3
Teoria Clássica dos Testes (TCT)

Josemberg Moura de Andrade

Germano Gabriel Lima Esteves

Jacob Arie Laros

Introdução

A teoria da medida ou medida em psicologia comumente é um assunto que causa certa estranheza aos estudantes iniciantes em cursos de graduação em psicologia. Ao contrário do que muitos imaginam, as medições existem para facilitar nossas vidas e para orientar a tomada de decisões. Ressalta-se que medir objetos físicos é uma tarefa relativamente fácil. Munidos de uma régua podemos medir o comprimento e a largura de uma mesa; também podemos fazer uma avaliação qualitativa do estado geral dessa mesa. Por outro lado, avaliar construtos psicológicos – atributos que existem, mas que não são observáveis a olho nu – é algo muito complexo e que exige um alto grau de especialização. Considerando essa problemática, o presente capítulo, escrito por três professores de medidas em psicologia, busca explicar em linguagem fácil e acessível os conceitos da área de medidas em psicologia, dirimindo possíveis dúvidas comuns entre estudantes e profissionais.

Podemos iniciar afirmando que no dia a dia, naturalmente, utilizamos diversos sistemas de medidas, seja para preparar uma comida

(Quantos mililitros de leite ou quantos gramas de farinha deveríamos adicionar em uma determinada receita?), para estudar (Quantas horas passaremos estudando medidas em psicologia?) ou para nos deslocarmos (Quantos quilômetros teremos que nos deslocar para chegar até a cafeteria mais próxima?). Esses são apenas alguns dos exemplos de sistemas de medida que fazemos uso para simplificar nossas tarefas diárias.

No entanto, comumente não nos perguntamos de onde surgiram esse ou aquele sistema, ou o quão sofisticado ele é para medir aquilo que pretendemos medir. Vamos fazer o exercício de imaginar se não tivéssemos um sistema de medida para distâncias tão sofisticado quanto o metro, o que responder se alguém perguntasse a distância entre o seu local de trabalho e a sua casa? Possivelmente, utilizaríamos uma comparação entre distâncias (*e. g.*: "daqui para o meu trabalho é duas vezes a distância entre o *shopping* e o centro da cidade") ou tentar criar um sistema menos sofisticado (*e. g.*: "meu trabalho fica a três quadras da minha casa"). Em ambas as respostas a precisão sobre a distância é uma incógnita por diferentes variáveis (*e. g.*: a percepção individual do quão distante é o *shopping* do centro da cidade ou os tamanhos de quadras residenciais que podem variar). Como consequência, não teríamos como saber o quanto de combustível precisaríamos para ir de casa ao trabalho. Este é um dos motivos pelo qual se fazem necessários sistemas de medidas mais sofisticados.

Esse mesmo tipo de tentativa de sistematização entre distâncias ocorre quando tentamos responder perguntas sobre aspectos emocionais e cognitivos de outras pessoas. Possivelmente, você já foi perguntado se uma terceira pessoa é tímida, ansiosa ou até mesmo se

tem inteligência acima da média. Para isso, utilizou uma comparação (*e. g.*: "ele é mais ansioso do que eu") ou indicou algo que essa pessoa fez como critério (*e. g.*: "ela é tão inteligente que conseguiu se graduar como a primeira de sua turma"). A informação fornecida nessas respostas pode ter sua utilidade para responder perguntas simples no dia a dia, no entanto traz pouca informação para a tomada de decisões com implicações diretas na vida das pessoas. Na prática, o quão confiável seria a tomada de decisão de um psicólogo clínico ao recomendar um procedimento ou tratamento específico com base unicamente na informação de que "o paciente é mais ansioso do que sua secretária" ou de um psicólogo organizacional ao recomendar determinado candidato para um cargo com base na informação de que "o candidato conseguiu se graduar como o primeiro de sua turma". Para as duas ocasiões, os elementos fornecidos trazem mais perguntas do que informações: Com base em quais informações pode-se afirmar que o paciente é mais ansioso do que a secretária? Essas informações são as mais representativas da ansiedade? Quais os aspectos avaliados no candidato que fez ele ser o primeiro de sua turma? Essas avaliações apresentam-se calibradas para avaliar esses aspectos? O conhecimento sobre esses e outros aspectos psicológicos são de responsabilidade da psicologia, que tem a psicometria como o ramo que cuida da sistematização e elaboração de medidas com a finalidade de testar hipóteses científicas.

É nesse contexto que este capítulo tem como objetivo apresentar a medida em psicologia com foco na Teoria Clássica dos Testes (TCT). Para tanto, serão apresentados detalhes sobre a psicometria e a TCT. Desse modo, serão descritos ainda os postulados que regem a TCT e sua importância na construção e na obtenção de evidências de validade

e precisão de testes. Prosseguiremos apontando as limitações da TCT, de modo a explicitar as implicações teóricas e práticas da utilização desse modelo. Assim, ao final deste capítulo espera-se que o leitor tenha adquirido conhecimentos necessários para refletir e analisar, tanto no nível teórico quanto prático sobre as medidas utilizadas em diversas áreas da psicologia.

A medida na psicologia e a Teoria Clássica dos Testes

Inicialmente, supondo que queremos medir qual o comprimento de uma barra de ferro, poderíamos simplesmente utilizar um instrumental baseado no sistema de medida métrico decimal (*e. g.*: uma fita métrica ou uma régua), comparando-o com a barra de ferro, correto? Agora, se queremos medir alguns aspectos psicológicos, como a agressividade ou extroversão de determinado indivíduo, como podemos fazer? Qual o instrumento que poderíamos utilizar? Como medir algo que não temos acesso direto? Afinal, você já viu uma agressividade ou uma extroversão, assim como vemos uma barra de ferro? Do mesmo modo que para medir uma barra de ferro se fez necessária uma sistematização prévia, para se estabelecer uma medida de aspectos psicológicos não é diferente. Em ambos os casos, a medida exige uma sistematização de relações entre o sistema numérico e aquilo que se quer medir, para que se possibilite o estabelecimento de operações matemáticas, em outras palavras: "medir consiste em assinalar números a objetos e eventos de acordo com alguma regra" (Stevens, 1946). Desse modo, podemos saber que ao juntar duas barras de ferro (soma) chegamos ao tamanho desejado ou que ao reduzir o nível de depressão de um paciente (subtração) temos que mudar o tipo de intervenção realizada. Podemos

estabelecer, portanto, que a medida tem como função permitir uma descrição mais precisa e objetiva de determinado fenômeno por meio da condensação de informações, possibilitando uma comparação e equivalência de fenômenos distintos e desse mesmo fenômeno em diferentes condições com uma elevada equivalência.

O estabelecimento de uma sistematização de relações entre o sistema numérico e o que se quer medir (*e. g.*: objeto empírico ou aspecto psicológico) é norteado pela teoria da medida, que se preocupa com a utilização do sistema numérico para descrever fenômenos naturais (Pasquali, 2010). Especificamente, a teoria da medida tem tido implicações diretas no delineamento, interpretação e resultados de pesquisas e na operacionalização de avaliações psicológicas e educacionais. Desse modo, o objetivo dessa teoria é descrever, categorizar e avaliar a qualidade das medidas, a fim de aperfeiçoar sua utilidade, precisão e significado. Além disso, a teoria da medida tem como foco propor novos e melhores métodos para desenvolver instrumentos de medida (Allen & Yen, 2002); como assinala Pasquali (2000, 2013), tal teoria desenvolve uma discussão epistemológica em torno da utilização dos números no estudo científico dos fenômenos naturais.

A mensuração – objeto da teoria da medida – pode ser definida como um conjunto de regras para representar o comportamento em categorias ou números (Mathison, 2005). Em outras palavras, mensurar significa designar ou especificar números para indivíduos de uma forma sistemática com a pretensão de representar as propriedades desses indivíduos (Allen & Yen, 2002). Duas abordagens teóricas são dominantes no campo da mensuração, a saber: a Teoria Clássica dos

Testes (TCT) e a Teoria de Resposta ao Item (TRI) (Mathison, 2005). A medida, nesse caso, só será possível se o objetivo que pretendemos medir considerar pelo menos duas de três das propriedades do sistema numérico. A medida, nesse caso, só será possível se a estrutura empírica daquilo que se quer medir é preservada nas relações com as propriedades do sistema numérico (Teorema da representação) (Hauck-Filho, 2014); e se essas relações são preservadas após a realização de operações matemáticas (Teorema da unicidade) (Rossi, 2007).

Desse modo, a estrutura empírica deve ser preservada nas seguintes propriedades do sistema numérico (Pasquali, 2010):

1) Identidade: indica que cada número é único, ou seja, cada número é igual a ele mesmo e somente a ele mesmo, ou seja, "a = a" e "a ≠ b" (axioma de reflexividade), se "a = b" então "b = a" (axioma de simetria) e se "a = b" e "b = c" logo "a = c" (axioma de transitividade). Caso apenas essa propriedade esteja presente naquilo que se quer medir, pode-se identificar um nível de medida chamado de nominal (Stevens, 1946). A escala nominal identifica elementos iguais a ele e somente ele mesmo ou que apresentam diferenças entre si (*e. g.*: o número na camisa de um time de basquete). Diferentemente de outros níveis de medida, essa escala não possibilita identificar qual elemento é quantitativamente maior ou a realização de operações matemáticas. Por esses motivos, a escala nominal não é considerada como uma medida, mas como uma classificação.

Na prática, a escala nominal pode ser utilizada, por exemplo, para coletar informações sobre o sexo biológico dos indivíduos. Nesse caso, com relação ao sexo biológico, alguém identificado como pertencente ao sexo masculino é qualitativamente comparável, e somente, a outro

indivíduo identificado como pertencente ao sexo masculino e diferente de um indivíduo pertencente ao sexo feminino. Observe que nesse nível de medida (ou classificação) não faz sentido dizer que o sexo masculino é melhor do que o sexo feminino (ou vice-versa) ou que a concatenação dos sexos acarretaria em um terceiro sexo biológico quantitativamente diferente dos anteriores. A primeira afirmação acarretaria perguntas como: Melhor em quê? Para responder a essa pergunta seria necessário definir a característica que se quer medir (*e. g.*: nível de organização) e demonstrar que esse sistema empírico pode preservar suas características nesta e em outras duas propriedades dos números, apresentados a seguir.

2) Ordem: Esta outra propriedade indica que os números diferem quantitativamente um do outro; indicando uma magnitude diferente para cada número. Desse modo, excetuando-se a igualdade dos números (1 = 1), os números podem ser dispostos em uma sequência, colocados em uma ordem crescente, logo, se "a > b" então "b < a" (axioma de assimetria), se "a > b" e "b > c" então "a > c" (axioma de transitividade) e ou "a > b" ou "b > a" (axioma de conectividade) (Pasquali, 2010, 2013). Assim, aquelas medidas que apresentam as propriedades de identidade e ordem são nomeadas de ordinal (Stevens, 1946). A escala ordinal indica elementos iguais a ele, e somente ele mesmo, ou que apresentam diferenças entre si e que, além disso, diferem quantitativamente entre si, o que já possibilita algumas operações matemáticas.

Na prática, uma escala ordinal poderia servir para ranquear seus quatro amigos mais íntimos (representados aqui pela letra "a") pelo nível de organização (do mais organizado ao menos organizado),

certamente você conseguirá distinguir e representá-los, como: a3, a3, a2, a1. Observe que no *ranking* apresentado dois dos seus amigos têm o mesmo nível de organização (a3 = a3) e que são quantitativamente diferentes dos outros (a3 ≠ a2; a3 ≠ a1), preservando a propriedade de identidade. Além disso, o *ranking* realizado também preserva a propriedade de ordem, uma vez que "a3" é mais organizado que "a2" (a3 > a2 > a1).

3) Aditividade: essa propriedade está diretamente relacionada ao teorema da unicidade, uma vez que representa a propriedade dos números serem somados e, excetuando-se o número zero, resultar em um terceiro número diferente. Essa propriedade aponta para as regras de operações matemáticas; "a + b = b + a" (axioma de comutatividade) e "(a + b) + c = a + (b + c)" (axioma de associatividade). Assim, essas operações devem ser possíveis de serem realizadas sem que as relações empíricas estabelecidas com as propriedades anteriormente estabelecidas se alterem (Rossi, 2007). Isso implicaria dizer, retomando nosso exemplo anterior, que a soma dos níveis de organização de um dos amigos "a3" com o amigo "a2" irá acarretar em um terceiro nível de organização maior do que o que eles têm separadamente (a3 + a2 = a4) e maior do que o dos outros amigos (a4 > a3 > a2 > a1). Na prática, atribuir atividades de organização para os amigos "a3" e "a2" em conjunto deve resultar uma organização mais eficaz do que se a atividade fosse atribuída individualmente a estes.

Nesse caso, se aquilo que se quer medir apresenta as propriedades de identidade, ordem e aditividade, pode-se sistematizar dois tipos de medida, a saber: "intervalar" e de "razão" (Stevens, 1946). Esses níveis

de medida apresentam--se como os mais sofisticados, diferindo-se pela existência de um zero absoluto no caso da escala de razão. Na escala de razão o numeral zero indica a ausência daquilo que está sendo medido, enquanto que na escala intervalar o numeral zero apresenta-se apenas como mais um ponto no intervalo daquilo que se está medindo. Isto implica dizer que em uma escala intervalar, como a escala Célsius, o numeral zero não significa ausência de temperatura, mas um ponto no qual ao grau de agitação das moléculas é baixo (frio); já em uma escala de razão, como a escala Kelvin, o numeral zero indica ausência total de agitação das moléculas (ausência de temperatura).

Nesse ponto, falando de medida em psicologia, poderíamos nos perguntar: É possível que alguém tenha ausência total de agressividade? Ou ausência total de ansiedade? Ou, ainda, ausência total de inteligência? Esses questionamentos apontam para a impossibilidade de se estabelecer um "zero absoluto" em aspectos psicológicos. Contudo, existem divergências sobre o nível de medida que é utilizado em escalas psicológicas, se são medidas ordinais ou intervalares (cf. Michel, 2002; Nunnally, 1995).

Sabemos então que é possível elaborar medidas quando se preserva a estrutura empírica em um sistema de relações com as propriedades dos números, mas como tais propriedades podem ser atribuídas àquilo que queremos medir? Como anteriormente mencionado, a forma de medida que possibilita a medição de uma barra de ferro (medida direta) é uma forma diferente da que ocorre em aspectos psicológicos (medida indireta), isso porque a barra de ferro é um objeto tangível, enquanto que um aspecto psicológico, como a agressividade, não é tangível ou visível. No entanto, ainda que se pense que a mensuração de algo

intangível seja uma idiossincrasia das ciências humanas, é plausível perguntar: Você consegue ver a velocidade de um carro, sem um velocímetro? Ou consegue ver a força gravitacional? Ainda que as formas de medida sejam diferentes, o processo de construção de um sistema de medida se utiliza sempre de um conjunto de regras para representar o comportamento em categorias ou números (Mathison, 2005), a fim de reunir evidências empíricas acerca do funcionamento de determinado fenômeno ou objeto. Para Pasquali (2010, 2013), essas formas de medida podem ser agrupadas em:

a) Medida fundamental: é uma forma de medida direta que pode ser sistematizada quando as dimensões do objeto empírico ao qual se quer medir permitem o estabelecimento de uma unidade-base, permitindo unidades múltiplas e divisoras dessa medida. O estabelecimento dessa medida é condicionado à existência da propriedade de aditividade nas dimensões do objeto. Um exemplo simples de medida fundamental é o sistema métrico decimal, que tem como unidade-base o metro, representado pela letra "m", possibilitando múltiplos (decâmetro, hectômetro, quilômetro) e divisores (decímetro, centímetro e milímetro). Assim, por exemplo, uma barra de ferro que tem várias dimensões (cor, peso, comprimento) possibilita o estabelecimento de uma unidade-base, como o metro, para medir o comprimento (dimensão), que por sua vez pode ser somado ao de outra barra de ferro. Entretanto, apesar de ser uma forma de medida bastante utilizada, a psicologia não se utiliza dessa forma de medida, pois é impossível se estabelecer uma unidade-base que possibilite múltiplos e divisores em aspectos psicológicos. Assim, como você pode imaginar, não existem unidades-base de inteligência ou unidades-base de ansiedade de estado, por exemplo.

b) Medida derivada: constitui-se como uma medida indireta, que se utiliza da relação de duas medidas fundamentais para medir determinado fenômeno. Em outras palavras, quando existem evidências empíricas de que duas dimensões de dois fenômenos afetam um terceiro fenômeno que se quer medir, é estabelecida a relação entre essas dimensões para se medir o terceiro fenômeno. Um exemplo de medida derivada que utilizamos com frequência é a velocidade média que relaciona duas medidas fundamentais, a saber: o espaço, medido pelo sistema métrico decimal, e o tempo, que tem como medida as horas. Desse modo, quando queremos saber a velocidade média de um carro, utilizamos a relação entre metros e segundos ou quilômetros por hora. Ainda, diferentemente da medida fundamental, a medida derivada não tem acesso direto ao fenômeno que se pretende medir.

c) Medida por lei: trata-se de uma forma de medida indireta estabelecida com base nas evidências científicas da relação entre duas ou mais variáveis. É importante destacar que para o estabelecimento de uma medida por lei, inicialmente deve existir evidências empíricas dessa relação, para depois se sistematizar a lei. Na psicologia, a medida por lei é uma medida possível, como é o caso da lei do efeito de Thorndike (1911), que dá base para a relação de estímulo (S) e reforço (R).

d) Medida por teoria: como o próprio nome já diz, nesse tipo de medida é necessário que se estabeleçam axiomas ou postulados que possam gerar hipóteses empiricamente testáveis, ou seja, inicialmente é elaborada uma teoria, sobre aquilo que se deseja medir, que permita a elaboração de hipóteses sobre o seu funcionamento. Esse é o tipo de medida que a psicologia se utiliza para medir aspectos psicológicos, sendo a psicometria o ramo responsável pelo estabelecimento desses

postulados. Especificamente, a psicometria apresenta duas teorias para se realizar a medida em psicologia. São elas: a Teoria Clássica dos Testes (TCT), que preconiza a relação entre a resposta em um dado teste e o comportamento, e a Teoria de Resposta ao Item (TRI), também conhecida como Teoria do Traço Latente, que sistematiza a relação entre a resposta dada a um item de um teste em função do traço latente (processo mental) e dos parâmetros (características ou propriedades) desse item (Pasquali, 2010, 2013; Urbina, 2007). A TRI é tema de um outro capítulo específico deste Compêndio.

Postulados da Teoria Clássica dos Testes (TCT)

A TCT é uma teoria psicométrica que se preocupa em explicar o resultado total das respostas dadas a uma série de itens ou teste. De acordo com a TCT, a pontuação obtida por um examinando em um teste representa o nível do atributo que está sendo avaliado como o somatório das respostas em cada um dos itens (Kline, 2005; Nunnally & Bernstein, 1995). Especificamente, o interesse da TCT não recai sobre o traço latente, e sim sobre o comportamento, ou melhor, o escore em um teste, o que equivale a um conjunto de comportamentos. O enfoque está no tau (τ) e não no teta (θ), sendo que esse primeiro é o escore em um determinado teste, enquanto o segundo, o traço latente (Pasquali, 2010, 2013). A TCT apresenta sete postulados que, de acordo com Grégoire e Laveault (2002), são: (1) O escore total (T – também nomeado de *tau*) de um indivíduo é resultante da soma do resultado verdadeiro (V) com o erro da medida associado a esse mesmo erro (E). Esse postulado pode ser expresso em uma fórmula simples T = V + E, logo, para se obter o escore verdadeiro de um indivíduo

bastaria subtrair o erro da medida do escore total (V = T – E). Isso implica dizer, por exemplo, que o desempenho (T) de um indivíduo em um teste de inteligência é resultante da medida real de sua capacidade cognitiva (V) acrescida de erros de medida (E). O erro é uma característica intrínseca de todas as formas de medida, independentemente de ser realizada de modo direto ou indireto. Não existe medida isenta de erro! Por exemplo, se vamos medir o comprimento de uma barra de ferro, o resultado dessa medição pode ser alterado por outras variáveis como a temperatura ou a falta de calibração do instrumento utilizado. Tais variáveis interferentes são compreendidas como erros da medida, uma vez que alteram o comprimento real da barra de ferro. Entretanto, quando falamos de medida em psicologia, identificar e/ou controlar o erro torna-se algo mais difícil; por exemplo, quando avaliamos um paciente no ambiente clínico ou um candidato no ambiente organizacional podem ocorrer erros relacionados a diversos fatores como, por exemplo, estado de humor transitório, desejabilidade social, fatores ambientais ou erros presentes no próprio instrumento de medida; (2) o segundo postulado é uma estratégia matemática para lidar com o erro, indicando que para se aumentar a precisão do resultado de um teste devemos aumentar o número de observações realizadas. Isso aconteceria, pois seriam gerados vários escores empíricos diferentes, nos quais essas variações se dão em decorrência dos erros ocorridos em diferentes magnitudes; logo, a média desses diferentes escores pode ser entendida como o escore verdadeiro; (3) o terceiro postulado afirma que não existe correlação entre o escore verdadeiro e o erro. Assim, mesmo que em diferentes aplicações os escores de um indivíduo aumente ou diminua, a quantidade de erro não apresenta relação com esse escore; (4) o quarto

postulado indica que os erros presentes nos escores de diferentes testes aplicados no mesmo sujeito não apresentam correlação entre si; (5) o quinto postulado indica a ausência de correlação entre o erro da medida e um teste e o resultado verdadeiro de outro teste do mesmo indivíduo; (6) o sexto postulado afirma que dois testes podem ser considerados paralelos apenas se a distribuição dos seus erros tem a mesma variância e os escores verdadeiros de um sujeito são iguais em ambos os testes. O conceito de teste paralelo é muito importante na TCT porque é um dos métodos para avaliar a fidedignidade; (7) o sétimo, e último postulado, afirma que um teste só é tau-equivalente quando seus resultados verdadeiros diferem por uma constante aditiva "k". De modo geral, os sete postulados da TCT admitem que os erros são aleatórios e independentes, em quaisquer circunstâncias (Sartes & Souza-Formigoni, 2013).

Como já assinalado, quando falamos de medida em psicologia, controlar o erro é algo difícil. Na prática da pesquisa o erro da medida é expresso pelo erro padrão de mensuração (EPM) que é a raiz quadrada da variância de erro (Hogan, 2006). Quanto menor for o EPM, menor será a variação dos escores em torno do escore verdadeiro. O EPM é utilizado para construir os intervalos de confiança (IC) de um escore observado. Quanto maior a fidedignidade, tanto menor será o intervalo de confiança.

Com base nesses pressupostos, a TCT apresenta-se focada no escore total do teste, obtido por meio da soma das respostas dos indivíduos a um conjunto de itens, em relação ao comportamento presente ou futuro (Pasquali, 2010, 2013). No caso de testes de desempenho (testes de inteligência, memória, raciocínio etc.) que possuem respostas certas

e erradas, o escore total é o somatório da quantidade de acertos. No caso de escalas de preferência, o escore total é o somatório dos pontos marcados na escala de respostas de todos os itens. Apesar desse enfoque no escore total, a TCT avalia dois parâmetros ou características dos itens, a dificuldade e a discriminação. O parâmetro de dificuldade está associado à quantidade de indivíduos que responderam corretamente o item, no caso de itens de desempenho, e pela proporção de respostas em um determinado ponto da escala, em itens de preferência (DeVellis, 2006; Grégoire & Laveault, 2002). O parâmetro de discriminação, por sua vez, é entendido como o quanto determinado item diferencia indivíduos com escores diferentes, o que pode ser realizado por meio de grupos de critério ou análises de correlação entre os itens. Quanto mais o item diferenciar sujeitos com magnitudes próximas, mais discriminativo será o item (DeVellis, 2006; Grégoire & Laveault, 2002). A avaliação das propriedades psicométricas dos itens também é particularmente importante em avaliações educacionais de larga escala (Andrade, Laros, & Gouveia, 2010). Prover avaliações justas com condições iguais de avaliação é uma forma de respeitar os direitos individuais. A propósito, a versão mais atual do *Standards for Educational and Psychological Testing* (Aera, APA, & NCME, 2014) – principal referência na área – possui um capítulo inteiro sobre justiça na testagem (*Fairness in testing*). Similarmente, a Resolução do Conselho Federal de Psicologia (CFP) n. 009/2018, que estabelece diretrizes para a realização de avaliação psicológica no exercício profissional do(a) psicólogo(a), possui uma seção intitulada de "Justiça e proteção dos direitos humanos na avaliação psicológica" (Andrade & Valentini, 2018; CFP, 2018). Importante destacar que tanto a TCT quanto a Teoria de Resposta ao Item (TRI) admitem que as evidências de

fidedignidade e validade são critérios fundamentais para a qualidade dos testes psicológicos.

Limitações da TCT e uso combinado com a Teoria de Resposta ao Item

Apesar da ampla utilização da TCT, a mesma apresenta algumas limitações teóricas. Exemplo disso é que, na TCT, os parâmetros psicométricos dos itens dependem estritamente da amostra de sujeitos utilizada para estabelecê-los (*group-dependent*). Em outras palavras, isto quer dizer que o teste será considerado fácil, mediano ou difícil, dependendo do desempenho do grupo de respondentes que se submeteu ao teste. Por exemplo, uma amostra de respondentes acima da média em termos de inteligência levará a acreditar que os itens de um teste de inteligência são mais fáceis do que realmente são. Ao contrário, uma amostra de respondentes abaixo da média em termos de inteligência levará a acreditar que os itens de um teste de inteligência são mais difíceis do que realmente são.

Outra crítica em relação à TCT é que os escores dos examinandos também dependem do tipo de teste utilizado (*test-dependent*). Com um teste difícil ou muito difícil, os examinandos tenderão a ter escores mais baixos. Ao mesmo tempo, caso os examinandos sejam avaliados com testes de desempenho fáceis ou muito fáceis tenderão a ter escores mais baixos. Esses problemas são solucionados quando temos amostras de respondentes representativas da população (Andrade et al., 2010; Andrade, Tavares, & Valle, 2000; Crocker & Algina, 1986; Hambleton, Swaminathan, & Rogers, 1991; Pasquali, 2007, 2013) e,

pelo menos na psicologia, esse tipo de amostragem não é facilmente obtido.

Importante destacar que, quando utilizamos a TCT, examinandos que acertam a mesma quantidade de itens, porém de propriedades psicométricas diferentes (discriminação, dificuldade, probabilidade de acerto ao acaso), apresentam o mesmo escore total ou desempenho. Imaginem, por exemplo, que o Alberto e a Cristina acertaram igualmente 7 itens de um total de 10 itens em um teste de matemática. De acordo com a TCT, ambos receberiam o escore bruto igual a 7,0. Acontece que Cristina acertou itens fáceis, medianos, difíceis e muito difíceis. Alberto, por sua, vez acertou apenas itens fáceis e medianos. Nesse caso, parece justo que Cristina recebesse um escore de desempeno mais alto. A TRI, metodologia utilizada, por exemplo, no Exame Nacional do Ensino Médio (Enem), estima tal proficiência considerando os parâmetros psicométricos dos itens.

Na literatura são comumente realizadas comparações da TCT com a TRI (Kohli, Koran, & Henn, 2015; Petrillo, Cano, McLeod, & Coon, 2015; Raykov, Dimitrov, Marcoulides, & Harrison, 2017; Raykov & Marcoulides, 2016; Sartes &
Souza-Formigoni, 2013). Spencer (2004), por exemplo, assinalou que uma vantagem da TRI em detrimento da TCT é que os valores dos parâmetros de dificuldade dos itens e as habilidades estimadas dos examinandos são colocados na mesma métrica, o que facilita a interpretação dos resultados. Além disso, itens podem ser adicionados ao banco de itens sem mudar a ordem relativa de itens já existentes ou de examinandos na escala de mensuração. A construção de bancos de itens é particularmente importante para a construção de Testagem

Adaptativa por Computador (*CAT – Computerized Adaptive Testing*). Kolen e Brennan (1995), por sua vez, assinalaram que o poder da TRI resulta da possibilidade de modelar as respostas dos examinandos no nível do item, ao invés do escore total do teste, como acontece na TCT. Ainda, Nunnally e Bernstein (1995) assinalaram que, basicamente, as vantagens do uso da TRI são: (1) diferentes pessoas ou a mesma pessoa em diferentes ocasiões podem ter suas habilidades comparadas (técnica da equalização); (2) a estimativa da habilidade de examinandos que acertaram o mesmo número de itens, porém itens diferentes, é diferenciada; e (3) os parâmetros obtidos por meio da TRI são medidas estatisticamente independentes da amostra de respondentes.

Essa última vantagem apresentada por Nunnally e Bernstein (1995) refere-se à propriedade de invariância dos parâmetros, considerada como uma das maiores distinções da TRI em relação à TCT. Essa propriedade refere-se à condição de que, quando um conjunto total de itens se adéqua satisfatoriamente a um modelo da TRI, os parâmetros desses itens são independentes da habilidade dos examinandos (Baker & Kim, 2017) e a habilidade dos examinandos pode ser estimada independente da dificuldade do teste utilizado. Ou seja, os parâmetros dos itens de discriminação (parâmetro *a*), dificuldade (parâmetro *b*) e probabilidade de acerto ao acaso (parâmetro *c*) independem do nível de habilidade dos examinandos que os responderam, e a habilidade dos examinandos independe dos itens utilizados para determiná-la (Embretson & Reise, 2000).

No estudo de Petrillo, Cano, McLeod e Coon (2015) foram realizadas análises psicométricas do *Visual Functioning Questionnaire* (VFQ-25) a partir de três modelos: TCT, TRI (a partir do modelo de

resposta gradual de Samejima) e modelo Rasch (*Rasch measurement theory*). Para isso foram utilizados dados de 240 participantes com edema macular diabético de um estudo clínico randomizado, duplo-cego e multicêntrico. Os autores concluíram que os resultados foram semelhantes entre os três métodos, com a TRI e o modelo Rasch fornecendo informações diagnósticas mais detalhadas sobre como melhorar o VFQ-25. A TCT, especificamente, identificou dois itens problemáticos que ameaçavam a validade da pontuação da escala global, conjuntos de itens redundantes e categorias de resposta distorcidas. A TRI, por sua vez, também identificou ajuste inadequado para um item, itens localmente dependentes, direcionamento inadequado e desordem em mais da metade das categorias de resposta.

Concluímos que mesmo considerando-se todas as vantagens da TRI, ressalta-se que a TCT continua sendo utilizada, sozinha ou em combinação com a TRI, a fim de oferecer informações adicionais sobre a qualidade do teste (Andrade et al., 2010; Bechger, Maris, Verstralen, & Béguin, 2003). As análises clássicas continuam sendo importantes ferramentas na validação de instrumentos, auxiliam na análise exploratória dos itens e possibilitam identificar inconsistências nos dados e itens problemáticos.

A importância da TCT para construção e obtenção de evidências de validade dos testes

Durante grande parte do século passado, a TCT foi a abordagem dominante para o desenvolvimento de instrumentos de medição na área educacional, das ciências humanas e sociais. Grande parte desse sucesso ocorreu devido a sua simplicidade metodológica que possibilitou uma

maneira muito útil e fácil de se pensar sobre os construtos psicológicos (Raykov & Marcoulides, 2016). De acordo com Kline (2005), pode-se afirmar que a TCT permitiu o desenvolvimento de escalas psicométricas bastante sólidas.

Importante destacar que na segunda metade do século passado, um interesse substancialmente maior voltou-se em direção à TRI e aos modelos de traços latentes (Raykov & Marcoulides, 2016). Por exemplo, Sartes e Souza-Formigoni (2013) afirmam que no século XX, o desenvolvimento e avaliação das propriedades psicométricas dos testes foram baseados principalmente na TCT. As autoras prosseguem afirmando que foram desenvolvidos muitos testes longos e redundantes, com medidas influenciáveis pelas características da amostra dos indivíduos avaliados durante seu desenvolvimento. Nesse contexto a TRI surgiu como uma possível solução para algumas limitações da TCT, melhorando a qualidade da avaliação da estrutura dos testes. Diante disso, seria justo dizer que a TCT está condenada ao fim? Nós, autores do presente capítulo, acreditamos que não!

No estudo anteriormente citado de calibração do VFQ-25 a partir de três modelos (TCT, TRI e modelo Rasch), Petrillo et al. (2015) concluíram que a seleção de uma abordagem psicométrica depende de muitos fatores. Segundo os autores, os pesquisadores devem justificar seu método de avaliação e considerar o público-alvo. Por exemplo, se o instrumento está sendo desenvolvido para fins descritivos e com um orçamento restrito, uma análise geral das propriedades psicométricas dos itens baseada na TCT pode ser tudo que é possível ser realizado. Problemas simples como identificação de dados omissos e efeitos de teto ou piso, por exemplo, são facilmente identificados pela TCT. Por

isso, não podemos subestimar o valor da TCT. No entanto, em uma avaliação de alto risco como o desenvolvimento de um instrumento para fins diagnósticos ou para seleção de pessoal, por exemplo, uma avaliação psicométrica completa do instrumento, incluindo análise dos itens por meio da TRI, deve ser incentivada.

A TCT também permite avaliar itens individuais de uma outra perspectiva. Isso pode ser particularmente útil para análises exploratórias. O objetivo da análise de itens é usar estatísticas detalhadas para determinar possíveis falhas no item e, em seguida, decidir se é necessário revisar, substituir ou retirar o item. A TRI é sem dúvida uma análise mais poderosa, mas só funciona com números amostrais maiores. Para conhecimento do impacto do tamanho amostral na calibração dos itens recomendamos a leitura de Nunes e Primi (2005). Dificilmente teremos um bom ajuste dos modelos com números amostrais pequenos. Isso torna extremamente importante o uso da TRI em testes de larga escala, mas completamente inadequado para amostras do tamanho de sala de aula ou outras situações de amostras pequenas ($n < 100$) (Thompson, 2016).

Considerações finais

Medir é um procedimento que diariamente realizamos, independente da nossa área de formação ou atuação, e que é fundamental para o estabelecimento de processos e intervenções mais válidos e precisos em todos os campos. Na psicologia, em específico, o teste é um instrumento de medida que funciona de forma semelhante a uma régua. Esse instrumento deve ser válido e preciso para que resultados inconsistentes não sejam emitidos. Em avaliações

psicológicas e educacionais, nas quais decisões são tomadas, resultados inválidos e imprecisos podem ser muito onerosos e sugerir caminhos desastrosos, seja para um indivíduo em particular, seja para uma rede educacional específica ou para o país como um todo (Andrade et al., 2010). A testagem psicológica faz parte de um processo mais amplo de avaliação psicológica (Andrade & Sales, 2017). Estas avaliações quando realizadas de forma inconsistente podem ser terminantemente prejudiciais. Por exemplo, detentos com uma elevada probabilidade de reincidência podem estar sendo recomendados para progressão de regime penal ou indivíduos com elevada impulsividade podem obter licença para a posse e/ou manuseio de armas de fogo. Em ambos os casos, o processo de medição e verificação dos parâmetros dessas medidas é fundamental.

Além disso, é de grande importância reforçar a premissa de que o teste psicológico é um instrumento de medida que se fundamenta em uma teoria (Pasquali, 2010, 2013); logo, se a teoria em que o instrumento está fundamentado não apresenta suporte empírico, o instrumento também não se apresentará como uma forma de medida válida.

Por último, Raykov e Marcoulides (2016) assinalam que uma meta de todos os psicometristas é, independentemente da abordagem, melhorar a metodologia de mensuração disponível, incluindo a combinação de informações qualitativas e quantitativas em avaliações de grande impacto. Nossa experiência profissional aponta que a junção de profissionais de diversas áreas é de grande importância, seja para conhecer melhor o fenômeno estudado, seja para pensar novas abordagens metodológicas e analíticas.

Referências

Allen, M.J. & Yen, W.M. (2002). *Introduction to measurement theory.* Illinois: Waveland.

American Educational Research Association – Aera, American Psychological Association – Apa, and National Council on Measurement in Education – NCME (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Andrade, D.F., Tavares, H.R., & Valle, R.C. (2000). *Teoria de resposta ao item: conceitos e aplicações.* São Paulo: ABE – Associação Brasileira de Estatística.

Andrade, J.M., Laros, J.A., & Gouveia, V.V. (2010). O uso da teoria de resposta ao item em avaliações educacionais: diretrizes para pesquisadores. *Avaliação Psicológica, 9*(3), 421-435.

Andrade, J.M. & Sales, H.F.S. (2017). A diferenciação entre avaliação psicológica e testagem psicológica: questões emergentes. In M.R.C. Lins & J.C. Borsa (Eds.). *Avaliação psicológica: Aspectos teóricos e práticos* (pp. 9-22). Petrópolis: Vozes.

Andrade, J.M. & Valentini, F. (2018). Diretrizes para a construção de testes psicológicos: a Resolução CFP n. 009/2018 em Destaque. *Psicologia: Ciência e Profissão, 38*(spe), 28-39 [https://doi.org/10.1590/1982-3703000208890].

Baker, F.B. & Kim, S. (2017). *The basics of item response theory using R.* Nova York: Springer International Publishing.

Bechger, T.M., Maris, G.F.H.H., & Béguin, A.A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*(5), 319-334 [https://doi.org/10.1177/0146621603257518].

Conselho Federal de Psicologia (2018). Resolução CFP n. 009/2018. [Recuperado de http://satepsi.cfp.org.br/docs/Resolu%C3%A7%C3%A3o-

CFP-n%C2%BA-09-2018-com-anexo.pdf].

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Nova York: Holt, Rinehart and Winston.

DeVellis, R.F. (2006). Classical test theory. *Medical Care*, *44*(11), 50-59.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists.* Nova Jersey: Lawrence Erlbaum Associates.

Grégoire, J. & Laveault, D. (2002). *Introdução às teorias dos testes em ciências humanas.* Porto: Porto Ed.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Califórnia: Sage.

Hauck-Filho, N. (2014). Medida psicológica: o debate entre as perspectivas conceituais representacionista e realista. *Avaliação Psicológica*, *13*(3), 399-408.

Hogan, T.P. (2006). *Introdução à prática de testes psicológicos.* Rio de Janeiro: LTC – Livros Técnicos e Científicos.

Kline, T.J.B. (2005). Classical test theory: Assumptions, equations, limitations, and item analyses. In J.T.B. Kline. *Psychological testing: A practical approach to design and evaluation* (pp. 91-106). Thousand Oaks: Sage.

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement*, *75*(3), 389-405.

Kolen, M.J. & Brennan, R.L. (1995). *Test equating: Methods and practices.* Nova York: Springer.

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

Mathison, S. (2005). *Encyclopedia of evaluation.* Thousand Oaks: Sage.

Michell, J. (2002). Stevens's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology*, *54*, 99-104 [doi: 10.1080/00049530210001706563].

Nunes, C.H.S.S. & Primi, R. (2005). Impacto do tamanho da amostra na calibração de itens e estimativa de escores por teoria de resposta ao item. *Avaliação Psicológica, 4*(2), 141-153.

Nunnally, J.C. & Bernstein, I.H. (1995). *Psychometric theory* (3a. ed.). Nova York: McGraw-Hill.

Pasquali, L. (2007). Validade dos testes psicológicos: será possível reencontrar o caminho? *Psicologia: Teoria e Pesquisa, 23*, 99-107.

Pasquali, L. (2010). *Instrumentação psicológica: fundamentos e práticas.* Porto Alegre: Artmed.

Pasquali, L. (2013). *Psicometria: teoria dos testes na psicologia e na educação* (5a. ed.). Petrópolis: Vozes.

Petrillo, J., Cano, S.J., McLeod, L.D., & Coon, C.D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health, 18*(1), 25-34 [http://dx.doi.org/10.1016/j.jval.2014.10.005].

Raykov, T., Dimitrov, D.M., Marcoulides, G.A., &
Harrison, M. (2017). On true score evaluation using item response theory modeling. *Educational and Psychological Measurement,* 1-12 [https://doi.org/10.1177/0013164417741711].

Raykov, T. & Marcoulides, G.A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement, 76*(2), 325-338.

Rossi, G.B. (2007). Measurability. *Measurement, 40*, 545-562.

Sartes, L.M.A. & Souza-Formigoni, M.L.O. (2013). Avanços na psicometria: da teoria clássica dos testes à teoria de resposta ao item. *Psicologia: Reflexão e Crítica, 26*(2), 241-250.

Spencer, S.G. (2004). *The strengrh of multidimensional item response theory in exploring construct space that is multidimensional and correlated* [Doctoral

dissertation, Brigham Young University – Recuperado de https://scholarsarchive.byu.edu/cgi/ viewcontent.cgi?article=1223&context=etd].

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, *103*(2.684), 677-680.

Thompson, N.A. (2016). *Introduction to classical test theory with Citas.* Minnetonka, MN: Assessment Systems Corporation.

Thorndike, E.L. (1911). *Animal intelligence.* Nova York: Macmillan.

Urbina, S. (2007). *Fundamentos da testagem psicológica.* Porto Alegre: Artmed.