6

FIDEDIGNIDADE

Cristian Zanon
Nelson Hauck Filho

Fidedignidade, ou precisão, de um teste refere-se à estabilidade com que os escores dos testandos conservam-se em aplicações alternativas de um mesmo teste ou em formas equivalentes de testes distintos (Anastasi & Urbina, 2000). Quanto mais similares forem os escores dos testandos em aplicações distintas, maior será a fidedignidade de um teste; quanto mais diferentes forem os escores dos participantes, menor será a fidedignidade do teste. Em outras palavras, a análise da fidedignidade dos escores de um teste permite estimar o grau de flutuação esperado dos escores em aplicações subsequentes. Muitas variáveis podem influenciar na flutuação dos escores ao longo do tempo (p. ex., aprendizado, lembrança das respostas anteriores), e, por essa razão, diferentes procedimentos para estimar a fidedignidade e considerações sobre suas limitações serão apresentados ao longo deste capítulo.

O avanço da ciência em várias áreas está atrelado ao aprimoramento dos instrumentos de medida, que possibilitam a adequada avaliação e o estudo do fenômeno de interesse. A criação de telescópios com maior alcance e precisão foi fundamental para a observação e o avanço do conhecimento sobre o universo, por exemplo. Na psicologia, a situação não é diferente. A replicação de estudos constitui um aspecto fundamental da ciência. Por isso, é esperado que o teste usado nas pesquisas psicológicas consiga diferenciar apropriadamente testandos com diferentes níveis no traço latente de interesse (p. ex., baixo, médio, alto), mas também consiga recuperar esses escores posteriormente – considerando que o traço medido apresente estabilidade ao longo do tempo (p. ex., habilidades, personalidade, psicopatologia). Se esse não fosse o caso, qualquer diferença encontrada entre aplicações distintas do mesmo teste no mesmo grupo de sujeitos poderia decorrer de um problema de mensuração inadequada do fe-

nômeno (ou inadequação do teste para a finalidade almejada). Isso, por sua vez, impediria ou limitaria o entendimento e o acúmulo de conhecimento sobre o fenômeno estudado – tudo o que não se deseja na ciência. O Capítulo 2, sobre questões básicas de mensuração, oferece mais exemplos sobre a importância de medição na ciência psicológica.

A fidedignidade é uma propriedade psicométrica fundamental para a validade de um teste, de modo que um teste com baixa fidedignidade não será válido, pois não mede apropriadamente o construto de interesse. Apesar de fundamental, a fidedignidade não é uma condição suficiente para o uso do teste (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), pois ele deve apresentar evidências de validade conjuntamente. Uma vez que a avaliação de evidências de validade requer procedimentos complexos e, muitas vezes, mais custosos, é comum pesquisadores avaliarem, inicialmente, a fidedignidade das escalas em desenvolvimento. Se não houver evidências de fidedignidade, também não haverá de validade.

O termo "fidedignidade" é substituído, muitas vezes, por outros, como "confiança", "consistência interna", "estabilidade" e "precisão". Todos esses termos referem-se ao quão bem o conjunto de itens do teste consegue produzir escores que diferenciam testandos com diferentes graus de habilidade, personalidade, entre outros. Ou seja, quanto maior a fidedignidade, maior a capacidade de distinguir participantes em termos de diferenças individuais. Para ilustrar, podemos pensar que um teste com baixa fidedignidade pode não conseguir distinguir um testando com altíssimas habilidades lógico-matemáticas (p. ex., superdotado) de outro testando com habilidades um pouco acima da média. Ou, ainda, o teste pode não diferenciar um testando incapaz de realizar operações matemáticas elementares de outro que as realiza de modo razoável. Tais limitações podem ser suficientes para decidir não usar um teste como esse para avaliar habilidades lógico-matemáticas de estudantes, por exemplo.

Um termo-chave e diretamente relacionado ao conceito de fidedignidade é o erro de medida. Ele está presente nas avaliações e representa uma variável que pode limitar ou impossibilitar o uso dos testes. Algumas explicações sobre o erro de medida serão fornecidas antes de voltarmos ao conceito de fidedignidade e seus procedimentos de avaliação.

Por que fidedignidade é tão importante para a psicometria? Porque a replicação de estudos é um quesito fundamental na ciência. Logo, o instrumento de medida usado em pesquisa deve ser o mais preciso possível, para garantir as condições necessárias para a adequada replicação de resultados.

ERROS DE MEDIDA

Qualquer tipo de medição está sujeito a erros, e, em psicometria, a presença do erro é sempre considerada existente nas testagens. O significado do erro pode ser pensado como equivalente à diferença entre escores observados e escores verdadeiros dos testandos.

Erro = escore observado – escore real

O escore observado é aquele obtido pelo participante durante a testagem (p. ex., valor bruto ou número fornecido pelo teste); o escore verdadeiro seria aquele que o participante deveria receber, mas é desconhecido (equivalente a seu real nível de habilidade); e o erro refere-se a fatores específicos (desconhecidos ou não mensurados) que podem fazer o testando apresentar desempenho superior ou inferior ao seu real nível de habilidade em um teste de inteligência, por exemplo. Como mencionado, o erro de medida pode aumentar ou diminuir o desempenho do testando. Por isso, se aplicarmos muitas vezes o mesmo teste no mesmo participante, a média dos escores brutos tenderá a representar seu escore real. Imagine que algumas vezes o participante apresenta melhor desempenho, e outras, pior. As diferenças para mais e para menos nos escores do teste tenderão a se anular. Apesar de o exemplo tratar de um teste de habilidade, o conceito é o mesmo para testes de personalidade, atitude e outros. Imagine que é possível alguém receber um escore mais alto (ou mais baixo) em extroversão devido ao erro, por exemplo.

A suposição de que alguém responderia muitas vezes o mesmo teste para que, com a média de seus escores brutos, fosse possível conhecer o seu escore verdadeiro é totalmente irrealista. Dificilmente, na prática, consegue-se que alguém responda ao mesmo teste duas ou mais vezes. Contudo, quando o interesse é conhecer o nível de habilidade de um grupo, por exemplo, a equação apresentada anteriormente também é válida para amostras aleatórias (conjunto de testandos sorteados da população de interesse). Assim, com uma aplicação, apenas se poderia conseguir uma aproximação do escore real do grupo. Imagine que muitos testandos terão sorte e acertarão questões por acaso, enquanto outros errarão questões que dominavam por falta de atenção ou outros fatores já mencionados. Na média, os aumentos e as diminuições devidos ao erro aleatório tendem a se anular, e, assim, pode-se chegar ao valor real dos escores dos grupos. A demonstração disso envolve cálculos complexos e caracteriza uma grande contribuição da psicometria para a Teoria Clássica dos Testes – o que respaldou a aplicação de testes em grupos. Para mais informações sobre a Teo-

ria Clássica dos Testes, sugere-se a leitura do Capítulo 2 sobre questões básicas de mensuração.

Dois tipos de erros podem influenciar os escores de um teste: erros aleatórios e erros sistemáticos. Fatores específicos, como fadiga, fome, sonolência, esquecimento temporário de informações, ruídos perturbadores, conversas desnecessárias na sala de aplicação, entre outros, podem reduzir o desempenho, enquanto sorte e conhecimento prévio de passagens contidas no texto base de algumas questões podem aumentar o desempenho dos participantes. Outros fatores, como itens e instruções inadequados dos testes, erros de digitação e cômputo errado dos escores dos participantes, também são classificados como erros aleatórios que enviesam os resultados. Por essa razão, todo cuidado possível é necessário antes, durante e depois das testagens. O seguimento das instruções do teste referente a todas as etapas de aplicação e correção é fundamental para a redução de erros de medida. Contudo, mesmo com todo esse cuidado, o erro estará presente em alguma medida – o que desejamos é que ele esteja reduzido.

Erros aleatórios podem trazer importantes consequências em nível individual. Por exemplo, um candidato pode ter acertado, por sorte, 3 de 10 questões que não dominava (em uma prova de múltipla escolha), enquanto outro pode ter errado 3 das 10 questões que dominava, por desatenção ou falta de motivação para terminar a prova. Em ambos os casos, os escores produzidos contêm grandes erros e podem trazer consequências importantes para os testandos, dependendo do propósito do teste. Quando o número total de itens do teste é maior, é menos provável a ocorrência de tais discrepâncias. Por essa razão, o julgamento sobre a habilidade numérica de um testando usando um teste com 50 itens tende a apresentar menos erros desse tipo do que com um teste de 10 itens. Ou seja, quanto mais itens no teste, maior sua fidedignidade.

Uma analogia do papel do erro na fidedignidade pode ser pensada se considerarmos uma balança desregulada que produz valores incorretos sobre o peso de objetos. Contudo, nesse caso, o erro (para mais ou para menos) seria sistemático, pois todos os objetos receberiam um valor constante superior ou inferior e, se alguém descobrisse o quanto ela está desregulada, seria possível corrigir a imprecisão dos pesos.

A presença do erro sistemático na testagem psicológica é pouco preocupante, especialmente quando o interesse recai sobre diferenças individuais, porque ele afeta todos os participantes igualmente. O problema do erro sistemático ocorre quando o interesse é comparar escores de um teste livre de erro sistemático com outro com tal erro (Kline, 1993). Na próxima seção, serão apresentadas as principais formas de avaliar a fidedignidade de um teste.

Se toda forma de testagem apresenta erros, como saber se o teste pode ser usado ou não? Toda testagem apresenta erros, mas, como podemos esti-

má-los, é possível usar essa informação para avaliar a pertinência de usar determinado teste. A quantia de erro aceita dependerá do propósito do teste. Por exemplo, para pesquisa se tolera mais erros do que em um contexto de seleção de pessoal.

Fidedignidade Teste-Reteste

A fidedignidade teste-reteste possivelmente se constitui como o método mais intuitivo para avaliação da consistência dos escores ao longo do tempo. Ela basicamente consiste em uma correlação dos escores dos mesmos testandos avaliados em momentos distintos. A correlação é uma análise estatística entre duas ou mais variáveis que produz um coeficiente que varia de -1 a +1, sendo que o valor "0" implica a ausência total de relação entre as variáveis. Quanto mais próximo a 0 for o valor do coeficiente, menor a relação das variáveis; quanto mais próximo a 1 for o valor do coeficiente, maior é a relação entre as variáveis. Imagine que um teste de habilidade motora foi aplicado a um grupo de crianças em um tempo T1 e reaplicado em um tempo T2 e que o resultado final do teste é V1 (no tempo 1) e V2 (no tempo 2). Vamos considerar V1 a variável 1 e V2 a variável 2. Se quisermos conhecer a fidedignidade desse teste, precisaríamos calcular a correlação entre V1 e V2. Detalhes sobre o cálculo de correlação podem ser encontrados em livros-texto básicos de estatística e este cálculo é facilmente implementado em pacotes estatísticos com SPSS, Excel ou R.

De forma geral, a reaplicação de um teste com alta fidedignidade revelará coeficientes de correlação com valores superiores a 0,80, que indicam considerável manutenção do valor dos escores entre T1 e T2. Mais especificamente, o valor 0,80 indica que 64% da variância dos escores de T1 e T2 é compartilhada. Se elevarmos o valor da correlação ao quadrado ($0,80^2 = 0,64$), obtemos a quantia de variância compartilhada entre as variáveis relacionadas. Os 36% restantes da variância total seriam atribuídos a erro ($1 - 0,64 = 0,36$). Se o valor da correlação fosse 0,90, teríamos um compartilhamento de 81% da variância de T1 e T2 (sendo apenas 19% da variância total atribuída a erro). Agora, imagine se o valor da correlação observada fosse 0,70. Aplicando o raciocínio anterior, perceberíamos que apenas 49% da variância total de T1 e T2 é compartilhada, sendo que 51% (maior parte) deve-se a erro. Quanto menor a correlação, menor a correspondência dos escores entre T1 e T2 e maior a parcela de variância de erro.

O período entre as aplicações desempenha um papel importante no coeficiente de correlação obtido. Se um teste é reaplicado no intervalo de dias ou semanas, ele tenderá a apresentar maiores coeficientes de correlação do que se ele for reaplicado em um período de meses ou anos. Mudanças, desenvolvimen-

to e aprendizagem podem ocorrer no intervalo entre as aplicações e alterar os escores de T2. Se um ditado é repetido em três semanas a crianças de ensino fundamental, é provável que apresente menores alterações do que ser for repetido no intervalo de um ano. É possível que a criança aprenda novas palavras e apresente desenvolvimento considerável na habilidade de escrita em um período de um ano.

Contudo, a reaplicação de testes em períodos curtos não é livre de problemas que incidirão sobre os erros. É possível que em períodos curtos de reaplicação os participantes lembrem de suas respostas anteriores e que seus escores reflitam muito sua memória às respostas, e não apenas suas habilidades. Ademais, a simples repetição de um problema (questão) que requer desenvolvimento analítico pode torná-lo mais fácil de ser respondido (e acertado) na segunda tentativa, por elaboração subsequente dos procedimentos necessários para a sua resolução (Anastasi & Urbina, 2000).

Eventos de vida positivos ou negativos podem influenciar consideravelmente os escores de uma reaplicação. Imagine que momentos antes da reaplicação de um teste que avalia neuroticismo um participante seja assaltado na rua. Seu escore de ansiedade muito provavelmente estará elevado e apresentará baixa relação com o escore anterior – considerando que não se trata de uma pessoa com altos escores de ansiedade. Outros fatores, como instruções diferentes nos momentos da aplicação, podem influenciar a motivação dos participantes e enviesar os escores. Por isso, a maioria dos testes psicológicos será influenciada por variáveis desconhecidas entre os períodos de aplicação, as quais podem produzir distorções consideráveis. Logo, não se recomenda a aplicação desse método para testes suscetíveis a tais influências. Testes motores e de discriminação sensorial são exceções (Anastasi & Urbina, 2000), pois estariam menos sujeitos às influências anteriormente mencionadas.

O método teste-reteste basicamente avalia a estabilidade dos escores do teste ao longo do tempo. O cálculo da correlação entre as aplicações produz um coeficiente que permite avaliar o nível de flutuação e estabilidade dos escores.

Fidedignidade de Formas Alternadas

Um procedimento similar ao teste-reteste é o de formas alternadas, que tem por objetivo avaliar a relação entre os escores dos testandos entre um período de tempo. O procedimento de formas alternadas difere do teste-reteste, pois as aplicações ocorrem com conjuntos de itens distintos. Para isso, é necessária a existência de duas formas equivalentes do mesmo teste. Por equivalência entende-se que os testes devem apresentar o mesmo número de itens, o mesmo formato (p. ex., mesmo número de alternativas falsas, mesmo número de pontos

na escala Likert), a mesma dificuldade ou atratividade, as mesmas instruções e cobrir os mesmos domínios.

O uso de formas alternadas é desejável porque reduz as chances de treinamento ou fraude (Anastasi & Urbina, 2000). Contudo, com raras exceções, conseguimos aplicar dois ou mais testes equivalentes a grupos de testandos. Como, então, podemos avaliar a fidedignidade de um teste aplicado uma única vez? Algumas soluções serão apresentadas a seguir.

DUAS METADES

Este procedimento consiste em separar o teste em duas partes e calcular a correlação entre elas. Se os valores de correlação forem elevados, então há evidências de fidedignidade pelo método das metades para o teste todo. Um problema, nesse caso, consiste em determinar como o teste deve ser dividido, pois a predominância de conteúdos relevantes em apenas uma das metades apenas pode comprometer a avaliação da fidedignidade por esse método.

A divisão entre itens pares e ímpares pode ser usada desde que os domínios de conteúdos cobertos pelo teste não se concentrem em uma das metades. Ademais, essa divisão é pertinente se os itens estiverem dispostos em ordem crescente de dificuldade. Contudo, uma das críticas endereçadas ao uso desse procedimento é que coeficientes distintos podem ser obtidos dependendo da forma como o teste é dividido (Cronbach, 1951).

Coeficiente Alfa – α

Um dos procedimentos provavelmente mais conhecidos e usados para avaliação da fidedignidade dos escores de um teste é o coeficiente alfa, também conhecido como alfa de Cronbach. Apesar de o nome "alfa de Cronbach" ter-se consolidado na literatura psicológica e educacional para designar o coeficiente, o próprio Lee J. Cronbach não aprovava tal nomenclatura (Hambleton, comunicação pessoal, 2011). O coeficiente alfa foi inicialmente proposto por Louis Guttman e, posteriormente, aprimorado por Cronbach (Maydeu-Olivares, Coffman, García-Forero, & Gallardo-Pujol, 2010). Hoje, a revista oficial da International Test Commission (ITC) – *International Journal of Testing* – recomenda o uso do termo "coeficiente alfa" em suas publicações.

O coeficiente alfa é a média de todos os coeficientes possíveis de duas metades de um teste e indica o valor esperado de uma divisão aleatória do conjunto de itens de um teste (Cronbach, 1951). Em outras palavras, se dividíssemos um teste usando todas as possibilidades, obteríamos um coeficiente de correla-

ção referente a cada divisão. Se, então, calcularmos a média de todos esses coeficientes, obteremos o coeficiente alfa.

Os valores de alfa geralmente vão de 0 a 1, sendo que, quanto mais próximos a 1, maior a fidedignidade do teste. Contudo, valores negativos de alfa podem ser produzidos – ainda que sem sentido prático. Esse é o caso em muitas situações em que se esquece de inverter itens negativos de um teste. George e Mallery (2002) sugerem alguns valores de referência para a interpretação dos coeficientes:

$\alpha > 0{,}90$ = excelente
$0{,}89 > \alpha > 0{,}80$ = bom
$0{,}79 > \alpha > 0{,}70$ = aceitável
$0{,}69 > \alpha > 0{,}60$ = questionável
$0{,}59 > \alpha > 0{,}50$ = ruim
$\alpha < 0{,}50$ = inaceitável

O coeficiente alfa é comumente usado em testes com itens politômicos, ou seja, aqueles cujas chaves de resposta estão dispostas em escalas Likert – que apresentam números representando o quanto o indivíduo concorda ou discorda com determinada afirmação, por exemplo. Testes compostos por itens politômicos geralmente são testes de personalidade, de atitude ou de psicopatologia. Contudo, o coeficiente alfa também pode ser usado com testes compostos por itens dicotômicos – aqueles que apresentam: a) uma alternativa apropriada ou certa, b) sim ou não, c) concordo ou discordo, entre outros. Testes com itens dicotômicos geralmente são aqueles que caracterizam testes de habilidades. Nesse caso, o coeficiente alfa equivalerá a outro índice de fidedignidade bastante usado em testes de desempenho – o coeficiente Kuder-Richardson. Em outras palavras, o coeficiente Kuder-Richardson pode ser pensado como um caso específico do coeficiente alfa aplicado em itens dicotômicos (Cronbach, 1951; Revelle, 2015).

Fidedignidade do Avaliador

Determinados tipos de testes, como os projetivos, podem estar mais sujeitos à avaliação subjetiva do avaliador e, por essa razão, podem necessitar de evidências adicionais de fidedignidade para seus escores (Anastasi & Urbina, 2000). Nesse caso, podem-se correlacionar os resultados dos protocolos de respostas dados a avaliadores diferentes e, então, verificar o grau de similaridade entre eles. O próprio valor da correlação pode ser o índice da fidedignidade do avaliador. Esse tipo de evidência complementa as evidências fornecidas por outros

procedimentos, como os mencionados anteriormente, e é desejável para testes em que a padronização da produção de escores pode estar mais sujeita a vieses do aplicador.

LIMITAÇÕES DOS MÉTODOS CLÁSSICOS DE AVALIAÇÃO DA FIDEDIGNIDADE

Os distintos procedimentos para avaliação de fidedignidade antes apresentados[1] tratam de diferentes fontes de erro (p. ex., temporal, conteúdos avaliados em formas distintas e subjetividade do avaliador) e apresentam formas variadas de estimar e lidar com a variância de erro (parcela total do escore do teste referente ao erro). Contudo, esses índices consideram que o erro é constante para todo o *continuum* do traço latente avaliado. Imagine, por exemplo, que temos um teste de extroversão que apresenta coeficiente alfa de 0,70. Portanto, poderíamos dizer que esse teste apresenta nível aceitável de fidedignidade. Mas o problema é que não sabemos qual é a quantia de erro para testandos que pontuam baixo, médio ou alto no teste.

Uma vez que a fidedignidade do teste tem por finalidade produzir informações sobre o quão bem o teste diferencia testandos com níveis baixo, médio ou alto, como podemos saber se ele está discriminando adequadamente participantes desses níveis? Imagine, agora, que estamos pensando em usar esse teste para selecionar participantes tímidos para participar de um curso de oratória. Como eu sei que o teste é capaz de identificar os mais tímidos mesmo? Se há considerável montante de erro na avaliação de participantes com baixos níveis de extroversão, é possível que pessoas tímidas pareçam não tímidas e pessoas não tão tímidas pareçam muito tímidas. Dependendo do propósito do teste, níveis de erro devem ser avaliados – como no caso de seleções. Todavia, saber em que parte do traço latente há maior concentração de erro também é fundamental.

A Teoria de Resposta ao Item (TRI) (Lord & Novick, 1968) constitui um conjunto de modelos estatísticos que suprem essa lacuna dos coeficientes tradicionais de fidedignidade. A aplicação da TRI nos testes permite avaliar especificamente em que parte do traço latente o teste está mensurando mais adequadamente os participantes e onde há mais erros de medida. A TRI constitui um

[1] Diversos outros coeficientes e métodos de avaliação de fidedignidade não foram contemplados neste capítulo, tampouco descritos detalhadamente com fórmulas. Interessados no tópico podem procurar por Revelle (2015) – Personality Project –, que reúne material abrangente sobre fidedignidade e psicometria.

poderoso recurso para avaliação de fidedignidade e contribui para o aprimoramento e o desenvolvimento de testes (ver, Zanon, Hutz, Yoo, & Hambleton, no prelo, para uma aplicação didática de um modelo de TRI em um teste psicológico, e Knijinick, Giacomoni, Zanon, & Stein, 2014, para uma aplicação em um teste de desempenho escolar).

QUESTÕES

1. Explique o que é fidedignidade. Por que ela é importante? Por que ela não constitui fonte suficiente de evidência para o uso de um teste?
2. O que são erros de medida?
3. Qual a relação entre erro de medida e fidedignidade?
4. Quais são os tipos de erro mencionados no capítulo?
5. Quais as principais formas de avaliar a fidedignidade de um teste?
6. Por que o teste-reteste não é apropriado para a maioria dos testes psicológicos?
7. O que significa o coeficiente alfa?

REFERÊNCIAS

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington: AERA, APA, NCME.

Anastasi, A., & Urbina, S. (2000). *Testagem psicológica* (7. ed.). Porto Alegre: Artmed.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

George, D., & Mallery, P. (2002). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Kline, P. (1993). *Handbook of psychological testing*. New York: Routledge.

Knijnik, L. F., Giacomoni, C. H., Zanon, C., & Stein, L. M. (2014). Avaliação dos subtestes de leitura e escrita do teste de desempenho escolar através da teoria de resposta ao item. *Psicologia: Reflexão e Crítica, 27*(3), 481-490.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Maydeu-Olivares, A., Coffman, D. L., García-Forero, C., & Gallardo-Pujol, D. (2010). Hypothesis testing for coefficient alpha: An SEM approach. *Behavior Research Methods, 42*(2), 618-625.