etnia influenciam os escores de teste e o desempenho no critério de diferentes grupos, bem como as correlações entre eles, nem sempre é possível estabelecer precisamente quais são esses fatores. Além disso, aparentemente, o grau de predição diferencial do desempenho universitário para grupos minoritários étnicos e raciais e mulheres tem diminuído ao longo dos últimos 25 anos (Young, 2001). Além disso, quaisquer que sejam essas variáveis, elas obviamente não se aplicam a todos os membros desses grupos – que são eles mesmos bastante heterogêneos – da mesma maneira. Não obstante, a possibilidade de validade diferencial para membros de diferentes grupos étnicos, grupos de gênero, falantes de inglês não-nativos e outras categorias de indivíduos tradicionalmente em desvantagem sempre precisa ser investigada para determinar que os escores de testes usados em decisões importantes sejam justos para todos.

Uma solução possível para o problema da predição diferencial de escores de teste seria usar diferentes equações regressivas e diferentes escores de corte para a seleção de indivíduos de grupos étnicos e gêneros diferentes. No caso dos escores de teste que superpredizem o desempenho de negros e hispânicos, por exemplo, isto significaria exigir escores de corte mais altos para membros destas minorias do que para brancos e asiáticos. No entanto, isso obviamente vai contra os objetivos do oferecimento de oportunidades iguais pela ação afirmativa e do aumento da diversidade na educação superior e no local de trabalho. Outra solução proposta, implementada nos anos de 1980 com a Bateria de Testes de Aptidão Geral (GATB) desenvolvida pelo Serviço de Emprego dos Estados Unidos (USES), é usar normas de subgrupo para garantir taxas comparáveis de encaminhamento ao emprego para negros, hispânicos e brancos. Esta prática gerou tanta oposição que levou à aprovação da Lei dos Direitos Civis de 1991 (P.L. 101-336), que baniu qualquer tipo de ajustamento de escores baseado em raça, cor, sexo, religião ou origem nacional. À luz desses obstáculos, a maioria dos trabalhos na área da igualdade na testagem agora se concentra em (a) identificar fatores que possam estar contribuindo para a predição diferencial entre grupos de raça e de gênero (Steele, 1997; Willingham, Pollack e Lewis, 2000) e (b) analisar como os itens de teste funcionam para dife-

rentes subgrupos, enquanto os testes estão em construção, para garantir que aqueles que funcionam diferentemente não sejam incluídos (ver Capítulo 6).

Metanálises

Desde o fim dos anos de 1970, uma boa dose de clareza e entusiasmo renovado tem influenciado a perspectiva um tanto pessimista resultante das antigas pesquisas sobre a validade dos escores de teste de seleção. Esta mudança se deve em grande parte ao uso de metanálises, que permitem aos investigadores cotejar dados de muitos estudos diferentes – especialmente em áreas onde abundam achados conflitantes – e chegar a conclusões mais definitivas do que as que resultavam das formas tradicionais de revisão bibliográfica. Em contraste com a natureza qualitativa das revisões bibliográficas tradicionais, as *metanálises* se valem de uma série de procedimentos quantitativos que possibilitam a síntese e a integração dos resultados obtidos pela literatura de pesquisa sobre um dado tema. Essas técnicas, que já eram usadas em outros campos científicos há algum tempo, foram introduzidas na pesquisa psicométrica por Schmidt e Hunter (1977) como uma forma de abordar o problema da generalização da validade. Nas duas últimas décadas, as técnicas metanalíticas têm demonstrado que a validade preditiva dos escores de teste não é tão situacionalmente específica como se pensava antes, e se tornaram um importante método para esclarecer achados conflitantes em outras partes da literatura psicológica (Hunter e Schmidt, 1990, 1996).

O amplo interesse e o uso das metanálises têm sido estimulados pela percepção de que muitos achados conflitantes na pesquisa psicológica – incluindo os de estudos de validação – podem ser atribuídos a imperfeições de estudos individuais. Quando a influência de artefatos como erro de amostragem, erro de mensuração, restrição da amplitude e dicotomização injustificada de variáveis são removidos com correções estatísticas, um quadro muito mais claro costuma surgir. Concomitantemente, tem havido um reconhecimento crescente de que a testagem de hipóteses nos estudos psicológicos tem dado uma ênfase exagerada a *níveis de significação estatística* que enfatizam a evitação de *erros do Tipo I* (isto é, rejeitar incorretamente a hipótese nula de ausência de diferença, quando esta é a verdadeira) ao mesmo tempo que negligencia a possibilidade de *erros do Tipo II* (isto é, aceitar incorretamente a hipótese nula, quando ela é falsa). Uma vez que a relação entre os erros do Tipo I e do Tipo II é inversa, a ênfase em evitar o Tipo I aumenta a probabilidade de erro do Tipo II. Como conseqüência, um número enorme de resultados de pesquisa que não alcança os níveis desejados de significância estatís-

Não esqueça

- Os estudos de generalização de validade (GV) estão em uso por mais de um quarto de século e devem aumentar em número e no impacto que têm na teoria e prática psicométrica.
- Leitores que desejem se aprofundar no tópico da GV podem consultar *Validity Generalization: A Critical Review*, uma obra organizada por Murphy, Fleishman e Cleveland (2003).

tica (p. ex., 0,05 ou 0,01), mas que mesmo assim poderia contribuir com informações valiosas, foi ignorado ou excluído da literatura. O quadro Consulta Rápida 5.10 lista algumas referências que fornecem informações adicionais sobre as desvantagens e vantagens dos testes de significância da hipótese nula. De qualquer forma, essas discussões levaram ao que a maioria dos investigadores vê como uma mudança salutar no modo como os achados de pesquisa são relatados. Em vez de simplesmente afirmar os níveis de significância ou probabilidade dos resultados, agora é considerado necessário incluir índices dos *tamanhos de efeito*, ou a força das relações encontradas por um estudo de pesquisa, juntamente com os intervalos de confiança para os tamanhos de efeito e para todas as estimativas de parâmetros resultantes de uma investigação (APA, 2001).

Embora ainda esteja evoluindo, a metodologia das metanálises já fez contribuições substanciais para as evidências de validade preditiva dos escores de teste e outros procedimentos – como entrevistas de emprego e inventários de dados biográficos – que são usados na seleção de pessoal (Hartigan e Wigdor, 1989; Schmidt e Hunter, 1998; Schmidt et al., 1993). Além disso, as metanálises ajudaram a esclarecer a literatura de pesquisa e a promover o desenvolvimento de teorias em diversas áreas da psicologia industrial-organizacional – como a relação entre a satisfação no trabalho e o desempenho – bem como em vários outros campos (Hunter e Schmidt, 1996; Kirsch e Sapirstein, 1998; Rosenthal e DiMatteo, 2001). Um exemplo da área de testagem para admissões na educação superior vai ilustrar o potencial inerente à pesquisa meta-analítica.

Um caso educacional. Os estudos sobre a validade dos escores do *Graduate Record Examination* (GRE) como preditores do desempenho em programas de pós-graduação têm um longo histórico – que remonta aos anos de 1940 – pontuado por

CONSULTA RÁPIDA 5.10 Fontes de informação sobre os prós e contras dos testes de significância

A discussão a respeito dos méritos e desvantagens inerentes ao uso dos testes da hipótese nula de significância estatística para a realização de inferências em pesquisas em ciências sociais acontece há décadas entre estatísticos e metodologistas. Alguns deles têm se mostrado tão convencidos de que esses testes têm um efeito prejudicial à iniciativa científica que sugeriram o banimento da testagem de significância nos relatos de pesquisa. Embora este banimento não tenha sido instituído, agora é comum que periódicos de psicologia, bem como periódicos da maioria das disciplinas relacionadas exijam estimativas de tamanho de efeito sempre que valores de probabilidade (*p*) são relatados, juntamente com os intervalos de confiança para os tamanhos de efeito, coeficientes de correlação e outras estimativas de parâmetros populacionais. Informações adicionais a respeito das questões envolvidas nesta discussão podem ser encontradas nas seguintes fontes:

- Abelson, R.P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*, 12-15.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.
- Thompson, B. (2002). What future quantitative science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*, 25-32.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

achados inconsistentes. Enquanto algumas investigações (p. ex., Briel, O'Neill e Scheuneman, 1993; Broadus e Elmore, 1983) consideraram os testes Geral e por Temas do GRE preditores bastante válidos do desempenho em cursos de pós-graduação, muitos outros – incluindo alguns estudos meta-analíticos limitados – concluíram que a relação entre os escores no GRE e vários índices de sucesso na pós-graduação era menos do que adequada (p. ex., Goldberg e Alliger, 1992; Marston, 1971; Morrison e Morrison, 1995; Sternberg e Williams, 1997). Muitos estudos sobre a validade do GRE produziram coeficientes que variavam de pequenas correlações negativas a correlações positivas na casa de 0,20 para os escores Verbal e Quantitativo e coeficientes um tanto mais altos para os escores por Temas. Embora alguns desses achados fossem criticados devido a artefatos metodológicos como amplitudes altamente restritas, tanto nos escores do GRE quanto nas medidas de critério, bem como pouca fidedignidade dos critérios, o teor geral da literatura sobre a validade dos escores GRE não parecia oferecer evidências substanciais para corroborar seu uso nas decisões de admissão em cursos de pós-graduação (Kuncel, Campbell e Ones, 1998).

Contra este pano de fundo, Kuncel, Hezlett e Ones (2001) recentemente conduziram uma metanálise meticulosa e abrangente dos dados de 1753 amostras independentes englobando um total de 82.659 estudantes de pós-graduação. Este estudo abordou sistematicamente aspectos teóricos, estatísticos e metodológicos da literatura sobre a validade preditiva dos escores GRE. Kuncel e seus colegas examinaram as relações entre cinco preditores – GRE verbal (GRE-V), quantitativo (GRE-Q), analítico (GRE-A) e escores por Temas, bem como a média de notas no curso de graduação (*undergraduate grade point average; UGPA*) – e oito critérios diferentes de sucesso em cursos de pós-graduação, incluindo *GPA* do primeiro ano (média das notas obtidas no primeiro ano da pós-graduação) e GPA geral da pós-graduação (*GGPA*), escores de exames abrangentes, avaliações do corpo docente, colação de grau e índices numéricos relacionados à produtividade em pesquisa. Eles conduziram análises separadas para a amostra total e para sub-amostras representativas de estudantes nas áreas de ciências humanas, ciências sociais, ciências biológicas e ciências físico-matemáticas, bem como para falantes não-nativos do inglês e estudantes mais velhos do que a idade tradicional dos alunos de pós-graduação. Entre os muitos refinamentos metodológicos que Kuncel e seus colegas empregaram na análise estavam correções para a restrição de amplitude e mudanças na variabilidade das distribuições de variáveis de preditor e de critério, bem como para a baixa fidedignidade das medidas de critério. Além disso, estes investigadores contemplaram diversas armadilhas potenciais inerentes à metanálise.

Kuncel e colaboradores (2001) relatam seus principais resultados em termos de correlações médias observadas, juntamente com seus desvios padrões, ponderadas pelo tamanho da amostra. Também relatam validades operacionais estimadas com seus desvios padrões e intervalos de confiança de 90%. Esses achados indicam que as quatro medidas GRE (GRE-V, GRE-Q, GRE-A e GRE por Temas) são preditores razoavelmente bons da maioria dos critérios empregados para a amostra total, bem como para as subamostras. Na verdade, na maioria dos casos, os escores GRE parecem ser preditores melhores do que o UGPA. Os escores GRE por Tema mostraram ser os melhores preditores isolados da GPA de pós-graduação em todas as

disciplinas, com validades operacionais estimadas variando de 0,40 a 0,49. Em contraste, os escores dos testes gerais (GRE-V, GRE-Q e GRE-A) – embora se correlacionassem substancialmente com os dos testes por Temas – tinham coeficientes de validade operacional variando entre 0,27 e 0,48. Kuncel e colaboradores concluíram que, embora os escores dos testes gerais do GRE contribuam com apenas um pequeno incremento na validade quando somados aos escores dos testes por Temas, eles ainda podem ser valiosos, especialmente para alunos que fizeram cursos de graduação em áreas diferentes daquelas nas quais estão tentando admissão.

Em suma, a metanálise de Kuncel e colaboradores (2001) claramente sugere (a) que muito da inconsistência nos estudos anteriores de validação do GRE era resultado de restrição de amplitude e erro de amostragem daqueles estudos, e (b) que os escores GRE merecem ter um papel no processo das admissões em escolas de pós-graduação. No entanto, esses autores não investigaram a questão da predição diferencial para mulheres e membros de minorias raciais ou étnicas, e admitem que ainda há muito espaço para melhorias na validade do processo de admissões a escolas de pós-graduação. Com respeito a este último ponto, as seguintes observações devem ser feitas:

- A finalidade dos escores GRE e da maioria dos outros preditores usados na decisões de seleção não é estimar a posição exata no critério de cada candidato, mas sim determinar se eles podem atingir o nível necessário de sucesso. Se os escores de critério tiverem que ser preditos exatamente, a margem de erro (EP_{esc}) para os coeficientes na casa dos 0,30 e 0,40 certamente seria considerável.
- O desempenho na maioria dos critérios, incluindo o sucesso na pós-graduação, é determinado por múltiplos fatores, incluindo características emocionais e atitudinais e hábitos comportamentais, bem como talentos criativos e práticos que não são medidos pelo GRE ou por outros testes cognitivos geralmente usados como preditores.
- A maioria das outras decisões de seleção, incluindo as admissões a escolas de pós-graduação, raramente são feitas apenas com base em um único preditor. Portanto, a questão crucial em relação aos testes de seleção diz respeito à sua utilidade. Isto significa que a pergunta que tem que ser feita é se o uso de escores de teste como parte de um processo de tomada de decisões aumenta o número de decisões válidas acima do que este seria com o uso de preditores de outra natureza, como as GPAs. Na maioria das situações, incluindo as decisões de admissão à educação superior, os dados sugerem que os escores de teste realmente contribuem para a eficiência preditiva na tomada de decisões (Hartigan e Wigdor, 1989; Kobrin, Camara e Milewski, 2002; Kuncel, Hezlett e Ones, 2001).

Além da seleção: usando escores de teste para outros tipos de decisões

Até aqui, a discussão dos procedimentos de validação relacionados ao critério tem se centrado primariamente no uso de testes para seleção ou triagem feita a partir

Não esqueça

Uma grande quantidade de informações a respeito da validade preditiva dos testes usados na educação superior, incluindo muitas questões relacionadas à predição diferencial para vários subgrupos, pode ser encontrada na Internet. Ver especialmente os seguintes sites:

- ACT (http://www.act.org)
- The College Board (http://www.college.board.com)
- Educational Testing Service (ETS: http://www.ets.org)

da validade de escores de teste concorrente ou preditiva. As decisões de *seleção* são aquelas que requerem uma escolha entre duas alternativas. No contexto do emprego e da educação, as alternativas habituais são aceitar ou rejeitar um candidato; no contexto clínico e forense, as decisões de seleção geralmente envolvem a determinação da presença ou ausência de uma síndrome ou condição em particular. O termo *triagem* se refere a um passo preliminar em um processo de seleção, geralmente realizado para separar indivíduos que justificam ou requerem uma avaliação mais extensa daqueles que não o fazem. Por exemplo, muitas clínicas usam periodicamente um questionário simples e curto para triar transtornos como depressão ou ansiedade na população geral; os empregadores podem triar candidatos com um instrumento breve para limitar o número de candidatos àqueles que satisfazem os requisitos mínimos de uma vaga.

Os escores de testes psicológicos também são usados para tomar decisões de colocação e classificação, ambas envolvem mais de duas opções. Dessas duas, as decisões de *colocação* são mais simples. Elas envolvem atribuir a indivíduos categorias ou tratamentos separados a partir de um único escore, ou de um escore composto calculado com uma única equação regressiva, em referência a um único critério. Embora as decisões de colocação não envolvam a opção de rejeitar indivíduos que não satisfazem um certo nível de desempenho em um teste ou preditor, elas não são substancialmente diferentes das decisões de seleção em termos das evidências que requerem, que são uma relação demonstrável entre um ou mais preditores e um critério. Escores em um teste de leitura, por exemplo, podem ser usados para colocar alunos em turmas adequadas aos seus níveis de habilidade. Da mesma forma, escores em uma escala de depressão podem ser usados para classificar pacientes psiquiátricos, em termos da severidade de seus sintomas depressivos, para ajudar a determinar tipos e níveis apropriados de intervenção terapêutica.

As decisões de *classificação*, por outro lado, são um tanto mais complicadas. Na classificação – assim como na colocação – ninguém é rejeitado, mas aos indivíduos devem ser atribuídos *diferencialmente* categorias ou tratamentos distintos com base em múltiplos critérios. Isso significa que múltiplos preditores são necessários e suas relações com cada critério devem ser determinadas independentemente, por equações regressivas separadas. A ferramenta mais apropriada para decisões de classificação é uma bateria de testes ou preditores, cujos resultados são validados em relação aos vários critérios a serem preditos e então combinados em equações que refletem seus pesos relativos para a predição de cada critério.

As decisões de classificação são necessárias no contexto do emprego, da educação, do aconselhamento e da clínica. No campo do emprego, incluindo o contexto militar ou industrial, essas decisões são necessárias, quando as aptidões de um grupo de funcionários disponíveis têm que ser avaliadas para designar indivíduos para os empregos ou programas de treinamento nos quais terão maior probabilidade de funcionar efetivamente. O aconselhamento vocacional de indivíduos que querem se decidir por um programa de estudos ou escolha de carreira também demanda decisões de classificação. No contexto clínico, as decisões de classificação devem ser tomadas em casos que requerem diagnósticos diferenciais. Um exemplo típico seria a necessidade de estabelecer se um paciente mais velho que exibe sintomas de depressão e problemas de memória pode estar sofrendo de um transtorno depressivo que afeta sua memória e concentração, de um processo incipiente de demência que está causando depressão ou de uma combinação dos dois.

As baterias de testes usadas para decisões de classificação devem ser avaliadas em termos de evidências de *validade diferencial*. Dentro desse contexto, o termo validade diferencial significa que uma bateria deve ser capaz de predizer ou estabelecer diferenças entre dois ou mais critérios. Em um problema de classificação de dois critérios, uma bateria ideal consistiria em preditores que se correlacionassem afirmativamente com um critério e não se correlacionassem ou se correlacionassem negativamente com o outro. No problema do diagnóstico diferencial da depressão *versus* demência, por exemplo, podem-se procurar diferenças na seqüência temporal de sintomas da depressão e comprometimento cognitivo, ou diferenças nos níveis relativos de desempenho em vários tipos de testes de memória.

Quando a situação de classificação envolve predições relativas a mais de dois critérios, tais como designar pessoal para qualquer um de diversos empregos ou programas de treinamento possíveis, o problema de estabelecer evidências de validade se torna ainda mais complexo. Nesse tipo de condição, um preditor que se correlaciona igualmente bem com todos os critérios envolvidos na decisão – como um teste de inteligência geral relacionado à maioria dos critérios de desempenho no emprego – tem relativamente pouco uso. Uma forma possível de lidar com problemas de classificação deste tipo é por meio do uso de múltiplas análises de função discriminante. As *funções discriminantes* envolvem a aplicação de combinações ponderadas de escores aos preditores – derivadas por meio de análise regressiva – para determinar o quanto o perfil de escores de um indivíduo corresponde aos perfis de indivíduos em diferentes grupos ocupacionais, diferentes especialidades ou diferentes categorias psiquiátricas. Embora as funções discriminantes sejam úteis em certos casos (p. ex., quando os critérios consistem simplesmente em afiliação a um grupo ou outro, ou quando existe uma relação não-linear entre um critério e um ou mais preditores), elas deixam a dever em termos dos requisitos de muitas situações porque não permitem a predição do nível de sucesso em um campo específico. Para um exemplo da aplicação da análise de função discriminante na diferenciação de perfis da WAIS-R entre pacientes com lesões cefálicas que não buscaram indenizações e sujeitos instruídos a simular traumas cefálicos, ver Mittenberg, Theroux-Fichera, Zielinski e Heilbronner (1995).

Outra estratégia tradicional que tanto pode ser usada para problemas de seleção quanto de classificação é a *validação sintética* (Balma, 1959). Esta técnica es-

sencialmente se vale de análises de cargo detalhadas que identificam componentes específicos dos cargos e seus pesos relativos em diferentes empregos. Com base nesta análise, coeficientes de regressão previamente estabelecidos para escores de testes que predizem esses elementos separados do cargo podem ser combinados em uma nova bateria sintética que vai predizer o desempenho nos cargos em questão. Procedimentos estatísticos associados a este método foram desenvolvidos por Primoff (1959; Primoff e Eyde, 1988) e têm sido expandidos por outros autores desde então. No entanto, para serem úteis nas decisões de classificação, as estratégias de validação sintética devem envolver preditores que mostram boa validade discriminante, a menos que os componentes do critério em si sejam substancialmente correlacionados (Guion, 1998, p.354-355).

Perspectivas adicionais sobre a validação relacionada ao critério

Diversos avanços metodológicos discutidos anteriormente neste capítulo, como a modelagem de equação estrutural e os estudos meta-analíticos de generalização da validade, foram aplicados em pesquisas recentes sobre os problemas da validação relacionada ao critério. Ao mesmo tempo, a disponibilidade dessas ferramentas estatísticas cada vez mais sofisticadas concentrou a atenção na conceitualização dos preditores e critérios de desempenho, bem como em suas inter-relações (J.P. Campbell, 1990). Um dos avanços recentes mais significativos em termos de preditores é o reconhecimento de que a inclusão de fatores relacionados a dimensões da personalidade – além daqueles relacionados às habilidades – pode aumentar a validade dos instrumentos usados para predizer o desempenho em várias áreas (Lubinski e Dawis, 1992). Com relação ao problema do critério, a natureza multifacetada do desempenho na maioria dos empregos e iniciativas educacionais agora é amplamente reconhecida e cada vez mais analisada. Isto requer um exame dos vários elementos que contribuem para o sucesso, avaliando seu valor relativo, reconhecendo quais estão sob controle do indivíduo, desenvolvendo métodos para avaliar cada elemento separadamente e, então, combinando-os em uma medida total do desempenho que leve em conta todos esses fatores e seus pesos relativos.

Um programa modelo de validação relacionada ao critério

Um exemplo notável da aplicação de muitas inovações nos métodos de pesquisa sobre avaliação pode ser encontrado no trabalho que John P. Campbell e colaboradores (J.P. Campbell, 1990, 1994; J.P. Campbell e Knapp, 2001) conduziram ao longo das duas últimas décadas em conjunção com um esforço abrangente para avaliar e melhorar os procedimentos do exército americano para a seleção e classificação de pessoal, conhecido como Projeto A. Um relato completo deste trabalho, que talvez seja o maior projeto na história da pesquisa de pessoal, não é viável neste contexto. No entanto, alguns de seus pontos altos são apresentados aqui para dar aos leitores uma idéia de seu alcance e significância.

O Projeto A

Usando uma base de dados extensa de mais de 50 mil pessoas, os investigadores do Projeto A selecionaram 21 especialidades ocupacionais militares (EOMs), de um total de mais de 200 cargos de nível inicial, com as quais conduziram estudos de validação concorrente e preditiva longitudinal. Os preditores estudados incluíam os 10 subtestes da Bateria de Aptidão Vocacional das Forças Armadas (*Armed Services Vocational Aptitude Battery;* ASVAB) listados no quadro Consulta Rápida 5.11. Vários escores compostos formados por diferentes combinações de subtestes da ASVAB têm sido usados tradicionalmente para seleção e classificação de indivíduos que se candidatam ao serviço das Forças Armadas. Além disso, diversos instrumentos novos – incluindo testes de habilidade psicomotora e espacial, bem como medidas de interesses e de personalidade – também foram incluídos no trabalho de validação.

Os investigadores do Projeto A realizaram extensas análises de cargo e prestaram especial atenção à padronização de medidas de proficiência no trabalho para cada uma das EOMs. Análises fatoriais exploratórias e confirmatórias, bem como outras técnicas estatísticas, foram empregadas em estudos de modelagem de

CONSULTA RÁPIDA 5.11

Bateria de Aptidão Vocacional das Forças Armadas (*Armed Services Vocational Aptitude Battery;* ASVAB)

A ASVAB é administrada como teste adaptativo computadorizado (CAT-ASVAB) nos postos de alistamento militar fixos e como teste de lápis e papel em unidades móveis. Os indivíduos submetidos ao ASVAB já foram pré-testados com instrumentos de triagem mais curtos administrados por recrutadores. Os 10 subtestes que compõem a ASVAB, usada por todas as Forças Armadas americanas na seleção e classificação de pessoal, são os seguintes:

- Raciocínio aritmético
- Operações numéricas
- Compreensão de parágrafo
- Conhecimento de palavras
- Velocidade de codificação
- Ciência geral
- Conhecimento matemático
- Informações de eletrônica
- Compreensão mecânica
- Informações automotivas

Os primeiros quatro subtestes compõem o Teste de Qualificação das Forças Armadas (*Armed Forces Qualification Test;* AFQT), que, juntamente com o histórico de graduação do ensino médio, é usado para determinar a elegibilidade para o alistamento. Os subtestes da ASVAB também são combinados em 10 escores compostos de aptidão diferentes que são usados como preditores do sucesso nos programas de treinamento da escola militar (J.P. Campbell e Knapp, 2001, p.14-18).

desempenho que levaram à especificação dos escores fatoriais usados como critérios em três diferentes estágios da carreira: desempenho ao final do treinamento, desempenho no primeiro cargo e desempenho no segundo cargo. A exploração das evidências de validade nos vários estágios da carreira era um aspecto significativo dessa pesquisa porque, além de desenvolver uma bateria de testes capaz de predição diferencial, outro objetivo importante da tarefa de gerenciamento de pessoal nas Forças Armadas voluntárias é minimizar atritos.

A análise de validação efetiva realizada ao longo de todo o Projeto A incluiu (a) correlações entre cada preditor e critérios de desempenho; (b) comparações da validade incremental fornecida pelas medidas experimentais em relação a cada critério, acima e além do poder preditivo dos escores da ASVAB previamente disponíveis; (c) desenvolvimento e comparação de várias equações ótimas para validade máxima na predição longitude do desempenho no primeiro cargo e (d) análises da validade de equações alternativas usando diferentes combinações de dados de testes e desempenho anterior para a sua predição no segundo cargo. Métodos de generalização de validade e validação sintética foram usados para investigar a eficiência de várias combinações de preditores para as 21 EOMs originalmente selecionadas para análise, bem como para muitas outras EOMs e grupos de ocupações. Estudos adicionais incluíram uma avaliação da utilidade e dos ganhos alcançados através de estratégias e condições de classificação alternativas.

Entre os muitos achados úteis do Projeto A, um dos mais significativos foi a corroboração do valor potencial de diversas novas medidas estudadas, incluindo aquelas relacionadas a dimensões da personalidade, bem como alguns testes novos de habilidades psicomotoras e espaciais. Estes preditores adicionais se encontram agora em vários estágios preliminares de implementação e estão sendo explorados mais aprofundadamente. Segundo os principais investigadores do Projeto A, essas medidas ainda não são plenamente operacionais devido a obstáculos criados pelas muitas partes envolvidas no sistema de seleção e classificação do exército, bem como pela inércia organizacional (J.P. Campbell e Knapp, 2001, p.570-574). Mesmo assim, o Projeto A e as investigações subseqüentes surgidas a partir dele já contribuíram com vários avanços metodológicos substanciais que certamente vão melhorar o calibre das pesquisas de validação para seleção e classificação, tanto nas Forças Armadas como em muitos outros contextos, ajudando a atingir a meta geral de maximizar a utilização de talentos humanos.

ASPECTOS ADICIONAIS DA VALIDADE: UTILIDADE E CONSEQÜÊNCIAS

Existem dois aspectos significativos a respeito do uso de escores de teste que estão muito ligados à sua validade, mas não são necessariamente à sua essência, quais sejam, sua utilidade e as conseqüências de seu uso. A complexidade desses tópicos impede uma discussão extensa no âmbito deste volume, mas sua importância crucial para as iniciativas de testagem psicológica justificam uma introdução neste ponto, a ser seguida por um tratamento mais aprofundado no Capítulo 7.

Avaliando a utilidade da testagem

A *utilidade* dos testes e seus escores se refere aos benefícios que eles trazem à tomada de decisões. A utilidade é contingente à extensão em que o uso de testes pode aumentar a taxa de precisão das inferências e decisões que desejamos fazer – acima e além do que seria se usássemos outras ferramentas disponíveis. Tipicamente, a utilidade é avaliada em termos econômicos, como as relações custo-benefício envolvidas no uso de testes *versus* uso de dados de outros procedimentos. Uma vez que o uso de testes sempre acontece dentro de um contexto, a análise de seus custos e benefícios necessariamente deve levar em conta dados adicionais pertinentes a cada situação particular na qual ele é contemplado. Dependendo do contexto, esses dados incluem questões como as probabilidades e riscos envolvidos em determinações falso-positivas e falso-negativas, a disponibilidade de ferramentas alternativas – e sua eficiência relativa comparada aos testes – bem como a relativa facilidade ou dificuldade nas determinações que precisam ser feitas. Este aspecto do uso de testes faz parte do tópico mais amplo da teoria das decisões, que não se aplica apenas à psicometria em particular e à psicologia em geral, mas também a campos tão diversos quanto medicina, economia, jurisprudência, ciência militar e jogos, bem como qualquer outra iniciativa humana na qual um planejamento estratégico seja necessário. Para uma amostra das numerosas contribuições que os especialistas em psicometria têm feito à teoria das decisões ao longo das últimas décadas, ver Boudreau (1991), Brogden (1946), Brown e Ghiselli (1953), Buchwald (1965), Cronbach e Gleser (1965), Hynter e Schmidt (1981), Schmidt, Hunter, McKenzie e Muldrow (1979) e Taylor e Russel (1939). Uma excelente discussão dos aspectos estatísticos da teoria das decisões dentro do contexto do melhoramento da precisão e da utilidade das decisões diagnósticas está disponível em um relato recente de Swets, Dawes e Monahan (2000).

Não esqueça

Os julgamentos a respeito da validade de escores de teste são relativos. Quando as evidências acumuladas da validade dos escores produzidos por um teste são consideradas abstratamente, elas podem ser ou não consideradas suficientes para os fins pretendidos do teste.

No entanto, quando qualquer escore ou conjunto de escores específicos de um indivíduo ou grupo é considerado, deve-se ter em mente que os escores podem ter sido afetados por fatores pertinentes apenas aos testandos, ao examinador, ao contexto no qual a testagem aconteceu e à interação entre esses fatores. Portanto, a situação de testagem sempre deve ser levada em conta quando os escores de teste são interpretados.

Além disso, fazer inferências com base em escores de teste requer informações a respeito de seus referenciais e sua fidedignidade, bem como evidências de validade de todas as fontes pertinentes.

Avaliando as conseqüências do uso de testes

As conseqüências individuais e sociais do uso de testes, que podem ser positivas ou negativas, devem ser avaliadas em termos de suas implicações de valor. Alguns teó-

ricos, especialmente Messick (1989, 1995), argumentaram que os julgamentos de validade são na verdade julgamentos de valor. Messick propôs a inclusão de aspectos *conseqüenciais* do uso e interpretação dos escores de teste – isto é, a avaliação de suas conseqüências sociais pretendidas e não pretendidas – dentro da noção de validade em seu sentido mais abrangente. Ao criticarem a proposta de Messick, alguns a rotularam como "validade conseqüencial", um termo que o próprio Messick nunca usou. De qualquer forma, esse aspecto em particular da imensa contribuição de Messick à teoria e à prática psicométrica não foi amplamente adotado pelos profissionais de testagem. A maioria deles poderia argumentar que, embora os aspectos conseqüenciais do uso de testes sejam de grande importância e devam ser investigados antes da implementação e documentados após esta, eles se inscrevem no campo da ética profissional, valores morais e considerações políticas, e não da determinação da validade como tal (Cole e Moss, 1989; Lees-Haley, 1996; Linn, 1998).

O *Ethical principles of psychologists and code of conduct* (APA, 2002), por exemplo, conclama todos os psicólogos – incluindo aqueles que usam testes e outras ferramentas de avaliação – a levar em conta as possíveis ramificações de seu trabalho, de modo a maximizar os benefícios e prevenir ou minimizar os danos. Os usuários de testes e ferramentas de avaliação, especificamente, devem se policiar contra interpretações e usos equivocados e obter o consentimento dos testandos em relação aos objetivos da testagem, a maneira como os escores serão usados, as pessoas que terão acesso aos escores e outras questões antes da testagem. Limitações semelhantes se aplicam ao trabalho dos autores e criadores de testes, editoras e revisores, bem como outros profissionais envolvidos na elaboração de testes (p. ex., Society for Industrial and Organizational Psychology [SIOP], 2003). Acrescentar à noção já complexa da validação de escores a faceta adicional de avaliar as ramificações do uso de testes em termos de preocupações sociais mais amplas – tais como o equilíbrio entre princípios morais como justiça e bem comum – colocaria um fardo indevido, que pertence à sociedade como um todo, unicamente sobre os profissionais de testagem.

Um exemplo de como os princípios e práticas éticas na testagem podem ser aplicados ao campo da educação pode ser encontrado em um guia de recursos, para educadores e elaboradores de políticas, que apresenta os padrões profissionais e os princípios legais pertinentes ao uso de testes na tomada de decisões educacionais de grande impacto para alunos (U.S. Department of Education, Office for Civil Rights, 2000). Conjuntos semelhantes de diretrizes para o uso de testes em outros campos de especialidade serão discutidos no Capítulo 7 (SIOP, 2003).

COMENTÁRIOS FINAIS

O quadro Consulta Rápida 5.12 delineia como várias fontes de evidências e estratégias de validação podem ser aplicadas na interpretação dos escores de um único teste, dependendo dos objetivos para os quais ele é usado. Na maioria dos casos, quando a interpretação proposta dos escores de um teste se afasta do objetivo original para o qual ele foi desenvolvido, as linhas de evidência para usos e interpretações alternativas se tornam menos diretas. Por exemplo, o teste usado como exemplo no quadro Consulta Rápida 5.12 é um exame final na disciplina de Cálcu-

Estratégias de validação em relação à interpretação de escores de teste

Teste cujos escores serão interpretados	Objetivo proposto da interpretação dos escores do teste	Tipo de estratégia de validação desejada	Possíveis fontes de evidências
Exame final da disciplina Cálculo I	Determinar se os estudantes serão aprovados na disciplina Cálculo I	Conteúdo	Relevância e representatividade do conteúdo do teste em relação aos temas abordados na disciplina Cálculo I
	Determinar se os estudantes estão prontos para a disciplina Cálculo II	Relacionada ao critério, tipo concorrente	Correlação positiva alta entre escores no teste de Cálculo I e notas na disciplina Cálculo II
	Predizer se os estudantes podem completar com sucesso a graduação em matemática	Relacionada ao critério, tipo preditivo	Correlação positiva alta entre escores no teste de Cálculo I e conclusão do curso de matemática
	Investigar a relação entre a habilidade matemática e o tipo de personalidade	Convergência	Suporte para a hipótese de que estudantes introvertidos vão ter escore mais alto do que estudantes extrovertidos no teste de Cálculo I

lo I. Podemos inferir que seu objetivo original era determinar se os testandos tinham dominado uma parte suficiente do conteúdo dessa disciplina para alcançar uma nota de aprovação. À medida que a interpretação dos escores deste teste é estendida para a determinação da prontidão para Cálculo II, a predição do sucesso como graduado em matemática ou o uso dos escores de teste como substituto para a habilidade matemática em uma investigação de correlatos do tipo de personalidade, a ligação entre as evidências de validação e a interpretação pretendida se torna mais e mais tênue. Isso não significa que o exame final de Cálculo I não deva ser usado para objetivos diferentes do original, mas sugere que evidências adicionais serão necessárias para esses outros fins.

No presente capítulo, discutimos as evidências da validade dos escores de teste do ponto de vista do teste como um todo. No próximo capítulo, vamos nos voltar para uma análise mais detalhada dos dados de teste na perspectiva das unidades de amostra de comportamento que compõem os escores, quais sejam, os itens de teste.

Teste a si mesmo

1. A *validade* é o grau em que
   (a) um teste mede o que pretende medir
   (b) as evidências corroboram as inferências feitas a partir de escores de teste
   (c) os escores de testes são consistentes em várias situações

2. Em décadas recentes, as várias formas de evidências de validade foram incluídas dentro da noção de *validade* _____.
   (a) de conteúdo
   (b) concorrente
   (c) preditiva
   (d) de constructo

3. O *intervalo nomotético* se refere a
   (a) uma rede de relações entre medidas
   (b) a decomposição de tarefas
   (c) a identificação de diferenças entre testandos
   (d) o alcance do constructo que está sendo medido

4. As evidências de validade que se baseiam no conteúdo do teste e processos de resposta são particularmente aplicáveis a
   (a) inventários de interesse
   (b) testes educacionais
   (c) testes de personalidade

5. A *validade de face* se refere primariamente a
   (a) a representatividade do conteúdo do teste
   (b) as evidências de validade da perspectiva psicométrica
   (c) as características superficiais de um teste
   (d) a quantidade de dados de validação empírica acumulada para um teste

6. Para coletar evidências de validade discriminante, devem-se correlacionar os escores de testes que pretendem avaliar constructos _____.

   (a) iguais
   (b) semelhantes, mas não iguais
   (c) diferentes

7. Um dos aspectos mais úteis da análise fatorial, em sua aplicação às pesquisas de validação de testes, é que os resultados da aplicação desta técnica podem

   (a) simplificar a interpretação e o relato dos escores de teste
   (b) revelar os aspectos essenciais dos constructos que os testes estão avaliando
   (c) ser prontamente generalizados para outras populações

8. Quais das seguintes afirmações a respeito de medidas de critério *não* é verdadeira?

   (a) as medidas de critério podem diferir em termos de fidedignidade e validade
   (b) diferentes medidas de critério nem sempre se correlacionam umas às outras
   (c) as medidas de critério podem ou não ser generalizadas para diferentes grupos
   (d) as melhores medidas de critério geralmente estão disponíveis no momento da testagem

9. Os erros padrões de estimativa são usados para medir

   (a) a fidedignidade dos critérios
   (b) a fidedignidade dos preditores
   (c) a precisão dos escores obtidos
   (d) a precisão com que os critérios são preditos

10. Do ponto de vista dos procedimentos de validação relacionados ao critério, quais dos seguintes tipos de decisões são os mais complexos?

    (a) seleção
    (b) colocação
    (c) classificação

Respostas: 1. b; 2. d; 3. a; 4. b; 5. c; 6. c; 7. a; 8. d; 9. d; 10. c.

[capítulo 6]

CONSIDERAÇÕES BÁSICAS SOBRE ITENS DE TESTE

Os *itens de teste* são as unidades que o compõem e os meios pelos quais as amostras de comportamento dos testandos são coletadas. Segue-se que a qualidade geral de um teste depende primariamente da qualidade dos itens que o compõem, embora o número de itens e seu seqüenciamento ou posição dentro do teste também sejam questões de fundamental importância. Assim como os testes são avaliados em relação ao grau em que satisfazem seus objetivos propostos, os itens individuais devem também ser avaliados com a mesma base para os fins do teste como um todo. *Análise de itens* é um termo geral que se refere a todas as técnicas usadas para avaliar as características de itens de teste e sua qualidade durante o seu processo de desenvolvimento e construção.

A análise de itens envolve procedimentos quantitativos e qualitativos. Os procedimentos de *análise qualitativa de itens* se apóiam em julgamentos de revisores a respeito das características substantivas e estilísticas dos itens, bem como de sua precisão e imparcialidade. Os principais critérios usados para avaliar itens qualitativamente são (a) adequação do conteúdo e do formato do item ao objetivo do teste e às populações para as quais ele se destina, (b) clareza de expressão, (c) correção gramatical e (d) aderência a algumas regras básicas para a redação de itens que evoluem com o tempo. Como será discutido mais adiante neste capítulo, o conteúdo dos itens também é examinado cuidadosamente para identificar e eliminar possíveis fontes de viés ou estereótipos ofensivos de subgrupos específicos da população. O quadro Consulta Rápida 6.1 lista livros que trazem informações a respeito do processo de desenvolvimento de itens e diretrizes práticas para a sua redação. A *análise quantitativa de itens* envolve uma variedade de procedimentos estatísticos para determinar as características psicométricas dos itens com base nas respostas obtidas a partir das amostras usadas no processo de desenvolvimento do teste. A maior parte do presente capítulo trata da análise quantitativa dos itens de teste.

CONSULTA RÁPIDA 6.1

Redigindo itens de teste

Para mais esclarecimentos sobre o processo de preparação de itens para testes de habilidade, bem como orientações claras sobre como redigi-los, os leitores podem consultar uma das seguintes fontes:

- Bennett, R.E., & Ward, W.C. (Eds.).(1993). *Construction versus choice: Issues in constructed response, performance testing and portfolio assessment*. Hillsdale, NJ: Erlbaum.
- Haladyna, T.M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn & Bacon.
- Haladyna, T.M. (1999). *Developing and validating multiple-choice test items* (2ª ed.). Mahwah, NJ: Erlbaum.

Embora não existam guias comparáveis para toda a ampla gama de abordagens ao desenvolvimento de instrumentos de avaliação da personalidade, alguns princípios básicos para a preparação de itens objetivos podem ser obtidos nas seguintes obras:

- Aiken, L.R. (1996). *Rating scales and checklists: Evaluating behavior, personality and attitudes*. New York: Wiley.
- Aiken, L.R. (1997) *Questionnaires and inventories: Surveying opinions and assessing personality*. New York: Wiley.
- Fink, A. (2002). *How to ask survey questions* (2nd ed., Vol. 2). Thousand Oaks, CA: Sage. [Este é um dos dez volumes do *Survey Kit*, organizado por Arlene Fink e publicado pela Sage.]

O CONTEXTO DA ANÁLISE DE ITENS: O DESENVOLVIMENTO DE TESTES

Fundamentalmente, os procedimentos envolvidos na geração, seleção e análise de itens dizem respeito ao tópico da teoria e delineamento de testes. Como tal, são de importância crítica para os autores e criadores de testes. Os usuários também precisam estar familiarizados com estes procedimentos para compreenderem a natureza das tarefas envolvidas em um teste e avaliarem os instrumentos que selecionam. No entanto, quando um teste se torna disponível para os usuários, ele já é um produto acabado. Do ponto de vista dos usuários, uma vez que um teste foi selecionado, seus itens são basicamente de interesse – especialmente no contexto da avaliação individual – primário como meio de observação e inspeção das respostas dos testandos na perspectiva singular da situação e das circunstâncias específicas nas quais o teste é administrado. Na avaliação individual, o modo como os examinandos respondem às tarefas e seus padrões particulares de resposta podem fornecer informações adicionais que complementam o processo de interpretação dos escores do teste. Naturalmente, os usuários também estão preocupados com as características práticas dos itens. As mais importantes são sua adequação a tipos específicos de contextos e examinandos, a facilidade com que podem ser administrados e avaliados, o tempo envolvido na sua administração e a quantidade de treinamento necessária para se dominar os procedimentos envolvidos na administração e avaliação.

Os procedimentos de análise de itens são implementados em vários pontos durante o desenvolvimento de um teste, um processo que inclui diversos outros passos. Para estabelecer o contexto para a discussão da análise de itens, os passos envolvidos no desenvolvimento de um teste são descritos de forma breve nos parágrafos seguintes. Explicações mais extensas deste processo estão disponíveis em muitas fontes (AERA, APA e NCME, 1999, Capítulo 3; DeVellis, 2003; Ramsay e Reynolds, 2000; Robertson, 1992).

Como esclarece Robertson (1992), desenvolver um teste padronizado engloba um considerável investimento de tempo e dinheiro e requer o trabalho de profissionais especializados em psicometria, bem como na área específica da qual o teste trata. Por isso, o desenvolvimento de testes com vistas à distribuição comercial – diferentemente das medidas experimentais a serem usadas primariamente para fins de pesquisa, testes de sala de aula ou testes desenvolvidos por empregadores para uso interno em empresas – é tipicamente realizado por editoras, que têm os recursos financeiros e os conhecimentos técnicos necessários. A iniciativa da criação de novos testes pode se originar dentro da equipe das próprias editoras ou partir de investigadores e autores independentes que oferecem suas idéias a elas.

A decisão de desenvolver um teste geralmente é tomada quando o criador em potencial se dá conta de que não existe um teste para um determinado fim, ou que os testes existentes para este fim não são adequados por um motivo ou outro. Considerações ligadas ao *marketing* também são centrais para o processo de tomada de decisões em uma editora de testes comerciais. De qualquer maneira, quando é tomada a decisão de se desenvolver um teste, seus objetivos e fundamentos teóricos devem ser articulados cuidadosamente em termos das inferências a serem feitas a partir de seus escores, e o desenvolvedor também deve elaborar um plano para o teste.

O planejamento de um teste envolve especificar (a) os constructos ou domínios de conhecimento que o teste vai avaliar, (b) o tipo de população em que o teste será usado, (c) os objetivos dos itens a serem desenvolvidos, dentro do referencial dos objetivos do teste e (d) os meios concretos por meio dos quais as amostras de comportamento serão coletadas e avaliadas. Este último ponto inclui decisões a respeito do método de administração, do formato dos estímulos e respostas do teste e dos procedimentos de avaliação a serem usados. Depois que estas questões são decididas e um plano preliminar do teste é elaborado, o processo de desenvolvimento geralmente envolve os seguintes passos:

1. Gerar um *pool* de itens, por escrito ou em outro formato, bem como determinar os procedimentos de administração e avaliação a serem usados.
2. Submeter o *pool* de itens a revisores para uma análise qualitativa e alterá-los ou substituí-los conforme necessário.
3. Testar os itens gerados e revisados com amostras representativas da população para a qual o teste se destina.
4. Avaliar os resultados das administrações do *pool* de itens pela análise quantitativa de itens e qualitativa adicional.

5. Adicionar, eliminar e/ou modificar itens conforme o necessário, com base nas análises qualitativa e quantitativa.
6. Conduzir administrações adicionais para verificar se as estatísticas dos itens se mantêm estáveis com diferentes grupos – um procedimento conhecido como *validação cruzada* – até que um conjunto satisfatório de itens seja obtido.
7. Padronizar ou fixar o tamanho do teste e o seqüenciamento dos itens, bem como os procedimentos de administração e avaliação a serem usados em sua forma final, com base nas análises anteriores.
8. Administrar o teste a uma nova amostra de indivíduos – cuidadosamente selecionados para representar a população de testandos para os quais o teste se destina – de modo a desenvolver dados normativos ou critérios de desempenho, índices de fidedignidade e validade dos escores, bem como estatísticas de item para a versão final do teste.
9. Publicar o teste em sua forma final, juntamente com um manual de administração e avaliação, documentação dos dados de padronização, estudos de fidedignidade e validade e os materiais necessários para a administração e a avaliação (AERA, APA e NCME, 1999, Capítulo 6).

Para um teste que vai ser publicado comercialmente, esses passos podem levar anos e podem ter que ser repetidos várias vezes se os resultados iniciais não forem plenamente adequados. Além disso, a maioria dos testes padronizados são revisados de tempos em tempos devido à obsolescência gradual das normas, critérios de desempenho e conteúdos. Alguns instrumentos padronizados usados em testagens em larga escala, como o SAT (antes conhecido como *Scholastic Aptitude Test*), são quase continuamente revisados e refinados. Obviamente, testes usados apenas em contextos limitados, como salas de aula ou estudos específicos de pesquisa, não passam por um processo tão rigoroso. Mesmo assim, precisam ser elaborados com cuidado segundo especificações pré-estabelecidas, e seus resultados devem ser psicometricamente defensáveis – em termos de características de itens, validade e fidedignidade – se quiserem atingir seus objetivos com sucesso.

Não esqueça

É de praxe categorizarmos os testes de modo amplo nas áreas de "habilidade" e "personalidade". Essa distinção tradicional – usada repetidamente neste capítulo e em todo este livro, por razões de conveniência – se baseia na noção de que alguns testes pretendem avaliar primariamente aspectos do comportamento cognitivo, enquanto outros buscam avaliar aspectos do comportamento relacionados ao funcionamento emocional. No entanto, ao considerarmos o tópico dos itens de teste, e dos testes como um todo, é importante lembrarmos que os fatores cognitivos e emocionais são inseparáveis, e que as amostras de comportamento refletem todos os aspectos do funcionamento de uma pessoa.

TIPOS DE ITENS DE TESTE

A variedade de itens que compõem os testes psicológicos é imensa e desafia uma categorização fácil. Os itens de teste, assim como os testes como um todo, podem diferir em termos de conteúdo e formato, bem como do meio pelo qual são administrados, a maneira como são avaliados e o tipo de processamento que demandam por parte dos testandos. Uma das distinções mais básicas entre os itens de teste diz respeito ao tipo de resposta que eles requerem, podendo ser classificados em duas categorias amplas, quais sejam, itens de resposta selecionada e itens de resposta construída. Os testes criados para avaliar habilidades e aqueles que pretendem avaliar a personalidade usam um ou ambos os tipos de itens, dependendo da natureza das amostras de comportamento necessárias para seu objetivo. Da mesma forma, itens de ambos os tipos podem ser usados em testagens individuais ou coletivas. O quadro Consulta Rápida 6.2 fornece informações sobre onde obter amostras de vários tipos de itens de teste.

CONSULTA RÁPIDA 6.2

Como localizar exemplos de vários tipos de itens de teste

O site do *Educational Testing Service* (http://www.ets.org) traz *links* para vários de seus principais programas de testagem. Amostras de itens estão disponíveis nas seções de preparação de testes para esses programas. Por exemplo:

- A página do *The College Board* (http://www.collegeboard.com) oferece mini-testes Verbais e Matemáticos do SAT – com o tipo de itens usados no verdadeiro SAT – que os usuários em potencial podem fazer gratuitamente.
- A página do *Graduate Record Examinations* (GRE) (http://gre.org) oferece uma versão prática do Teste Geral GRE.
- Perguntas das várias seções do *Test of English as a Foreign Language* (TOEFL) podem ser baixadas da página do TOEFL (http:/www.toefl.org).

Amostras de itens de vários testes de habilidade e de personalidade estão disponíveis em http://www.schuhfried.co.at, página da Schuhfried Company, uma organização austríaca que comercializa programas para a administração de testes por computador.

Imagens dos instrumentos usados em muitos testes de sensibilidade e desempenho podem ser encontradas no catálogo *Evaluation and Assessment* da Lafayette Instrument Company, disponível em http://www.licmef.com/downloads.htm#cat

Os catálogos impressos das editoras contêm descrições do conteúdo de testes e, muitas vezes, amostras de itens. Os catálogos disponíveis na Internet tendem a listar e descrever os testes, mas a não incluir amostras. Os usuários podem obter catálogos impressos contatando as editoras (ver Apêndice B).

Itens de resposta selecionada

Os *itens de resposta selecionada*, também conhecidos como *itens objetivos* ou *de resposta fixa*, são de natureza fechada. Eles apresentam um número limitado de alternativas entre as quais o testando deve escolher. Nos testes de habilidade, os itens desse tipo incluem os de múltipla escolha, verdadeiro-falso, ordenamento e combinação e itens que pedem o rearranjo das opções apresentadas. Tipicamente, os itens objetivos em testes de habilidade são avaliados de maneira simples como certo-errado, embora também seja possível atribuir um crédito parcial para certas opções de resposta. Exemplos de vários tipos de itens de habilidade de resposta selecionada, tanto de testes padronizados quanto de testes elaborados por professores, podem ser citados por qualquer pessoa que tenha sido escolarizada nos Estados Unidos nas últimas décadas. Os itens de resposta selecionada nem sempre foram usados com tanta freqüência nos Estados Unidos como hoje, nem são tão freqüentes em todos os países. Na verdade, muitas críticas à testagem padronizada na educação nos Estados Unidos giram em torno do uso generalizado de itens de teste de resposta selecionada – especialmente itens de múltipla escolha – e sua fraqueza percebida do ponto de vista pedagógico. Muitos críticos da "testagem padronizada" usam este termo de modo vago e incorreto, como sinônimo para testes que empregam o formato dos itens de múltipla escolha (Mitchell, 1992; Sacks, 1999).

Nos testes de personalidade, os itens objetivos podem ser dicotômicos ou politômicos. Os itens *dicotômicos* requerem uma escolha entre duas alternativas (p. ex., verdadeiro-falso, sim-não, gosto-não gosto, etc), enquanto os itens *politômicos* apresentam três ou mais (geralmente um número ímpar, como 5 ou 7) respostas alternativas a uma afirmação. Estas alternativas são tipicamente escalonadas em termos de grau de aceitação (p. ex., *gosto, indiferente, não gosto*), intensidade de concordância (p. ex., de *concordo totalmente* a *discordo totalmente*), freqüência (p. ex., de *nunca* a *quase sempre)*, etc. –, com o ponto intermediário geralmente significando uma posição neutra, incerta ou de meio termo.

Escolha forçada

Itens objetivos que requerem que os testandos escolham qual entre duas ou mais alternativas é a mais ou menos característica deles são denominados *itens de escolha forçada*. Cada alternativa em um conjunto de escolhas forçadas representa um constructo diferente, mas elas são pareadas em termos de desejabilidade social, de modo que pareçam igualmente atraentes ou indesejáveis aos testandos. Este tipo de item é usado principalmente em inventários multidimensionais de personalidade (isto é, inventários que avaliam diversos constructos de personalidade) para controlar a tendência dos testandos a responder na direção que eles percebem como socialmente mais desejável. No entanto, as alternativas de escolha forçada freqüentemente são pareadas de tal forma que cada escolha feita pelo testando limita a possível amplitude de seus escores em outro dos constructos ou traços avaliados pelo teste multidimensional. Quando isso acontece, os escores resultantes são de natureza ipsativa e não podem ser interpretados de modo normativo.

Escores *ipsativos* são essencialmente números ordinais que refletem as classificações dos testandos nos constructos avaliados pelas escalas dentro de um formato de teste de escolha forçada. Isso significa que a magnitude relativa dos escores em cada uma das escalas de um teste desses pode ser medida apenas em comparação com os outros escores obtidos pelo mesmo indivíduo nas outras escalas do teste, e não com escores obtidos pelos grupos normativos. Além disso, o formato de escolha forçada não pode eliminar totalmente a influência da desejabilidade social e pode até mesmo interferir no *rapport* (ver Capítulo 7 para uma definição de *rapport*). Apesar destes problemas, os itens de escolha forçada ainda são usados, especialmente em inventários de interesse e em testes – como o *Myers-Briggs Type Indicator (MBTI)* – cujo objetivo primário é classificar os indivíduos em categorias mutuamente exclusivas. Alguns testes no formato de escolha forçada (p. ex., o *Jackson Vocational Interest Survey*) evitam o problema da ipsatividade pareando alternativas derivadas de dois conjuntos diferentes de escalas paralelas, de modo que a amplitude de escores em cada escala não seja restrita.

Vantagens dos itens de resposta selecionada

Os itens objetivos são sem dúvida o tipo de item de teste mais popular e mais usado. Suas vantagens derivam da facilidade e da objetividade com que podem ser avaliados, que resultam em economia significativa de tempo e melhoram a fidedignidade dos testes; a questão do erro de avaliação é virtualmente inaplicável a itens deste tipo, exceto por meio de erros de verificação. Além disso, os itens de resposta selecionada fazem um uso eficiente do tempo de testagem, porque um número maior de itens pode ser administrado em qualquer intervalo, o que não acontece com os itens de resposta construída. Embora também possam ser administrados individualmente, a maioria dos testes que usa itens de resposta selecionada é destinada a testagens em grupo.

Todas as respostas a itens objetivos podem ser transformadas em uma escala numérica para fins de avaliação de maneira fácil e fidedigna, um fato que simplifica muito a análise quantitativa destes itens. Nos testes de habilidade, as respostas corretas e incorretas geralmente recebem valores de 1 e 0, respectivamente; às vezes, variações como 2, 1 ou 0 estão disponíveis para crédito parcial. Nos testes de personalidade, os itens dicotômicos também são pontuados com 1 ou 0, dependendo se a resposta do testando vai ou não na direção do constructo que o teste quer avaliar. As alternativas apresentadas em itens politômicos ou de múltiplas respostas podem ser traduzidas em várias escalas numéricas, como 5, 4, 3, 2, 1 ou +2, +1, 0, –1, –2, ou reduzidas a um formato binário de pontuação (1 ou 0) pela fusão de categorias.

Desvantagens dos itens de resposta selecionada

Apesar de suas vantagens, os itens de resposta selecionada são mais suscetíveis a certos problemas do que os itens de resposta construída. Nos testes de habilidade,

o principal problema ligado aos itens objetivos gira em torno da questão da adivinhação, ou "chute". A possibilidade de se adivinhar corretamente está sempre presente quando as respostas simplesmente têm que ser selecionadas. Nos sistemas dicotômicos, como nos de verdadeiro-falso, a probabilidade de se adivinhar corretamente é 50% substancial. Quando os testandos "chutam" as respostas corretas de itens objetivos, a quantidade de erro introduzida em seus escores varia dependendo de que fatores (p. ex., puro acaso, conhecimento parcial, formulação dos itens, etc.) foram responsáveis pelas respostas corretas. Da mesma forma, respostas incorretas para itens objetivos podem facilmente ocorrer como resultado de pressa, desatenção, descuido, simulação, ou outros fatores do acaso, sem qualquer relação com o nível de conhecimento do testando ou sua habilidade na área explorada pelo item.

Nos testes de personalidade, os objetivos dos itens de resposta selecionada podem ser facilmente subvertidos por um número ainda maior de motivos. Estes incluem não apenas respostas aleatórias ou negligentes, mas também conjuntos de respostas enganosas, sejam intencionais ou não. Dependendo do contexto no qual a testagem acontece e a disposição mental particular do testando, suas respostas a um teste de personalidade podem ser enganosas em direções negativas ou positivas. Por exemplo, indivíduos submetidos a um inventário de personalidade no contexto de uma seleção de emprego naturalmente vão optar por se apresentarem de forma muito mais favorável do que pessoas que estão sendo testadas para determinar se uma doença psiquiátrica pode ser usada como fator atenuante de culpa no julgamento de um crime. Claramente, as respostas a itens objetivos de testes de personalidade podem ser manipuladas mais facilmente pelos testandos do que as respostas a itens de testes de habilidade, que não podem ser simuladas a não ser por fraude (para uma análise minuciosa de muitos aspectos das fraudes em testes, ver Cizek, 1999). Devido à sua vulnerabilidade à distorção, muitos inventários de personalidade usam conjuntos especiais de itens, escalas de validade ou outros dispositivos criados especificamente para detectar respostas enganosas ou negligentes.

Todas as possibilidades expostas acima podem diminuir a fidedignidade e a validade dos escores de teste. Embora os itens de resposta construída também sejam suscetíveis a alguns desses problemas, a adivinhação em testes de habilidade de resposta construída é mais difícil e, portanto, menos provável. Responder de forma enganosa em técnicas projetivas e outras ferramentas de avaliação da personalidade de resposta construída representa um desafio maior para os testandos. Além disso, a natureza relativamente não-estruturada desses instrumentos é tal que, mesmo quando tentam conscientemente enganar, os testandos podem estar fornecendo alguma informação útil.

Preparar itens de resposta selecionada é uma tarefa difícil e demorada, que requer habilidades especializadas de desenvolvimento de testes e redação de itens, além de grande familiaridade com o constructo ou tema de que trata o teste. Itens objetivos mal preparados podem inadvertidamente dar pistas aos testandos ou ser formulados de maneira que beneficie ou prejudique um subconjunto deles. Itens de múltipla escolha mal-escritos, em particular, muitas vezes incluem alternativas

que são (a) gramaticalmente incompatíveis com o corpo do item, (b) suscetíveis a várias interpretações devido à formulação imprecisa ou (c) tão ridículas que podem facilmente ser descartadas.

Por fim, os itens de resposta selecionada são claramente menos flexíveis do que os itens de resposta construída, em relação à possível gama de respostas. Por isso, não oferecem a oportunidade de avaliar características de um testando que podem ser especiais ou singulares, ou que ficam de fora da gama de alternativas apresentadas.

Itens de resposta construída

A característica essencial dos *itens de resposta construída*, também conhecidos como *itens de resposta livre*, é que eles são abertos. Sua variedade é ilimitada, porque as respostas construídas podem envolver amostras de redação, respostas orais livres, desempenhos de qualquer tipo e produções de qualquer espécie.

Nos testes de habilidade, o tipo mais comum de item de resposta construída são as perguntas relativas a dissertar e completar lacunas. As únicas restrições pertinentes aos itens de resposta livre nos testes psicológicos são as condições impostas pelas instruções do teste. Instruções minuciosas e regras de procedimento são indispensáveis para a administração e a avaliação padronizada de todos os testes, incluindo os de resposta livre. As instruções para a administração de testes de resposta construída devem incluir estipulações sobre temas como (a) limites de tempo; (b) meio, modo ou tamanho exigido da resposta e (c) se o acesso a materiais ou instrumentos pertinentes ao teste (p. ex., obras de consulta, calculadoras, computadores, etc) é permitido.

Entrevistas, questionários de dados biográficos e observações comportamentais são ferramentas para a avaliação da personalidade que muitas vezes se valem de respostas abertas. Na testagem da personalidade propriamente dita, o uso de respostas construídas se limita principalmente às *técnicas projetivas*. Estes métodos geralmente requerem que os testandos respondam a estímulos ambíguos na forma de imagens (incluindo manchas de tinta) ou materiais verbais, como palavras ou frases incompletas. Algumas técnicas projetivas demandam auto-expressão por meio de desenhos ou de outros tipos de desempenho. A idéia básica em todos estes métodos – que tiveram origem e são usados principalmente no contexto clínico – é apresentar aos testandos tarefas com um mínimo de estrutura para que eles possam responder com o máximo de liberdade e, nesse processo, revelar aspectos significativos de suas personalidades. Em contraste com inventários, questionários e outros instrumentos objetivos que avaliam constructos ou constelações de traços específicos relacionados à personalidade, as técnicas projetivas fornecem uma abordagem menos focal, mais global da avaliação.

De modo geral, as vantagens e desvantagens dos itens de resposta construída são opostas àquelas apresentadas pelos itens de resposta selecionada. Não obstante, merecem ser mencionadas.

Vantagens dos itens de resposta construída

Mesmo quando não são administrados individualmente, os itens de resposta construída fornecem amostras mais ricas do comportamento dos examinandos e permitem a observação de suas características singulares. Os itens abertos oferecem uma gama mais ampla de possibilidades e abordagens mais criativas da testagem e avaliação do que os itens de resposta selecionada. Além disso, as tarefas de resposta construída produzem amostras autênticas do comportamento dos testandos em domínios específicos, e não meras escolhas entre alternativas pré-fabricadas. Se o que se deseja é avaliar habilidades de escrita, memória, conhecimento matemático, habilidades mecânicas, capacidade de liderança ou qualquer outro tipo de desempenho, amostras reais do que um indivíduo pode fazer são o único padrão incontestável.

Desvantagens dos itens de resposta construída

As principais desvantagens dos itens de resposta construída estão relacionadas à fidedignidade dos escores e, como conseqüência, também à validade (ver a seção sobre a relação entre fidedignidade e validade no Capítulo 4). Estas desvantagens têm origem no modo como as respostas construídas são pontuadas e nas limitações práticas que as respostas deste tipo impõem à duração dos testes.

A avaliação de respostas construídas, tanto em testes de habilidade quanto nos de personalidade, sempre é uma questão mais demorada e complexa do que a de respostas selecionadas, porque um certo grau de subjetividade invariavelmente é necessário. Mesmo quando as *rubricas de avaliação* (instruções que especificam os critérios, princípios e regras a serem usados na avaliação e fornecem exemplos ilustrativos) são preparadas e aplicadas com cuidado, sempre existe a possibilidade de que uma resposta vá ser avaliada diferentemente por diferentes avaliadores devido à sua singularidade ou a algum outro fator. Verificar a fidedignidade de escores atribuídos por diferentes avaliadores é um aspecto indispensável e oneroso do uso de testes com itens de resposta construída. Embora as diferenças entre avaliadores não possam ser completamente eliminadas, elas certamente podem ser minimizadas por meio de procedimentos de avaliação minuciosos, explícitos e testados, bem como pelo treinamento adequado dos avaliadores.

A pontuação de respostas construídas coletadas com as ferramentas projetivas usadas na avaliação da personalidade representa um desafio especial, uma vez que a subjetividade do avaliador pode entrar em jogo de mais formas do que na pontuação de respostas construídas em testes de habilidade. Além disso, as técnicas projetivas se prestam mais ao uso de métodos informais e muitas vezes idiossincráticos de administração e avaliação, o que pode enfraquecer ainda mais sua integridade psicométrica (Lanyon e Goodstein, 1997, Capítulo 4).

A duração do teste é outro fator que afeta a fidedignidade dos escores de testes que usam respostas construídas. Como estas respostas requerem mais tempo de administração e avaliação, o número de itens que podem ser incluídos nos testes de resposta construída geralmente é muito menor do que nos de resposta seleciо-

nada. Como foi discutido no Capítulo 4, não havendo outras falhas, os testes mais curtos são mais propensos a erros de amostragem de conteúdo e produzem escores menos consistentes do que os testes mais longos. Portanto, do ponto de vista da consistência interna, os testes de resposta construída também tendem a ser menos fidedignos do que os de resposta selecionada.

Uma outra complicação pertinente aos itens de resposta construída diz respeito ao *tamanho da resposta*. Como as respostas mais longas contêm mais material do que as mais curtas, variações no tamanho das respostas construídas podem afetar consideravelmente os escores. Isso é especialmente pertinente às técnicas projetivas, porque respostas projetivas mais longas – ou mais elaboradas – provavelmente contêm mais elementos passíveis de serem avaliados (isto é, psicologicamente significativos) do que as mais curtas. Além disso, instrumentos projetivos que permitem variabilidade no número e no tamanho das respostas representam um fator complicador adicional na investigação de suas propriedades psicométricas devido à falta de uniformidade entre os testandos. O teste de Rorschach é o exemplo mais proeminente deste problema, conforme evidenciado pela persistente polêmica a respeito do impacto da produtividade de respostas na sua avaliação e interpretação (Groth-Marnat, 1997, p.399; Meyer, 1992).

ANÁLISE DE ITENS

Nas últimas décadas, o campo do desenvolvimento e delineamento de testes e as técnicas de análise de itens têm passado por uma transição gradual que vem alterando fundamentalmente a natureza dos testes psicológicos. Esta transição se deve, em parte, à facilidade e à eficácia com que os dados de teste podem ser coletados, armazenados, analisados, recuperados e disseminados com o uso de computadores. Além disso, desde os anos de 1960, a metodologia fornecida pelas novas abordagens à construção de testes psicológicos, conhecida coletivamente como teoria da resposta ao item (TRI) ou teoria do traço latente, tem complementado de forma eficaz – e em alguns casos substituído – os métodos tradicionais de construção e delineamento de testes. Embora os métodos da TRI também sejam usados no desenvolvimento de testes de lápis e papel e testes computadorizados de tamanho fixo, sua vantagem mais destacada em relação à metodologia tradicional é que eles permitem um formato mais flexível e eficiente pela *testagem adaptativa computadorizada* (TAC). Na TAC, as seqüências de itens podem ser adaptadas individualmente em nível de habilidade do testando ou de sua posição naquele traço que o teste pretende avaliar, com base em suas respostas anteriores. A seguir, apresentaremos inicialmente os procedimentos tradicionais de análise de itens, seguidos de uma discussão sobre a metodologia da TRI.

Métodos quantitativos de análise de itens

Para os testes psicológicos em geral, o aspecto mais importante da análise quantitativa de itens está centrado nas estatísticas que abordam a *validade de item*. A

questão que os índices de validade de item tentam responder é se específico é importante dentro de um teste por coletar as informações pertinentes à finalidade deste. Os psicometristas geralmente se referem às estatísticas de validade de item como índices de *discriminação de item,* porque seu papel é revelar o grau em que ele diferencia com precisão os testandos em relação aos traços ou comportamentos que o teste pretende avaliar. Para os testes de habilidade, em particular, a análise de itens inclui procedimentos para medir duas outras características dos itens que influenciam sua validade, quais sejam, a *dificuldade de item* e *justiça (imparcialidade) de item*. Todas estas características podem ser avaliadas quantitativa e qualitativamente. A avaliação qualitativa geralmente é realizada por especialistas que examinam o conteúdo dos itens com atenção à sua adequação e nível de dificuldade, bem como à demonstração verdadeira dos objetivos que foram especificados para o teste. O conteúdo também é examinado do ponto de vista do potencial de injustiça ou ofensa a qualquer grupo específico de testandos. A avaliação quantitativa da dificuldade e discriminação dos itens é realizada por estatísticas que avaliam o desempenho dos testes quando administrados ao tipo de testandos para os quais foram idealizados.

Advertência

O termo *discriminação* adquiriu uma conotação negativa no uso diário devido à associação freqüente com o tratamento injusto dado às mulheres e minorias raciais.
Em contraste, no campo da psicometria, a discriminação é considerada uma característica desejável dos itens de teste. Ela se refere ao grau em que os itens produzem respostas que diferenciam com precisão os testandos, ao longo das dimensões que os testes pretendem avaliar.

Dificuldade de item

O papel da dificuldade de item na testagem de habilidades

Dada a proposição auto-evidente de que o nível de dificuldade de um teste como um todo é uma função dos níveis de dificuldade dos itens individuais que o compõem, segue-se que um teste fácil é composto de itens fáceis, e um teste difícil é composto de itens difíceis. Esta premissa aparentemente simples torna-se um pouco mais complicada quando consideramos que a dificuldade é uma questão relativa. A dificuldade de um item de teste não depende apenas de sua simplicidade ou acessibilidade intrínsecas, mas também do nível de habilidade do testando. Por exemplo, o uso correto do verbo *être* (ser/estar) – o verbo mais comum no francês – é uma tarefa muito mais fácil para um aluno de um curso avançado de francês do que para um iniciante no estudo dessa língua. Portanto, para calibrarmos adequadamente o nível de dificuldade de um teste, são necessários índices da dificuldade *relativa* dos itens para um ou mais grupos relevantes de testandos. Os criadores de testes usam esses índices para determinar a propriedade dos itens para a

Não Esqueça

Da mesma forma que os referenciais para a interpretação de escores de teste, discutidos no Capítulo 3, podem ser normativos ou referenciados no critério, a dificuldade dos itens de teste pode ser determinada em bases absolutas ou relativas. Ambos os aspectos precisam ser considerados no processo de construção e desenvolvimento de testes com referência específica à população pretendida de testandos e finalidade de cada instrumento.

população e a finalidade para a qual o teste se destina, bem como para decidir em que posição os itens devem ser colocados dentro do teste.

Como é medida a dificuldade de um item?

Durante os estágios iniciais do desenvolvimento de um teste, quando o *pool* de itens é gerado, os autores podem medir a dificuldade dos itens criada com base em padrões mais ou menos objetivos definidos nas especificações estabelecidas para o teste ou com base em critérios de consenso entre especialistas no tema ou habilidade cognitiva abordada. Por exemplo, um padrão que pode ser aplicado para calibrar a dificuldade das palavras é a freqüência com que elas são usadas em uma determinada língua. Assim, em um teste de vocabulário, os itens fáceis são palavras empregadas com freqüência pelos falantes da língua em questão, enquanto que os itens mais difíceis consistem em palavras que ocorrem raramente e com as quais a maioria dos testandos não estaria familiarizada. Da mesma forma, em um teste de aritmética, itens individuais podem ser alinhados em termos de dificuldade com base na complexidade evidente das operações que exigem, como multiplicação de números inteiros ou frações, etc.

Quando um conjunto de propostas é administrado a um ou mais grupos, índices quantitativos da dificuldade de item, que contempla esta questão de uma perspectiva normativa, também podem ser obtidos. Ao analisar itens de teste do ponto de vista normativo, a informação essencial usada para determinar sua dificuldade é a percentagem de testandos que respondem corretamente, também conhecida como *proporção* (ou *percentagem*) *de acertos*, ou p para abreviar. Quanto maior a percentagem dos que acertam, mais fácil é o item. Como os valores p dos itens dependem inteiramente do nível de habilidade dos grupos aos quais eles são administrados, a composição desses grupos é muito importante e deve refletir a composição da população para a qual o teste se destina.

Os valores p são números ordinais que, como os postos de percentil, não representam unidades iguais. Por esta razão, desde que se pressuponha que o traço medido por um item tem distribuição normal, os valores p com freqüência são transformados em valores z, usando-se a Tabela de Áreas da Curva Normal (ver Apêndice C). Depois que os valores p são convertidos em valores z, os níveis relativos de dificuldade dos itens podem ser comparados entre vários grupos administrando-se um conjunto comum de itens – denominados *itens-âncora* – a dois ou

mais grupos. Fórmulas para estimar os níveis de dificuldade de itens adicionais entre os grupos em questão podem então ser derivadas com base nas relações estabelecidas entre os itens âncora. Este procedimento, conhecido como *escalonamento absoluto*, foi desenvolvido por Thurstone (1925). Ele permite que a dificuldade dos itens seja colocada em uma escala numérica uniforme para amostras de testandos em diferentes níveis de habilidade, como estudantes em várias séries escolares. O quadro Consulta Rápida 6.3 apresenta um exemplo numérico simples de como isto é feito, usando os resultados de cinco itens administrados a dois grupos. Como todos os cinco itens do exemplo têm valores *p* mais altos (e valores *z* mais baixos) para o Grupo B do que para o Grupo A, podemos concluir que o Grupo B está funcionando em um nível mais avançado do que o Grupo A na habilidade ou área de conteúdo explorada por esses itens. A Figura 6.1 retrata a dificuldade relativa dos cinco itens para os dois grupos e demonstra graficamente que os níveis de dificuldade dos cinco itens para os dois grupos se correlacionam forte e positivamente. O tipo de dados apresentado no quadro Consulta Rápida 6.3 e na Figura 6.1 pode ser usado para estimar os níveis de dificuldade de itens adicionais para um grupo, com base em seus valores de dificuldade para o outro, por meio de análise regressiva (ver Capítulo 5). Este tipo de procedimento é aplicado no equacionamento de testes e escores de testes por meio de testes-âncora e grupos de referência fixa (ver Capítulo 3).

CONSULTA RÁPIDA 6.3

Conversão da dificuldade de item de proporção dos que acertam (*p*) para unidades da curva normal (*z*)

A dificuldade de item pode ser representada em unidades da curva normal (valores z), desde que o traço medido por um item tenha distribuição normal.

O valor z para um item é derivado localizando-se a proporção de testandos que o acertam (isto é, seu valor p) na Tabela de Áreas da Curva Normal (ver Apêndice C): *valores p acima de 0,50* são encontrados na coluna 3 da tabela e recebem os valores z correspondentes com um sinal *negativo*; *valores p abaixo de 0,50* se localizam na coluna 4 da tabela e os valores z correspondentes têm valor *positivo*. Se p = 0,50, o valor z para o item é zero.

Exemplo numérico para cinco itens administrados a dois grupos:

Número do item	Grupo A Valor p^a	Grupo A Valor z^b	Grupo B Valor p^a	Grupo B Valor z^b
1	0,841	-1,00	0,894	-1,25
2	0,50	0,00	0,691	-0,50
3	0,067	+1,50	0,159	+1,00
4	0,023	+2,00	0,067	+1,50
5	0,977	-2,00	0,994	-2,51

[a]Os valores p representam a proporção de indivíduos dos grupos A e B que acertou cada item.

[b]Os valores z para os itens mais fáceis são grandes e negativos, enquanto que aqueles para os itens mais difíceis são grandes e positivos.

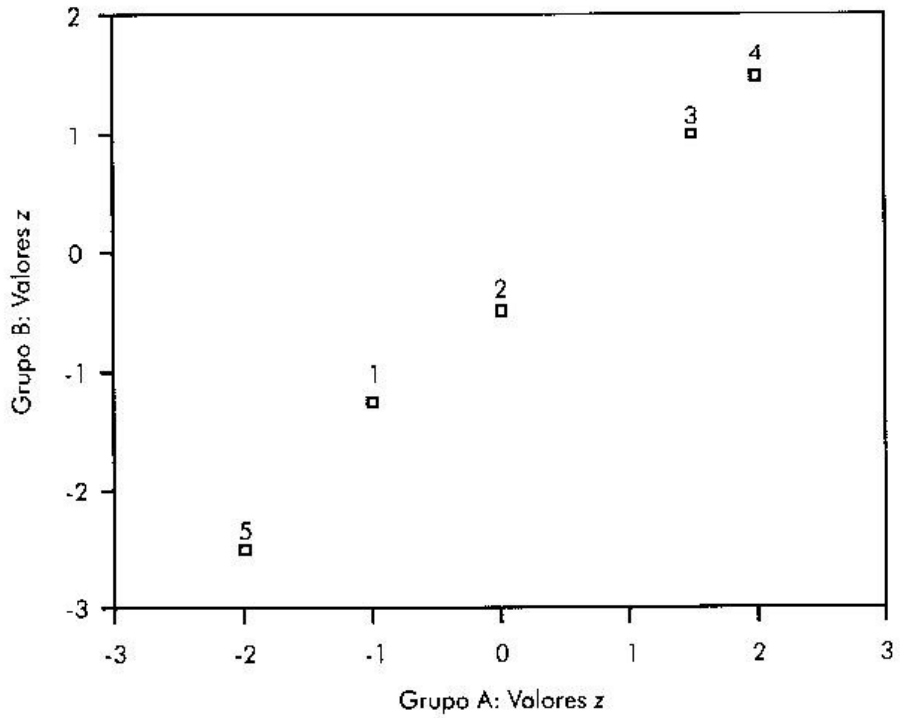

Figura 6.1 Diagrama de dispersão da dificuldade relativa de cinco itens para os Grupos A e B (ver quadro Consulta Rápida 6.3).

Níveis de dificuldade de item, níveis de dificuldade de teste e finalidade de teste

Para qualquer grupo de testandos, o escore médio em um teste é o mesmo que a dificuldade média de seus itens. Portanto, se a percentagem *média* dos que acertam (p) os itens de um teste for de 80%, o escore médio no teste também vai ser de 80%. A significância da relação entre a dificuldade de item, a finalidade do teste e o nível de habilidade da população de testandos para os quais o teste se destina pode ser esclarecida por alguns exemplos.

- Os testes de realização de sala de aula têm a finalidade de avaliar o grau em que os alunos de uma turma dominaram o conteúdo de uma disciplina. Na maioria dos contextos acadêmicos, uma nota na faixa de 70 a 79% é considerada média. Para atingir esta média, a maioria dos itens dos testes de sala de aula deve estar ao alcance da maioria dos alunos, desde que estes tenham dominado o conteúdo da disciplina em um nível que o instrutor considere médio. Esses testes podem incluir alguns itens – envolvendo conceitos que foram enfatizados no curso – que a turma toda vai responder corretamente (p = 1,00), embora tais itens não diferenciem entre os testandos (ver a Tabela 6.1 mais adiante neste capítulo). Porém, um item que ninguém responde corretamente (p = 0) não é desejável em

um teste de sala de aula, pois indica que nem mesmo os melhores alunos conseguiram compreendê-lo.

- Por outro lado, um teste com o objetivo de triar um grande grupo de candidatos para selecionar os melhores 10% deles pode ter itens que se agrupam em torno de um valor *p* de 0,10, ou 10%. Este teste seria considerado difícil demais por todos, exceto pelos candidatos mais qualificados que possuíssem uma grande quantidade dos conhecimentos ou das habilidades que o teste buscasse avaliar.
- Muitos testes de habilidade são delineados de modo a diferenciar ao máximo os indivíduos de uma determinada população em termos de um traço cognitivo que se supõe ter distribuição normal, como a inteligência geral ou habilidades verbais. Nestes testes, a gama de dificuldade dos itens deve ser suficientemente ampla para acomodar ao mesmo tempo os indivíduos mais e menos capazes da população potencial de testandos, e o valor *p* dos itens deve se agrupar em torno de 0,50 (ou 50%) para fornecer o máximo de diferenciação entre os testandos. Itens com valores *p* extremos (isto é, próximos de 0 ou 1,00) devem ser evitados porque não diferenciam os testandos e, portanto, são "excesso de bagagem". Além disso, se os indivíduos que pertencem à população para a qual um teste desta natureza é destinado são capazes de acertar todos os itens ou nenhum deles, seus escores são indeterminados. Como foi discutido no Capítulo 3, quando os itens são fáceis demais para um determinado grupo, diz-se que o teste tem *teto insuficiente*, e sua distribuição de escores terá declividade negativa; quando os itens são difíceis demais para um grupo, o teste têm um *solo inadequado*, e sua distribuição de escores tem declividade positiva. A Figura 2.4 mostra exemplos de distribuições assimétricas.

Distratores e dificuldade

Em testes que usam itens de múltipla escolha, as alternativas incorretas, ou *distratores*, podem ter uma influência grande na dificuldade dos itens em dois aspectos importantes. Em primeiro lugar, o número de distratores afeta diretamente os índices de dificuldade de item porque a probabilidade de se adivinhar a resposta correta é mais alta quando o número de opções é menor. Além disso, a dificuldade de item também é afetada pelo calibre dos distratores. Um item de múltipla escolha ideal é aquele em que (a) a alternativa correta é óbvia para o testando que conhece a resposta e (b) os distratores parecem igualmente plausíveis para aqueles que não a conhecem. Itens como este são difíceis de construir. Quando parecem obviamente errados, são mal formulados ou são muito mais longos ou curtos do que a alternativa correta, os distratores fornecem pistas que examinandos espertos podem usar para eliminar alternativas e selecionar a resposta correta, mesmo sem conhecê-la realmente. Para evitar este e outros problemas na criação de itens de múltipla escolha, os autores de testes devem seguir as diretrizes para a elaboração de itens fornecidas em obras como a de Haladyna (1999). Depois que um teste é administrado, uma análise dos distratores também deve ser conduzida, começan-

do pelo cálculo do número de testandos que selecionou cada distrator. O exame cuidadoso da freqüência com que os vários distratores foram escolhidos por testandos de diferentes níveis de habilidade serve para detectar possíveis falhas nos itens. Se o teste ainda está em desenvolvimento, os distratores que não estão funcionando adequadamente (p. ex., aqueles que não são escolhidos por ninguém ou que são escolhidos com maior freqüência por testandos com altos níveis de habilidade) devem ser descartados e substituídos.

A dificuldade de item é um conceito relevante na testagem da personalidade?

As tarefas que compõem os testes de personalidade podem não ser delineadas para avaliar o funcionamento cognitivo, mas envolvem processos cognitivos. Em instrumentos de resposta selecionada, como inventários e questionários de personalidade, os processos cognitivos relevantes estão relacionados à capacidade do testando de compreender os itens. Portanto, os níveis de vocabulário e habilidade de leitura dos testandos em potencial precisam ser levados em conta ao se elaborar esses itens. Tarefas projetivas, por outro lado, envolvem uma certa quantidade de proficiência na modalidade em que as respostas deverão ser expressas. A maioria dos instrumentos projetivos requer alguma habilidade em expressão verbal, desenho ou algum outro tipo de desempenho. Assim sendo, a relativa dificuldade ou facilidade das tarefas projetivas para vários tipos de examinados também deve ser considerada no desenvolvimento, administração e interpretação desses instrumentos.

Discriminação de item

A *discriminação de item* se refere ao grau em que os itens produzem respostas que diferenciam com precisão os testandos em termos dos comportamentos, conhecimentos ou outras características que um teste – ou subteste – pretende avaliar. Na vasta maioria dos casos, o poder discriminatório é a qualidade mais básica que os itens devem ter para serem incluídos em um teste. No processo de desenvolvimento, os índices de discriminação de item – também conhecidos como índices de validade de item – são obtidos usando-se algum critério ou indicador da posição do testando em relação ao constructo que o teste avalia. Os critérios empregados para este fim podem ser (a) critérios internos com respeito ao teste em desenvolvimento (isto é, escore total do teste), (b) critérios externos do mesmo tipo que os usados para validar testes como um todo descritos no Capítulo 5 (p. ex., idade, escolaridade, pertencimento a grupos diagnósticos, ou ocupacionais contrastados, etc) ou (c) combinações de critérios internos e externos.

Critérios de validação de item

A escolha dos critérios contra os quais os itens de um teste são validados depende de seus objetivos. Os testes de habilidade requerem critérios relacionados às áreas

de conteúdo ou habilidades avaliadas; os testes de personalidade requerem critérios pertinentes aos traços ou aspectos do comportamento com que lidam. A qualidade e a propriedade dos critérios usados na validação dos itens de teste têm importantes conseqüências para a seleção dos itens que serão mantidos em um teste e, conseqüentemente, para a fidedignidade e a validade de seus escores.

Quando critérios externos ao teste são usados para validar itens, a validade dos escores como um todo é melhorada; quando o critério interno do escore total do teste é usado na validação, a homogeneidade do teste aumenta, e, com isso, os índices de fidedignidade baseados na consistência entre itens são melhorados. No desenvolvimento de testes que avaliam um único traço unidimensional, como vocabulário ou depressão, o escore total pode ser usado para validar itens. Esta prática se baseia na premissa de que todos os itens dentro desses testes devem se correlacionar altamente com o escore total do teste e uns com os outros. Por outro lado, no desenvolvimento de testes que avaliam constructos complexos e multifacetados, como a inteligência, os itens são validados contra critérios externos que também são mais globais. Como podem estar avaliando diferentes aspectos de um constructo complexo, os itens desses testes não precisam necessariamente se correlacionar altamente uns com os outros, e seu grau de correlação com o escore total pode variar. A maioria das escalas de inteligência, por exemplo, inclui uma mistura de itens que explora os vários tipos de habilidades associadas a esse constructo – como habilidades verbais, numéricas, espaciais e de raciocínio lógico – e fornece escores compostos que incorporam o desempenho em todos os tipos de itens e são validados contra critérios externos, como realização educacional. Em instrumentos deste tipo, os itens geralmente são agrupados em subtestes com conteúdo homogêneo que são avaliados separadamente (ver Tabela 4.1 e Figura 4.1 no Capítulo 4).

Como vimos, embora a validade externa e a consistência interna sejam metas desejáveis na construção de um teste, a natureza dos constructos avaliados por ele pode não permitir que ambas sejam atingidas concomitantemente. Além das limitações impostas pela finalidade do teste, a validação externa dos itens também pode não ser prática, devido à indisponibilidade ou inacessibilidade de dados de critério externo. Um exemplo típico deste tipo de situação é fornecido pelos itens de testes de sala de aula elaborados por professores, como aqueles apresentados na Tabela 6.1. Ao conduzirem análises de itens destes testes, os professores não podem usar outro critério que não o do escore total no teste, por uma questão de justiça. Os testes de sala de aula têm por finalidade avaliar o domínio das habilidades e dos conteúdos abordados dentro de uma disciplina, e seus escores não devem ser atrelados a qualquer outro fator além daquele do domínio obtido pelos alunos dos objetivos especificados.

Estatísticas de discriminação de item

Todos os procedimentos estatísticos usados para medir o grau em que os itens discriminam em termos de um critério requerem informações sobre (a) o desempenho do item e (b) a posição no critério para os indivíduos das amostras das quais as estatísticas de discriminação de item são extraídas. As estatísticas tradicionais usa-

das para este fim são de dois tipos: a estatística do índice de discriminação (*D*) e uma variedade de índices correlacionais.

O *índice de discriminação (D)* é usado primariamente para itens de testes de habilidade que são avaliados como certo-errado, mas também pode ser aplicado na análise de itens de outros testes que usam avaliação binária. Para calcular o *D*, os testandos devem ser classificados em grupos distintos de critério com base em seu escore total no teste ou em algum indicador externo de sua posição no constructo avaliado por ele. É costume criar grupos de critério separando-se os testandos usados para análises de validade de item em dois grupos extremos, como, por exemplo, aqueles que ficam nos terços inferior e superior da medida de critério. Depois de criados os grupos superior e inferior do critério, a percentagem de indivíduos (*p*) dentro de cada grupo que acertou o item – ou respondeu na direção indicativa do constructo avaliado pelo teste – é calculada. O *índice de discriminação* é simplesmente a diferença na percentagem ou proporção de testandos nos grupos inferior e superior do critério que acerta um determinado item ou responde na direção esperada; *D* pode variar de +100 a –100 (ou de +1,00 a –1,00). Para os testes de habilidade, índices de discriminação positivos indicam que mais indivíduos no grupo superior do critério do que no grupo inferior acertaram o item, e que os valores mais desejáveis de *D* são aqueles mais próximos de +100. Valores negativos de *D* indicam que os itens em questão discriminam na direção oposta e precisam ser corrigidos ou descartados.

A Tabela 6.1 mostra os índices de discriminação de item de seis itens de um teste administrado a uma turma de testagem psicológica. Todos os alunos passaram no Item 1, o mais fácil dos seis (p = 100%), e somente 13% passaram no Item 6, o mais difícil. O Item 3, no qual passaram 38% dos alunos, foi um item relativamente difícil e o mais discriminante entre eles, com um valor *D* de 100. Os itens 4 e 5 foram relativamente fáceis (p = 75%), mas de valor questionável. O Item 4 não

Tabela 6.1 Dados de amostra de análise de item de um teste de sala de aula

	Percentagem que acertou (valor *p*)				
Número do item	Total do grupo	Grupo superior[a]	Grupo inferior[a]	Índice D (Superior-Inferior)	Correlação ponto-bisserial (r_{pb})[b]
1	100%	100%	100%	0	0,00
2	88%	100%	50%	50	0,67
3	38%	100%	0%	100	0,63
4	75%	50%	50%	0	0,13
5	75%	50%	100%	-50	-0,32
6	13%	50%	0%	50	0,43

[a]Os grupos superior e inferior do critério são compostos por alunos cujos escores no teste como um todo ficaram em 27% das extremidades superior e inferior, respectivamente, da distribuição de escores.

[b]Ponto-bisserial é um índice da correlação entre o desempenho de cada testando no item avaliado de forma dicotômica (certo-errado) e seus escores totais no teste.

Não Esqueça

A maioria dos índices de discriminação de item favorece itens de dificuldade intermediária. Por exemplo, se a percentagem que acerta (valor p) um item na amostra total for extrema (100% ou 0%), não poderá haver diferença nos valores p dos grupos inferior e superior do critério para aquele item, e seu índice *D* será 0. Por outro lado, quando o valor p para o grupo total for 50%, será possível ao índice *D* atingir seu valor máximo de +100, se todos os integrantes do grupo superior do critério e nenhum dos integrantes do grupo inferior acertarem. Portanto, para todos os testes cujo objetivo for determinar diferenças entre indivíduos em termos de alguma habilidade, itens centrados em torno de um nível de dificuldade de 50% são preferíveis.

discriminou entre os dois grupos extremos do critério, e o Item 5 teve que ser descartado porque seu índice *D* de –50 indicou que ele discriminava na direção errada.

Coeficientes de correlação de vários tipos também podem expressar a relação entre o desempenho em um item e a posição no critério, e, com isso, fornecer índices de discriminação de item. O tipo de coeficiente de correlação escolhido para calcular esses índices depende da natureza das duas variáveis que devem ser correlacionadas, que são os escores do item e as medidas de critério. Por exemplo, quando os escores são dicotômicos (p. ex., certo-errado) e a medida de critério é contínua (p. ex., escore total no teste), o coeficiente de correlação ponto-bisserial (r_{pb}) é o mais usado. Por outro lado, quando os escores de item e a medida de critério são ambos dicotômicos, é usado o coeficiente phi (F). Os coeficientes ponto-bisserial e phi podem variar de –1,00 a +1,00 e são interpretados da mesma forma que o *r* Pearson. Fórmulas para o cálculo dos coeficientes ponto-bisserial e phi e de diversos outros tipos de coeficientes de correlação usados na análise de discriminação de item estão disponíveis na maioria das obras básicas de estatística. Em todos os casos, correlações positivas altas indicam uma relação direta e forte entre item e critério; correlações negativas altas indicam uma relação inversa e forte entre item e critério, e correlações baixas indicam uma relação baixa entre os dois. A Tabela 6.1 também lista as correlações ponto-bisseriais entre itens e escores de teste para cada um dos seis itens discutidos anteriormente.

Uma observação a respeito da velocidade

Sempre que os testes de habilidade têm limites de tempo, a velocidade de desempenho afeta os escores em algum grau. Este tópico foi discutido no Capítulo 4, em conexão com os problemas que os testes com limites de tempo representam para o cálculo dos coeficientes de fidedignidade pelo método das metades, e também precisa ser considerado em relação às estatísticas de item. Com respeito à velocidade, os testes podem ser classificados em três tipos: testes de pura velocidade, testes de pura potência e testes que mesclam velocidade e potência.

- Os *testes de pura velocidade* simplesmente medem a velocidade com que os testandos conseguem realizar uma tarefa. Neles, a dificuldade é manipula-

da principalmente pelo controle do tempo. O nível de dificuldade de seus itens é uniforme e tende a estar dentro das capacidades dos indivíduos que se submetem a ele, mas os limites de tempo são tão curtos que a maioria dos testandos não consegue completar todos os itens. Por isso, na maioria dos casos, o escore total em um teste de pura velocidade é simplesmente o número de itens completados pelo testando. Se os testandos terminam todos os itens em um teste de pura velocidade, sua capacidade efetiva não foi determinada porque não há meios de saber quantos itens a mais poderiam ter sido completados se estivessem disponíveis.

- Os *testes de pura potência,* por outro lado, não têm limites de tempo. Neles, a dificuldade é manipulada aumentando ou diminuindo a complexidade dos itens. Sua faixa de dificuldade precisa ser suficientemente ampla para acomodar os níveis de habilidade de todos os testandos em potencial. Nos testes de potência, os itens são dispostos em ordem crescente de dificuldade para que todos os testandos sejam capazes de completar pelo menos alguns itens, mas os itens mais difíceis geralmente estão fora do alcance da maioria dos testandos. Um escore perfeito em um teste de pura potência sugere que o nível de habilidade do testando excede o nível de dificuldade dos itens mais difíceis. Nestes casos, o nível efetivo de habilidade do testando é indeterminado devido ao teto insuficiente do teste.
- A maioria dos testes de habilidade se encaixa em algum ponto entre os extremos do contínuo de pura velocidade/pura potência. Seus limites de tempo geralmente permitem que os testandos tentem completar todos ou a maioria dos itens. Como discutido anteriormente, a amplitude e a média específicas dos níveis de dificuldade dos itens de testes de habilidade dependem dos objetivos para os quais estes são empregados.

Em qualquer teste com limite rígido de tempo, os valores *p* e índices de discriminação dos itens são uma função de sua *posição* dentro dos testes mais do que de sua dificuldade ou validade discriminante intrínsecas. Isso acontece porque os itens na parte final de um teste no qual a velocidade desempenha um papel significativo são testados por menos testandos, e aqueles que tentam completar estes itens tendem a ser os mais capazes ou os que se apressam respondendo aleatoriamente. Como resultado, os índices de dificuldade e discriminação para os itens que ocorrem mais adiante nos testes de velocidade tendem a ser enganosos, e estratégias especiais precisam ser implementadas para se obter uma compreensão dos papéis específicos da velocidade e da dificuldade nesses testes.

Combinando dificuldade de item e discriminação de item

À luz da inter-relação entre a dificuldade de item e sua discriminação, o desenvolvimento da maioria dos testes de habilidade requer uma análise que combine ambas as características. Existem duas abordagens para esta finalidade. Os métodos mais antigos consistem na análise de regressão item-teste, e os mais recentes envolvem a teoria da resposta ao item (TRI).

Regressão item-teste

Para realizar análises de regressão item-teste, é necessário calcular a proporção de indivíduos, em cada nível de escore total, que acertou um determinado item. A Tabela 6.2 apresenta amostras de dados deste tipo para 2 itens de um teste hipotético de habilidade de 10 itens no qual os escores totais variam de 1 a 10. A Figura 6.2 mostra as regressões item-teste para ambos os itens, dispostas em gráfico a partir dos dados da tabela. Os gráficos de regressão item-teste combinam informações sobre dificuldade e discriminação de item e nos permitem visualizar como cada item funciona dentro do grupo que foi testado. Se pressupomos que as estatísticas de item apresentadas na Tabela 6.2 se basearam em uma amostra grande e representativa de testandos, esses dados possibilitariam avaliar os dois itens e tirar certas conclusões a respeito deles, conforme é descrito nos parágrafos a seguir.

- *O Item 1 é mais fácil que o Item 2, porque seu limiar de 50% é mais baixo.* Os limiares de 50% são representados na Figura 6.2 pelas linhas tracejadas perpendiculares traçadas a partir dos pontos onde os gráficos de regressão para cada item encontram a linha horizontal em $p = 0{,}50$ até a linha de base que mostra os escores totais no teste. Para o Item 1, o limiar de 50% se localiza no ponto onde o escore total é igual a 4, enquanto para o Item 2 ele se localiza onde o escore total é 7. Estes dados sugerem que o nível de habilidade necessário para se obter uma chance de 50-50 de acerto no Item 1 é mais baixo do que o nível de habilidade necessário para uma chance igual de sucesso no Item 2.
- *O Item 2 discrimina melhor que o Item 1.* A regressão item-teste é mais íngreme para o Item 2 do que para o Item 1 e não mostra inversões de direção na proporção de acertos no item em cada ponto do escore total. Em contraposição, o Item 1 mostra uma regressão item-teste mais gradual e quatro inversões em sua direção (nos pontos de escore total 2, 6, 7 e 9). Uma vez que se supõe que o escore total no teste reflete o nível de habilidade de um testando, as regressões item-teste na Figura 6.2 sugerem que a relação entre habilidade e desempenho no item é mais direta e estável para o Item 2 do que para o 1.
- *O Item 1 tem maior probabilidade que o Item 2 de ser respondido corretamente por adivinhação ou "chute".* Esta inferência se baseia no fato de que a proporção de respostas corretas ao Item 1 é bastante alta (.35) mesmo para aqueles indivíduos que obtiveram um escore total de 1, que foi o escore mais baixo no teste. Em contraste, ninguém com escore abaixo de 5 foi capaz de responder (ou adivinhar) o Item 2 corretamente.
- *Conclusão.* De modo geral, o exame dos dados de regressão item-teste apresentados na Tabela 6.2 e na Figura 6.2 sugere que (a) o Item 2 é mais difícil do que o Item 1, (b) o Item 2 parece funcionar melhor do que o Item 1 em termos de capacidade de discriminar entre indivíduos com escores altos e baixos no conjunto hipotético de 10 itens de habilidade e (c) o Item 2 é mais imune a "chutes" do que o Item 1.

Tabela 6.2 Dados de regressão item-teste para dois itens

	Proporção de examinandos que responderam cada item corretamente	
Escore total	Item 1	Item 2
10	1,00	1,00
9	0,60	0,85
8	0,75	0,70
7	0,65	0,50
6	0,70	0,45
5	0,80	0,30
4	0,50	0,00
3	0,40	0,00
2	0,30	0,00
1	0,35	0,00

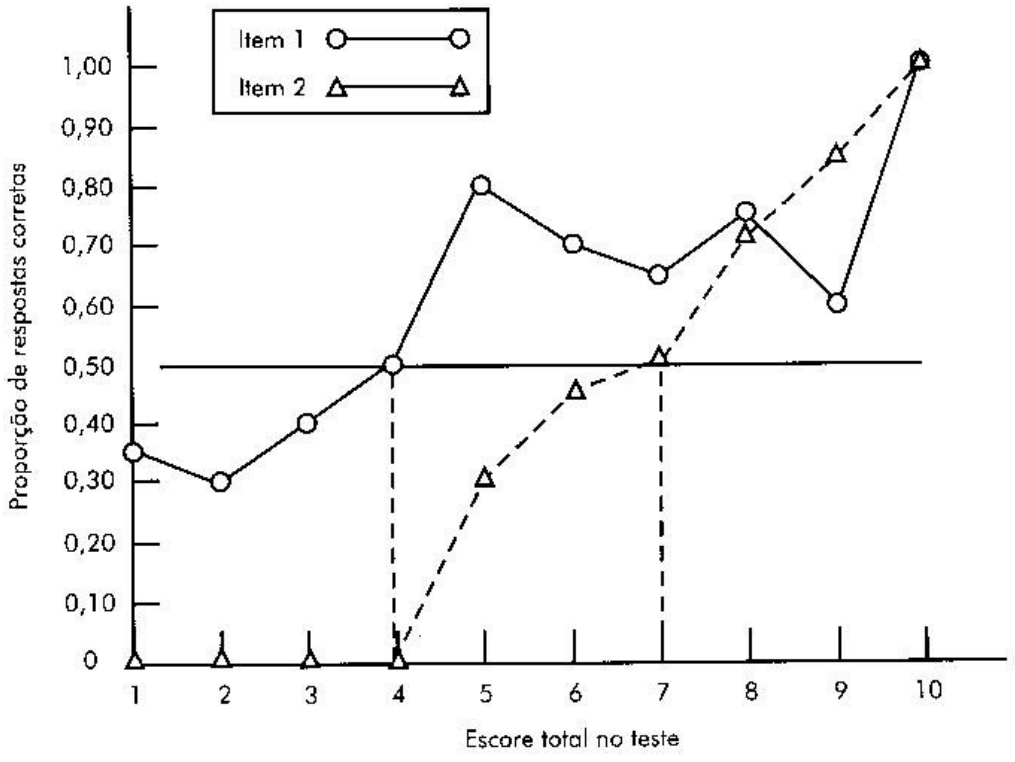

Figura 6.2 Regressão item-teste para os itens 1 e 2 (ver Tabela 6.2).

Embora essas análises de regressão item-teste sejam informativas, elas são um tanto "cruas" e dependem muito das amostras e conjuntos de itens dos quais os dados são obtidos. A teoria da resposta ao item usa os mesmos tipos de dados empíricos envolvidos na análise de regressão item-teste como ponto de partida para formas muito mais sofisticadas de análise de itens e estratégias mais ambiciosas de desenvolvimento de testes.

TEORIA DA RESPOSTA AO ITEM

A denominação *teoria da resposta ao item (TRI)* se refere a uma ampla e crescente variedade de modelos que podem ser usados para delinear ou desenvolver novos testes e para avaliar instrumentos já existentes. Os modelos da TRI diferem nas fórmulas matemáticas que empregam, no número de características de item que podem explicar e no número de dimensões de traços ou habilidades que especificam como objetivos de mensuração. Além disso, diferentes métodos são usados, dependendo se os dados de itens são dicotômicos (certo-errado, verdadeiro-falso, etc) ou politômicos (isto é, consistindo em múltiplas categorias de respostas). Os procedimentos englobados pela TRI são extensos e complexos, e até muito recentemente as apresentações publicadas desses métodos eram muito difíceis de compreender sem uma sólida base em matemática e estatística. Felizmente, nos últimos anos, vem sendo publicado um grande número de materiais mais acessíveis e de qualidade sobre técnicas da TRI. O quadro Consulta Rápida 6.4 lista uma seleção de alguns dos recursos mais úteis disponíveis atualmente.

Teoria clássica dos testes *versus* teoria da resposta ao item

O termo *teoria clássica dos testes (TCT)* é usado, em contraste com a TRI, em relação a todos os métodos psicométricos tradicionais de desenvolvimento e avaliação de testes que a antecedem. Os métodos fundamentais da TCT foram desenvolvidos no início do século XX e já estavam bem estabelecidos em meados daquele século. Seu melhor resumo talvez esteja na obra clássica de Gulliksen (1950) sobre a teoria dos testes mentais, mas eles já haviam sido descritos antes e o foram desde então em diversas outras fontes. Os princípios e procedimentos psicométricos da TCT têm sido continuamente refinados e expandidos; ainda são amplamente usados e continuarão a sê-lo no futuro próximo. Na verdade, a maioria dos livros sobre a

Não Esqueça

As seções sobre a teoria da resposta ao item (TRI) e a testagem adaptativa computadorizada (TAC) do Capítulo 3 oferecem uma introdução básica a algumas características distintivas destas abordagens relativamente novas à mensuração psicológica. Os leitores podem achar útil revisar as seções anteriores antes de se aprofundarem nos tópicos apresentados no presente capítulo.

CONSULTA RÁPIDA 6.4

Fontes de informação sobre a teoria da resposta ao item e testagem adaptativa computadorizada

Livros

- Bond, T.G. e Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.
- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Wainer, H. (2000). *Computer adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.

Internet

Muitos recursos disponíveis anteriormente a respeito da teoria da resposta ao item (TRI) no agora extinto *Educational Resource Information Center (ERIC) Clearinghouse on Assessment and Evaluation Online* podem ser encontrados em http://edres.org/irt, uma página mantida por Lawrence Rudner, ex-diretor do *ERIC Clearinghouse on Assessment and Evaluation*. Por intermédio desta página podem-se acessar muitos materiais relacionados à TRI, como os seguintes:

- Um excelente tutorial sobre a teoria da resposta ao item, disponibilizado pelo Laboratório de Modelagem de TRI da University of Illinois at Urbana-Champaign;
- A segunda edição do livro clássico de Frank Baker *The Basics of Item Response Theory* (2001);
- *Links* para programas de computador de TRI gratuitos e comercializados, bem como coletâneas de trabalhos e livros sobre a TRI.

A *CAT Central* é uma página com uma variedade de recursos para pesquisa e aplicações de testagem adaptativa computadorizada (TAC), incluindo informações básicas sobre o tema, uma bibliografia extensa, listagens dos principais programas de testagem que empregam a TAC e *links* para outros recursos relacionados. A *CAT Central* pode ser encontrada no seguinte endereço:

– http://www.psych.umn.edu/psylabs/CATCentral

testagem psicológica – incluindo este – trata em grande parte da TCT. Alguns dos principais contrastes entre a TCT e a TRI foram mencionados brevemente no Capítulo 3 em conexão com o tópico do equacionamento de testes. No entanto, a amplitude e a significância das mudanças acarretadas pela transição dos procedimentos convencionais da TCT para a abordagem baseada no modelo de mensuração que caracteriza a TRI vão muito além deste tópico.

Atualmente, os métodos da TRI são empregados em uma gama mais limitada de instrumentos do que os métodos tradicionais da TCT. Isso se deve em parte às premissas significativas que a TRI requer – em relação a respostas a itens, traços latentes e suas relações – e em parte aos esforços mais extensos de coleta de dados necessários para calibrar itens com os modelos da TRI. Além disso, em contraste com as técnicas bem-estabelecidas, comparativamente simples e amplamente usadas da TCT, os métodos da TRI ainda estão em evolução, são consideravelmente mais sofisticados do ponto de vista matemático e ainda são desconhecidos para muitos profissionais da testagem. Como Embretson (1996, 1999) deixou claro, embora a TCT e a TRI compartilhem alguns fundamentos conceituais e haja uma

certa reciprocidade entre as duas abordagens, muitas regras tradicionais de mensuração implícitas na TCT devem ser revisadas ou abandonadas quando os modelos da TRI são aplicados a tarefas de mensuração. O quadro Consulta Rápida 6.5 apresenta uma das várias características contrastantes das duas abordagens.

Uma das diferenças mais básicas entre a TCT e a TRI se origina no fato de que na TCT o interesse está centrado principalmente no escore total do examinando no teste, que representa a soma dos escores nos itens, enquanto que na TRI – como o nome já sugere – o foco principal está em seu desempenho nos itens individuais. Na TRI, o desenvolvimento e a calibração cuidadosa dos itens em termos das informações que eles fornecem a respeito de um constructo psicológico específico é uma preocupação primária. Para realizar essa calibração, a TRI se vale de modelos matemáticos das relações entre habilidades – ou outros constructos não-observáveis (isto é, traços latentes) que o teste deve avaliar – e respostas a itens individuais.

Definidos de modo amplo, os objetivos da TRI são (a) gerar itens que forneçam o máximo de informações possíveis sobre os níveis de habilidade ou traço dos examinandos que respondem a eles de uma forma ou de outra, (b) propiciar aos examinandos itens sob medida para seus níveis de habilidade ou traço, e, com isso, (c) reduzir o número de itens necessários para identificar a posição de qualquer testando na habilidade ou traço latente, ao mesmo tempo em que se minimiza o erro de mensuração. Reduzir o número de itens de um teste selecionando aqueles

CONSULTA RÁPIDA 6.5

Teoria clássica dos testes *versus* teoria da resposta ao item: Um contraste na questão do tamanho e fidedignidade dos testes

As novas regras da mensuração, descritas por Embretson (1996, 1999), enfatizam algumas diferenças cruciais entre a teoria clássica dos testes (TCT) e a teoria da resposta ao item (TRI). Entre elas está o contraste entre a antiga regra de que "testes mais longos são mais fidedignos do que testes mais curtos" e a nova regra de que "testes mais curtos podem ser mais fidedignos do que testes mais longos" (p.343). A saber:

- Como foi discutido em conexão com a fidedignidade pelo cálculo das metades e a fórmula Spearman-Brown (Capítulo 4), a TCT afirma que, não havendo outras falhas, um número maior de observações vai produzir resultados mais fidedignos do que um número menor de observações. Se o tamanho de um teste aumenta pela soma de itens paralelos, a proporção de variância verdadeira para variância de erro também aumenta e, assim sendo, o mesmo acontece com a fidedignidade dos escores. Por isso, para dois testes comparáveis de tamanho fixo (p. ex., 50 *vs.* 40 itens), os escores do teste mais longo serão mais fidedignos do que os do teste mais curto.
- Na testagem adaptativa computadorizada (TAC) permitida pelos métodos da TRI, a seleção dos itens é adequada de forma ótima em termos do testando no traço que está sendo avaliado. Itens inapropriados (p. ex., fáceis ou difíceis demais para o testando) são eliminados, resultando em um teste mais curto. Como os métodos da TRI também calibram as informações obtidas de cada resposta com mais precisão, o erro de mensuração pode ser reduzido e escores confiáveis podem ser obtidos com um número menor de respostas, mas respostas mais informativas.
- Para explicações mais detalhadas dessas noções, juntamente com ilustrações numéricas e gráficas, ver Embretson (1996, 1999) e Embretson e Reise (2000).

que são mais apropriados ao nível de habilidade do testando – sem perda de fidedignidade – é uma meta importante na testagem em grupo. Isso se aplica especialmente a programas de testagem realizados em escala maciça, como o SAT. Uma redução no número de itens administrados poupa tempo e dinheiro e minimiza a frustração dos testandos quando confrontados com itens inadequados a seus níveis de habilidade. De fato, em programas de testagem em larga escala, as TACs desenvolvidas pelos métodos da TRI estão gradualmente substituindo os testes de tamanho fixo em formatos de computadores e de lápis e papel (Embretson e Reise, 2000).

Deficiências da teoria clássica dos testes

A teoria clássica dos testes e a TRI diferem em muitos outros aspectos. Embora uma discussão completa destas diferenças esteja além do alcance deste livro, algumas precisam ser mencionadas devido à sua importância para o desenvolvimento de testes e a análise de itens. Uma forma de contrastar as duas metodologias é descrever as deficiências da TCT que a TRI busca superar. Embora alguns desses pontos já tenham sido mencionados anteriormente, eles serão reiterados em referência específica à comparação entre TCT e TRI.

- Os índices de dificuldade e discriminação de item da TCT são *grupo-dependentes:* Seus valores podem se alterar quando calculados para amostras de testandos que diferem das usadas para a análise inicial de itens em algum aspecto do constructo que está sendo medido. Em contraste, pressupõe-se que as estimativas de características de item obtidas por métodos da TRI são invariantes e fornecem uma escala uniforme de mensuração que pode ser usada com diferentes grupos.
- Para testes de tamanho fixo desenvolvidos com métodos da TCT, as estimativas de traço ou habilidade (isto é, os escores) dos testandos são *teste-dependentes*. Em outras palavras, os escores são uma função dos itens específicos selecionados para inclusão no teste. Assim sendo, a comparação de escores derivados de diferentes testes ou conjuntos de itens não é possível a menos que sejam usados procedimentos de equacionamento de teste, que muitas vezes não são viáveis (ver Capítulo 3). Além disso, mesmo quando são aplicados procedimentos de equacionamento, as comparações possíveis se limitam aos testes que foram equacionados. No caso da TRI – desde que os dados se encaixem no modelo, e que certas premissas sejam satisfeitas – as estimativas de habilidades ou traços são independentes, em particular do conjunto de itens administrado aos examinandos. Em vez disso, as estimativas estão ligadas às probabilidades dos padrões de resposta aos itens dos examinandos.
- Na metodologia da TCT, a fidedignidade dos escores (isto é, estimativas de traços ou habilidades) é medida por meio do erro padrão de mensuração (*EPM*), que se pressupõe ser de magnitude igual para todos os examinandos (ver Capítulo 4). Na verdade, como a fidedignidade dos escores de-

pende da adequação dos itens de teste aos níveis de traço ou habilidade dos examinandos, e como os níveis de traço não são iguais entre todos eles, essa premissa não é plausível para os testes tradicionais. Por outro lado, quando a metodologia da TRI é combinada a procedimentos de testagem adaptativa, os erros padrões de estimativas de traço ou habilidade resultantes da administração de um teste dependem do conjunto específico de itens selecionados para cada examinando (ver quadro Consulta Rápida 6.5). Como conseqüência, essas estimativas variam apropriadamente em diferentes níveis das dimensões de traço e transmitem informações mais precisas a respeito da fidedignidade da mensuração.

Características essenciais da teoria da resposta ao item

Como a maioria dos modelos da TRI usada atualmente é unidimensional, nossa discussão se limitará a ela. Os *modelos unidimensionais da TRI* partem da premissa de que (a) os itens que compõem um teste ou um segmento de um teste medem um único traço e (b) as respostas dos testandos aos itens dependem somente de sua posição em relação ao traço que está sendo medido. Como é sugerido no quadro Consulta Rápida 6.6, de um ponto de vista realista, nenhuma dessas premissas pode ser plenamente satisfeita. No entanto, quando todos os itens de um teste ou segmento de teste são delineados de modo a medir um único traço predominante, as premissas dos modelos unidimensionais podem ser adequadamente satisfeitas o bastante para torná-los funcionais.

CONSULTA RÁPIDA 6.6

O que torna as amostras de comportamento coletadas pelos itens de teste tão complexas?

- *Independentemente de quais constructos os itens de teste devem avaliar, eles sempre envolvem múltiplas dimensões.* Para começar, alguma capacidade de atentar para os estímulos do teste é necessária para se responder a *qualquer* tarefa, assim como uma certa quantidade de memória de curto prazo. Além disso, todos os itens de teste envolvem conteúdo, formato e meio específicos, e requerem um conjunto específico de habilidades cognitivas. Por exemplo, dependendo de seu modo de apresentação e resposta, um item simples de vocabulário pode envolver leitura, escrita, ortografia, compreensão oral, expressão verbal, capacidade de raciocínio lógico ou conhecimento de etimologia, para não falar de atenção, memória e possivelmente velocidade.
- *Os testandos são seres complexos e únicos.* Eles carregam consigo uma combinação de fatores – como dotação genética, histórico de experiências, habilidades desenvolvidas, hábitos e atitudes, bem como estados fisiológicos e emocionais transitórios – que influenciam as tarefas de teste. Como as respostas aos itens são uma função da mescla singular de todos os elementos que o testando traz para as tarefas, elas nunca são equivalentes em todos os aspectos. Por exemplo, itens que requerem uma série de cálculos aritméticos apresentam um problema maior para um testando que experimenta ansiedade em relação à matemática do que para outro que não, mesmo que ambos sejam igualmente capazes de realizar os cálculos em uma situação que não envolva testagem.

Nos parágrafos a seguir, algumas características comuns à maioria dos modelos da TRI usados atualmente são resumidas para dar aos leitores uma idéia geral de como esta metodologia é aplicada à calibração de dados de itens de teste. Por uma questão de brevidade e simplicidade, a apresentação evita o uso de fórmulas matemáticas e conceitos que não sejam essenciais para uma compreensão básica das idéias fundamentais da TRI. Os leitores interessados poderão encontrar exposições mais extensas desses métodos, bem como explicações de seus fundamentos matemáticos, nas fontes listadas no quadro Consulta Rápida 6.4.

- Na TRI, os modelos se baseiam na premissa de que o desempenho de uma pessoa em qualquer item de teste é uma função de um ou mais traços ou habilidades e pode ser predito por eles. Os modelos buscam especificar as relações esperadas entre as respostas (observáveis) dos examinandos aos itens e os traços (não-observáveis) que as governam. Como acarretam predições, os modelos da TRI podem ser avaliados (isto é, confirmados ou rejeitados) dependendo de como se ajustam aos dados derivados das respostas aos itens do teste.
- Os métodos da TRI empregam dados de testes e respostas a itens de amostras grandes que reconhecidamente diferem quanto à habilidade ou traço de personalidade que o teste em desenvolvimento deve avaliar. Essas amostras não precisam ser representativas de uma população definida, mas devem incluir grupos de indivíduos em diferentes níveis no contínuo de traço ou habilidade. Além disso, os itens do *pool* inicial precisam ser construídos ou selecionados cuidadosamente por seu potencial como indicadores do traço a ser avaliado.
- Depois de coletados, os dados de escore de item e de teste são usados para derivar estimativas de parâmetros de item que vão posicionar os examinandos e os itens ao longo de uma escala comum para a dimensão de traço ou habilidade. Os *parâmetros de item* são os valores numéricos que especificam a forma das relações entre as habilidades ou traços medidos e a probabilidade de certas respostas. Por exemplo, os parâmetros de *dificuldade de item* expressam a dificuldade de um item em termos da posição na escala de habilidade onde a probabilidade de acertar o item é de 0,50. A Tabela 6.3 mostra um pequeno conjunto hipotético de dados brutos de 10 itens de escore dicotômico administrados a 10 indivíduos (de A a J). Embora um exemplo realista devesse incluir uma amostra muito maior e variada de testandos – possivelmente agrupados em categorias baseadas em seus escores totais, e não individualmente – os dados da Tabela 6.3 ilustram o tipo de informação que pode ser usada para estimar parâmetros de dificuldade de item em relação a níveis de habilidade. Os parâmetros de item são obtidos por meio de uma variedade de procedimentos que requerem o uso de programas de computador especializados (ver, p. ex., Embretson e Reise, 2000, Capítulo 13). Estes procedimentos empregam funções matemáticas *não-lineares*, como funções logísticas, que produzem curvas de características de item (ver a seguir). Os modelos matemáticos não-lineares são necessários porque o modelo da regressão linear, discuti-

Tabela 6.3 Exemplo hipotético de dados brutos de itens e pessoas usados em estimativa de parâmetros na TRI

	Item										
Pessoa	1	2	3	4	5	6	7	8	9	10	Total
A	1	1	1	1	0	0	1	1	1	1	8
B	0	0	1	1	1	1	0	0	0	0	4
C	0	0	1	1	1	1	0	0	0	0	4
D	1	1	0	0	1	0	0	0	1	1	5
E	1	1	1	1	1	1	1	1	1	1	10
F	1	1	1	1	0	0	0	0	1	1	6
G	1	1	0	1	0	1	0	1	1	1	7
H	0	1	1	1	0	0	0	0	0	0	3
I	0	0	1	0	0	0	0	1	1	1	4
J	0	1	1	1	0	0	0	0	1	1	5
Total	5	7	8	8	4	4	2	4	7	7	

Nota: Os números 1 e 0 em cada célula indicam se as pessoas de A a J acertaram ou erraram cada um dos 10 itens. Os escores totais dos testandos, na última coluna, podem ser usados para calcular estimativas de habilidades; os escores totais dos itens, na última linha, podem ser usados para calcular a dificuldade dos itens.

do nos Capítulos 2 e 5, não é adequado para descrever como as mudanças nos níveis de traços se relacionam às mudanças na probabilidade de responder a um item de uma forma específica.

- Uma *curva de característica de item* (CCI) é a representação gráfica de uma função matemática que relaciona a probabilidade de resposta a um item em nível de traço, dados os parâmetros de item que foram especificados. Por exemplo, a CCI de um item de teste de habilidade dicotômico expressa visualmente a relação esperada entre o nível de habilidade e a probabilidade de acertar o item. No caso dos testes de personalidade, as CCIs mostram a relação esperada entre os níveis de traço e a probabilidade de se responder a um item de uma forma específica. As CCIs hipotéticas apresentadas na Figura 6.3 exemplificam os três modelos logísticos unidimensionais mais comuns para dados de resposta a itens dicotômicos. O Painel A da Figura 6.3 mostra CCIs para o modelo logístico de um parâmetro, também conhecido como *modelo Rasch*, em homenagem ao matemático que o desenvolveu (Rasch, 1960/1980). Os painéis B e C da Figura 6.3 retratam os modelos logísticos de dois e três parâmetros, respectivamente.
  - *O Painel A da Figura 6.3* mostra as CCIs para dois itens que diferem apenas em relação à dificuldade. O Item 1 é mais fácil que o Item 2. A localização do parâmetro de dificuldade (isto é, o nível de habilidade associado a 0,50 ou 50% da probabilidade de sucesso) é mais baixa para o Item 1 (X_1) do que para o Item 2 (X_2). Uma vez que as inclinações das duas curvas são iguais, podemos inferir que os dois itens funcio-

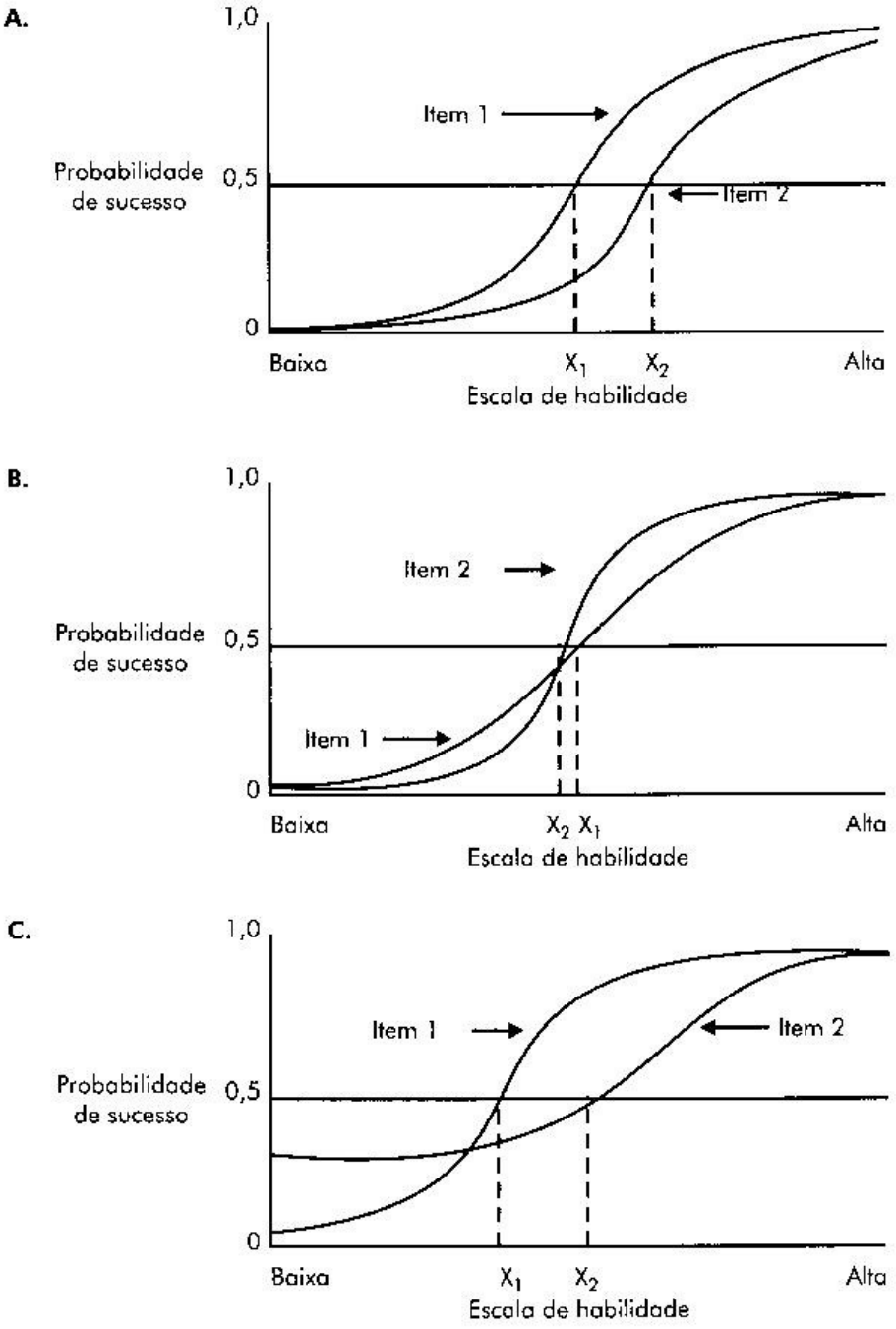

Figura 6.3 Curvas de características de item: *A*, modelo de um parâmetro; *B*, modelo de dois parâmetros; e *C*, modelo de três parâmetros.

nam igualmente bem em termos das relações entre habilidade e probabilidade de sucesso em toda a escala de habilidade.

– *O Painel B da Figura 6.3* mostra CCIs para dois itens que diferem em dois parâmetros, quais sejam, dificuldade e discriminação. Neste exem-

plo, o nível de habilidade associado a uma probabilidade de sucesso de 50% é um tanto mais alto para o Item 1 (X_1) do que para o Item 2 (X_2). Além disso, as inclinações das duas curvas – que mostram a razão entre mudança na habilidade e mudança na probabilidade de sucesso para cada item – são diferentes e se cruzam em um certo ponto. Esta configuração sugere que os dois itens funcionam diferentemente em termos de sua relação com o traço de habilidade e não discriminam igualmente bem em todos os pontos na escala de habilidade. Em um teste realmente unidimensional, itens com CCIs que se interceptam – como os do Painel B – são indesejáveis.

- *O painel C da Figura 6.3* mostra as CCIs para dois itens que diferem ao longo de três parâmetros, quais sejam, dificuldade, discriminação e probabilidade de sucesso por acaso (ou "chute"). A CCI para o Item 1 tem inclinação mais acentuada do que a curva para o Item 2, e mostra uma elevação contínua na probabilidade de sucesso à medida que os níveis de habilidade aumentam, até um certo ponto. Em contraste, o Item 2 claramente não discrimina entre indivíduos em diferentes níveis de habilidade tão bem quanto o Item 1: sua CCI mostra uma relação menos pronunciada entre nível de habilidade e probabilidade de sucesso. Observe também que a CCI para o Item 2 mostra uma probabilidade de sucesso bastante alta, até mesmo no extremo inferior do espectro de habilidade. Isso sugere que testandos com níveis baixos de habilidade são capazes de adivinhar corretamente a resposta ao Item 2. Além disso, a probabilidade de sucesso de 50% está associada a um nível mais alto de habilidade (X_2) para o Item 2. Claramente, um item com uma CCI como a do Item 2 do Painel C seria menos eficiente do que a do Item 1, do ponto de vista da mensuração.

- Como acontece com qualquer modelo teórico, o grau em que as premissas dos modelos da TRI tem a possibilidade de serem satisfeitas pode ser avaliado comparando-se suas predições com dados empíricos e avaliando-se a magnitude e a significância de quaisquer discrepâncias encontradas entre os dados observados e as predições dos modelos. Se o ajuste entre as expectativas baseadas no modelo CCI e o desempenho dos examinandos em um item de teste for suficientemente próximo, os parâmetros da TRI são usados para derivar a função de informação do item.
- A *função de informação de um item* reflete a contribuição que este faz à estimativa de um traço ou habilidade em diferentes pontos no contínuo de

Não Esqueça

Na teoria da resposta ao item (TRI), pressupõe-se que os parâmetros de item sejam *invariantes* para a população, o que significa que eles devem ser estáveis, mesmo quando são calculados para grupos que diferem quanto ao traço ou habilidade que está sendo medido. Portanto, ao contrário das estatísticas de análise de item descritas anteriormente neste capítulo, os parâmetros da TRI não estão ligados ao desempenho de um determinado grupo de referência.

mensuração. As funções de informação dos itens ajudam a decidir se e onde incorporar itens a um teste à luz de seus objetivos. As estimativas de nível de traço que localizam os testandos na dimensão do traço são derivadas de seus padrões específicos de sucesso ou fracasso em uma série de itens. Na TRI, a *função de informação do teste*, que é a soma das funções de informação dos itens, corresponde à noção de fidedignidade de escore da TCT (ver Capítulo 4). As funções de informação dos testes são calculadas e usadas para se obter erros padrões de estimativa em cada nível na escala de traço ou habilidade. Estes erros padrões, por sua vez, criam faixas de confiança para as estimativas de habilidade de modo semelhante à maneira como os erros padrões de mensuração são usados na TCT para criar intervalos de confiança para escores obtidos.

Como se pode depreender da discussão sobre as CCIs apresentadas na Figura 6.3, mesmo os modelos unidimensionais da TRI exemplificados naquela figura podem se tornar bastante complexos à medida que o número de parâmetros incluídos nos modelos aumenta. Aplicar modelos da TRI ao desenvolvimento de instrumentos voltados para a avaliação de constructos intelectuais e personalísticos mais amplos e controversos é uma proposta muito mais difícil (Reise e Henson, 2003). Modelos multidimensionais da TRI – que pressupõem que dois ou mais traços contribuem para as respostas aos itens – agora estão sendo usados para explorar e explicar constructos mais complexos e multifacetados. Alguns desses modelos mais novos e mais complicados são descritos por Embretson e Reise (2000).

Justiça dos itens

De modo geral, existem muitas maneiras como os itens de testes, bem como os testes, podem ser tendenciosos ou injustos para testandos individuais ou grupos de testandos. No que diz respeito aos testes, a possibilidade de viés pode ser investigada determinando-se se os escores têm o mesmo significado para membros de diferentes subgrupos da população (ver Capítulo 5). A questão da *justiça* ou imparcialidade dos testes, por outro lado, é mais complexa e mais polêmica. Embora haja um consenso de que o uso injusto de testes deve ser evitado, ainda há muita controvérsia sobre exatamente o que constitui a justiça na testagem (AERA, APA, NCME, 1999, p.74-76). Não obstante, os usuários de testes têm uma grande responsabilidade na implementação de práticas justas de testagem por meio de uma consideração criteriosa da propriedade dos instrumentos para seus fins pretendidos e para os testandos em potencial (Capítulo 7).

No nível dos itens de teste, as questões relativas ao viés e à injustiça estão mais circunscritas e geralmente são tratadas enquanto um teste está em desenvolvimento. Para esse fim, os itens são analisados qualitativa e quantitativamente ao longo de todo processo de construção do teste. Naturalmente, o grau em que os itens são submetidos a essas revisões está relacionado à finalidade do teste. Um cuidado especial é tomado para eliminar qualquer viés ou injustiça possível nos

itens de testes de habilidade que serão usados na tomada de decisões com conseqüências significativas para os testandos.

Análise qualitativa do viés de item

A avaliação qualitativa dos itens de teste, do ponto de vista da imparcialidade, se baseia em procedimentos de julgamento conduzidos por grupos de indivíduos demograficamente heterogêneos, qualificados em virtude de sua sensibilidade para essas questões e, preferencialmente, também por seus conhecimentos nas áreas abordadas pelos testes. Tipicamente, essas revisões ocorrem em dois estágios. Durante a fase inicial de construção de um teste, quando os itens são elaborados ou gerados, eles são examinados de modo a (a) eliminar descrição estereotipada de qualquer subgrupo identificável da população, (b) eliminar itens cujo conteúdo possa ser ofensivo a membros de minorias e (c) garantir que diversos subgrupos sejam representados apropriadamente nos materiais contidos no *pool* de itens. Nessa revisão inicial, indivíduos familiarizados com os hábitos lingüísticos e culturais dos subgrupos específicos com chance de serem encontrados entre os potenciais testandos também devem identificar o conteúdo de itens que podem funcionar em benefício ou detrimento de qualquer grupo específico, para que possam ser revistos. O segundo estágio da revisão qualitativa ocorre mais adiante no processo da construção do teste, depois que os itens foram administrados e seus dados de desempenho foram analisados separadamente para diferentes subgrupos. Neste estágio, os itens que exibem diferenças de subgrupo em índices de dificuldade, discriminação, ou ambos, são examinados para identificar os motivos dessas diferenças, sendo revisados ou descartados conforme necessário.

Análise quantitativa de viés de item

A avaliação quantitativa do viés de item por vezes tem sido relacionada simplesmente às diferenças na dificuldade relativa dos itens de teste para indivíduos de diferentes grupos demográficos. No entanto, esta interpretação é vista como ingênua pelos profissionais da testagem, que não consideram as diferenças na dificuldade relativa de um item para diferentes grupos uma evidência suficiente de seu viés (Drasgow, 1987). Ao invés disso, do ponto de vista psicométrico, considera-se que um item contém viés somente se indivíduos de diferentes grupos que têm a mesma posição em um traço diferem na probabilidade de responderem ao item de uma forma específica. Em testes de habilidade, por exemplo, o viés pode ser inferido quando pessoas que possuem níveis de habilidade iguais, mas pertencem a grupos demográficos diferentes, têm probabilidades de sucesso diversas em um item. Por isso, na literatura de testagem, o viés de item é descrito mais propriamente como *funcionamento diferencial de item (FDI)*, um termo que denota melhor os casos em que as relações entre o desempenho do item e o constructo avaliado pelo teste diferem em dois ou mais grupos.

Os procedimentos clássicos para a análise quantitativa do FDI envolvem comparações das estatísticas de dificuldade e discriminação de item para diferentes grupos. Por exemplo, se um item tem correlação baixa com o escore total do teste (isto é, baixa discriminação) e é mais difícil para mulheres do que para homens, ele obviamente é suspeito e deveria ser descartado. No entanto, a análise do FDI por meio de comparações simples das correlações item-teste e valores *p* para diferentes grupos é complicada pelo fato de que grupos de vários tipos (p. ex., grupos de gênero, de raça e de nível socioeconômico, etc.) muitas vezes diferem em termos de seu desempenho médio e variabilidade, especialmente em testes de habilidade. Quando diferenças grupais deste tipo são encontradas nas distribuições de escores de teste, (a) as estatísticas de dificuldade de item são confundidas por diferenças válidas entre grupos na habilidade que um teste mede e (b) os índices correlacionais de discriminação de item são afetados pelas diferenças na variabilidade dentro dos grupos que estão sendo comparados. Devido a estes fatores complicantes, as estatísticas tradicionais de análise de itens não se mostraram muito úteis para detectar o funcionamento diferencial de itens.

Funcionamento diferencial de itens

A avaliação e o estudo apropriado do FDI requer métodos especializados, grande número dos quais já foi proposto. Um dos mais usados é a técnica Mantel-Haenszel (Holland e Thayer, 1988), que expande os procedimentos analíticos tradicionais. Neste tipo de análise, cada um dos grupos em questão (p. ex., grupos de maiorias e minorias) é dividido em subgrupos baseados no escore total no teste, e o desempenho dos itens é avaliado entre subgrupos comparáveis. Embora este método seja mais refinado do que a simples comparação de estatísticas de análise de itens entre grupos, o procedimento Mantel-Haenszel ainda se vale de um critério interno (o escore total) que pode ser insensível a diferenças no funcionamento dos itens entre grupos, e sua capacidade de detectar o FDI é substancialmente dependente do uso de grupos muito grandes (Mazor, Clauser e Hambleton, 1992).

A teoria da resposta ao item oferece uma base muito melhor para a investigação do FDI do que os métodos da teoria clássica dos testes. Para estabelecer se indivíduos de diferentes grupos com níveis iguais de um traço latente têm desempenho diferente em um item, é necessário localizar pessoas de dois ou mais grupos em uma escala comum de habilidade. Os procedimentos da TRI para atingir este objetivo começam pela identificação de um conjunto de itens-âncora que não exibem qualquer FDI entre os grupos de interesse. Depois que isto é feito, outros itens podem ser avaliados em termos de FDI comparando-se as estimativas de parâmetro de item e as CCIs obtidas separadamente para cada grupo. Se os parâmetros e CCIs derivados dos dois grupos para um dado item forem substancialmente os mesmos, pode-se inferir com segurança que o item funciona igualmente bem para ambos os grupos. Não surpreende que os procedimentos da TRI estejam se tornando os métodos de escolha para a detecção de FDI (ver Embretson e Reise, 2000, Capítulo 10).

Aplicações da teoria da resposta ao item

Conforme observado no Capítulo 3, o uso de métodos da TRI no desenvolvimento de testes e calibração de itens não impede a interpretação normativa ou referenciada no critério dos escores de teste. Na verdade, por causa de seus métodos mais refinados para a calibração de itens e avaliação de erros de mensuração, a TRI pode melhorar a interpretação dos escores de teste. Embora não forneça soluções para todos os problemas de mensuração psicológica, a TRI já ajudou a criar uma abordagem mais disciplinada e objetiva para o desenvolvimento de testes nas áreas em que foi aplicada.

No presente momento, os métodos da TRI estão sendo aplicados mais extensamente no desenvolvimento de testes adaptativos computadorizados usados em programas de testagem de larga escala, como o SAT e o ASVAB. O desenvolvimento deste tipo de teste requer a colaboração de indivíduos com conhecimentos técnicos consideráveis em matemática e programação de computadores, além do conhecimento na área de conteúdo explorada pelos testes. Aplicações mais limitadas dos métodos da TRI estão em uso há algum tempo. Por exemplo, a avaliação de parâmetros de dificuldade de item pelos métodos da TRI tornou-se bastante comum no desenvolvimento de baterias de habilidade e realização, como as *Differential Ability Scales*, as escalas Weschler, os *Wide Range Achievement Tests* e os testes Woodcock. Os modelos da teoria da resposta ao item também estão sendo cada vez mais usados na avaliação do FDI em testes cognitivos. Embora os métodos da TRI sejam promissores no campo da testagem da personalidade, sua aplicação nesta área tem sido muito mais limitada do que na testagem de habilidades (Embretson e Reise, 2000, Capítulo 12).

Explorações no desenvolvimento e avaliação de itens

A revolução na tecnologia da informática e o ritmo acelerado dos avanços na teoria e metodologia da ciência psicológica permitem uma exploração quase ilimitada de técnicas inovadoras que podem ser aplicadas a problemas de mensuração. Para concluirmos este capítulo, são apresentadas duas aplicações promissoras da informática no campo dos itens de teste.

Avanços recentes na geração de itens

Os métodos de decomposição de tarefas e análise de protocolo de testes introduzidos pela psicologia cognitiva levaram a avanços recentes na exploração e no esclarecimento dos processos, estratégias e reservas de conhecimento envolvidos nos desempenhos dos itens de teste (ver Capítulo 5). Na verdade, desde a década de 1980, esses avanços – juntamente com progressos concomitantes na TRI e na tecnologia – foram aplicados à geração computadorizada de itens para testes de

habilidade e realização. Esta metodologia ainda está engatinhando, em termos comparativos, pois as especificações necessárias para se criarem regras a partir das quais os computadores possam gerar itens de teste precisam ser consideravelmente mais detalhadas do que nos métodos tradicionais de geração de itens, e requerem um nível mais alto de fundamentação teórica. Não obstante, a geração computadorizada de itens já foi implementada no desenvolvimento de ferramentas para áreas – como avaliações de matemática e testagem de aptidões para aviação – nas quais (a) existem modelos cognitivos de desempenho, (b) os constructos a serem examinados podem ser representados em termos de uma sintaxe lógica e (c) a dificuldade dos itens pode ser medida por meio de referentes objetivos. Sem dúvida alguma, as técnicas de geração de itens vão continuar a ser ativamente pesquisadas pelos estudiosos, devido às muitas vantagens que apresentam em termos de eficiência e economia, bem como por seu potencial em aplicações pedagógicas (Irvine e Kyllonen, 2002).

Avaliação automatizada de questões dissertativas

Uma inovação significativa no esforço para padronizar a avaliação de questões dissertativas é o desenvolvimento de tecnologia informatizada para a *avaliação automatizada de dissertações (AAD)* A tentativa de avaliar a prosa escrita por meio de programas de computador está em progresso nas últimas décadas e, como a geração computadorizada de itens, também foi facilitada por avanços na psicologia cognitiva e na ciência da computação. Embora ainda esteja em seus estágios iniciais, a AAD já demonstra grandes promessas, não apenas como meio de aumentar a fidedignidade e a validade dos escores, mas também como ferramenta educativa (Shermis e Burstein, 2003).

Teste a si mesmo

1. Os procedimentos envolvidos na análise de item dizem respeito primariamente aos _____ de testes.
   (a) criadores
   (b) usuários
   (c) testandos
   (d) administradores

2. A análise qualitativa de item tipicamente ocorre
   (a) depois que um teste é padronizado
   (b) ao mesmo tempo que um teste passa pela validação cruzada
   (c) depois que o *pool* de itens é gerado
   (d) um pouco antes do teste ser publicado