7. *Noções básicas sobre mensuração*

Para realizar qualquer pesquisa no campo das relações sociais, devemos estar aptos a observar os constructos que desejamos estudar. *Constructos* são as abstrações qe os cientistas sociais consideram nas suas teorias, tais como "*status* social", "poder" e "inteligência". Freqüentemente, devemos não só estar aptos a observar os constructos, mas também a medi-los. "Mensuração" é a atribuição de algarismos a objetos ou eventos de acordo com regras (Stevens, 1951). Para medir um constructo, precisamos primeiramente identificar uma *variável* que represente, de maneira mais concreta, abstração. Por exemplo, "renda" pode ser considerada como uma variável que representa "*status* social". Nenhuma variável , tomada isoladamente, serve como representação completa de um constucto, por razões que discutiremos mais adiante, neste capítulo. Apesar disso, se pretendemos medir um constructo, precisamos identificar pelo menos uma variável. As próprias variáveis exigem maior especificação, sob a forma de uma *definição operacional,* isto é, sob a forma de um conjunto de *valores* da variável, acompanhados de um conjunto de instruções que permitam atribuir um valor determinado a uma unidade de análise determinada. Por exemplo, a "renda" pode ser medida de muitas maneiras diferentes, e precisamos portanto de uma definição operacional que estabeleça especificamente de que maneira a renda será medida em uma pesquisa: serão considerados exclusivamente os rendimentos pessoais dos indivíduos? Serão incluídos apenas salários, ou serão incluídos também outros tipos de rendimentos, como juros e aluguéis? A informação será obtida através de uma pergunta feita durante uma entrevista, através de um questionário remetido pelo correio, ou através de alguma outra forma? Será solicitado ao indivíduo que forneça um valor exato, ou lhe serão apresentadas categorias mais amplas de renda, solicitando-lhe que aponte aquela na qual ele estaria incluído? O conjunto de todos os valores possíveis de uma variável forma uma *escala,* ou conjunto de categorias. Uma escala pode ter apenas duas categorias, tais como "renda alta", "renda baixa", ou pode ter um grande número de categorias ou gradações. Quando associamos um conjunto de algarismos ao conjunto de valores de uma variável, de acordo com regras especificadas de correspondência, dizemos que estamos *medindo*

a variável em questão. À medida que avançamos de constructos para escalas, devemos ir nos tornando progressivamente mais específicos em nossa linguagem, e nosso trabalho pode ir se tornando progressivamente mais aberto ao escrutínio público. A mensuração científica deve ser aberta á inspeção e à replicação por parte de outros cientistas e leigos bem informados.

Definições operacionais

Uma definição operacional deve especificar a seqüência de passos que você dá para obter uma medida. Essa seqüência deve ser replicável, de forma que você possa instruir outra pessoa para obter as mesmas medidas. Se alguém proclamar que descobriu uma forma de medir a saúde das pessoas "sentindo a aura" que as envolve e se esta técnica for tão mística que ninguém mais possa usá-la e obtiver os mesmos resultados, esta será uma definição operacional inaceitável. A medição de temperatura de uma pessoa através da leitura de um termômetro, em contraste, é uma definição operacional aceitável, porque é um procedimento direto que pode ser repetido facilmente. Mas a leitura de um termômetro, é, na realidade, uma definição operacional de quê? É uma medida de "saúde"?

Tecnicamente, uma leitura de um termômetro é uma medida do quanto o mercúrio subiu num tubo, o que, por sua vez, é uma medida de quão quente está o mercúrio, o que finalmente, é uma medida de quão quente está a boca da pessoa que colocou o termômetro sob a língua. Isto é uma medida de saúde? Geralmente aceitamo-la como uma medida que nos diz se uma pessoa tem ou não tem febre, mas mesmo isto é questionável. Todos nós apresentamos diferentes marcações no termômetro em diferentes momentos do dia; quando diremos que determinada temperatura é "febre"? A 37,1°C? Ou a 37,7°C? A relação entre temperatura e febre não é completamente clara. A leitura no termômetro, portanto, é uma medida imperfeita de se alguém tem ou não febre. Ela é uma medida pouco adequada da saúde, porque a relação entre temperatura e saúde é mais tênue que a relação entre temperatura e febre. Uma pessoa pode ter temperatura normal mas sofrer de pressão alta, diabetes ou artrite e, portanto, não estar em perfeita saúde.

"Saúde" é um constructo abstrato bastante complexo, com muitos componentes: pressão sanguínea, níveis de açúcar no sangue, contagem de glóbulos brancos, contagem de glóbulos vermelhos, grau de obesidade, nível de colesterol, história de câncer e muitos outros detalhes que constituem o estado de saúde de uma pessoa. A leitura da temperatura, portanto, mede apenas um componente de saúde. Para se obter uma avaliação adequada do estado de saúde de alguém precisamos fazer muitas leituras e muitas perguntas, como um médico faz durante um *check-up* geral. Qualquer medida isoladamente, como, por exemplo, contagem de glóbulos vermelhos, é uma definição operacional de um componente de saúde, por exemplo, grau de anemia. E cada definição operacional do tipo é mais aceitável como uma medida científica de saúde do que a leitura da "aura", porque "auras" não são publicamente acessíveis. Quase todo mundo pode aprender a medir a pressão sanguínea ou a contar glóbulos vermelhos, e duas pessoas podem concordar em suas medidas. O mesmo não acontece com a leitura de auras. A mensuração científica é obtida com definições operacionais que podem ser usadas e replicadas por qualquer número de pessoas. Isto é o que faz uma definição operacional objetiva.

Os primeiros positivistas lógicos eram superzelosos em sua aplicação das definições operacionais. Eles faziam afirmações circulares; tais como: "inteligência é o que os testes de inteligência medem", o que não deixa espaço para debate. Tais tautologias não nos ajudam a compreender nem o conceito de inteligência nem o teste usado para medi-la. Donald Campbell critica a abordagem tautológica à qual denomina "operacionismo definicional" (Campbell, 1969). Segundo ele toda observação é afetada por uma variedade de fatores que não guardam nenhuma relação com o constructo que desejamos medir. Por exemplo, respostas ao censo referentes à renda familiar ou ao número de adultos que moram numa casa são determinadas apenas parcialmente pela renda verdadeira ou tamanho da família do informante. Estas respostas são também determinadas pela "interação social da entrevista... a aparência do entrevistador... o medo que o informante possa ter de visitantes semelhantes, tais como cobradores, investigadores do bem-estar social e a lei... "(Campbell, 1969, p.15). Medidas de papel e lápis são tão vulneráveis a influências irrelevantes como o são as entrevistas face a face ou observações:

> Uma resposta a um questionário construído para medir a ansiedade manifesta pode ser em parte função da ansiedade, mas é também uma função da compreensão do vocabulário, de diferenças individuais e de classe social quanto ao uso de adjetivos eufóricos ou disfóricos, de definições idiossincráticas dos termos-chaves freqüentemente repetidos, das expectativas dos informantes relativas às conseqüências pessoais de se descrever a si mesmo como estando bem ou doente etc. (Campbell, 1969, p. 15).

Portanto, nenhuma definição operacional, isoladamente, pode oferecer "a" medida, ou a única medida verdadeira, porque envolve também fatores irrelevantes. Na ausência de outras definições operacionais do mesmo constructo, não sabemos quanto da mensuração reflete esses fatores irrelevantes e quanto reflete o que pretendemos medir.

Uma segunda razão para não aceitarmos definições isoladas como medidas definitivas de conceitos é que isto logicamente impossibilitaria nossos esforços para aprimorar a mensuração na ciência. Um crítico antigo do "operacionismo definicional" considerou-o como "um obstáculo para o progresso científico, porque exclui a crítica" (Adler, 1947, p. 441).

Uma terceira razão para suspeitar de qualquer afirmativa de que a inteligência (ou qualquer outra qualidade) é o que qualquer teste isoladamente mede é que, sem um segundo teste independente para medir a mesma quantidade não há garantia de que o teste esteja relacionado com qualquer outra coisa que não consigo mesmo. Adler (1947) projetou um teste para medir o que ele chamou de "C_N", como se segue:

O Teste C_N

1. Quantas horas você dormiu a noite passada? _____
2. Faça uma estimativa do comprimento de seu nariz, em polegadas, e multiplique por 2. _____
3. Você gosta de fígado frito? (Marque +1 para Sim e -1 para Não). _____
4. Quantos pés há numa jarda? _____
5. Faça uma estimativa do número de copos de Ginger Ale que o inventor deste teste bebeu enquanto o elaborava. _____

Soma = Escore C_N bruto.

Suas instruções diziam: "Este teste deve ser realizado diariamente, sempre à mesma hora, pelo tempo que você conseguir. Então, você poderá proceder ao cálculo de seu C_N elaborado... "(p. 439), e ele fornece uma fórmula elaborada para o cálculo do índice a partir do escore bruto. Se você desejar saber o que seu escore C_N representa, ele responde que o "teste mede C_N e C_N é o que o teste mede" (p. 439).

Adler criou este teste para mostrar a futilidade de tentar se esconder atrás do operacionismo definicional. Ele acrescenta:

> Defrontamo-nos aqui com um sistema aparentemente fechado. Toda crítica é excluída. C_N é o que a definição afirma que é, e o teste é o que o define. Ainda assim, C_N não faz sentido, não somos capazes de formar um conceito a respeito (Adler, 1947, p. 439).

A menos que haja pelo menos duas definições operacionais de um conceito, não temos maneira de saber se uma definição particular é apropriada. No caso do teste C_N, precisamos ter alguma idéia do que

seja C_N para construir uma segunda definição operacional, e não é suficente dizer que "C_N é o que o teste C_N mede". Isto é petição de princípio.

Fidedignidade

A teoria clássica de mensuração parte da suposição de que toda mensuração comporta algum erro (Guilford, 1954). Qualquer escore observado tem dois componentes:

Escore Observado = Escore Verdadeiro + Erro.

Uma medida fidedigna é aquela cuja componente de erro é pequeno e, portanto, não flutua aleatoriamente de um momento para outro. Para compreender fidedignidade da mensuração, considere o que ela significa quando se trata de uma pessoa. Se você diz que alguém é fidedigno provavelmente você quer dizer que a pessoa é consistente — se ela diz uma coisa hoje, ela dirá a mesma coisa amanhã. Se uma pessoa fidedigna lhe disser que o encontrará no dia seguinte na hora do almoço, quando você chegar ao lugar marcado, ela estará lá. Uma pessoa fidedigna é também aquela que, ao narrar a ocorrência de um acontecimento mantém um relato consistente e não dá diferentes versões a cada momento. Estas várias definições de uma pessoa fidedigna implicam que ela não irá dizer uma coisa e fazer outra ou dar diferentes versões da "verdade" em diferentes momentos.

Um instrumento fidedigno também "mantém a mesma história" de um momento para outro. Em comparação considere um instrumento não fidedigno, uma régua *elástica*. Se você tentasse medir sua altura pisando numa ponta da régua e segurando a outra ponta no alto da cabeça, você obteria medidas ligeiramente diferentes a cada vez, porque você puxaria a régua um pouco mais ou um pouco menos e ela se esticaria ou se encolheria. Uma régua elástica tem um elevado componente de erro em relação ao escore verdadeiro e isto torna os escores observados não fidedignos. Quanto maior for o componente de erro em relação ao escore verdadeiro, menos fidedigno será o instrumento.

Fidedignidade é calculada de várias maneiras e a régua elástica se sairá mal em todas elas. Dois dos tipos mais comuns de cálculo de fidedignidade são a correlação teste-reteste e a correlação de duas metades.

Correlação teste-reteste

Usando um instrumento duas vezes com as mesmas pessoas ou grupos, podemos calcular a correlação entre seus dois escores. Essa correlação é uma medida da fidedignidade do instrumento, fidedignidade interpretada como "estabilidade". Se o instrumento for fidedigno, as pessoas deverão manter as mesmas posições relativas no instrumento. Se não ocorrerem modificações importantes na vida daquelas pessoas ou grupos, as pessoas que obtiveram escores elevados na primeira testagem deveriam obter escores também elevados na segunda. Não se espera encontrar uma correlação perfeita para a maioria das medidas nas ciências sociais porque todas elas contêm um elemento de erro, que fez os escores observados flutuarem de uma sessão de testagem para outra. Mesmo uma régua de madeira poderia não produzir medidas idênticas de cultura de uma vez para outra, porque pode haver certo deslize ao posicionar a régua, porque as posturas das pessoas mudam e influenciam sua altura, porque elas podem mudar de calçado de uma vez para outra. Contudo, leituras repetidas feitas com uma régua de madeira seriam mais consistentes de uma vez para a seguinte do que leituras feitas com uma régua elástica. Quanto menor for o componente de erro, quanto mais consistentes forem as leituras de uma mensuração para outra, mais alta será a fidedignidade teste-reteste do instrumento.

Correlação de duas metades

Correlacionando-se os resultados obtidos em duas metades quaisquer do mesmo instrumento, podemos, calcular a "fidedignidade de duas metades" *(split-half)*. Se nossa régua elástica tiver 5 metros de comprimento, podemos medir a altura das pessoas usando a primeira metade da régua uma vez, de 0 a 2,5 metros e, então, a segunda metade, de 2,5 a 5 metros. Deveremos encontrar aproximadamente a mesma medida em centímetros para uma única pessoa utilizando as duas extremidades da régua, mas elas não serão idênticas, porque a elasticidade cria erros. O mesmo ocorreria se tivéssemos uma escala de atitudes com 200 itens. Poderíamos dar a cada pessoa dois escores, um baseado, por exemplo, nos primeiros 100 ítens, e outro baseado nos últimos 100 ítens. Os escores seriam similares mas não idênticos, porque os ítens diferem e ambas as metades da escala incluem algum erro em sua mensuração. A fidedignidade de duas metades, seja da régua, seja da escala de atitudes, é a correlação entre os escores obtidos nas duas metades. Quanto mais semelhantes forem os escores das duas metades, maior ser a correlação e mais fidedigno o instrumento, fidedignidade interpretada, aqui, como "consitência interna"

Validade

Uma medida válida é aquela que abrange o constructo que queremos abranger. Um instrumento pode ser muito fidedigno e abranger um constructo com grande precisão, mas não ser válido para nosso objetivo por medir o constructo errado. Por exemplo, se desejássemos medir inteligência e déssemos um teste de QI padrão em inglês para um grupo de colegiais franceses, poderíamos encontrar uma elevada fidedignidade teste-reteste mas ter uma medida não válida da inteligência daqueles estudantes. Ao invés disso, teríamos uma medida de seu conhecimento de inglês. Para estudantes franceses, um teste de QI em inglês é uma medida da proficiência na língua inglesa, e não uma medida de inteligência. Este teste seria um teste inadequado e, portanto, *não válido* para este grupo.

A utilização de um teste de QI em inglês para medir a inteligência de estudantes franceses é um caso óbvio em que se mede outro constructo que não o desejado. A maioria das medidas nas ciências sociais não apresenta erros tão grandes, mas todas elas compartilham deste problema em algum grau. Definições operacionais inevitavelmente incluem componentes que não deveriam ser incluídos e excluem partes do constructo subjacente que deveriam ser medidas. Como o constructo subjacente não pode ser abordado diretamete, mas apenas indiretamente, através de definições operacionais, nunca podemos estar certos de que porção dele a definição operacional abrange, e que porções não são medidas. Sabemos, contudo, que qualquer medida tomada isoladamente inclui componentes irrelevantes e exclui partes relevantes do constructo subjacente, pelo fato de outra medida produzir resultados ligeiramente diferentes. A figura 7.1 ilustra isso.

Suponha que a variável subjacente que desejamos medir seja a inteligência de uma amostra de colegiais franceses. Poderíamos usar um teste de inteligência em inglês como definição operacional, uma tradução francesa do mesmo teste como definição operacional $_2$ e uma entrevista face a face realizada por uma banca de educadores franceses como uma terceira definição operacional.

As porções dos círculos da figura 7.1 que não se superpõem consistem de dois componentes: erro e aspectos do constructo que não medidos por uma única definição operacional. As discordâncias entre as medidas de inteligência obtidas na versão inglesa do teste, na versão francesa e nas entrevistas com educadores provêm tanto de erros como de componentes exclusivos da variável subjacente contidos em cada definição ope-

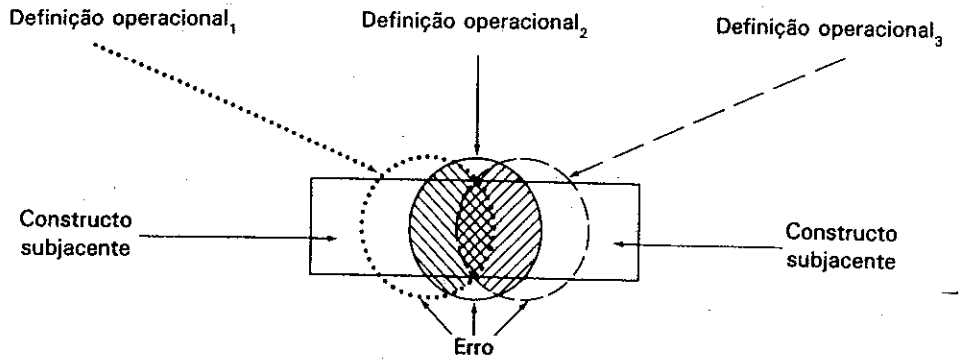

Figura 7.1. Definições operacionais incluem componentes irrelevantes e não incluem todas as partes relevantes do constructo subjacente.

racional. O componente de erro envolve tanto as variações aleatórias como as sistemáticas. As qualidades sistemáticas incluem conhecimento de inglês para a versão inglesa do teste de inteligência e habilidades de conversação e equilíbrio para a entrevista com educadores. Elas não são parte do que concebemos como inteligência, pois um estudante francês pode ser muito inteligente, mas ser incapaz de responder um único item da versão inglesa, por nunca ter estudado inglês. Outro estudante pode obter escores elevados em ambos os testes escritos, mas ficar paralisado durante a entrevista com educadores franceses. Facilidade com o inglês e postura de ator são variáveis irrelevantes que estão incluídas nestas duas medidas de inteligência. Se usarmos uma única definição operacioal, não sabemos quanto do escore observado representa o constructo que pretendemos medir — inteligência — e quanto representam as qualidades irrelevantes, como conhecimento de inglês ou habilidades para conversação. Quando usamos duas ou mais definições operacionais, podemos correlacioná-las, calcular a variância comum e saber quanto as medidas abrangem o que queremos e quanto de erro elas contêm. Quanto menos semelhantes forem as definições operacionais — como um teste escrito e uma entrevista — mais heterogêneos serão os componentes irrelevantes. Se você usar definições operacionais as mais diferentes possíveis, os componentes irrelevantes também serão diferentes e a variância compartilhada refletirá a variável desejada. Quanto menos semelhantes forem os dois métodos de mensuração, mais significativa será a correlação entre eles, pois, neste caso, a variância comum refletirá a variável subjacente e não erro ou componentes irrelevantes.

Fidedignidade ⟶ *Validade*

Correlação entre medidas o mais semelhantes possível.

Correlação entre medidas o mais diferentes possível.

Figura 7.2. Continuum fidedignidade-validade.

Os coeficientes de fidedignidade e de validade situam-se num *continuum,* conforme ilustra a figura 7.2. Num extremo do *continuum* estão as correlações entre métodos idênticos de mensuração. Correlações teste-reteste baseiam-se no mesmo teste aplicado duas vezes. No outro extremo do *continuum* estão correlações entre métodos muito diferentes de mensuraão da mesma variável. Um teste de inteligência do tipo lápis-papel e uma entrevista podem ser medidas bastante diferentes entre sí. Cálculos de fidedignidade são correlações entre métodos idênticos ou semelhantes; cálculos da validade são correlações entre métodos diferentes.

Há várias formas de se avaliar a validade de um unstrumento, cada uma das quais baseada na concordância entre duas avaliações diferentes da mesma variável.

Validade aparente

Validade aparente é avaliada por um grupo de juízes, algumas vezes especialistas que lêem ou examinam técnica de mensuração e decidem se, em sua opinião, ela mede o que seu nome sugere. Por exemplo, fonoaudiólogos poderiam examinar um teste elaborado para medir graus de retardo do desenvolvimento da linguagem e decidir se o teste mede o que se pretente. A avaliação da validade aparente é um processo subjetivo, mas podemos calcular um índice de validade calculando a quantidade de concordância entre os juízes. Quanto maior for a porcentagem dos que afirmam que o teste mede o que pretente medir, maior será a validade aparente. Todo instrumento deve passar pela avaliação da validade aparente, quer de maneira formal, quer informalmente. Todo pesquisador que escolhe um instrumento é um juíz que decidiu que o teste mede o constructo que ele deseja estudar. Sem esta validade aparente mínima, um instrumento não seria usado.

Validade simultânea

Validade simultânea é a capacidade de um teste em distinguir entre indivíduos sabidamente diferentes. Por exemplo, se você estivesse desenvolvendo um teste para medir o conservadorismo político das pessoas, o teste deveria distinguir entre pessoas que pertencessem a grupos que supomos diferentes, tais como "Jovens Americanos pela Liberdade" e "Estudantes por uma Sociedade Democrática". Se membros desses grupos obtiverem os mesmos escores, o teste seria uma medida não do conservadorismo político mas de algo que estes grupos compartilhassem, como, por exemplo, uma descrença na administração política atual.

Validade preditiva

Validade preditiva é a capacidade de um teste em identificar diferenças futuras. Por exemplo, a validade preditiva dos vestibulares é a capacidade destes testes identificarem que se formaria na faculdade e quem a abandonaria, de preverem quem tiraria notas altas e quem tiraria notas baixas ou de preverem quem prosseguiria os estudos e quem não. Dependendo de qual destes critérios escolhemos, o teste pode ter uma validade preditiva alta, média ou baixa. Validade preditiva é uma avaliação do valor prático de um teste em antever o futuro. É uma abordagem pragmática à validade.

Validade de constructo

Validade de constructo é uma avaliação do quanto um instrumento mede o constructo teórico que o investigador deseja medir. Ao mesmo tempo, é uma avaliação desse próprio constructo e da teoria que o sustenta. Ao contrário da validade aparente, a validade de constructo requer mais do que a opinião de especialistas: requer uma demonstração de que o constructo em questão existe, do que é distinto de outros constructos e, portanto, digno de seu nome, e de que o instrumento mede este constructo em particular e não outros. A validação de constructo requer *concordância* entre escores obtidos com dois instrumentos que presumivelmente medem o *mesmo constructo* e *discordância* entre dois instrumentos que presumivelmente medem *diferentes* constructos. A concordância fornece evidências da validade *convergente,* que é aspecto da validação de constructo. Discordância é uma evidência da validade *discriminatória,* o segundo aspecto da valida-

ção de constructo. Por exemplo, se você desejasse elaborar um questionário para medir atitudes em relação às mulheres a validade convergente, no caso, consistiria na concordância entre os escores de pessoas no questionário e a classificação de suas atitudes em relação às mulheres feita por seus amigos. Para demonstrar que a entrevista mede atitudes em relação às mulheres e não liberalismo político ou atitudes em relação a pessoas em geral, é necessária fazer outras comparações: é necessário demonstrar que as atitudes das pessoas em relação às mulheres *não* estão muito fortemente correlacionadas com um conjunto diferente de atitudes, como, por exemplo, atitudes em relação aos homens. Baixas correlações com testes que medem diferentes constructos demonstram a validade *discriminatória* do instrumento.

Não há critérios sobre quão altas devem ser as correlações para demonstrar validade convergente e quão baixas devem ser para demonstrar validade discriminatória. O que importa é o padrão de correlações — as primeiras devem ser mais altas que as últimas.

Matriz multitraços-multimétodos

A matriz multitraços-multimétodos, uma tabela de correlações que demonstra o papel tanto da validade convergente como da validade discriminatória na validação de constructos (Campbell e Fiske, 1959), requer pelo menos dois métodos, para medir pelo menos duas variáveis diferentes.

A matriz se baseia no princípio de que quanto mais características duas mensurações têm em comum, maior será sua correlação. Mensurações podem compartilhar dois tipos de características: traços e métodos. *Traço* é o constructo subjacente que se supõe a mensuração abranja; é o *conteúdo. Metodo* é a forma de mensuração — questionários envolvendo o uso de lápis e papel, entrevistas face a face, observações não-reativas, registos de recenseamento e assim por diante. Idealmente, os escores deveriam refletir apenas os traços pretendidos, e não deveriam ser influenciados pelo método. Na realidade, entretanto, a forma ou métodos de mensuração também afeta o escore, e parte da variação nos escores observados é um produto do método utilizado para obter escores. Uma entrevista face a face acerca de atitudes em relação às mulheres, por exemplo, mediria não apenas atitudes em relação às mulheres mas também o desejo das pessoas de parecerem liberais ou atualizadas em suas opiniões. Um teste de inteligência em inglês mediria não apenas a inteligência dos estudantes, mas também sua proficiência em inglês. Cada tentativa para medir um constructo, portanto, está contaminada por características do método que são irrelevantes para o constructo, porém inevitáveis na mensuração.

Alguns aspectos irrelevantes dos métodos são conhecidos e podem ser levados em conta. Por exemplo, entrevistadores podem tentar estabelecer uma relação com seus informantes de forma a eliciar respostas honestas e não respostas producentes ou socialmente desejáveis (vide cap. 8). Pesquisadores também podem compensar vieses conhecidos nos questionários. Por exemplo, as pessoas freqüentemente desenvolvem preferência para assinalar "sim" ou "não" nas suas respostas, independentemente do conteúdo dos itens que lhes são apresentados. Os pesquisadores podem controlar mais vieses, formulando os itens de tal forma que metade contenha afirmações positivas, e a outra metade afirmações negativas acerca de assuntos.

Além desses vieses e fontes de contaminação conhecidos, as técnicas de mensuração contêm outras características que o investigador não pode controlar e os escores, portanto, inevitavelmente contêm um componente de traço e um componente do método. A matriz multitraços-multimétodos permite ao pesquisador determinar em que medida os escores refletem o traço e o método contidos em toda mensuração.

Uma vez que todo escore é constituído de dois elementos — traço e método — a correlação entre dois conjuntos de escores depende do quanto eles compartilham traço e método. Coeficientes de fidedignidade são correlações entre escores que refletem o mesmo traço e o mesmo método. Coeficientes de validade convergente são correlações entre escores que refletem o mesmo traço medido por diferentes métodos. Os coeficientes de fidedignidade de um instrumento deveriam, portanto, logicamente ser maiores que seus coeficientes de validade, pois os primeiros se baseiam em mais elementos comuns. A matriz multitraços-multimétodos introduz dois coeficientes de correlação adicionais para avaliar a validade de um instrumento. Ambos são correlações entre *diferentes traços*. Uma delas é uma correlação de validade discriminatória entre diferentes traços medidos pelo mesmo método e a outra é uma correlação sem sentido entre diferentes traços medidos por diferentes métodos. A tabela 7.1. apresenta estas correlações e seus elementos.

As duas primeiras correlações deveriam ser altas; as duas últimas deveriam ser baixas. Se a terceira e a quarta correlações forem tão altas quanto as correlações de validade convergente, isto significa que os dois traços não são diferentes, mas são o mesmo ou muito semelhantes. Por exemplo, se um investigador montar dois testes distintos um denominado Atitudes em Relação às Mulheres e o outro Atitudes em Relação aos Homens, os dois não deveriam manter correlações altas se estas forem, realmente, atitudes distintas. Se as intercorrelações entre estes testes fossem como aquelas mostradas na figura 7.3, estaríamos satisfeitos, pois eles mediriam dois traços ou atitudes distintas.

Os coeficientes de correlação "mesmo traço-métodos diferentes" (0,70) são maiores que os coeficientes de correlação "traço diferente mesmo método" (0,30) e que os coeficientes "traço diferente-método diferente" (0,10). Isto justifica a afirmação de que os testes medem dois traços diferentes. Se, por outro lado, a matriz fosse semelhante à da figura 7.4, os testes seriam invalidados, não porque as correlações seriam muito baixas, mas justamente pelo fa-

Tabela 7.1 Coeficientes de correlação numa matriz multitraços-multimétodos.

Coeficientes	*Elementos contidos nos escores que são correlacionados*	
	Traços	Métodos
1. Correlação de fidedignidade	Mesmos	Mesmos
2. Correlação de validade convergente	Mesmos	Diferentes
3. Correlação de validade discriminatória	Diferentes	Mesmos
4. Correlação sem significado	Diferentes	Diferentes

		Questionário		*Observações do comportamento*	
		Atitudes em relação às mulheres (ARM)	Atitudes em relação aos homens (ARH)	Atitudes em relação às mulheres (ARM)	Atitudes em relação aos homens (ARH)
Questionário	ARM	(0,90)*	0,30		
	ARH	0,30	(0,90)		
Comportamento	ARM	0,70	0,10	(0,90)	0,30
	ARH	0,10	0,70	0,30	(0,90)

* As correlações entre parênteses são correlações de fidedignidade.

Figura 7.3 Matriz multitraços-multimétodos de correlações entre atitudes em relação aos homens e atitudes em relação às mulheres (medidas hipotéticas).

to de que algumas das correlações seriam muito altas.

		Questionário		*Observações do comportamento*	
		Atitudes em relação às mulheres (ARM)	Atitudes em relação aos homens (ARH)	Atitudes em relação às mulheres (ARM)	Atitudes em relação aos homens (ARH)
Questionário	ARM	(0,90)*	0,80		
	ARH	0,80	(0,90)		
Comportamento	ARM	0,40	0,30	(0,90)	0,80
	ARH	0,30	0,40	0,80	(0,90)

* As correlações entre parênteses são correlações de fidedignidade.

Figura 7.4. Matriz multitraços-multimétodos de correlações entre atitudes em relação às mulheres e atitudes em relação aos homens.

As correlações na figura 7.4 mostram que as duas atitudes são muito semelhantes, porque estão altamente correlacionadas uma com a outra. A correlação entre dois traços diferentes medidos pelo mesmo método (0,80) é maior que a correlação entre o mesmo traço medido por dois métodos diferentes (0,40). Estas duas atitudes não apresentam validade discriminatória por estarem tão estreitamente intercorrelacionadas. Se efetivamente fossem duas atitudes diferentes, a correlação entre elas não deveria ser maior que as correlações de validade convergente do mesmo traço medido por diferentes métodos (0,40).

Se obtivéssemos as correlações mostradas na figura 7.4, concluiríamos que os dois testes medem aproximadamente a mesma atitude e, em vez de falarmos em "Atitudes em Relação às Mulheres" e "Atitudes em Relação aos Homens", redenominaríamos as escalas, chamando-as, a ambas, "Atitudes em Relação às Pessoas".

Escalas

Uma escala, nas ciências sociais, é um conjunto de categorias para diferenciar entre pessoas em relação a qualquer variável. Pode haver tão pouco como duas categorias numa escala ou tanto quanto 100, ou mais.

Escalas nominais

Escalas nominais contêm categorias qualitativamente diferentes às quais atribuímos nomes em vez de significado numérico. As mais simples são dicotômicas, com apenas dois valores, tais como: "masculino" e "feminino", ou "proprietário" e "inquilinos". "As categorias diferem qualitativa e não quantitativamente. Se, para poder codificar e perfurar os cartões com os dados usamos números como 1 e 2 em lugar de "masculino" e "feminino", respectivamente, os número não têm valor aritmético, servindo exclusivamente como "nomes," para identificar a categoria. O número "2" não significa que os casos colocados nesta categoria têm mais da qualidade do que os casos colocados na categoria número "1". Outros exemplos de escalas nominais são os seguintes:

Tipos de tensões urbanas:
1 = ruído de tráfego,
2 = poluentes atmosféricos,
3 = multidões,
4 = aborrecimentos burocráticos,
5 = outros.

Possibilidades de moradia para idosos:
1 = casa ou apartamento próprio,
2 = casa de parente,
3 = asilo,
4 = outros.

A lista de alternativas não precisa cobrir todas as categorias possíveis, mas deveria incluir aquelas relevantes para a teoria e a população testada e deveria possibilitar ao codificador classificar todos os casos. Por exemplo, há muito mais possibilidades de moradia para idosos do que as três que acabamos de listar. Contudo, se planejássemos um estudo para testar os efeitos de morar em sua própria casa, na casa de outra pessoa ou em uma instituição, então as três categorias mais a categoria não-especificada "outros" seriam suficientes para os propósitos do estudo. A inclusão da categoria "outros" permite-nos classificar todos os casos.

Escalas ordinais

Uma escala ordinal contém categorias que podem ser ordenadas em seqüência num *continuum*. As categorias têm um significado artimético rudimentar, tal como "mais" ou "menos" da quantidade que está sendo medida. Por exemplo, podemos ordenar as ocupações em termos de quanta autonomia os trabalhadores têm em seus empregos:

1 = pouca autonomia: (por exemplo, trabalhadores em linha de montagem, perfuradores e caixas numa grande loja de departamentos);
2 = autonomia média: (por exemplo, trabalhadores de construção, enfermeiras e motoristas de táxi);
3 = muita autonomia: (por exemplo, artistas independentes, joalheiros, médicos, advogados).

A escala afirma que 1 significa uma ocupação que permite menos autonomia que 3 e que 2 está situado entre as duas. Uma escala ordinal nos dá apenas esta informação e não fornece nenhuma informação acerca das distâncias entre os valores. O intervalo entre 1 e 2 pode ser maior ou menor que entre 2 e 3. Uma escala ordinal não implica absolutamente nada mais acerca dos valores aritméticos além de que eles estão em ordem.

Escalas intervalares

Quando os números atribuídos a uma variável implicam não somente que 3 é mais que 2 e 2 é mais que 1, mas também que o tamanho do intervalo entre 3 e 2 é o mesmo do intervalo entre 2 e 1, eles compõem uma escala intervalar. Apenas pelo fato de uma escala conter números de 1 a 100 não decorre automaticamente que a diferença entre 60 e 70 seja a mesma diferença que entre 90 e 100. Por exemplo, se construíssemos um teste de vocabulário com 100 itens no qual a maioria das pessoas definisse entre 60 e 70 palavras corretamente, apenas duas pessoas definissem 90 e somente uma pessoa definisse 100 palavras corretamente, o intervalo entre 90 e 100 provavelmente representaria uma diferença maior no nível de vocabulário que o intervalo entre 60 e 70.

Se os intervalos representarem quantidades *iguais* da variável medida, eles comporão uma escala intervalar. A cada aumento de unidade na escala corresponde um aumento de unidade na variável. A escala Celsius mede temperatura em intervalos iguais. A diferença de temperatura entre 33 e 34 graus é a mesma diferença que entre 36 e 37 graus. Se isto parece óbvio é porque estamos acostumados com a escala Celsius e assumimos como certo que ela representa intervalos iguais de calor e frio físicos. Não podemos assumir como certo que as escalas das ciências sociais representam intervalos iguais.

A maioria dos constructos das ciências sociais é medida em escalas ordinais e não intervalares. Por exemplo, se usarmos a renda anual das famílias para medir o constructo subjacente *"status* social," não po-

demos assumir que a escala de cruzeiros represente intervalos iguais de *status* social. A diferença de *status* no intervalo entre Cz$ 20.000,00 e Cz$ 40.000,00 de renda anual é muito maior que a diferença de *status* no intervalo entre Cz$ 120.000,00 e Cz$ 140.000,00. À medida que subimos na escala de renda, a diferença de Cz$ 20.000,00 vai gradativamente fazendo menos diferença em termos de *status* social. É mais provável que duas famílias com rendas anuais de Cz$ 120.000,00 e Cz$ 140.000,00 sejam vizinhas do que duas famílias cujas rendas são Cz$ 20.000,00 e Cz$ 40.000,00 porque o intervalo de Cz$ 20.000,00 representa uma diferença maior em *status* social na extremidade inferior da escala do que na extremidade superior. Renda anual, portanto, não é uma medida intervalar de *status* social.

Os números numa escala intervalar podem ser somados ou subtraídos porque as propriedades de tal escala são tais que 20-10 = 40-30, mas não podem ser multiplicados ou divididos, porque a escala não tem um "zero absoluto". O "zero" da escala intervalar é arbitrário. Só podemos multiplicar e dividir valores quando temos uma escala de razão.

Escalas de razão

As escalas de razão efetivamente têm um zero absoluto isto é, uma origem "natural," ou não-convencional, e, em conseqüência, seus valores representam quantidades multiplicáveis. Escalas físicas que medem comprimento e peso são escalas de razão: um quadro de quatro metros é duas vezes mais comprido que um de dois metros; 10 quilogramas de plumas pesam duas vezes mais que 5 quilogramas, pois, nestas escalas físicas, o zero é real e não arbitrário. Embora não possamos apontar algo que tenha zero metros ou zero quilos, sabemos o que isto significa em nossas réguas ou escalas e não colocamos o zero arbitrariamente em qualquer ponto da escala.

Algumas variáveis usadas para medir constructos sociais se parecem superficialmente com as medidas de razão porque têm o zero como seu escore mais baixo. Dinheiro como medida de *status* social, por exemplo, tem a aparência de uma escala de razão porque a variável tem um zero absoluto. Uma pessoa pode estar sem um tostão e não possuir nenhum dinheiro. Contudo, isto não significa que a pessoa tenha zero *status* social. Um monge, que faz voto de pobreza, por exemplo, não tem dinheiro mas tem *status* entre as pessoas que respeitam ordens religiosas. O máximo que podemos assumir a respeito de *status* social medido em cruzeiros é que mais dinheiro representa mais *status*, desde que todos os outros bens permaneçam iguais.

Mesmo medidas padronizadas, como as escalas de QI, não têm zeros absolutos. Nenhum psicólogo tentaria argumentar que uma pessoa com QI 150 seja duas vezes mais inteligente do que outra com QI 75. Embora os números possam ser somados ou multiplicados e embora uma escala possa começar do zero, isto não significa que o constructo subjacente tenha estas propriedades. É difícil imaginar qualquer constructo social, tal como "*status* social" ou "poder," para o qual haja um zero absoluto, porque é sempre possível imaginar um caso com um pouco menos do constructo. Os constructos das ciências sociais têm esta característica de regressão infinita para a extremidade inferior da escala. É sempre possível imaginar um caso com um pouco menos de *status*, ou um pouco menos de poder, e a escala, portanto, não tem um zero verdadeiro. Sem um zero absoluto a escala não tem razões — não é possível dizer que um escore 10 represente duas vezes mais do constructo do que um escore 5.

Resumo

Começamos este capítulo dizendo que as definições operacionais são indispensáveis — sem elas não haveria mensuração científica. Elas possibilitam um processo público para reproduzir e replicar medidas. Reproduzindo e replicando mensurações podemos determinar quão fidedignas elas são. Contudo, as definições operacionais também contém erros. Sempre incluem elementos irrelevantes que não mantém relação com o constructo que desejamos medir e excluem partes do constructo que teríamos gostado de medir. Por esta razão, nenhuma operação isoladamente é definicional do constructo. Cada uma é apenas uma aproximação, um substituto parcial do constructo subjacente. Em reconhecimento ao fato de nenhuma medida ser fidedigna 100% ou 100% válida, defendemos o uso de múltiplas definições operacionais de um único constructo. "Quando operações múltiplas fornecem resultados consistentes, a possibilidade de defasagem entre a definição conceptual e a especificação operacional é bastante reduzida" (Webb, Campbell, Schwartz e Sechrest, 1966, p. 5). A evidência mais convincente provém de uma triangulação de processos de mensuração. Se uma proposição, pode sobreviver à investida rigorosa de uma série de medidas imperfeitas, com todo seu erro irrelevante, deve-se confiar nela" (Webb *et al.*, 1966, p. 3).

O principal problema da mensuração é que nunca podemos estar certos de que medimos o que pretendíamos. Este é um problema que tanto apresenta uma regressão infinita como uma solução. A regressão infinita ocorre porque nunca podemos estar certos de que qualquer escala mede o que pretendíamos medir, a menos que a comparemos com outra, a qual, por sua vez, admitimos, também é falível. A solução está em conceder crédito temporário às outras medidas, enquanto testamos a qualidade de uma. Campbell descreve esta situação como se segue: "Somos como marinheiros que precisam consertar um navio podre em pleno mar. Confiamos na maior parte da madeira enquanto substituímos uma tábua particularmente fraca. Cada uma das tábuas em que agora confiamos pode ser, por sua vez, substituída. A proporção das tábuas que estamos substituindo em relação às que imaginamos em bom estado deve sempre ser pequena" (1974, p. 6). Podemos viver com o conhecimento de que cada uma de nossas medidas é imperfeita enquanto houver alguma concordância entre as medidas imperfeitas. Desistimos de buscar a certeza e, em vez disso, aceitamos o consenso.