[capítulo 3]

FUNDAMENTOS EM INTERPRETAÇÃO DE ESCORES

Independentemente de quantas funções estatísticas forem usadas na testagem psicológica, na análise final o significado dos escores dos testes deriva dos referenciais que usamos para interpretá-los e do contexto no qual eles são obtidos. Sem dúvida alguma, os escores também precisam ser fidedignos, e os itens dos testes cuidadosamente desenvolvidos e avaliados para que contribuam para o sentido dos escores, questões das quais trataremos nos Capítulos 4 e 6. Neste capítulo, vamos analisar os referenciais para a interpretação de escores, um tópico intimamente relacionado à validade das inferências que podemos fazer a partir dos testes, discutidas mais detalhadamente no Capítulo 5. O contexto no qual a testagem acontece, uma questão de importância central que está relacionada ao processo de seleção e administração dos testes, é discutido no Capítulo final. O quadro Consulta Rápida 3.1 lista três excelentes fontes de informação nas quais muitos tópicos discutidos neste capítulo são abordados mais detalhadamente.

CONSULTA RÁPIDA 3.1

Para informações mais extensas sobre os aspectos técnicos de muitos tópicos discutidos neste capítulo, ver qualquer uma das seguintes fontes:

- Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, NJ: Educational Testing Service.
- Petersen, N. S., Kolen, M. J., Hoover, H. D. (1989). Scaling, Norming, and equating. In R. L. Linn. (Ed.), *Educational Measurement* (3rd ed., pp. 221-262). New York: American Council on Education/Macmillan.
- Thissen, D., Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

ESCORES BRUTOS

Um *escore bruto* é um número (X) que resume ou representa alguns aspectos do desempenho de uma pessoa nas amostras de comportamento cuidadosamente selecionadas e observadas que configuram os testes psicológicos. Por si só, um escore bruto não transmite qualquer significado: escores altos podem ser um resultado favorável em teste de habilidade, mas desfavorável em testes que avaliam algum aspecto de psicopatologia. No Inventário Multifásico Minnesota de Personalidade (MMPI), por exemplo, escores elevados geralmente indicam algum tipo de desajustamento, embora escores baixos não necessariamente indiquem bom ajustamento. Mesmo se soubermos em que tipo de teste um escore foi obtido, ainda podemos nos equivocar. Alguns testes de habilidade cognitiva – em particular muitos instrumentos neuropsicológicos – são pontuados em termos de número de erros ou velocidade de desempenho, de modo que quanto mais alto o escore, menos favorável o resultado. Além disso, não podemos sequer saber o quão alto é um escore "alto" sem algum tipo de referencial. Um escore que parece alto – como um QI de 130, por exemplo – pode ter sentidos bastante diferentes dependendo do teste do qual foi derivado, das áreas abordadas pelo teste e da atualização de suas normas, bem como de aspectos específicos da situação na qual o escore foi obtido e de características do testando.

REFERENCIAIS PARA A INTERPRETAÇÃO DE ESCORES

Subjacente a todas as outras questões relativas à interpretação de escores, de uma forma ou de outra, está a questão dos referenciais usados para interpretá-los. Dependendo de seu objetivo, os testes se valem de uma ou ambas das seguintes fontes de informação para derivar referenciais para o seu significado:

1. *Normas.* A *interpretação de testes referenciada em normas* usa padrões baseados no desempenho de grupos específicos de pessoas para fornecer informações para a interpretação de escores. Esse tipo de interpretação de teste é útil primariamente quando precisamos comparar indivíduos ou uns com os outros, ou com um grupo de referência, para avaliar diferenças entre eles nas características medidas pelo teste. O termo *normas* se refere ao desempenho no teste ou ao comportamento típico de um ou mais grupos de referência. As normas geralmente são apresentadas na forma de tabelas com estatísticas descritivas – como médias, desvios padrões e distribuições de freqüência – que resumem o desempenho do grupo ou grupos em questão. Quando as normas são coletadas em testes de desempenho de grupos de pessoas, estes grupos de referência são denominados *amostras normativas* ou *de padronização*. Coletar normas é um aspecto central do processo de padronização de um teste referenciado em normas.

2. *Critérios de desempenho.* Quando a relação entre os itens ou tarefas de um teste e os padrões de desempenho é demonstrável e bem definida, os escores podem ser avaliados através de uma *interpretação referenciada em critérios*. Esse tipo de interpretação faz uso de procedimentos, como amostragem de domínios de conteúdo ou comportamentos relacionados ao trabalho, delineados para avaliar se e em que grau os níveis desejados de competência ou os critérios de desempenho foram satisfeitos.

INTERPRETAÇÃO DE TESTES REFERENCIADA EM NORMAS

As normas são, sem dúvida, o referencial mais amplamente usado para a interpretação de escores de teste. O desempenho de grupos definidos de pessoas é usado como base para a interpretação de escores tanto em testes de habilidade como de personalidade. Quando as normas são o referencial, a pergunta que elas tipicamente respondem é: como o desempenho deste testando se compara ao de outros? O escore em si é usado para localizar o desempenho do testando dentro de uma distribuição preexistente de escores ou dados obtidos a partir do desempenho de um grupo adequado de comparação.

Normas desenvolvimentais

Escalas ordinais baseadas em seqüências comportamentais

O desenvolvimento humano se caracteriza por processos seqüenciais em uma série de campos do comportamento. Um exemplo clássico é a seqüência seguida pelo desenvolvimento motor normal durante a infância. No primeiro ano de vida, a maioria dos bebês progride da posição fetal do nascimento até finalmente o caminhar sozinhos, passando por sentar e ficar de pé. Sempre que uma seqüência universal de desenvolvimento envolve uma progressão ordenada de um estágio comportamental para outro mais avançado, a seqüência em si pode ser convertida em uma escala *ordinal* e usada normativamente. Nesses casos, o referencial para a interpretação de escores deriva da observação de certas uniformidades na ordem e no momento das progressões comportamentais em muitos indivíduos. O pioneiro no desenvolvimento desse tipo de escalas foi Arnold Gesell, um psicólogo e pediatra que publicou as Escalas de Desenvolvimento de Gesell (*Gesell Developmental Schedules)* em 1940, baseado em uma série de estudos longitudinais conduzidos por ele e seus colegas em Yale ao longo de várias décadas (Ames, 1989).

Um exemplo atual de um instrumento que usa escalas ordinais é o Perfil Desenvolvimental Provence do Nascimento aos Três Anos (*Provence Birth-to-Three Developmental Profile*) ("Perfil Provence"), que faz parte da *Infant-Toddler Developmental Assessment* (IDA; Provence, Erikson, Vater e Palmieri, 1995). O IDA é um sistema integrado criado com o objetivo de ajudar na identificação precoce de crianças com risco de problemas de desenvolvimento e possíveis necessidades de monitoramento ou intervenção. Por meio de observações naturalísticas e rela-

tos parentais, o Perfil Provence oferece informações a respeito da adequação temporal com que uma criança atinge marcos evolutivos em oito domínios, em relação à sua idade cronológica. Os domínios evolutivos são: Comportamento Motor Amplo, Comportamento Motor Fino, Relação com Objetos Inanimados, Linguagem/Comunicação, Autonomia *(Self-Help)*, Relação com Pessoas, Emoções e Estados de Sentimento (Afetos) e Comportamento Adaptativo. Para cada um desses domínios, o perfil agrupa itens em faixas etárias que variam de 0 a 42 meses. As faixas etárias podem ser de apenas 2 meses nas idades mais precoces e até 18 meses em alguns domínios em idades mais avançadas, mas a maioria delas varia entre 3 e 6 meses. O número de itens em cada faixa etária também difere, assim como a quantidade de disposições que precisa estar presente ou ser realizada de forma competente para satisfazer os critérios de cada faixa. A Tabela 3.1 lista quatro itens de cada domínio evolutivo do Perfil Provence do IDA. Os escores nos itens em cada faixa etária são somados para se obter uma *idade de desempenho*, que é avaliada em comparação com a *idade cronológica* da criança. As eventuais discrepâncias entre os níveis etários de desempenho e de cronologia podem então ser usadas para determinar a possível presença e o grau de atraso evolutivo da criança.

Escalas ordinais baseadas em teorias

As escalas ordinais podem basear-se em fatores outros que não os da idade cronológica. Diversas teorias, como a dos estágios do desenvolvimento cognitivo da in-

Tabela 3.1 Amostras de itens do Perfil Provence da *Infant-Toddler Developmental Assessment*

Domínio	Faixa etária (em meses)	Item
Comportamento motor amplo	4 a 7	Senta-se sozinho por pouco tempo
	7 a 10	Apóia-se para ficar em pé
	13 a 18	Caminha bem sozinho
	30 a 36	Sobe e desce escadas
Linguagem/comunicação	4 a 7	Ri em voz alta
	7 a 10	Responde ao "não"
	13 a 18	Mostra um sapato quando solicitado
	30 a 36	Conhece rimas ou canções
Autonomia	4 a 7	Recupera o bico ou mamadeira perdidos
	7 a 10	Afasta a mão do adulto
	13 a 18	Alimenta-se parcialmente sozinho com colheres ou dedos
	30 a 36	Calça os sapatos

Fonte: Adaptado de *Infant-Toddler Developmental Assessment (IDA) Administration Manual* por Sally Provence, Joana Erikson, Susan Vater e Sara Palmieri e reproduzido com permissão da editora. Copyright 1995, The Riverside Publishing Company. Todos os direitos reservados.

fância até a adolescência propostas por Jean Piaget ou a teoria de Lawrence Kohlberg do desenvolvimento moral, postulam uma seqüência ou progressão ordenada e invariável derivada pelo menos em parte de observações comportamentais. Algumas destas teorias geraram escalas ordinais delineadas para avaliar o nível atingido por um indivíduo dentro da seqüência proposta, mas estas ferramentas são usadas primariamente para fins de pesquisa e não de avaliação individual. Os exemplos desse tipo de instrumento incluem escalas padronizadas baseadas no delineamento piagetiano da ordem em que as competências cognitivas são adquiridas durante a infância, como as Escalas Ordinais de Desenvolvimento Psicológico, também conhecidas como Escalas de Desenvolvimento Psicológico do Bebê (Užgiris e Hunt, 1975).

Escores de idade mental

A noção de escores de idade mental foi discutida no Capítulo 2 em conexão com o QI-razão das primeiras Escalas de Inteligência Stanford-Binet. Os escores de idade mental derivados destas escalas eram calculados a partir do desempenho da criança, que obtinha créditos em termos de anos e meses dependendo do número de testes dispostos em ordem cronológica que completasse satisfatoriamente. À luz das dificuldades apresentadas por este procedimento, descritas no Capítulo 2, este modo particular de obter escores de idade mental foi abandonado. No entanto, diversos testes atuais ainda oferecem normas que são apresentadas como *escores equivalentes à idade* e se baseiam na média dos escores brutos de desempenho de crianças de diferentes faixas etárias em amostras de padronização.

Os escores equivalentes a idades, também conhecidos como *idades de teste*, simplesmente representam uma forma de equacionar o desempenho do testando com o desempenho médio da faixa etária normativa à qual corresponde. Por exemplo, se o escore bruto de uma criança for igual à média do escore bruto das crianças de 9 anos na amostra normativa, seu escore equivalente à idade no teste será de 9 anos. Apesar desta mudança nos procedimentos usados para se obter escores equivalentes a idades, as desigualdades na taxa de desenvolvimento em diferentes idades continuam a ser um problema quando este tipo de norma etária é usada, porque as diferenças de progressão comportamental que podem ser esperadas a cada ano diminuem muito desde o início da infância até a adolescência e a vida adulta. Se isso não for compreendido, ou se o sentido de uma *idade de teste* for aplicado a outros domínios além daqueles do comportamento específico medido – como acontece, por exemplo, quando um adolescente que obtém um escore de idade de teste de 8 anos é descrito como o que tem "a mente de uma criança de 8 anos" – o uso destes escores pode ser bastante enganoso.

Escores equivalentes a séries escolares

A progressão seqüencial e a relativa uniformidade dos currículos escolares, especialmente na escola fundamental, oferecem outra base para a interpretação de

escores em termos de normas evolutivas. Por isso, o desempenho em testes de realização no contexto escolar muitas vezes é descrito com base nas séries escolares. Estes *escores equivalentes a séries* são derivados da localização do desempenho dos testandos dentro das normas dos estudantes de cada série – e frações de séries – na amostra de padronização. Se dizemos, por exemplo, que uma criança pontuou na 7ª série em leitura e na quinta série em aritmética, isto significa que seu desempenho no teste de leitura equivale ao desempenho médio dos alunos da 7ª série na amostra de padronização, e que no teste de aritmética seu desempenho equivale ao de alunos da 5ª série.

Apesar de seu apelo, os escores equivalentes a séries também podem ser enganosos por várias razões. Em primeiro lugar, o conteúdo dos currículos e a qualidade do ensino varia entre as escolas, distritos escolares, estados, etc., e, portanto, os escores equivalentes a séries não oferecem um padrão uniforme. Além disso, o avanço esperado nas primeiras séries escolares em termos de desempenho acadêmico é muito maior do que nas séries mais avançadas do ensino fundamental ou médio, e, por isso, assim como acontece com as unidades de idade mental, uma diferença de um ano em atraso ou aceleração é muito mais significativa nas primeiras séries do que nos últimos anos de escola. Também, se uma criança que está na 4ª série pontua na 7ª série em aritmética, isto não significa que ela dominou a aritmética da 7ª série, mas sim que seu escore está significativamente acima da média para crianças de 4ª série em aritmética. Por fim, os escores equivalentes a séries às vezes são vistos erroneamente como padrões de desempenho que todas as crianças em uma determinada série devem satisfazer, enquanto que simplesmente representam níveis médios de desempenho que – devido à inevitável variabilidade entre os indivíduos – alguns estudantes vão satisfazer, outros não, e outros vão exceder.

Não esqueça

Todas as normas desenvolvimentais são relativas, exceto as que refletem uma seqüência ou progressão comportamental que é *universal* em humanos.

- As *escalas ordinais baseadas em teorias* são mais ou menos úteis dependendo se os pressupostos nos quais se baseiam são sólidos e aplicáveis a um dado segmento de uma população ou à população como um todo.
- As *normas de idade mental ou escalas de escores equivalentes a idades* refletem nada mais do que o *desempenho médio de certos grupos de testandos de faixas etárias específicas* em um determinado momento e local em um teste específico. Estão sujeitas a mudanças temporais, tanto como culturais e subculturais.
- As *normas baseadas em séries escolares ou escalas de escores equivalentes a séries* também refletem o desempenho médio de certos grupos de alunos em séries específicas, em um dado momento e local. Também estão sujeitas a variações temporais, bem como curriculares em diferentes escolas, distritos escolares e países.

Normas intragrupo

A maioria dos testes padronizados usa algum tipo de *norma intragrupo*. Essas normas essencialmente oferecem um meio de avaliar o desempenho de uma pessoa em comparação com o de um ou mais grupos de referência apropriados. Para a adequada interpretação referenciada em normas dos escores de teste é necessário compreender os procedimentos numéricos nos quais os escores brutos são transformados na grande variedade de *escores derivados* que são usados para expressar normas intragrupo. Não obstante, é bom ter em mente que todos os tipos de escores revisados nesta seção servem ao simples fim de localizar o desempenho de um testando em uma distribuição normativa. Por isso, a questão mais importante em relação a este referencial diz respeito à constituição exata do grupo ou grupos dos quais as normas são derivadas. A composição da amostra normativa ou padronizada é de importância essencial nesse tipo de interpretação de escore de teste, porque os membros desta amostra determinam o padrão em relação ao qual todos os outros testandos serão medidos.

A amostra normativa

À luz do importante papel do desempenho da amostra normativa, o principal requisito destas amostras é que sejam representativas do tipo de indivíduos para os quais os testes estão voltados. Por exemplo, se um teste vai ser usado para avaliar as habilidades de leitura de alunos do ensino fundamental de 3ª a 5ª série de todo o país, a amostra normativa para o teste deve representar a população nacional de alunos de 3ª, 4ª e 5ª séries em todos os aspectos pertinentes. A constituição demográfica da população nacional em variáveis como gênero, etnia, linguagem, *status* socioeconômico, residência urbana ou rural, distribuição geográfica e matrícula em escolas públicas ou privadas deve estar refletida na amostra normativa para este teste. Além disso, a amostra precisa ser suficientemente grande para garantir a estabilidade dos valores obtidos a partir de seu desempenho.

O tamanho das amostras normativas varia tremendamente dependendo do tipo de teste que está sendo padronizado e da facilidade com que as amostras podem ser reunidas. Por exemplo, testes de habilidade grupal usados no contexto escolar podem ter amostras normativas de dezenas ou até mesmo centenas de milhares, enquanto que testes individuais de inteligência, administrados a uma única pessoa de cada vez por examinadores altamente treinados, são normatizados com amostras muito menores – tipicamente formadas por 1 a 3 mil indivíduos – obtidas da população geral. Testes que requerem amostras especializadas, como membros de um certo grupo ocupacional, podem ter amostras normativas até menores. O fato de as informações normativas serem recentes também é importante se os testandos tiverem que ser comparados com padrões contemporâneos, como costuma acontecer.

Os fatores relevantes a serem considerados na constituição da amostra normativa variam dependendo do objetivo do teste e da população na qual será usado. No caso de um teste delineado para detectar comprometimentos cognitivos em adultos mais velhos, por exemplo, variáveis como estado de saúde, situação de moradia independente *versus* institucional e uso de medicações seriam pertinentes, além das variáveis demográficas de gênero, idade, etnia, etc. O quadro Consulta Rápida 3.2 lista algumas das perguntas mais comuns que os usuários de testes devem se fazer a respeito da amostra normativa quando estão no processo de avaliar a adequação de um teste para seus objetivos.

Os grupos de referência podem ser definidos em um contínuo de amplitude ou especificidade, dependendo do tipo de comparação que os usuários do teste precisam fazer para avaliar seus escores. Em um extremo, o grupo de referência pode ser a população geral de um país inteiro ou mesmo uma população multinacional. No outro extremo, os grupos de referência podem se originar de populações definidas de forma restrita em termos de *status* ou contexto.

CONSULTA RÁPIDA 3.2

Informações necessárias para avaliar a aplicabilidade de uma amostra normativa

Para avaliarmos a adequação de um teste referenciado em normas para um objetivo específico, os usuários do teste precisam ter o maior número possível de informações a respeito da amostra normativa, incluindo respostas para perguntas como:

- Qual o tamanho da amostra normativa?
- Quando a amostra foi montada?
- Onde a amostra foi reunida?
- Como os indivíduos foram identificados e selecionados para a amostra?
- Quem testou a amostra?
- Como o examinador ou examinadores se qualificaram para a testagem?
- Qual era a composição da amostra normativa em termos de
  - idade?
  - sexo?
  - etnia, raça ou origem lingüística?
  - escolaridade?
  - *status* socioeconômico?
  - distribuição geográfica?
  - qualquer outra variável pertinente, como estado de saúde física e mental ou afiliação a um grupo atípico, que possa influenciar o desempenho no teste?

Os usuários de testes somente podem avaliar a adequação de um teste referenciado em normas para seus objetivos específicos, quando as perguntas a essas questões forem fornecidas pelo manual do teste ou por documentos relacionados.

Normas de subgrupo. Quando amostras grandes são coletadas para representar populações amplamente definidas, as normas podem ser relatadas no agregado ou podem ser separadas em *normas de subgrupo*. Desde que tenham tamanho suficien-

Advertência

Embora os três termos muitas vezes sejam usados de forma equivalente – neste e em outros textos – e possam efetivamente se referir ao mesmo grupo, estritamente falando o significado preciso de *amostra de padronização*, *amostra normativa* e *grupo de referência* é um tanto diferente:

- A *amostra de padronização* é o grupo de indivíduos com o qual o teste é originalmente padronizado tanto em termos de procedimentos de administração e pontuação, como no de desenvolvimento de normas. Os dados para este grupo geralmente são apresentados no manual que acompanha o teste em sua publicação.
- A *amostra normativa* muitas vezes é usada como sinônimo de amostra de padronização, mas pode se referir a qualquer grupo a partir do qual as normas são coletadas. Normas adicionais para um teste, coletadas após sua publicação, para uso com um subgrupo distinto, podem aparecer na literatura periódica ou manuais técnicos publicados posteriormente. Ver, por exemplo, o estudo sobre americanos idosos feito por Ivnik e seus colegas (1992) na Clínica Mayo, no qual foram coletados dados para viabilizar normas para pessoas com idade superior à da faixa etária mais alta da amostra de padronização da Escala de Inteligência Wechsler para Adultos – Versão Revisada (WAIS-R).
- O *grupo de referência*, em contraposição, é um termo usado de forma menos estrita para identificar qualquer grupo de pessoas com o qual os escores de um teste são comparados. Pode ser aplicado ao grupo de padronização, a uma amostra normativa desenvolvida subseqüentemente, a um grupo testado para fins de desenvolvimento de normas locais ou qualquer outro, como os alunos de uma única turma ou os participantes de um estudo de pesquisa.

te – e sejam suficientemente representativos de suas categorias – os subgrupos podem ser formados em termos de idade, sexo, ocupação, etnia, escolaridade ou qualquer outra variável que possa ter um impacto significativo nos escores de um teste ou produzir comparações de interesse. As normas de subgrupo também podem ser coletadas depois que o teste foi padronizado e publicado para complementar e expandir sua aplicabilidade. Por exemplo, antes do MMPI ser revisado para a criação do MMPI-2 e se criar um formulário separado para adolescentes (MMPI-A), os usuários do teste original – que havia sido normatizado exclusivamente com adultos – desenvolveram normas especiais de subgrupo para adolescentes em várias faixas etárias (ver, p. ex., Archer 1987).

Normas locais. Por outro lado, existem algumas situações nas quais os usuários de testes desejam avaliar os escores a partir de grupos de referência derivados de um contexto geográfico ou institucional específico. Nestes casos, podem optar por desenvolver um conjunto de normas locais para membros de uma população definida de forma mais estrita, como os funcionários de uma empresa, em particular, ou os alunos de uma certa universidade. Normas locais podem ser usadas para avaliar o desempenho de alunos ou empregados dentro de um determinado contexto ou para se tomarem decisões a respeito de candidatos a cursos ou empregos levando em conta os padrões de uma certa empresa ou instituição.

Normas de conveniência. Ocasionalmente, por questões de conveniência ou limitações financeiras, os criadores de testes usam normas baseadas em um grupo de

pessoas que simplesmente está disponível no momento em que o teste está sendo construído. Estas *normas de conveniência* têm uso limitado porque não são representativas de qualquer população definida, e sim, muitas vezes, compostas de indivíduos que podem ser facilmente acessados pelos criadores do teste, como os alunos de uma turma universitária ou os residentes de um asilo de idosos, em particular. Em casos como esses, a natureza da amostra normativa deve ser explicitada para os potenciais usuários do teste.

Escores usados para expressar normas intragrupo

Percentis

Os escores de postos de percentil, já discutidos no Capítulo 2, são o método mais direto e disseminado para transmitir resultados de testes referenciados em normas. Suas principais vantagens são a facilidade de compreensão pelos testandos e a aplicabilidade à maioria dos testes e populações. Um *escore de percentil* indica a posição relativa de um testando comparada a um grupo de referência, como a amostra de padronização; especificamente, representa a percentagem de pessoas no grupo de referência que teve escore igual ou inferior a um determinado escore bruto. Assim, escores de percentil mais altos indicam escores brutos mais altos naquilo que está sendo medido pelo teste. O 50º percentil (P_{50}), ou mediana, corresponde ao ponto de escore bruto que separa as metades inferior e superior da distribuição de escores do grupo de referência. Em uma distribuição normal, o 50º percentil também é o nível médio de desempenho do grupo.

Uma outra vantagem dos escores de postos de percentil se revela quando existe mais de um grupo normativo para o mesmo teste ou quando os grupos normativos são subdivididos em categorias como gênero, idade ou etnia. Quando estão disponíveis normas adicionais, um escore bruto pode ser localizado dentro das distribuições de dois ou mais grupos ou subgrupos diferentes e facilmente con-

Advertência

Devido à semelhança dos dois termos, os escores de percentil muitas vezes são confundidos com escores percentuais. Estes dois tipos de escores na verdade são bastante diferentes e usam referenciais inteiramente distintos:

- Os *percentis* são escores que refletem o posto ou posição do desempenho de um indivíduo em um teste em comparação a um grupo de referência; seu referencial são os outras pessoas.
- Os escores *percentuais* refletem o número de respostas corretas que um indivíduo obtém em meio ao número total possível em um teste; seu referencial é o conteúdo do teste como um todo.

Uma forma de evitar confusão é adquirir o hábito de usar o símbolo de percentagem (%) estritamente para escores de percentuais, e usar uma abreviação diferente, como *PR* ou *%'il*, para designar escores de percentil.

vertido em postos de percentil. Por exemplo, escores de inventários de interesses para vários grupos ocupacionais muitas vezes são relatados para homens e mulheres como normas separadas por sexo, para que os testandos possam ver suas posições em um determinado interesse e escala ocupacional comparadas a ambos os grupos. Esta informação é particularmente útil para aqueles que estão considerando uma ocupação significativamente segregada por questões de gênero, como engenharia ou enfermagem. A separação das normas permite aos indivíduos medirem a força relativa de seus interesses expressos em comparação com membros de ambos os sexos.

Se os escores fossem distribuídos uniformemente ao longo de sua amplitude, resultando em um polígono de freqüência retangular, os percentis provavelmente seriam os escores de escolha de quase todas as situações. No entanto, como vemos na Figura 2.2 e na Figura 3.1 mais adiante neste capítulo, em uma distribuição normal a maioria dos escores tende a se agrupar em torno de um valor central e se dispersar mais nas extremidades. Este fato, que também se aplica a muitas distribuições não-normais, significa que as unidades de escores de percentil em geral são acentuadamente desiguais em diferentes pontos da amplitude. Em uma distribuição normal ou quase normal, como a obtida na maioria dos testes, a percentagem de pessoas que obtem escores próximos do meio é muito maior do que a das extremidades. Por isso, qualquer diferença nas unidades de escore de postos de percentil amplia a discrepância aparente no desempenho relativo dos indivíduos cujos escores estão na faixa central e comprime a extensão aparente da diferença no desempenho relativo dos indivíduos nas extremidades superior e inferior das distribuições.

Outra desvantagem dos escores de postos de percentil diz respeito àqueles mais extremos de uma distribuição. Com certeza, para qualquer grupo normativo sempre existe um escore mais alto e um mais baixo. Desde que sejam interpretados estritamente em referência a uma amostra normativa específica, pode-se dizer que o escore mais alto está no 100º percentil, porque todos os casos são iguais ou inferiores a este escore. Tecnicamente, poderíamos até mesmo descrever o escore abaixo do mais baixo obtido por todos em uma amostra específica como o percentil zero, embora isto não aconteça normalmente. Em termos da população mais ampla que a amostra normativa representa, no entanto, a interpretação destes escores é problemática.

Teto de teste e solo de teste. A questão de como acomodar indivíduos nas extremidades superior e inferior do espectro de habilidade para a qual um teste é aplicado é muito relevante no contexto do desenvolvimento de testes, discutido no Capítulo 6. Não obstante, neste ponto é importante observarmos que os indivíduos empregados na padronização de um teste efetivamente determinam os limites inferior e superior do desempenho no mesmo. Se um testando alcança o escore mais alto obtenível em um teste já padronizado, isto significa que o *teto do teste*, ou seu nível máximo de dificuldade, é insuficiente: não se pode saber quão mais alto poderia ter sido o escore do testando se houvesse itens adicionais ou mais dificultosos. Da mesma forma, se uma pessoa é reprovada em todos os itens apresentados em um teste ou tem escore mais baixo do que qualquer uma das pessoas da amostra

Pondo em prática

Usando escores de postos de percentil: A desvantagem das unidades desiguais

O subteste de Vocabulário da Escala de Inteligência Wechsler para Adultos – Terceira Edição (WAIS-III), um teste para indivíduos entre 16 e 89 anos, consiste em 33 palavras. Cada definição de palavra pode resultar em um escore de 0, 1 ou 2. Assim, os escores brutos do subteste podem variar entre 0 e 66, dependendo da qualidade das respostas.

O manual do WAIS-III (Wechsler, 1997) exibe o desempenho das amostras de padronização em tabelas para várias faixas etárias. Para indivíduos entre 45 e 54 anos de idade, a tabela mostra que escores brutos variando de 45 a 48 pontos estão localizados no 50º percentil, e escores brutos de 49 a 51 pontos estão no 63º percentil. Em contraste, escores brutos entre 15 e 19 pontos estão no 2º percentil, e aqueles entre 20 e 24 pontos estão no 5º percentil.

Isso enfatiza claramente o problema da desigualdade de unidades de escores de percentil: Para pessoas na faixa etária de 45 a 54 anos, uma diferença de apenas 6 pontos (45 a 51) na metade da distribuição dos escores brutos resulta em uma diferença de 13 unidades de postos de percentil (do 50º ao 63º percentil), enquanto que uma diferença de escore bruto de 9 pontos (de 15 para 24) na extremidade inferior da faixa resulta em uma diferença de apenas 3 unidades de postos de percentil (do 2º ao 5º percentil).

normativa, o problema é de *solo de teste* insuficiente. Nesses casos, os indivíduos em questão não foram adequadamente testados.

Escores padrões

Uma forma de contornar o problema da desigualdade das unidades de percentil, e ainda assim transmitir o sentido dos escores de teste em relação a um grupo normativo ou de referência, é transformar os escores brutos em escalas que expressem a posição dos escores em relação à média em unidades de desvio padrão. Isso pode ser feito por meio de transformações lineares simples. Uma *transformação linear* altera as unidades nas quais os escores são expressos, ao mesmo tempo que deixa inalteradas as inter-relações entre eles. Em outras palavras, o formato da distribuição dos escores de uma escala derivada linearmente para um determinado grupo de testandos é o mesmo que o da distribuição original dos escores brutos. Uma grande vantagem desse procedimento é que normalmente os escores distribuídos de testes com média, desvio padrão e amplitude de escores diferentes podem

Advertência

Assim como nos capítulos anteriores, as fórmulas e os procedimentos estatísticos apresentados aqui são os *essencialmente* necessários para uma compreensão básica das transformações de escore. Embora as operações estatísticas descritas neste livro possam ser e sejam realizadas rotineiramente por programas de computador como o SPSS e o SAS, as fórmulas e os passos envolvidos devem ser compreendidos de modo a se obter um entendimento básico do sentido de vários escores.

ser comparados de forma significativa depois de serem transformados linearmente em uma escala comum, desde que seja usado o mesmo grupo de referência.

A primeira transformação linear realizada em escores brutos é convertê-los em desvios do escore padrão, ou escores *z*. Um escore *z* (ver Apêndice C) expressa a distância entre um escore bruto e a média do grupo de referência em termos do seu desvio padrão. Vamos recordar que a média e o desvio padrão dos escores *z* são 0 e 1, respectivamente, e que a distribuição dos escores *z* é bilateralmente simétrica, com metade dos casos em cada lado da média. A posição dos escores *z* relativa à média é indicada pelo uso de um sinal positivo (ou nenhum sinal) para os escores *z* que estão acima da média e um sinal negativo para os que estão abaixo dela. Por isso, o sinal de um escore *z* indica a direção na qual este se desvia da média de um grupo, enquanto que seu valor reflete a distância entre o escore e a média em unidades de desvio padrão. Por exemplo, um escore *z* de +1,25 indica que o escore bruto original está 1¼ unidade de desvio padrão acima da média do grupo, enquanto que um escore bruto que fica ¾ de unidade de desvio padrão abaixo da média se converte em um escore *z* de 0,75. Se a distribuição dos escores para a amostra de referência for normal, os escores *z* podem ser facilmente transformados em percentis consultando-se a Tabela de Áreas da Curva Normal apresentada e explicada no Apêndice C.

O quadro Consulta Rápida 3.3 mostra as formas de transformação linear utilizadas para derivar escores *z* de escores brutos e transformá-los em outros tipos de escores padrões. Como a transformação de escores brutos em escores *z* geralmente é a primeira no processo de transformações, os escores *z* são considerados o tipo mais básico de escore padrão, e muitas vezes são identificados simplesmente como escores padrões. Isso também os distingue de outros tipos familiares de escores

CONSULTA RÁPIDA 3.3

Fórmula para transformar escores brutos em escores z:

$$z = \frac{X - \bar{X}}{DP_x}$$

onde

X = Escore bruto
$\bar{X}$ = Média do grupo de referência
DP_x = Desvio padrão (DP) do grupo de referência

Fórmula para transformar escores z em outros escores padrões derivados:

Novo escore padrão = (Escore *z*)(Novo *DP*) + Nova média

Exemplo: Para transformar um escore *z* de +1,00 em um escore de QI como $M = 100$ e $DP = 15$,

Escore de QI = (+1,00)(15) + 100 = 115

derivados ou padrão, como os QIs, que se tornaram associados a testes específicos e dos quais trataremos a seguir.

Sistemas adicionais para derivação de escores padrões. Embora os escores z nos permitam saber imediatamente a magnitude e a direção da diferença entre qualquer escore e a média de sua distribuição, eles envolvem valores negativos e decimais. Por isso, geralmente sofrem outras transformações lineares. O objetivo dos sistemas de escore padrão que resultam destas transformações subseqüentes é simplesmente expressar os resultados de teste de forma mais conveniente.

Os números escolhidos como médias e desvios padrões para transformar os escores z em vários outros formatos de escore padrão são arbitrários. No entanto, em virtude de seu uso freqüente nos contextos em que são empregados, esses formatos de escore se tornaram familiares e adquiriram certos significados comumente compreendidos – que podem nem sempre ser justificados – por aqueles que os utilizam. A Figura 3.1 mostra a curva normal com as linhas de base para percentis, escores z e os seguintes sistemas de escore padrão de uso disseminado:

- *Escores-T* ($M = 50$, $DP = 10$), usados em muitos inventários de personalidade, como o Inventário Multifásico Minnesota de Personalidade (MMPI) e o Inventário Psicológico da Califórnia (CPI; *California Psychological Inventory*).
- *Escores da Junta de Seleção para Admissão Universitária (CEBB, College Entrance Examination Board* ($M = 500$, $DP = 100$), usado pelo SAT da Junta Universitária e pelo Serviço de Testagem Educacional para muitos programas de testagem na admissão em escolas técnicas e de graduação, como o *Graduate Record Exam* (GRE).
- *Escores de subtestes das escalas Wechsler* ($M = 10$, $DP = 3$), usados para todos os subtestes das escalas Wechsler, bem como para os subtestes de vários outros instrumentos.
- *QIs de desvio da escala Wechsler* ($M = C$, $DP = 15$), usados para os escores resumidos de todas as escalas Wechsler e outros testes, incluindo muitos que não denominam seus escores "QI".
- *Índices de Habilidade Escolar Otis-Lennon* ($M = 100$, $DP = 16$), usados no Teste de Habilidade Escolar Otis Lennon (OLSAT, *Otis-Lennon School Ability Test*), que é o título atual da série de testes grupais que começou com a Escala Grupal de Inteligência Otis.

Os índices OLSAT estão incluídos na Figura 3.1, como exemplo de um sistema de escore padrão com média de 100 e desvio padrão diferente de 15, para ilustrar a arbitrariedade da escolha de unidades nos sistemas de escore padrão. Embora a maioria dos sistemas que usa média de 100 opte por 15 como desvio padrão, existem alguns que usam outros desvios padrões, como 12 ou 18. Estas escolhas alternativas podem fazer uma diferença significativa na interpretação dos escores. Por exemplo, se dois testes de distribuição normal – ambos com média 100 – têm desvios padrões de 12 e 15, respectivamente, um escore padrão de 112 no primei-

Pondo em prática

Transformando escores brutos em escores z

1. Suponha que todos os alunos de uma turma de 8º série foram submetidos a testes de desempenho em estudos sociais, gramática e matemática. Os escores da turma nos três testes tiveram *distribuição normal*, mas os testes apresentaram as seguintes diferenças:

	Número total de itens	Média	DP
Teste de estudos sociais	50	35	5
Teste de gramática	20	15	3
Teste de matemática	100	70	10

2. Suponha também que você tem motivos para desejar comparar os escores de três alunos da turma uns com os outros e com a turma toda. Os três alunos em questão – Alfred, Beth e Carmen – tiveram os seguintes escores:

	Escores brutos		
	Alfred	Beth	Carmen
Teste de estudos sociais	49	38	48
Teste de gramática	15	12	18
Teste de matemática	68	95	75

3. Estes escores não podem ser comparados entre os testes, nem se pode tirar sua média, porque estão em escalas diferentes. Para comparar os escores, mesmo em um único teste, devemos consultar as médias, desvios padrões (*DPs*) e número de itens de cada teste. Uma forma mais fácil de compará-los é converter os escores brutos em escores *z* – subtraindo a média do teste respectivo de cada escore bruto e dividindo o resultado pelos desvios padrões correspondentes, como é mostrado na Fórmula do quadro Consulta Rápida 3.3 –, com os seguintes resultados:

	Escores z		
	Alfred	Beth	Carmen
Teste de estudos sociais	+2,80	+0,60	+2,60
Teste de gramática	0,00	-1,00	1+1,00
Teste de matemática	-0,20	+2,50	+0,50
Nota média	+0,87	+0,70	+1,37

4. As transformações lineares dos escores brutos em escores *z* nos permitem tirar a média das três notas para cada aluno e comparar o desempenho dos alunos em todos os testes entre si e com a classe toda. Além disso, uma vez que supomos distribuições normais para todos os testes, podemos usar a tabela do Apêndice C para traduzir cada escore *z* em um escore de postos de percentil.

ro teste vai se transformar em um escore *z* de +1,00 e ficar no 84º percentil, enquanto que no segundo teste um escore padrão de 112 vai se transformar em um escore *z* de +0,80 e se localizar apenas no 79º percentil.

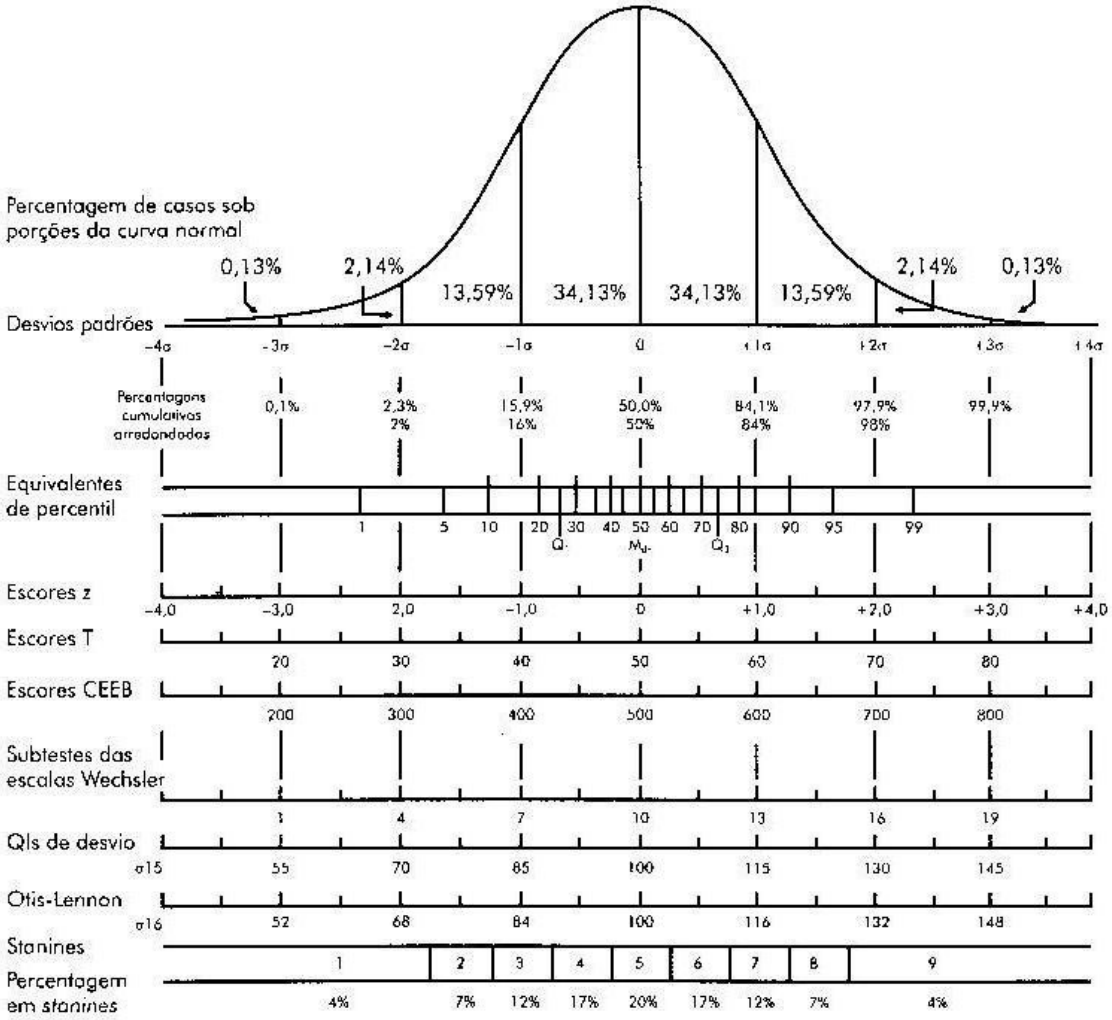

Figura 3.1 A curva normal, percentis e escores padrões selecionados.
Nota: Adaptado de *Test Service Notebook # 148* de The Psychological Corporation

Uma observação sobre QIs de desvio. Os escores conhecidos como *QIs de desvio* foram introduzidos por David Wechsler em 1939 para serem usados com sua primeira escala de inteligência, a Wechsler Bellevue I, que mais tarde se transformou na Escala de Inteligência Wechsler para Adultos (WAIS). Estes escores passaram a ser mais amplamente usados depois que a primeira edição da Escala de Inteligência Wechsler para Crianças (WISC) foi publicada em 1949, e são chamados de QIs *de desvio* para diferenciá-los dos QIs-*razão*, originais usados na Stanford-Binet e outras escalas. Os QIs de desvio são obtidos somando-se os escores de escala que o testando obtém em vários testes e localizando esta soma na tabela normativa apropriada, em vez de se usar a fórmula IM / IC x 100.

A escala de escores do tipo QI de Wechsler foi adotada por muitos outros criadores de testes para expressar os escores totais de vários instrumentos, incluindo a edição mais recente da Stanford-Binet, que usa 15 como unidade de desvio

Pondo em prática

Transformando escores z em diferentes escores padrões

1. Para ilustrar a conversão de escores z em diferentes tipos de sistemas de escore padrão, retornamos ao exemplo anterior dos três alunos de uma turma de 8ª série cujos escores z em três testes de distribuição normal foram os seguintes:

	Escores z		
	Alfred	Beth	Carmen
Teste de estudos sociais	+2,80	+0,60	+2,60
Teste de gramática	0,00	–1,00	+1,00
Teste de matemática	–0,20	+2,50	+0,50

2. Para transformar estes escores z em uma escala mais conveniente, aplicamos a fórmula para convertê-los em escores padrões, apresentada no quadro Consulta Rápida 3.3:

> Novo escore padrão = (Escore z)(Novo *DP*) + Nova média

3. Usando as médias e desvios padrões apropriados para cada sistema, obtemos os seguintes escores:

	Alfred	Beth	Carmen
Escores T (*M* = 50, *DP* = 10)			
Teste de estudos sociais	78	56	76
Teste de gramática	50	40	60
Teste de matemática	48	75	55
Escores do tipo CEEB (*M* = 500, *DP* = 100)			
Teste de estudos sociais	780	560	760
Teste de gramática	500	400	600
Teste de matemática	480	750	550
QIs da escala Wechsler (*M* = 100, *DP* = 15)			
Teste de estudos sociais	142	109	139
Teste de gramática	100	85	115
Teste de matemática	97	138	108

4. Observe que, como essas são transformações lineares, os escores permanecem exatamente nas mesmas posições em relação uns aos outros, independentemente das diferenças de unidade. Isso pode ser observado mais claramente na relação entre os escores T e os escores CEEB, já que a média e o desvio padrão do CEEB são iguais à média e o desvio padrão do escore T vezes 10; portanto, em nosso exemplo, cada escore CEEB equivale ao escore T correspondente vezes 10.

padrão em vez do desvio padrão original de 16. Entre a variedade de testes que emprega escores padrões com média de 100 e desvio padrão de 15 estão todos os principais instrumentos delineados para avaliar o funcionamento cognitivo geral,

como a série de testes Kaufman (*Kaufman-Assessment Battery for Children, Kaufman Adolescent and Adult Intelligence Test*, etc), as Escalas de Habilidade Diferencial, o Sistema de Avaliação Cognitiva Das-Naglieri e muitos outros. Embora usem o mesmo tipo de unidade para seus escores globais ou resumidos que os testes Wechsler, todos estes testes mais recentes descartaram o uso do termo QI para designar seus escores. Esta é uma manobra sensata, porque os chamados QIs de desvio *não* são quocientes. Além disso, como pode se notar nos títulos de alguns instrumentos mais novos, os autores dos testes estão abandonando o uso da palavra *inteligência* para designar o constructo avaliado por seus testes, em favor de outros termos mais neutros.

Transformações não-lineares

Nem todas as transformações de escore são lineares. As *transformações não-lineares* são aquelas que convertem uma distribuição de escore bruto em uma distribuição com um formato diferente do da original. Isso pode ser feito por meio de métodos que permitem ao criador de testes maior flexibilidade ao lidar com distribuições de escores brutos do que com as conversões lineares. Embora algumas transformações não-lineares de escores envolvam operações complexas que estão muito além do âmbito deste livro, outras são simples. Por exemplo, a transformação de escores brutos de distribuição normal em escores de postos de percentil, que já consideramos, é uma conversão não-linear. Ela é feita transformando-se cada escore bruto em um escore *z* e localizando o escore *z* na Tabela de Áreas da Curva Normal para derivar a proporção ou percentagem da área da curva normal que está abaixo daquele ponto. Nesta seção, discutiremos outros dois tipos muito usados de escores padrões derivados não-linearmente.

Escores padrões normalizados são usados quando uma distribuição de escores se aproxima da distribuição normal mas não chega a se igualar a ela. Para normalizar escores, primeiramente é necessário encontrar a percentagem de pessoas na amostra de referência que se localiza exatamente em cada escore bruto ou abaixo dele (ver, p. ex., a coluna de Percentagem Cumulativa *[PC]* para a distribuição de 60 escores na Tabela 2.3). A seguir, as percentagens são convertidas em proporções. Essas proporções são então localizadas na Tabela de Áreas da Curva Normal para se obter os escores *z* correspondentes àquelas áreas (ver Apêndice C). Os escores padrões derivados dessa forma são indistinguíveis daqueles obtidos com a fórmula da transformação linear, mas sempre devem ser identificados como escores padrões normalizados a fim de alertar o usuário do teste para o fato de que tiveram origem em uma distribuição que não era normal. Depois que os escores padrões normalizados são obtidos, podem ser transformados em qualquer um dos outros sistemas de escore padrão convenientes que já discutimos, como os escores T, QIs de desvio ou escores CEEB, usando-se o mesmo procedimento empregado para os escores *z* linearmente derivados descritos no quadro Consulta Rápida 3.3.

As estaninas foram criadas originalmente pela Força Aérea Americana durante a Segunda Guerra Mundial. A escala do "*standard nine*" (padrão nove), ou estanina,

transforma todos os escores de uma distribuição em números de um único dígito de 1 a 9. Este artifício tem a vantagem distinta de reduzir o tempo e o esforço necessários para se digitar os escores no computador para armazenamento e processamento. As transformações para estanina também fazem uso de distribuições de freqüência cumulativa e percentagem cumulativa como as da Tabela 2.3; os escores estanina são alocados a partir da percentagem de casos em faixas de escore determinadas. A Tabela 3.2, Parte A, mostra as percentagens da curva normal dentro de cada unidade de estanina e as percentagens cumulativas usadas para converter escores brutos em estaninas. A Parte B contém alguns exemplos de transformações para estaninas usando alguns escores da Tabela 2.3. A Figura 3.1 mostra a posição dos escores estaninas dentro da curva normal. Como pode ser visto na figura, o estanina 5 engloba os 20% do meio dos escores. A média da escala de estanina é 5 e seu desvio padrão é aproximadamente 2. Embora a escala

Pondo em prática

Escores padrões normalizados

Para demonstrar o processo de normalização dos escores de uma distribuição que não se conforma à curva normal, mas se aproxima dela, podemos usar cinco dos escores brutos da distribuição de 60 escores de teste apresentada na Tabela 2.3 do Capítulo 2.

Os escores brutos selecionados arbitrariamente para este exercício são 49, 40, 39, 36 e 29.

Os passos envolvidos nesta conversão são os seguintes:

Escore bruto → Percentagem cumulativa (PC) → Proporção cumulativa (pc) → Escore z normalizado

1. O escore bruto e a percentagem cumulativa são localizados na distribuição.
2. A proporção cumulativa é a percentagem cumulativa dividida por 100.
3. Um escore z normalizado é obtido na Tabela de Áreas da Curva Normal do Apêndice C encontrando-se a proporção da área da curva que mais se aproxima da proporção cumulativa para um determinado escore. Para escores com proporções cumulativas acima de 0,50, as áreas da porção maior – da coluna (3) da tabela – devem ser usadas para obter os escores z normalizados, que vão portar um sinal negativo. Para escores com proporções cumulativas abaixo de 0,50, as áreas da porção menor – da coluna (4) da tabela – devem ser usadas, e os escores z normalizados resultantes vão ter sinais negativos.
4. Esses procedimentos produzem os seguintes resultados:

Escore bruto	% Cumulativa	cp	Escore z normalizado[a]
49	98,3	0,983	+2,12
40	53,3	0,533	+0,08
39	45,0	0,450	–0,13
36	23,3	0,233	–0,73
29	1,7	0,017	–2,12

[a]Os escores padrões normalizados têm o mesmo significado que os escores padrões derivados linearmente, em termos da posição dos escores brutos originais que representam, em relação a suas distribuições. Eles podem ser transformados em vários outros sistemas de escore padrão, mas sempre devem ser identificados como derivados não-linearmente de um procedimento de normalização.

Tabela 3.2 Convertendo escores brutos em escores estanina

A. Percentagem da curva normal para uso na conversão em estaninas

	Estaninas								
	1	2	3	4	5	6	7	8	9
Percentagem de casos dentro de cada estanina	4	7	12	17	20	17	12	7	4
Percentagem cumulativa em cada estanina	4	11	23	40	60	77	89	96	100

B. Escores de testes selecionados da Tabela 2.3 convertidos em estaninas

Escores selecionados[a]	% cumulativa	Escores estanina
49	98,3	9
47	93,3	8
45	83,3	7
43	71,7	6
41	60,0	5
39	45,0	5
37	30,0	4
35	20,0	3
33	11,7	3
31	5,0	2
29	1,7	1

[a]Ver Tabela 2.3 no Capítulo 2 para a distribuição completa dos 60 escores de teste.

de estanina seja econômica e simples, sua brevidade e simplicidade também resultam em uma certa perda de precisão.

COMPARAÇÕES INTERTESTES

Na maioria das circunstâncias, os escores de testes referenciados em normas não podem ser comparados, a menos que tenham sido obtidos do mesmo teste, usando-se a mesma distribuição normativa. Outro motivo para a falta de comparabilidade dos escores tem origem nas diferenças de unidades das escalas, como os vários tamanhos de unidade de DP discutidos anteriormente em conexão com os QIs de desvio. Além disso, mesmo quando os testes, as normas e as unidades de escala empregados são os mesmos, os escores não necessariamente têm o mesmo significado. Quando são usados no contexto da avaliação individual, deve-se ter em mente que muitos outros fatores alheios ao teste também podem influenciar seus resul-

Não esqueça

1. Escores de teste não podem ser comparados significativamente se
   - os testes ou versões dos testes são diferentes,
   - os grupos de referência são diferentes,
   - as escalas de escore diferem,

   exceto quando os testes, grupos ou escalas foram equacionados intencionalmente. A monografia de Angoff (1984) sobre *Escalas, normas e escores equivalentes* é uma das melhores fontes de informação sobre procedimentos de equacionamento.
2. Mesmo quando os escores de teste se tornaram comparáveis, tanto o contexto no qual a testagem acontece como o histórico dos testandos precisam ser levados em consideração na interpretação dos resultados.

tados (p. ex., a origem e a motivação do testando, a influência dos examinadores e as circunstâncias em que os testes são aplicados).

Procedimentos de equacionamento

Independentemente das advertências já feitas na seção anterior, existe uma série de situações nas quais é necessário ou desejável comparar os escores de indivíduos ou grupos ao longo do tempo ou em várias funções psicológicas, em relação a uma norma uniforme. Para estas situações, os criadores e editoras de testes criaram vários meios de se alcançar *alguma* comparabilidade de escores entre testes. Essencialmente, eles foram delineados para colocar os escores em um referencial comum, com o benefício adicional de reduzir o custo e o tempo consideráveis envolvidos na padronização de testes. Muitos procedimentos de equacionamento envolvem detalhes altamente técnicos, descritos na monografia de Angoff (1984) e outras fontes (Petersen et al., 1989). As seguintes descrições abreviadas de algumas abordagens usadas com maior freqüência fornecem explicações básicas sobre o que elas envolvem.

- *Formas alternativas* consistem em duas ou mais versões de um teste que podem ser usadas de forma intercambiável, para os mesmos fins, e administradas de forma idêntica. Criar formas alternativas semelhantes em conteúdo, mas que variam nos itens específicos, é um dos meios mais simples de produzir testes comparáveis. Uma forma mais estrita de comparabilidade pode ser alcançada com o tipo de versão alternativa conhecida como *forma paralela*. Estas formas são equacionadas não apenas em conteúdo e procedimentos, mas também em algumas características estatísticas, como médias dos escores brutos e desvios padrões, bem como índices de fidedignidade e validade. As formas alternativas são especialmente úteis quando uma pessoa tem que ser submetida ao mesmo teste mais de uma vez. Os *efeitos da aprendizagem* (isto é, o aumento do escore que pode ser atribuí-

do à exposição anterior aos itens do teste ou itens semelhantes) entram em jogo quando uma forma alternativa de um teste já realizado é administrada, mas não são tão grandes quanto seriam se a mesma versão fosse administrada mais uma vez.

- *Testes-âncora* consistem em conjuntos comuns de itens administrados a diferentes grupos de examinandos no contexto de dois ou mais testes, e oferecem uma solução diferente ao problema da comparabilidade dos escores de teste. Ter as respostas de mais de um grupo normativo para conjuntos comuns de itens em um mesmo intervalo de tempo permite o uso de procedimentos de equacionamento de escores – baseados em estatísticas derivadas dos itens comuns – que possibilitam extrapolar e comparar os escores de um teste aos de outro, tanto para indivíduos quanto para grupos. Esta técnica pode ser usada quando os criadores de testes desejam realizar comparações de níveis de desempenho em diferentes áreas de habilidade – como compreensão de leitura e expressão escrita – de dois testes diferentes em uma escala uniforme. Mais recentemente, no entanto, esse tipo de comparabilidade passou a ser obtido mais facilmente por meio da normatização simultânea ou de técnicas de teoria da resposta ao item, que serão discutidas mais adiante.
- *Grupos de referência fixos* oferecem um meio de se obter alguma comparabilidade e continuidade nos escores ao longo do tempo. Este método faz uso de testes-âncora embutidos em cada forma sucessiva de um teste, para criar um elo com uma ou mais formas anteriores do mesmo. Deste modo, uma série de testes permanece unida, por meio de uma corrente de itens comuns, aos escores do grupo selecionado como referência fixa para manter a continuidade da escala de escores ao longo do tempo. O SAT da Junta Universitária é o exemplo mais conhecido de um teste que faz uso de grupos fixos de referência. Até abril de 1995, todos os escores SAT eram expressos em termos da média e do desvio padrão dos 11 mil vestibulandos que fizeram o teste em 1941. Nesta escala, um escore de 500 correspondia à média do grupo de referência fixo de 1941, o escore de 400 ficava 1 *DP* abaixo desta média, e assim por diante. Depois de abril de 1995, os escores SAT relatados refletem um *recentramento* da escala de escores nos vestibulandos contemporâneos, de tal modo que um escore de 500 representa o nível médio atual de desempenho no teste. O uso de um grupo de referência fixo no SAT ao longo de várias décadas permitiu a avaliação de aumentos ou diminuições no calibre do desempenho dos vestibulandos em diferentes épocas, como pode ser visto no quadro Consulta Rápida 3.4. Além dos escores padrões recentrados, os escores de postos de percentil dos vestibulandos no SAT ainda podem ser, e ainda são, relatados usando-se o grupo mais recente de formandos do ensino médio como grupo de referência.
- A *normatização simultânea* de dois ou mais testes com a mesma amostra de padronização, muitas vezes referida como *co-normatização*, é outro método usado para se obter a comparabilidade de escores. Ao normatizar

CONSULTA RÁPIDA 3.4

Alterando os padrões dos grupos de referência por meio do recentramento do SAT

Entre as décadas de 1940 e 1990, os escores dos vestibulandos nas seções de Expressão Verbal e de Matemática do SAT vinham demonstrando um declínio significativo. Por isso, depois do recentramento realizado na década de 1990, o escore Verbal recentrado de 500 passou a ser equivalente a um escore de 420, quando comparado ao grupo de referência de 1941. Esta mudança representa um declínio de quase 1 unidade de *DP*. O escore recentrado de Matemática de 500 foi considerado equivalente a um escore de 470 para o grupo de referência de 1941. Vários motivos foram propostos para estas mudanças. O mais plausível diz respeito a maior diversidade socioeconômica e étnica dos vestibulandos e a mudanças na qualidade dos currículos escolares do ensino médio entre o início dos anos de 1940 e início dos anos de 1990.

testes ao mesmo tempo e com o mesmo grupo de pessoas, podemos comparar o desempenho de indivíduos ou subgrupos em mais de um teste, usando o mesmo padrão. Esta possibilidade é particularmente útil quando se deseja contrastar níveis relativos de desempenho em duas ou mais funções psicológicas, como níveis de vocabulário expressivo e receptivo ou memória de curto e longo prazo, para o mesmo indivíduo ou subgrupo. O Woodcock-Johnson III (WJ III) oferece um excelente exemplo de co-normatização de duas baterias de testes. Os Testes WJ de Habilidades Cognitivas (WJ III COG) são uma bateria criada para medir funções cognitivas gerais e específicas, enquanto que a bateria Testes de Desempenho WJ III (WJ III ACH) tem como objetivo avaliar os pontos fortes e fracos de uma pessoa em termos acadêmicos. Estas duas baterias foram normatizadas com a mesma amostra grande de indivíduos, cujas faixas etárias iam de pré-escolares a adultos mais velhos, representativa da população dos Estados Unidos, e por isso oferece amplas oportunidades para a comparação de níveis intra-individuais de desempenho em diversos índices de funcionamento cognitivo e habilidades acadêmicas.

Teoria da resposta ao item (TRI)

Uma variedade de procedimentos sofisticados baseados em modelos matemáticos estão cada vez mais substituindo as técnicas tradicionais de equacionamento descritas acima. Estes procedimentos, que remontam à década de 1960 e também são conhecidos como *modelos de traços latentes*, com freqüência são agrupados sob o nome de *teoria da resposta ao item* (TRI). O termo *traço latente* reflete o fato de que estes modelos buscam estimar os níveis de várias habilidades, traços ou constructos psicológicos não-observáveis, subjacentes ao comportamento observável dos indivíduos, demonstrados por suas respostas aos itens dos testes. Ao contrário das técnicas discutidas anteriormente para o equacionamento de testes e escores, os métodos TRI aplicam modelos matemáticos a dados de *itens* de teste derivados de

amostras grandes e diversificadas, daí o nome *teoria da resposta ao item*. Estes dados são usados para calibrar os itens de teste em relação a um ou mais parâmetros e derivar estimativas de probabilidade da quantificação de habilidade, ou nível de traço, necessária para responder a cada item de uma certa maneira. Essencialmente, estes modelos colocam pessoas e itens de teste em uma escala comum (Embretson e Reise, 2000).

Outros elementos e procedimentos básicos da TRI serão discutidos mais aprofundadamente no Capítulo 6. No contexto atual, no entanto, o aspecto a ser destacado é que, se satisfizerem certas condições, os modelos TRI podem produzir estimativas de parâmetros de item que são *invariantes* entre populações. Isto significa que estas estimativas não estão necessariamente atreladas ao desempenho de qualquer grupo de referência específico. Em vez disso, os dados das respostas aos itens podem ser interpretados em termos de uma dimensão de habilidade ou traço. Por isso, quando os modelos TRI são aplicados aos conjuntos de respostas aos itens e dados de escores de teste de várias amostras e as premissas dos modelos são satisfeitas, eles podem ser usados de duas formas recíprocas: (a) para estimar a probabilidade de que pessoas com níveis específicos da habilidade ou traço em questão vão responder ao item corretamente ou de uma determinada forma e (b) para estimar os níveis de traço necessários para se obter uma probabilidade especificada de responder ao item de uma maneira específica.

Testagem adaptativa computadorizada

Uma das principais vantagens da metodologia TRI é que ela é idealmente adequada para uso na testagem adaptativa computadorizada (CAT, *computerized adaptive testing*). Quando os indivíduos se submetem a testes adaptativos computadorizados (CATs), seus níveis de habilidade podem ser estimados a partir das respostas aos itens dos testes durante o processo de testagem, e estas estimativas são usadas para selecionar o conjunto subseqüente de itens apropriados aos níveis de habilidade do testando. Embora muitos CATs tenham um número fixo de itens, outros são delineados de tal forma que a testagem pode ser interrompida sempre que uma regra específica de interrupção for satisfeita e níveis de traço ou habilidade tiverem sido estabelecidos com precisão suficiente. Em qualquer caso, os procedimentos de CAT diminuem a duração dos testes e o tempo de testagem significativamente, e também podem reduzir a frustração que muitas pessoas submetidas a testes de papel e lápis podem experimentar quando são expostas a itens consideravelmente acima ou abaixo de seus níveis de capacidade. Programas de testagem em larga escala, como os do Serviço de Testagem Educacional e do Departamento de Defesa dos Estados Unidos têm testado e usado a metodologia CAT por vários anos (Campbell e Knapp, 2001; Drasgow e Olson-Buchanan, 1999). Embora tenha vantagens claras em relação aos testes de papel e lápis de duração fixa – e sua gama de aplicações tenda a se expandir – esta metodologia apresenta alguns problemas novos relacionados à segurança e aos custos do teste e à impossibilidade dos examinandos de revisarem e alterarem suas respostas (Wainer, 2000).

Revisões de testes

Os nomes dos testes podem facilmente apresentar uma fonte de confusão para os usuários que não são suficientemente bem-informados. Julgar o conteúdo do teste apenas por seu título raramente se justifica. Algumas discrepâncias entre os títulos dos testes e o que eles efetivamente avaliam são bastante óbvias: os inventários de personalidade não fazem um verdadeiro inventário da personalidade. Outras não ficam tão aparentes: os testes de aptidão podem ou não testar aptidões, e assim por diante.

Mesmo que tenham o mesmo nome, dois testes podem não ser equivalentes. Muitos passam por revisões de tempos em tempos e conservam o mesmo título, exceto pela adição de um número ou letra que identifica uma versão específica (p. ex., WISC, WISC-R, WISC-III, WISC-IV). Em geral, os testes mais bem-sucedidos e mais amplamente usados têm maior probabilidade de serem revisados do que outros. Os objetivos e a magnitude das revisões podem variar de pequenas mudanças na formação dos itens até grandes reorganizações do conteúdo, forma de pontuação ou procedimentos administrativos.

Um teste que foi modificado de alguma forma não pode ser considerado comparável a uma versão anterior, a menos que a semelhança seja estabelecida empiricamente. Quando a revisão de um teste é pequena e não afeta os escores, a comparabilidade das versões antiga e revisada pode ser estabelecida de forma relativamente fácil. Isso geralmente é feito administrando-se ambas as versões do teste ao mesmo grupo de pessoas. Se as duas versões forem altamente correlacionadas e tiverem médias, desvios padrões e distribuições de escore semelhantes, para a maioria dos objetivos práticos pressupõe-se que sejam equivalentes. Um exemplo típico desta prática é quando testes de lápis e papel, como o MMPI-2, são transferidos para um formato de administração computadorizada sem qualquer alteração em seu conteúdo ou forma de pontuação.

Grandes revisões de testes referenciados em normas, por outro lado, requerem a padronização do teste com uma nova amostra normativa. Por isso, quando as alterações são significativas o bastante para justificar diferenças na escala ou forma de pontuação do teste, na verdade se está lidando com um teste novo, embora este possa manter alguma semelhança ou o mesmo título de versões anteriores. Um exemplo proeminente é a Escala de Inteligência Stanford-Binet (S-B), que foi publicada pela primeira vez em 1916. A quarta e a quinta edições da S-B, publicadas em 1986 e 2003, respectivamente, são totalmente diferentes de suas versões anteriores em quase todos os aspectos, e ao longo do tempo se tornaram mais semelhantes em forma e conteúdo às escalas Wechsler do que à S-B original.

Alterações longitudinais em normas de testes

Quando um teste é revisado e padronizado com uma nova amostra após um período de vários anos, mesmo que as revisões do conteúdo sejam pequenas, as normas de pontuação tendem a se desviar em uma direção ou outra devido a mudanças na

população em diferentes períodos de tempo. Uma dessas mudanças, discutida no quadro Consulta Rápida 3.4, é o declínio nos escores médios do SAT entre o grupo de referência fixo testado em 1941 e os vestibulandos da década de 1990. Uma tendência longitudinal surpreendente na direção oposta, conhecida como *efeito Flynn*, foi bem documentada em revisões sucessivas dos principais testes de inteligência (como as escalas Wechsler e S-B) que invariavelmente envolvem a administração das versões antigas e novas com um segmento da nova amostra de padronização para fins comparativos. Dados de revisões de vários testes de inteligência nos Estados Unidos e em outros países – extensamente analisados por J. R. Flynn (1984, 1987) – demonstram uma tendência pronunciada de elevação, a longo prazo, no nível de desempenho necessário para se obter qualquer escore de QI. O efeito Flynn presumivelmente reflete os ganhos populacionais ao longo do tempo no tipo de desempenho cognitivo avaliado pelos testes de inteligência. Uma variedade de fatores – como melhoras na nutrição, cuidados pré-natais e maior complexidade do ambiente – foram propostos como razões para este achado. Ainda assim, o grau em que o efeito Flynn se generaliza para populações no mundo todo – bem como possíveis causas onde ele aparece – continuam a ser objeto de considerável controvérsia (Neisser, 1998).

Um exemplo pertinente de alterações longitudinais no desempenho em testes de personalidade foi observado na renormatização do MMPI (originalmente publicado na década de 1940), que aconteceu na década de 1980 como parte do desenvolvimento do MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen e Kaemmer, 1989).

Pondo em prática

Como o *Efeito Flynn* pode afetar mudanças aparentes nos escores de testes de inteligência: Um exemplo usando a Escala de Inteligência Wechsler para Crianças (WISC)

John obteve um escore de QI Wechsler de 117 na WISC-R aos 9 anos, e um escore de QI Wechsler de 107 na WISC-III aos 13 anos. Este declínio aparentemente significativo em seu desempenho não pode ser aceito sem uma análise.

Um fator que pode explicar em parte o declínio aparente é o efeito Flynn. Este efeito se refere ao nível mais alto de desempenho tipicamente observado nos grupos normativos de versões mais novas dos testes de inteligência geral comparados a suas versões mais antigas – por exemplo, o WISC-III, publicado em 1991, comparado ao WISC-R, publicado em 1974. Segundo o manual do WISC-III, uma amostra de 206 crianças entre 6 e 16 anos de idade foi submetida a ambas as versões em ordem contrabalançada, com intervalos que variaram de 12 a 70 dias entre os dois testes. Seus QIs globais médios foram 108.2 na WISC-R e 102.9 na WISC-III (Wechsler, 1991).

Embora a correlação entre as duas versões da WISC para esta amostra fosse alta (.89), os escores das crianças na WISC-III foram em média mais baixos – com uma diferença de um pouco mais de cinco pontos – do que no WISC-R. A diferença nos escores médios de QI deste grupo indica que as normas dos dois testes são tais que, no todo, qualquer escore de QI representa um nível superior de desempenho no teste mais novo do que no mais antigo.

Naturalmente, muitos fatores, além do efeito Flynn, podem ter contribuído para a diferença entre os dois QIs de John. Os principais são o erro de mensuração em cada escore (ver Capítulo 4) e a possibilidade de que John tenha passado por um declínio efetivo em seu funcionamento intelectual geral devido a alguma doença ou evento de vida.

As alterações no conteúdo do inventário foram relativamente pequenas. Mesmo assim, a mudança das normas originais para as do MMPI-2 resultou em modificações nos níveis de escore consideradas clinicamente significativas em várias escalas, devido a mudanças substanciais no modo como as pessoas dos dois diferentes períodos de tempo respondiam aos itens.

O tipo de mudanças descritas acima são lugar-comum quando se atualizam as amostras normativas. Elas enfatizam a necessidade desta atualização para que se mantenha a atualidade das normas intragrupo. Estas mudanças também explicam por que a documentação que acompanha qualquer tipo de teste referenciado em normas deve oferecer uma descrição completa e precisa da composição das amostras usadas para sua padronização e normatização, incluindo as datas em que as amostras foram montadas. Os documentos dos testes devem incorporar respostas para todas as perguntas listadas no quadro Consulta Rápida 3.2, incluindo uma exposição clara dos passos usados no desenvolvimento do teste e nos procedimentos de administração e pontuação usados durante sua padronização. Os usuários de testes, por sua vez, precisam estar atentos a estas informações e levá-las em consideração ao selecionar testes e interpretar seus escores.

INTERPRETAÇÃO DE TESTES REFERENCIADOS EM NORMAS

No campo da avaliação educacional e ocupacional, os testes costumam ser usados para ajudar a determinar se uma pessoa alcançou um certo nível de competência em um campo de conhecimento ou habilidade no desempenho de uma tarefa. Nestes casos, o referencial para a interpretação do escore do teste deve mudar. Em vez de comparar o desempenho de uma pessoa ao de outras, o desempenho de um indivíduo ou grupo é comparado a um critério ou padrão predeterminado. Quando usado neste contexto, o termo *critério* pode se referir tanto ao conhecimento de um domínio específico de conteúdo, como à competência em algum tipo de ação. Os padrões pelos quais os testes referenciados em critérios são avaliados permanecem tipicamente definidos em termos de níveis especificados de conhecimento ou especialização necessários para se obter a aprovação em um curso, um diploma ou licença profissional, e também podem envolver a demonstração de competência suficiente para realizar um trabalho ou criar um produto. Muitas vezes, mas não sempre, a aplicação dos testes referenciados em critérios envolve o uso de escores de corte, ou faixas de escore que separam a competência da incompetência ou demarcam diferentes níveis de desempenho. Nestes casos, a validade das inferências feitas a partir dos escores precisa ser estabelecida por meio de ligações empíricas entre os escores no teste e o desempenho no critério.

Variedades de interpretação de testes referenciados em critérios

O termo *testagem referenciada em critérios*, popularizado por Glaser (1963), às vezes é usado como sinônimo de *testagem referenciada em domínios*, *em conteúdos* ou *em objetivos*, ou *testagem de competência*. Esta situação produz confusão, em

parte devido ao fato de que a interpretação de testes referenciados em critérios faz uso de pelo menos dois conjuntos subjacentes de padrões: (a) aqueles baseados na *quantidade de conhecimento* sobre um domínio de conteúdo, demonstrados em testes objetivos padronizados e (b) os que se baseiam no *nível de competência* em uma área de habilidade, demonstrados pela qualidade do desempenho em si ou do produto que resulta do exercício da habilidade. Ocasionalmente, o termo *testagem referenciada em critérios* também é usado para se referir a interpretações baseadas na relação preestabelecida entre os escores de um teste e os níveis esperados de desempenho em um critério, como uma ação futura ou mesmo outro teste. Neste uso em particular, o "critério" é um resultado específico e pode estar relacionado ou não às tarefas propostas pelo teste. Isto contrasta acentuadamente com os testes referenciados em conteúdos ou baseados no desempenho, nos quais as tarefas são essencialmente amostras de comportamento diretamente relacionadas ao critério. O quadro Consulta Rápida 3.5 lista alguns dos principais aspectos nos quais a testagem referenciada em normas difere da testagem referenciada em critérios, bem como algumas diferenças entre os tipos de testagem referenciada em critérios.

Além dessas distinções, o modo em particular como os testes referenciados em critérios são usados também varia. Às vezes, os critérios são estritamente quantitativos, como quando são determinadas certas percentagens de respostas corretas (p. ex., 80 ou 90%) necessárias para estabelecer a maestria adequada. Em outros casos, os critérios são mais qualitativos e subjetivos. Além disso, às vezes, o desempenho nestes testes é avaliado na base do tudo-ou-nada em relação a se um certo nível de competência foi atingido, e, às vezes, pode haver graduações para níveis intermediários de competência.

Apesar das diferenças de ênfase e nomenclatura entre os testes referenciados em critérios, estes instrumentos têm algumas características em comum. Tipicamente, os testes referenciados em critérios (a) têm por objetivo avaliar o grau em

CONSULTA RÁPIDA 3.5

Interpretação de testes referenciados em normas *versus* testes referenciados em critérios

- Os testes *referenciados em normas* buscam localizar o desempenho de um ou mais indivíduos em relação ao constructo que o teste avalia, em um contínuo criado pelo desempenho de um grupo de referência.
- Os testes *referenciados em critérios* buscam avaliar o desempenho dos indivíduos em relação a padrões relacionados ao constructo em si.
- Enquanto na interpretação de testes referenciados em normas o referencial sempre são pessoas, na interpretação de testes referenciados em critérios o referencial pode ser
  - o conhecimento sobre um domínio de conteúdo, demonstrado em testes padronizados objetivos;
  - o nível de competência exibido na qualidade do desempenho ou de um produto.
- O termo *testagem referenciada em critérios* às vezes também é aplicado para descrever a interpretação de testes que usam a relação entre os escores e níveis esperados de desempenho ou posicionamento em um critério como referencial.

que os testandos são proficientes em certas habilidades ou domínios de conhecimento e (b) são pontuados de tal forma que o desempenho de uma pessoa não influencia o resultado relativo das outras. Enquanto os testes referenciados em normas buscam ordenar ou localizar um ou mais indivíduos em relação a outros com respeito ao constructo que avaliam, os testes referenciados em critérios buscam avaliar o desempenho de indivíduos em relação ao constructo em si.

No presente contexto, somente serão discutidos aqueles aspectos da interpretação de testes referenciados em critérios necessários para uma compreensão básica de suas premissas e terminologia. Diversos desses conceitos são revisitados mais detalhadamente no Capítulo 5 em conexão com o tópico da validade, com o qual estão intimamente relacionados. O quadro Consulta Rápida 3.6 lista algumas fontes de informação retiradas da extensa literatura que está disponível sobre vários tipos de testagem referenciada em critérios.

Testando o conhecimento de domínios de conteúdo

Para se usar o conhecimento de domínios de conteúdo como referencial para a interpretação de escores de testes, deve haver um campo ou tema cuidadosamente definido e claramente demarcado a partir do qual derivam-se amostras (isto é, itens ou tarefas) para medir o conhecimento do testando. Além disso, os objetivos a serem avaliados tanto em termos do conhecimento de um domínio de conteúdo e das aplicações deste conhecimento, como dos padrões a serem usados na avaliação desses objetivos devem ter origem em um consenso de especialistas na área. Esta situação é encontrada basicamente no contexto de educação ou de treinamento, onde as matérias e disciplinas tendem a ser divididas em lições, cursos, programas

CONSULTA RÁPIDA 3.6

Leituras selecionadas sobre testagem referenciada em critérios

A literatura sobre a testagem referenciada em critérios, que remonta aos anos de 1960, é abundante. Para se aprofundar neste tópico, os leitores podem consultar uma ou mais das seguintes fontes:

- O livro de Cizek (2001), *Setting performance standards*, que se concentra nos aspectos teóricos e práticos do estabelecimento de padrões e suas muitas ramificações.
- O artigo clássico de Popham e Husek (1969) sobre as implicações da mensuração referenciada em critérios.
- O capítulo do livro de Hambleton e Rogers (1991) sobre "Avanços na mensuração referenciada em critérios", que oferece uma introdução útil a este campo de estudos e uma descrição dos avanços técnicos ocorridos entre as décadas de 1960 e 1980.
- A obra organizada por Wigdor e Green (1991) sobre os vários aspectos da avaliação de desempenho no local de trabalho.
- A exposição breve, porém esclarecedora, de Linn (1994) sobre parte da confusão nas várias interpretações do sentido da testagem referenciada em critérios.

As referências completas destes trabalhos estão listadas na seção de referências no final deste livro.

de estudo e outras unidades curriculares às quais os alunos são expostos e das quais se podem colher amostras de áreas de conteúdo e resultados de aprendizagem. Estes testes geralmente são descritos como medidas de "desempenho" e tendem a ter itens – como perguntas de múltipla escolha – que exigem que os testandos selecionem uma resposta ou completem uma tarefa altamente estruturada (como escrever um parágrafo curto sobre um tópico ou resolver um problema matemático).

Quando domínios de conhecimento são o referencial para a interpretação de testes, a questão a ser respondida é "quanto do domínio especificado o testando conhece?", e os escores com freqüência são apresentados na forma de percentagens de respostas corretas. Esse tipo de interpretação de teste referenciado em critérios muitas vezes é descrito como *testagem referenciada em conteúdo ou domínio*. Na verdade, alguns consideram esses dois termos sinônimos de testagem referenciada em critérios. O planejamento desses testes requer o desenvolvimento de uma *tabela de especificações* com células que especifiquem o número de itens ou tarefas a serem incluídas no teste para cada um dos objetivos de aprendizagem e áreas de conteúdo que este quer avaliar. A proporção de itens do teste alocados para cada célula reflete o peso ou a importância designada a cada objetivo e área. O quadro Consulta Rápida 3.7 mostra exemplos de vários objetivos e itens em duas áreas de conteúdo típicas dos testes referenciados em domínios. Exemplos de tabelas de especificações para testes referenciados em conteúdos e informações sobre como prepará-las podem ser encontrados em Gronlund (2003) e Linn e Gronlund (1995, p.119-125).

Avaliação de desempenho

Para fins de tomada de decisões no local de trabalho e no campo da educação, muitas vezes existe a necessidade de determinar ou certificar a competência no desempenho de tarefas que são mais realistas, mais complexas, mais demoradas ou mais difíceis de avaliar do que aquelas típicas da testagem referenciada em conteúdos ou domínios. Este tipo de situação demanda a avaliação do desempenho através de amostras de trabalho, produtos do trabalho ou alguma outra demonstração comportamental de competência e habilidade em situações que simulem o contexto da vida real.

Quando o objetivo de uma avaliação é determinar níveis de competência no contexto em que as habilidades são aplicadas na vida real, o critério na interpretação de testes *referenciados em critérios* é a qualidade do próprio desempenho ou do produto que resulta da aplicação de uma habilidade. Neste referencial, as questões típicas a serem respondidas são "o testando demonstra competência na habilidade em questão?" ou "Qual é a proficiência deste testando ou grupo de testandos no contínuo de competência relevante para esta tarefa em particular?".

Avaliação e pontuação na avaliação de desempenho

À luz das questões que a avaliação de proficiência deve responder, a avaliação do desempenho acarreta um conjunto diferente de procedimentos daqueles usados

CONSULTA RÁPIDA 3.7

Exemplos de objetivos e itens de testes referenciados em domínios

I. Domínio: Aritmética
   A. Área de conteúdo a ser avaliada: Multiplicação de frações
   B. Objetivos a serem avaliados:
      1. Conhecimento dos passos envolvidos na multiplicação de frações
      2. Compreensão dos princípios básicos envolvidos na multiplicação de frações
      3. Aplicação dos princípios básicos à resolução de problemas de multiplicação de frações
   C. Amostras de itens de teste para cada objetivo:
      Item 1. Listar os passos envolvidos na multiplicação de frações.
      Item 2. Desenhar um diagrama para mostrar ¼ de ½ de uma torta.
      Item 3. Quanto é ¾ x ½?

II. Domínio: Vocabulário
   A. Área de conteúdo a ser avaliada: Conhecimento de palavras
   B. Objetivos a serem avaliados:
      1. Definição de palavras
      2. Compreensão do sentido das palavras
      3. Aplicação do conhecimento de palavras na expressão escrita
   C. Amostras de itens de testes para cada objetivo:
      Item 1. O que significa marinheiro?_____
      Item 2. Qual palavra tem o sentido mais aproximado de "marinheiro"?
         a. marinar
         b. marimba
         c. pescador
         d. pirata
         e. andarilho
      Item 3. Faça uma frase usando "marinheiro" em um contexto significativo.

quando se testa o conhecimento em domínio de conteúdo. Em geral, a avaliação de desempenho tende a se concentrar mais pesadamente no julgamento subjetivo. Uma exceção a esta regra ocorre quando os critérios podem ser quantificados em termos de velocidade de desempenho, número de erros, unidades produzidas ou algum outro padrão objetivo. Um exemplo clássico e simples de um tipo objetivo de avaliação de desempenho é o teste de digitação aplicado a pessoas que se candidatam a empregos de escritório, o que exige muita digitação. Mesmo nesse tipo de teste, no entanto, o critério numérico efetivo usado como escore de corte para o desempenho aceitável – por exemplo, 65 palavras por minuto com menos de cinco erros – provavelmente será determinado arbitrariamente. A maioria dos outros tipos de avaliação de desempenho envolve (a) identificar e descrever critérios qualitativos para avaliar um desempenho ou produto e (b) desenvolver um método para aplicar os critérios. Os métodos habituais para avaliar critérios qualitativos envolvem *escalas de mensuração* ou *rubricas de pontuação* (isto é, guias para a pontuação), que descrevem e ilustram as regras e princípios a serem aplicados na

mensuração da qualidade de um desempenho ou produto. Um exemplo bem-conhecido deste tipo de procedimento é a pontuação de desempenhos atléticos por juízes especializados em eventos como competições de patinação artística ou salto ornamental.

Testagem de competência. Procedimentos que avaliam o desempenho baseados na demonstração ou não de um nível preestabelecido de maestria do testando individual são conhecidos como *testes de competência.* Muitos destes testes produzem escores do tipo tudo-ou-nada, como *aprovado* ou *reprovado*, baseados em algum nível de critério que separe a competência da não-competência. Um exemplo típico com o qual a maioria dos leitores vai estar familiarizado é o dos testes para a obtenção da carteira de motorista. Neles, o que importa é se os indivíduos são capazes de demonstrar que conhecem as regras do trânsito e sabem lidar com um automóvel em várias situações. Além disso, espera-se que a vasta maioria das pessoas que se submete ao teste de direção será capaz de passar, mesmo se for necessária mais do que uma tentativa, e não há necessidade de fazer distinções entre os testandos em termos de seus níveis de desempenho, além de simplesmente atestar a aprovação ou a reprovação.

Existem algumas situações e habilidades – como pousar um avião em um porta-aviões ou realizar uma neurocirurgia – em que a competência incompleta simplesmente não entra em cogitação. Por outro lado, nos círculos educacionais, a noção de testagem de competência pode ser interpretada de várias formas. Alguns educadores e outros interessados argumentam que todos os estudantes capazes devem atingir a competência completa dos objetivos institucionais prescritos em um nível antes de se graduarem ou passarem para o nível seguinte, independentemente do tempo que isso leve. No entanto, a maioria das pessoas está disposta a admitir que quando a testagem tem por objetivo avaliar a competência das habilidades básicas que são ensinadas nas escolas – como ler, escrever e calcular – claramente existe espaço para vários níveis de realização entre a competência e a não-competência. Nestes casos, o problema passa a ser a determinação de alvos apropriados que se espera que os estudantes atinjam, dentro de um contínuo de desempenho, para serem promovidos ou graduados.

Predição de desempenho

Às vezes, o termo *interpretação de testes referenciados em critérios* é usado para descrever a aplicação de dados empíricos que dizem respeito à ligação entre escores de teste e níveis de desempenho a um critério como o desempenho no trabalho ou o sucesso em um programa acadêmico. Neste contexto, o termo *critério* é usado em um sentido diferente, mais de acordo com as práticas psicométricas tradicionais do que nos exemplos anteriores. Aqui o critério é um resultado a ser estimado ou predito por meio de um teste. Esse tipo de informação constitui a base para o estabelecimento da validade preditiva dos testes, a ser discutida de maneira mais detalhada no Capítulo 5. Não obstante, ela é mencionada neste capítulo porque, quando a relação entre os escores dos testes e os critérios é usada para a seleção ou colocação profissional de indivíduos no contexto de educação, de treinamento ou

de emprego, esta relação também pode ser entendida como um referencial para a interpretação do escore. Neste referencial, as perguntas a serem respondidas com a ajuda dos escores de teste são "que nível de desempenho no critério podemos esperar de uma pessoa que obtém este escore?" ou "o desempenho do testando neste teste é suficiente para determinar o nível desejado de desempenho no critério em uma dada tarefa?".

As informações sobre a relação entre os escores e os critérios podem ser apresentadas de várias formas, incluindo coeficientes de correlação e equações de regressão, que são descritos mais extensamente no contexto da validade no Capítulo 5. Para os objetivos atuais, no entanto, dois procedimentos que são especialmente relevantes para o tipo de interpretação de testes referenciados em critérios discutido no parágrafo anterior – quais sejam, as tabelas de expectativa e os gráficos de expectativa – vão servir para esclarecer esta abordagem.

As *tabelas de expectativa* mostram a distribuição dos escores de teste para um ou mais grupos de indivíduos, com tabulação cruzada contra seu desempenho no critério. Pressupondo que exista um grau substancial de correlação entre os escores no teste e as medidas no critério, esta informação pode ser usada para estimar a posição provável no critério de indivíduos que pontuaram em diferentes níveis. Por exemplo, usando informações das turmas anteriores, um professor pode fazer uma tabulação cruzada dos escores nos testes do meio do semestre como preditora e da nota final do curso como critério, como é mostrado na Tabela 3.3. A tabela resultante, baseada em seus escores no meio do semestre, pode informar aos futuros alunos quais poderão ser suas notas finais.

Os *gráficos de expectativa* são usados quando o desempenho no critério no emprego, programa de treinamento ou programa de estudos pode ser classificado como bem-sucedido ou malsucedido. Estes gráficos apresentam a distribuição dos escores para um grupo de indivíduos juntamente com a percentagem de pessoas, em cada intervalo de escore, que tiveram sucesso (ou fracassaram) em termos do critério. Quando a tendência é tal que a percentagem de indivíduos bem-sucedidos é muito maior entre os que tiveram escore alto do que entre os que tiveram escore baixo, gráficos desse tipo podem ser extremamente úteis na tomada de decisões envolvendo seleção.

Tabela 3.3 Tabela de expectativas: Relação entre escores de testes no meio do semestre e notas finais

		Percentagem que recebe cada nota final			
Escore no meio do semestre	Número de casos	D ou F	C	B	A
90 ou mais	9		11	22	67
80 a 89	12	8	25	50	17
70 a 79	13	23	46	31	
60 a 69	3	33	67		
59 ou menos	3	67	33		

Relação entre Referenciais

Não é possível distinguir entre testes referenciados em normas e em critérios simplesmente olhando para exemplares de cada tipo. Na verdade, em grande parte as distinções entre os referenciais para a interpretação dos testes – bem como entre as variedades dentro de cada referencial – são questões de ênfase. Muito embora os escores possam ser expressos de uma variedade de maneiras, fundamentalmente toda a testagem se vale de um referencial normativo.

Os padrões usados na interpretação de escores de testes referenciados em critérios devem se basear em expectativas realistas ou viáveis para a população de testandos na qual o teste é aplicado. Dependendo dos objetivos do teste, pode ser que um número muito pequeno ou muito grande de pessoas seja capaz de satisfazer os critérios, caso em que o teste pode se mostrar pouco útil. Em outras palavras, os critérios se baseiam ao mesmo tempo no objetivo da testagem e também, em certo grau, no desempenho que as pessoas podem obter em uma dada situação.

O uso de escores de corte e outros padrões de desempenho na interpretação referenciada em critérios não significa que as diferenças no desempenho dos indivíduos serão eliminadas ou desconsideradas, e nem impede comparações de escores entre os indivíduos. Por exemplo, mesmo se duas candidatas a um emprego de secretária satisfaçam o critério de 65 palavras por minuto em um teste de digitação, tendo desempenho igual em todos os outros aspectos, aquela que conseguir digitar 90 palavras por minuto terá maior probabilidade de ser escolhida do que aquela que digitar apenas 65.

Da mesma forma, o uso de normas não impede o exame do desempenho no teste, do ponto de vista do conteúdo ou de critérios comportamentais. Ao corrigir provas de sala de aula, por exemplo, professores que escolheram cuidadosamente as questões aplicadas entre o material indicado para um teste podem optar por atribuir notas em forma de letras baseadas nas normas para a turma (na curva das notas) ou nos critérios de desempenho ligados à percentagem de itens respondidos corretamente. Alguns testes padronizados usados no contexto educacional – como os *Iowa Tests Of Basic Skills (ITBS)*, o *Stanford Diagnostics Mathematics Test (SDMT)* e o *Stanford Diagnostic Reading Test (SDRT)* – também tentam fornecer interpretações referenciadas em normas e em critérios. No entanto, as revisões desses testes sugerem que esta tentativa tem sucesso apenas em relação a um determinado tipo de interpretação, mas não a ambas, porque testes referenciados em normas e em critérios requerem uma ênfase um tanto diferente no modo como são construídos. Para mais informações sobre esta questão, consulte as revisões do ITBS, SDMT e SDRT no *13th Mental Measurements Yearbook*, organizado por Impara e Plake (1998).

Como, então, a interpretação de testes referenciada em normas difere da interpretação referenciada em critérios? A diferença fundamental entre as duas está em seus objetivos primários:

- Na testagem referenciada em normas, o objetivo primário é fazer distinções entre os indivíduos em termos da capacidade ou traço avaliado por um teste.

- Na testagem referenciada em critérios, o objetivo primário é avaliar o grau de competência de uma habilidade ou conhecimento em termos de um padrão preestabelecido de desempenho.

Como vimos, estes dois objetivos não são sempre ou necessariamente exclusivos entre si e, em algumas situações, o mesmo instrumento pode ser usado para ambos. Qual dos dois objetivos é o primário é determinado pela meta específica do usuário do teste. Esse objetivo, por sua vez, deve ajudar o usuário a determinar qual abordagem de desenvolvimento de testes é a mais adequada.

Em relação às variedades de interpretação de teste referenciada em critérios, deve estar claro que a distinção entre a avaliação baseada em domínios e a baseada no desempenho também é arbitrária em certo grau. O conhecimento dos domínios de conteúdo deve ser demonstrado através de um comportamento observável no qual uma ou mais habilidades desempenham um papel. Da mesma forma, todo tipo de desempenho requer algum tipo de conhecimento. Além disso, enquanto que em matérias que são elementares e relativamente estruturadas os domínios de conteúdo podem ser mapeados, quando se trata de áreas mais avançadas e menos estruturadas – como conhecimentos que englobam várias disciplinas – esse tipo de divisão se torna difícil ou impossível. Da mesma forma, os limiares de competência podem ser preestabelecidos facilmente em relação a habilidades básicas, mas em campos que envolvem habilidades de nível superior esta determinação de padrões pode não ser aplicável, porque a gama das realizações possíveis é muito mais ampla.

A Interpretação de teste referenciada em critérios na avaliação clínica

Critérios como a competência de uma habilidade ou o conhecimento em um dado campo claramente não são aplicáveis em conexão com instrumentos delineados para avaliar a personalidade. Portanto, o termo *interpretação referenciada em critérios* geralmente não é usado para testes deste tipo. Não obstante, alguns testes relevantes para o funcionamento emocional ou cognitivo são usados por clínicos para ajudar a estabelecer se certos critérios diagnósticos foram satisfeitos. Estes testes usam escores de corte – estabelecidos a partir de dados normativos – para identificar a presença de certos transtornos baseados em ligações estabelecidas por critérios clínicos. Esta aplicação em particular constitui uma interpretação referenciada em critérios, no mesmo sentido que quando a relação entre os escores de teste e os critérios é usada para selecionar ou colocar indivíduos no contexto educacional ou profissional. Estas práticas envolvem a estimativa ou predição de certos resultados. Da mesma forma, esse tipo de interpretação dos testes clínicos depende de evidências de validade relevantes para o critério diagnóstico (ver Capítulo 5). Exemplos de testes que são usados desta forma incluem instrumentos como o Inventário de Depressão de Beck, que pode ajudar a avaliar a intensidade dos transtornos depressivos, e o miniexame do estado mental, que ajuda a identificar o comprometimento cognitivo. Estes testes e outras ferramentas clínicas – como listas de verificação e escalas de mensuração comportamental –, que envolvem o uso

de escores de corte e faixas de escore para avaliar o comportamento sintomático de transtornos mentais, também podem ser definidos como referenciados em critérios.

A teoria da resposta ao item como base para a combinação de referenciais

Como seu objetivo é estimar a posição do testando em um traço ou dimensão de habilidade latente, os métodos TRI são bem adequados ao desenvolvimento de testes cujos escores possam ser interpretados em bases normativas e referenciadas em critérios. Embora os dados usados na modelagem TRI sejam derivados do desempenho de amostras de referência, eles podem ser combinados com outros tipos de análise de itens para criar escalas que ofereçam informações de natureza tanto comparativa (isto é, referenciada em normas) quanto substantiva (isto é, referenciada em critérios). Um exemplo recente de como isso pode ser feito é encontrado no trabalho de Primi (2002), que integra a teoria cognitiva e a TRI na construção de uma medida do tipo de habilidade necessária para resolver problemas de matriz geométrica. Outro é o modelo hierárquico da TRI proposto por Janssen, Tuerlinckx, Meulders e De Boeck (2000), que é aplicado a um teste que mede metas de maestria em compreensão de leitura no nível escolar fundamental.

Considerações sociais na testagem referenciada em normas ou critérios

As pressões prevalentes pelo uso da testagem referenciada em critérios para certificar competências e tomar decisões com conseqüências no contexto educacional e ocupacional surgiram de uma insatisfação com a fraqueza percebida na testagem referenciada em normas. Parte desta insatisfação tem origem na visão de que o uso da referência em normas na educação é uma forte causa de declínio dos padrões, uma vez que por pior que seja o desempenho de uma população de estudantes como um todo, em relação a suas próprias normas pelo menos metade deles sempre vai estar acima da média.

Outra fonte de insatisfação com a testagem referenciada em normas se deve ao fato de que, quando usada como base para decisões nos campos educacional e ocupacional, ela muitas vezes coloca membros de grupos de minorias em desvantagem, comparados com indivíduos que podem ter tido mais oportunidades educacionais. No entanto, em uma virada irônica, nas últimas décadas o uso da testagem referenciada em critérios para certificar competências tornou-se objeto de tanta ou mais controvérsia que a testagem referenciada em normas nas arenas profissional e política. Sem dúvida alguma, grande parte da controvérsia a respeito dos tipos de testagem se deve à compreensão equivocada do papel dos testes como ferramentas – e não como árbitros –, bem como a políticas que parecem oscilar entre os dois extremos de uma ênfase exagerada ou oposição total à testagem padronizada naquilo que chamamos de decisões de "críticas" (Jaeger, 1989; Mehrens, 1992; U.S. Department of Education, Office for Civil Rights, 2000; Wigdor e Green, 1991).

Teste a si mesmo

1. Se não for acompanhado de outras informações, um escore bruto alto é
   (a) sem sentido
   (b) mesmo assim sempre melhor do que um escore baixo

2. _____ constituem o referencial mais disseminado para a interpretação de escores de testes.
   (a) domínios de conteúdo
   (b) amostras de trabalho
   (c) critérios
   (d) normas

3. De todas as seguintes normas desenvolvimentais, quais têm aplicação mais universal?
   (a) escalas ordinais baseadas em teorias
   (b) normas de idade mental
   (c) seqüências naturais
   (d) normas baseadas em séries escolares

4. Em relação às amostras usadas para estabelecer normas intragrupo, o requisito mais importante é que elas sejam
   (a) reunidas localmente pela instituição ou organização que vai usá-las
   (b) muito grandes, com milhares de sujeitos
   (c) representativas do grupo para o qual serão usadas
   (d) convenientes de obter no processo de padronização

5. Os conceitos de teto e solo de teste estão mais intimamente relacionados à questão de
   (a) validade dos testes
   (b) dificuldade dos testes
   (c) tipo de escores padrões usados em um teste
   (d) tipo de itens usados em um teste

6. Quando transformado no QI de desvio do tipo da escala de Wechsler, um escore *z* de –1,00 se tornaria um QI Wechsler de
   (a) 85
   (b) 95
   (c) 105
   (d) 115

7. Qual dos seguintes procedimentos de transformação de escores é o único que se qualifica como uma transformação linear?
   (a) escores padrões normalizados
   (b) de percentis para estaninas
   (c) de escores brutos para escores de percentil
   (d) de escores *z* para escores *T*