8

Psicometria e avaliação por testes: um marco metodológico*

Jorge Luis Bazán

Introdução

No presente capítulo, discutiremos o desenvolvimento de instrumentos (testes, escalas, questionários ou provas) e a melhor definição de propósitos de pesquisa (objetivos e resultados) com melhores modelos matemáticos e estatísticos e melhores sistemas de cômputo para bases de dados, análises e aplicações de provas, bem como com o desenvolvimento de critérios para propor políticas usando os seus resultados. Isso induz ao desenvolvimento metodológico da avaliação e a um debate a respeito de sua aplicabilidade, especialmente quando a comunidade interessada em avaliação desconhece os seus aspectos técnicos.

Dentre os diferentes objetos da avaliação está a avaliação educativa, reconhecida como um componente importante da qualidade educativa (APARICIO; BAZÁN ABDOUNOR, 2013). Artiles, Mendoza e Yera (2008) também sinalizam que a avaliação é importante para as instituições de ensino. Nesse sentido, o processo de avaliação educacional está relacionado à produção de informações sobre o aprendizado. Este processo é algo que está bastante presente no cotidiano escolar e na educação superior: usualmente, os professores aferem o aprendizado de seus alunos por meio de diversos instrumentos (observações, questionários, escalas, listas, registros, provas etc.) e indicam, a partir daí, o que precisa ser feito para que seus alunos possam avançar no sistema escolar.

* Agradeço a José Carlos Rothen pela leitura atenta e pelas sugestões dadas a este texto.

Nas últimas décadas, junto às avaliações tradicionais nas salas de aula, um outro tipo de avaliação educacional tem ganhado espaço: são as avaliações externas, geralmente em larga escala – isto é, instrumentos são aplicados simultaneamente a grandes amostras ou censos, de forma padronizada, incluindo, às vezes, alunos, professores, diretores e coordenadores. Exemplos destas avaliações são o Exame Nacional do Ensino Médio (Enem), o Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (Saresp) e o *Programme for International Student Assessment* (do inglês, "Programa Internacional de Avaliação de Alunos", o Pisa). Estes exemplos de avaliações têm objetivos e procedimentos diferenciados das avaliações tradicionais, vindas das salas de aula. Elas podem, por exemplo, ser instrumentos para certificação, credenciamento de instituições, diagnóstico, bem como para prestação de contas.

Em geral, as avaliações educacionais privilegiam a avaliação do sistema cognitivo-produtivo, deixando de lado outros sistemas da personalidade dos estudantes – como o sistema afetivo emotivo e o sistema conativo volitivo, na classificação proposta por Ortiz (1994) e discutida por Bazán e Aparicio (2006). Os aspectos técnicos da avaliação educacional se apoiam fortemente na *Psicometria*, campo de estudo relacionado com a teoria e técnica da medição psicológica, incluindo a medição de conhecimentos, habilidades, atitudes, traços de personalidade e a medição educacional.* A Psicometria está associada principalmente com a construção e validação de instrumentos de medição, tais como questionários, provas, escalas, inventários, testes, entre outros. Ela possui duas tarefas de pesquisa principais: (i) a construção de instrumentos e procedimentos de medição; e (ii) o desenvolvimento e aperfeiçoamento de abordagens teóricas e práticas para a medição.

Podemos dizer que psicometristas são cientistas envolvidos no planejamento de testes para tentar medir diferentes características humanas, sendo que a área sofreu um rápido crescimento e especialização desde a sua criação – o que requer uma formação interdisciplinar tanto de aspectos quantitativos (principalmente formação estatística) quanto de aspectos qualitativos (principalmente Psicologia e Educação). Os testes psicométricos são utilizados em escolas, organizações, empresas, governos, forças armadas e em ambientes hospitalares e clínicos. Cada vez são mais requeridos testes dessa natureza, e não há especialistas suficientes em Psicometria para atender a esta demanda.

---

* Cf., por exemplo, Pasquali (2004).

O objetivo do presente capítulo é introduzir um marco metodológico para avaliação educacional, o qual pode ser usado para a elaboração e revisão de testes. Adicionalmente, são destaques os dois principais modelos de medição: a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI). Assim, ele está organizado da seguinte maneira: primeiramente, são introduzidas algumas noções a respeito das medidas psicométricas. Em seguida, apresenta-se o marco metodológico da avaliação assumido neste trabalho, com destaque para o modelo de medição como parte do chamado modelamento do construto para a elaboração e revisão de testes. Na seção seguinte são apresentados, brevemente, os dois principais modelos de medição: Teoria Clássica dos Testes (TCT) e Teoria de Resposta ao Item (TRI). Algumas reflexões são apresentadas ao final do capítulo.

Noções a respeito das medidas psicométricas

Como é a construção de um instrumento psicométrico?

Imagine, por um momento, que nós não tivéssemos inventado uma forma de medir a estatura de uma pessoa, isto é, não temos uma fita métrica para medir quão alta ela é. Seria possível, neste caso, medir a sua altura? A resposta é: sim, e nós poderíamos usar um instrumento psicométrico. Ou seja, podemos usar um conjunto de questões associadas com atividades vinculadas à altura de uma pessoa (por exemplo: "Eu literalmente olho para meus colegas de cima para baixo") e, logo, usar um formato de resposta que pode ser, por exemplo, dicotômica ("Sim" ou "Não") ou politômica ("Nunca", "Raramente", "A metade do tempo", "Muitas vezes", "Sempre"). Em seguida, sumarizamos os resultados numa medida geral, por exemplo, usando uma soma simples das respostas, de modo que altos valores expressem maior estatura e valores baixos expressem uma estatura menor. As correspondentes medidas dos itens podem, finalmente, ser levadas a uma escala que corresponde à medida física da estatura de uma pessoa.

É claro que esta não será uma medida precisa, pois temos um instrumento de medida *físico* para a altura. Mas, por outro lado, quando aquilo que nos interessa medir não possui um instrumento de medida físico – por exemplo, o desempenho de um estudante em Matemática –, construir um instrumento psicométrico é uma alternativa viável.

O que mede um teste?

Um teste (ou medida) pode ser visto com um conjunto de questões ou perguntas de autorrelato (também chamado de "itens"), cujas respostas são pontuadas e, de alguma forma, agregadas para se obter um escore total. As características essenciais são: a) uma série de perguntas às quais os indivíduos respondem; e b) um escore composto que surge a partir da pontuação das respostas para as perguntas.

O conjunto resultante de perguntas é referido como uma "escala", "teste" ou "medida". O importante não é tanto o formato da pergunta, mas sim o formato da resposta ou da pontuação, pois é a partir da pontuação que se obtém um escore que representa o que está sendo avaliado.

Em geral, um instrumento psicométrico tem dois tipos de resultados disponíveis:

a) Pontuações *binárias* ou *dicotômicas*: por exemplo, "sim" ou "não", itens que estão qualificados como resposta "correta" ou "incorreta" em testes de rendimento ou itens que são classificados dicotomicamente de acordo com um tipo de pontuação em escalas de personalidade, como "verdadeiro" ou "falso" e "de acordo com" ou "em desacordo com".

b) Pontuações *ordinais* ou *politômicas*: por exemplo, uma escala de cinco pontos, que vai desde "em total acordo" até "em total desacordo" (comumente conhecida como Escala Likert, geralmente usada para medir atitudes), ou ainda uma escala de três pontos ("pouco", "medianamente" ou "muito") para quando é requerido avaliar, por exemplo, a frequência de uma determinada característica da personalidade.

Como é determinada a qualidade dos instrumentos psicométricos?

Associações profissionais e usuários possuem, muitas vezes, dentro de contextos mais amplos, preocupações a respeito do desenvolvimento de critérios para avaliar a qualidade de qualquer teste em determinado contexto. Um destes critérios são as Normas para Testagem Educacional e Psicológica (AERA; APA; NCME, 1999), conjunto de critérios de avaliação desenvolvidos pela *American Educational Research Association* (AERA), pela *American Psychological Association* (APA) e pelo *National Council on Measurement in Education* (NCME). No quadro a seguir são mostrados os tópicos cobertos

nesta publicação e que fazem parte das preocupações pela melhora da qualidade dos instrumentos.

Quadro 1 Tópicos para avaliar a qualidade dos instrumentos psicométricos.

Construção de testes, avaliação e documentação	Erros de medida e confiabilidade
	Desenvolvimento de teste e revisão
	Escalas, normas e comparabilidade dos escores
	Administração de teste, qualificação e relatórios
	Documentação de apoio para os testes
Equivalência dos testes	Teste de equidade e uso do teste
	Os direitos e as responsabilidades dos examinadores
	Testes individuais de pessoas de diversas procedências linguísticas
	Testes individuais para pessoas com deficiência
Aplicações dos testes	As responsabilidades de usuários de teste
	Avaliação e medição psicológica
	Avaliação e medição educacional
	Avaliação e certificação do trabalho
	Teste de avaliação de programas e políticas públicas

Fonte: elaboração própria, com base em AERA, APA e NCME (1999).

As considerações de validade e confiabilidade dos instrumentos psicométricos, pelo geral, são vistas como elementos essenciais para determinar a qualidade de qualquer teste. A *confiabilidade* (ou *fidedignidade*) é uma característica da medida que faz referência ao grau de consistência ou reprodutibilidade das medidas quando os procedimentos das avaliações são replicados sob as mesmas condições. Por outro lado, a *validade* da medida faz referência ao grau pelo qual a evidência e a teoria suportam interpretações a partir dos valores das medidas.

Marco metodológico da avaliação

Para medir o quanto de habilidade uma pessoa tem, é preciso ter uma escala de medição, ou seja, uma regra com uma métrica. Esta regra deve ser utilizada para quantificar qual capacidade uma determinada pessoa possui. A prática habitual é definir uma medida da capacidade e desenvolver um teste, que consiste num determinado número de itens sob a definição de perguntas. Por exemplo: definimos o "desempenho em leitura de textos escritos" como "habilidade" para responder perguntas associadas à leitura

de diferentes tipos de texto e, então, baseando-se em quatro leituras de quatro diferentes tipos de textos, são apresentados quatro itens para cada texto indagando a respeito do contexto de cada leitura – os personagens envolvidos, a localização temporal do texto, sua ideia principal –, sendo que cada um desses itens mede alguma faceta de uma particular habilidade de interesse. Assume-se que cada examinando que responde a um item de um teste possui certa quantidade da capacidade subjacente. Assim, podemos considerar que cada examinando possui um valor numérico que toma o lugar da sua posição na escala de habilidade.

Para elaborar testes que meçam uma determinada característica de interesse podemos considerar, como marco metodológico, uma versão adaptada da proposta de Duckor, Draney e Wilson (2009) e também discutida em Wilson (2005).

Estes autores apresentam uma proposta para a construção de medidas com base em quatro etapas e princípios do sistema de avaliação, os quais são apresentados na figura a seguir.

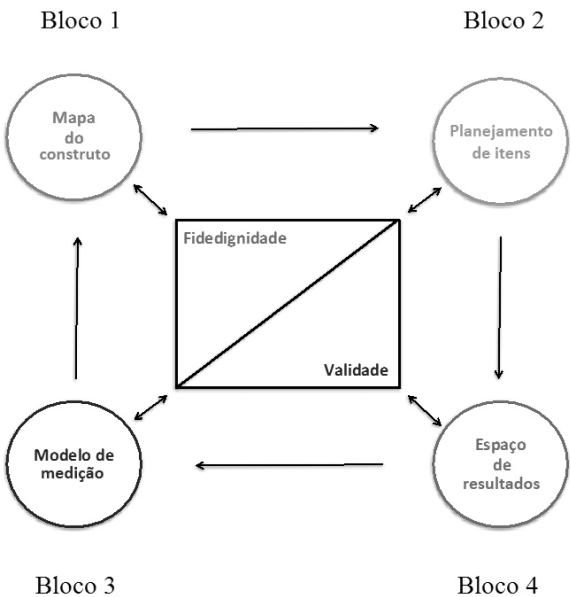

Figura 1 Relações entre os quatro blocos para a construção de medidas.

Fonte: extraída e traduzida de Duckor, Draney e Wilson (2009).

O processo da construção de uma *medida* ou *modelamento do construto* é uma estratégia para o desenvolvimento de um instrumento usando cada

um dos quatro tijolos da construção: inicia-se com 1) a definição do mapa do construto, segue-se com 2) o planejamento de itens e com 3) a definição do espaço de resultados, e finaliza-se com 4) o modelo de medição a ser considerado. Neste processo, as etapas 1 e 3 envolvem a fidedignidade ou confiabilidade da medida, enquanto que as etapas 2 e 4 envolvem a validade desta. Isto é, como sinalizado na Figura 1, as etapas 1 a 3 fazem referência ao grau de consistência ou reprodutibilidade das medidas quando os procedimentos das avaliações são replicados sob as mesmas condições, enquanto que as etapas 2 a 4 se concentram em determinar o grau pelo qual a evidência e a teoria suportam interpretações a partir dos valores das medidas.

Na prática, esse processo não é necessariamente explícito, isto é, os elaboradores ou construtores de medidas não seguem necessariamente esse processo no nível de detalhe discutido na proposta dos autores. Entretanto, quando se deve analisar uma medida, é requerido avaliar cada uma dessas etapas. Assim, a análise de toda a medida enfatiza tópicos que envolvem diferentes objetivos (como revisão, descrição, crítica ou proposta). Neste texto, propomos a classificação destes diferentes objetivos, dependendo do foco em que eles se centram. Por exemplo, alguns trabalhos podem enfatizar a revisão, a descrição, a crítica ou a proposta da definição do mapa de construto, mas outros podem ser melhor classificados centrando seus objetivos na análise do modelo de medição adotado.

Note-se que, em nossa proposta, trocamos a ordem das etapas 3 e 4 (ver a figura a seguir). Ou seja, uma vez que as medidas já estão definidas, primeiro revisamos o modelo de medição e, em seguida, o espaço de resultados adotado como mostrado na Figura 2.

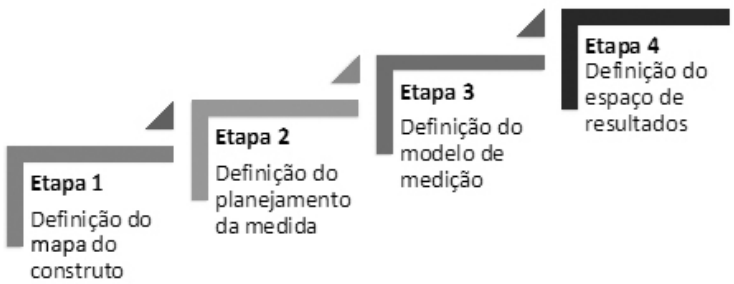

Figura 2 Etapas para a avaliação de construção de medidas.

Fonte: adaptada de Duckor, Daney e Wilson (2009).

A sequência de etapas apresentada na Figura 2 fica semelhante a um planejamento estatístico, como mostramos a seguir no Quadro 2. Ou seja, uma vez que as medidas já estão definidas, primeiro revisamos o modelo de medição e em seguida o espaço de resultados adotado. Todas as etapas respondem a uma pergunta dada no processo de construção de medidas.

Quadro 2 Etapas de avaliação do processo de construção de medidas como etapas do planejamento estatístico.

Etapa	Nome da etapa	Definição	Pergunta	Planejamento estatístico
I	Definição do mapa do construto	Definição daquilo que está sendo medido	O que vai ser medido?	Definição da pesquisa ou interação entre o(a) pesquisador(a) e o(a) analista de dados
II	Definição do planejamento da medida	Definição do formato de avaliação ou instrumento e das unidades de observação ou fontes de informação (alunos, diretores etc.), do processo, amostragem ou instrumentos e das bases de dados	Como vai ser medido?	Definição dos instrumentos, amostragem, processo de captura de dados e elaboração de base de dados
III	Modelo de medição	Definição do modelo de medição (modelo estatístico) que é aplicado na Etapa II	Como vão ser analisadas as medidas?	Definição do modelo estatístico ou técnica de análise de dados a ser adotada
IV	Definição da apresentação do espaço de resultados	Definição da forma de apresentação dos resultados finais e sua interpretação e uso (que é aplicado ao processo na Etapa III)	Como serão comunicados os resultados da medida?	Definição do modelo de reporte de resultados

Fonte: elaboração própria.

Definição do mapa do construto

No proceso de construçao de medidas, o mais importante é estabelecer a finalidade para a qual é necessário um instrumento e um o contexto em que ele será utilizado, ou seja, a medição do conceito envolvendo algum tipo de decisão. Aquela ideia (ou conceito, que é o objeto teórico de nosso interesse e que precisamos medir) é conhecida comumente como *construto*.

O construto pode ser parte de um modelo teórico de uma cognição pessoal – como o desempenho em Matemática ou a habilidade em resolver problemas, ou ainda a capacidade de compreender textos escritos – ou pode ser alguma outra variável psicológica ou atitude, uma característica de determinado domínio emocional etc. Ele pode, ainda, estar relacionado com um grupo em vez de estar relacionado com um indivíduo, ou também ser um objeto inanimado complexo. Há uma infinidade de teorias: o importante, aqui, é ter uma estrutura para proporcionar motivação e uma estrutura para o construto a ser medido. A ideia de *construir um mapa* é, portanto, um conceito mais preciso do que falar em *construto*.

Supõe-se que o construto a ser medido tem uma forma particularmente simples, estendida de um extremo a outro, de um alto para um baixo valor, a partir de um pequeno a grande escore, do positivo para o negativo, ou do forte para o fraco. Há alguma complexidade entre os valores extremos, mas estamos interessados sobretudo na localização de um entrevistado entre um extremo e outro. Em particular, podem ser definidos níveis qualitativos entre os extremos – estes níveis são importantes e úteis para a interpretação. Este ponto ainda é uma ideia latente antes de se tornar algo claro. Embora os níveis qualitativos sejam definíveis, presume-se que os entrevistados podem estar em qualquer lugar no "continuum" do construto subjacente.

Em resumo, um mapa do construto pode ser considerado algo que será medido e, desde um ponto de vista mais estatístico, é definido como uma variável. No entanto, dado que esta variável não é observável diretamente se não através do uso de um teste, ela é denominada *variável latente* e pode ser considerada unidimensional, no sentido de que é única. Uma discussão mais aprofundada a respeito da unidimensionalidade e dos modos empíricos de como avaliá-la pode ser revisada em Burga (2005) e Abedi (1997).

Muitos outros construtos são mais complexos do que isso. Por exemplo, um construto pode ser multidimensional – mas isso não configura uma barreira para a modelagem que fazemos, porque cada uma dessas dimensões pode ser considerada unidimensional e, portanto, podemos ter um mapa de construto para cada uma delas. Estas dimensões podem ser organizadas mediante uma matriz de referência.

O que é uma matriz de referência?

É o consenso do *que* e do *quanto* deve conhecer o examinado ou aquilo que envolve o construto a ser medido. Este consenso é representado numa tabela de especificações ou matriz de referência que, geralmente, é de dupla

entrada. De modo geral temos, nas linhas, conteúdos, e nas colunas, níveis cognitivos, para estabelecer os pesos destes.

Quadro 3 Exemplo de matriz de referência para uma prova educacional cognitiva.

	Níveis cognitivos			Total
	Nível 1	Nível 2	Nível 3	
Conteúdo 1				
Conteúdo 2				
Conteúdo 3				
Conteúdo 4				
Conteúdo 5				
Total				

Fonte: elaboração própria.

Quando um mapa do construto é postulado pela primeira vez, muitas vezes ele é menos desenvolvido do que será, de fato, no final. A melhora do mapa é obtida por meio de vários processos, à medida que o instrumento é desenvolvido. Esses processos incluem: a) explicar o construto a outras pessoas usando o mapa do construto; b) criar itens que ajudem o entrevistado a responder segundo os níveis do mapa de construto definido pelo(a) pesquisador(a); c) testar esses itens com uma amostra de respondentes; e d) analisar os dados resultantes para verificar se os resultados são consistentes, coerentes com as suas intenções (expressas por meio do mapa do construto).

Elaboração de itens ou planejamento da medida

Em seguida, ao ter o mapa do construto, o medidor deve pensar em alguma maneira para que este construto teórico possa se manifestar em uma situação do mundo real. No início, não será mais do que um palpite, um contexto no qual se acredita que o construto deva estar envolvido de fato – aquele contexto em que o construto deverá desempenhar um papel decisivo na situação. Este palpite se tornará ainda mais cristalizado e se transformará em certos padrões na medida em que o processo de construção de medidas se consolide.

A relação apresentada, em que primeiro é definido o construto e em seguida são escritos os itens, nem sempre ocorre necessariamente desta forma. Muitas vezes, os itens podem ser pensados primeiro e o construto pode ser elucidado mais tarde. Um item também pode assumir muitas formas ou formatos de resposta, tais como múltipla escolha (tipos de itens de escolha

forçada entre várias alternativas, com uma resposta considerada correta) e Escalas Likert, ordenadas de modo que nenhuma das alternativas é correta. Há muitas variações neste sentido: o entrevistado também pode produzir uma resposta livre, tal como um teste, uma entrevista ou uma representação de desempenho (redação, experimento científico, desenho etc.), que pode ter escores atribuídos posteriormente, dependendo até que nível do mapa do construto o entrevistado consegue responder.

Os itens variam em *conteúdo* e *modo*: perguntas de entrevista normalmente apresentam uma ampla gama de aspectos para um tópico; questões ou tarefas de um desempenho cognitivo podem ser apresentadas dependendo das respostas dadas aos itens iniciais; questões em uma pesquisa podem usar diferentes conjuntos de opções e algumas respostas podem ser de caráter forçado e/ou livre.

Em geral, o medidor ou o responsável pela elaboração do teste assume que o examinado ou entrevistado "possui" uma certa quantidade de construto. Por exemplo, em um teste de Matemática para alunos de 4ª série, entende-se que o aluno tem a capacidade "de diferenciar e usar operações matemáticas para resolver problemas", e que este valor, no construto do aluno, é a causa e a explicação para que ele consiga (ou não) dar respostas aos itens do instrumento utilizado para medir esta capacidade – isto é, a capacidade (ou construto) mencionada anteriormente é a causa para que ele acerte ou erre determinados itens do teste proposto pelo examinador. No entanto, este agente causal é latente no sentido de que não é diretamente observável, mas pode ser medido indiretamente usando-se o teste, ou seja, o medidor não pode observar diretamente o construto. Em vez disso, são observadas as respostas aos itens considerados no teste e, então, infere-se o construto subjacente a estas observações. Note-se que a ideia de causalidade é uma suposição, e a análise não fornece evidência dessa causalidade. Na verdade, esta relação pode ser mais complexa, como verificado em Borsboom, Mellenbergh e Van Heerden (2003).

Modelos de medição

Destaca-se, no presente capítulo, o modelo de medição utilizado para relacionar as variáveis observadas, registradas e medidas (respostas aos itens) com as variáveis latentes ou não observadas (habilidade) e com o construto a ser medido. Existem dois principais modelos de medição: o modelo clássico dos testes e o modelo de resposta ao item. Embora estes modelos não sejam os únicos, eles são os mais consolidados.

Modelo de testes clássicos ou Teoria Clássica dos Testes

O modelo clássico dos testes, chamado também de Teoria Clássica dos Testes (TCT), é um enfoque da Psicometria que prediz as respostas dos testes, tais como a dificuldade dos itens ou a habilidade dos respondentes, sendo o principal propósito a compreensão e melhoria da confiabilidade dos testes. Ela também é considerada como sinônimo para a Teoria do Escore Verdadeiro, formulada por Spearman em 1904 e, posteriormente, sistematizada em Novick (1966), descrita no clássico livro de Lord e Novick (1968).

A TCT se baseia em três ideias principais: a do reconhecimento da presença de erros de medida, a de que o erro é uma variável aleatória e, por último, a concepção de que, através de uma determinada medida de correlação (isto é, através de uma medida da associação entre o valor verdadeiro e o valor observado), é possível estimar a pontuação verdadeira. Especificamente, postula-se o seguinte:

Escore observado = pontuação verdadeira + erro

A pontuação expressa uma relação linear entre o verdadeiro valor de habilidade – por exemplo, o verdadeiro desempenho em Matemática – e o escore de habilidade observado ou o desempenho no teste em questão. O resultado do teste (ou escore de linha) é a soma das pontuações recebidas sobre os itens do teste. Tradicionalmente, a teoria da medição foi estabelecida baseada numa análise de escala ou de nível do teste, fundado, por sua vez, em métodos de correlação. Os resultados ou escores de linha são, claro, não segmentados, ou seja, não se tem ideia de como uma pessoa, com determinado escore no teste, apresenta uma determinada habilidade para responder a um particular conjunto de itens; há uma única e simples medida geral do desempenho, e podemos dizer que esta pessoa é mais hábil do que outra, mas não necessariamente podemos estabelecer em que consiste esta habilidade. Outra característica é que os itens contribuem com igual importância no escore.

A principal ferramenta estatística é a chamada metodologia ANOVA dos efeitos aleatórios, chamada também de análise de componentes de variância ou variabilidade associada – técnicas que podem ser revisadas, por exemplo, em De Gruijter e Van Der Kamp (2008), e cujo principal objetivo é medir a quantidade de erro na medida. No entanto, identificando-se diferentes fontes de explicação, na prática, há o envolvimento de um conjunto de índices que fazem parte da chamada *análise de itens*, tais como: a) proporção de acerto;

b) porcentagem de omissão; c) discriminação do item; d) correlação pergunta-prova; e) Alfa de Cronbach, se o item é desconsiderado; f) média e variância do item; entre outros índices. Uma medida geral de consistência interna, como estimado pela confiabilidade da prova, baseando-se no Alfa de Cronbach,[1] é vista como uma medida apropriada neste contexto. Há uma extensa bibliografia sobre TCT[2] em que podem ser consultadas estas definições. Também há diversos softwares (livres e pagos) que implementam estas análises.[3]

Esta teoria é válida para qualquer formato de pontuação dos itens, e pode ser aplicada tanto para itens dicotômicos quanto para itens politômicos (ou qualquer subtipo destes itens). Adicionalmente, sobre a base desta teoria, há um conjunto de técnicas que fazem parte da chamada análise de itens.

A maior parte dos vestibulares de universidades privadas e de algumas universidades estaduais, bem como diferentes concursos em nível federal, estadual e municipal, faz uso da TCT.

Modelos de resposta ao item

Os modelos de resposta ao item, entendidos também por Teoria de Resposta ao Item (TRI) na Psicometria, apresentam como interesse primário saber se o examinando assinala um determinado item (correto ou não), em vez de saber a sua pontuação total. A TRI especifica como a variável latente de uma pessoa – chamada de traço latente nesta aproximação – e/ou as características do item – chamadas de parâmetros do item nesta aproximação – estão relacionadas com as respostas dadas aos itens, através da especificação de um determinado modelo probabilístico para estas respostas. O modelo mais simples é o chamado Modelo Rasch, que se aplica para o caso de as respostas aos itens serem binárias. Neste caso, a probabilidade de uma resposta "correta" para um item é modelada como função do traço

1 O Alfa de Cronbach é a medida mais usada para avaliar a confiabilidade de um teste, sob abordagem da TCT. O coeficiente tem recebido críticas e propostas de melhorias. Cf., por exemplo: Sijtsma (2009); Zumbo, Gadermann e Zeisser (2007); e Gadermann, Guhn e Zumbo (2012).

2 Algumas publicações relevantes na área são as de Kline (1986), Nunnally (1987) e Magnusson (1990).

3 Entre os softwares psicométricos especializados pagos que implementam as análises da TCT, podemos citar os seguintes: IRTPRO, Winstep, ITEMAN, Bilog, Conquest, Quest, Winmira, RUMM2020, Logimo, MSP, LPCM-WIN, RSP, T-Rasch, ICL-WIN, LEM, Multilog e Xcalibret. Por outro lado, softwares estatísticos pagos (tais como SPSS, SAS, STATISTIC, Stata, Systat, OpenStat) apresentam um módulo para estas análises. Há também o software R (livre), que apresenta os seguintes pacotes para estimação de diferentes análises da TCT: CTT, Psychometric, Cocron, CMC, Psy, Psych, ICC e ltm. Finalmente, algumas macros para Excel podem obter alguns índices da análise de itens.

latente do examinando, do que se pretende medir e dos parâmetros do item ao qual ele responde. Imagine, por exemplo, uma prova de Matemática num determinado concurso público, em que há um conjunto de questões sendo respondidas por um grupo de candidatos. Cada pessoa apresenta dois possíveis cenários para cada questão: responder corretamente ou ter um acerto em um item específico ($Y_{ij} = 1$), e responder erroneamente ou ter um fracasso no item ($Y_{ij} = 0$), sendo que a probabilidade p_{ij} de acertar o item depende da habilidade da pessoa, que denotamos por θ_i, e da dificuldade do item, que denotamos por b_j, o qual pode ser visualizado na figura a seguir.

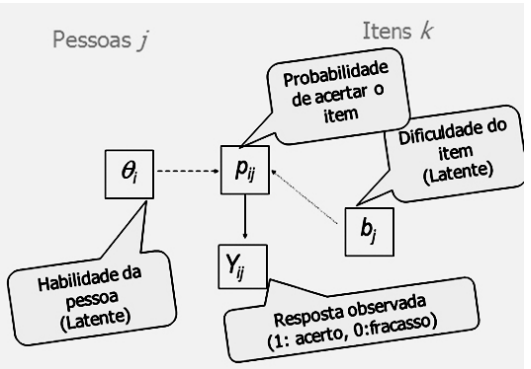

Figura 3 O Modelo de Rasch, as variáveis observáveis e os parâmetros latentes.

Fonte: elaboração própria.

Neste modelo, em termos de posição, os examinados e os itens se encontram na mesma escala, o que facilita a interpretação. Por exemplo: se o desempenho em Matemática é maior do que a dificuldade do item examinado, ele está "acima" do item, ou seja, o examinado possui uma habilidade que lhe permite acertar este item; por outro lado, se o desempenho está próximo da dificuldade do item, considera-se que o examinado possui uma habilidade "próxima" ou semelhante ao item em questão; e, finalmente, se o desempenho em Matemática é menor do que a dificuldade do item, considera-se que o examinado possui uma habilidade que está "abaixo" do item. Assim, diferente do caso da TCT, na TRI interpretamos adequadamente os níveis de habilidade, como se o nível de habilidade fosse o que permite responder a determinados itens (aqueles com dificuldade menor do que essa habilidade) e não responder outros (aqueles com dificuldade maior do que essa habilidade).

Portanto, baseando-se nas respostas dos itens de um teste, o propósito da TRI é estimar os parâmetros dos itens (com o chamado "processo de calibração"), assim como estimar os traços latentes dos examinados (com o chamado "processo de *escoring*"), incluindo alguns parâmetros da população ("distribuição dos traços latentes"): média, desvio padrão etc. Por exemplo: o processo de calibração no Enem se refere ao processo de determinação dos parâmetros do item, que acontece num estudo piloto, geralmente um ano antes da aplicação do Exame. Por outro lado, o processo de *escoring* no Enem se refere ao processo de determinação das habilidades de cada estudante que se submete à prova. Devemos notar que ambos os processos – de calibração e *escoring* – podem ser obtidos separadamente, como se faz no Enem, ou ao mesmo tempo, como em outras aplicações.

A figura a seguir mostra as chamadas *curvas características dos itens*, isto é, a curva das probabilidades para um valor fixo de dificuldade, considerando variações nos valores dos traços latentes. Neste caso, são apresentados três itens com diferentes níveis de dificuldade. A curva à esquerda corresponde a um item fácil, e a curva à direita equivale a um item difícil. Note que, quando fixado um valor do traço latente, obtemos uma probabilidade maior ou menor, respectivamente, o que é interpretado como a dificuldade do item.

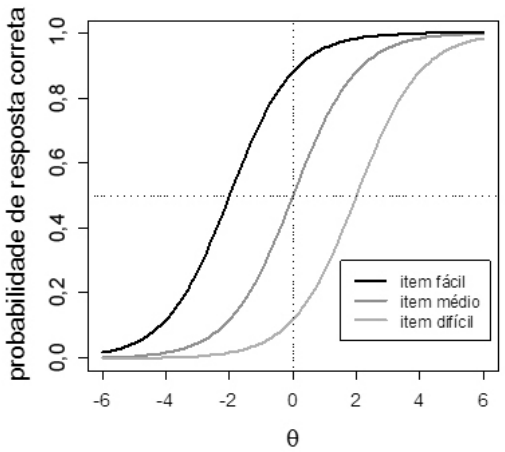

Figura 4 Curvas características de três itens com diferentes níveis de dificuldade: fácil (b = -2), médio (b = 0) e difícil (b = 2).

Fonte: elaboração própria.

Embora as bases teóricas da TRI tenham surgido entre 1950 e 1960, os métodos não foram amplamente utilizados até os anos 1970 devido à complexidade dos cálculos – solucionada com a chegada/uso dos computadores. Assim como para a TCT, há uma extensa (e cada vez maior) bibliografia sobre a TRI,[4] bem como diferentes softwares (livres e pagos).[5]

No quadro a seguir, apresentamos um resumo comparativo de ambos os modelos de medição.

Quadro 4 Comparação entre os modelos de medição.

Modelo de testes clássicos	Modelo de resposta ao item
O modelo é expresso ao nível de teste	O modelo é expresso ao nível do item
As características do item são dependentes da amostra	As características do item são independentes da amostra (invariância de item)
Estimativas da habilidade dependem dos itens	Estimativas da habilidade são independentes dos itens (invariância de pessoas)
O erro de medição é o mesmo para todos os examinados	O erro de medição é para cada nível de habilidade
Testes mais longos são mais confiáveis do que testes mais curtos	Pequenos testes podem ser mais confiáveis do que testes longos

Fonte: elaboração própria.

Na prática, a TCT ainda é usada. Contudo, a TRI é cada vez mais adotada nas avaliações, dado que esta Teoria oferece não só tratamento para novos problemas (como a multidimensionalidade das medidas), mas também possíveis vieses das medidas usando o conceito de diferenciabilidade, diferentes testes – testes de adaptação, de velocidade (*speedness*), *testlet*, testes longitudinais etc. –, uma solução para o problema de tornar diferentes medidas comparáveis (usando processos de equalização ou equiparação) e o uso de testes adaptativos computorizados, entre outros tópicos, associados com a elaboração das medidas.

4 Algumas publicações relevantes na área são as de Baker e Kim (2004), Bond e Fox (2001), De Boeck e Wilson (2004), Embretson e Reise (2000), Fox (2010), Hambleton, Swaminathan e Rogers (1991), Lord (1980) e Van der Linden e Hambleton (1997).

5 Dentre os softwares psicométricos especializados pagos que implementam alguns dos modelos da TRI, destacamos os seguintes: IRTPRO, Winstep, Rascal, Bilog, Conquest, Quest, Winmira, RUMM2020, Logimo, MSP, LPCM-WIN, RSP, T-Rasch, ICL-WIN, LEM, Multilog e Xcalibret. Por outro lado, softwares estatísticos (como SAS, Stata, WinBUGS, Systat e OpenStat) já incluem diferentes métodos para o processo de calibração. O software R também apresenta os seguintes pacotes para estimação de diferentes modelos da TRI: eRM, ltm, TAM, mirt, IRTShiny, mcIRT, irt, pcIRT, kcirt, MultiCIRT, mRm, Psychomix, mixRasch, PP, plRasch e mokken fwdmsa.

Reflexões finais

Apresentamos, neste capítulo, algumas noções da Psicometria que se baseiam nos objetivos da avaliação educacional, especialmente da avaliação em larga escala. A principal contribuição deste trabalho é a de apresentar um marco metodológico da avaliação, que pode ser usado tanto para o processo de elaboração de testes quanto para o processo de revisão de testes já disponíveis. Os principais modelos de medição que fazem parte deste marco foram apresentados: por um lado, o tradicional modelo de testes clássicos, conhecido comumente na literatura psicométrica como Teoria Clássica dos Testes (TCT) e, por outro lado, o modelo de resposta ao item conhecido na literatura psicométrica como Teoria de Resposta ao Item (TRI).

Sob a base do marco metodológico apresentado, diferentes ferramentas, técnicas e métodos para cada uma das etapas deste marco ainda podem ser desenvolvidos com mais detalhes em futuros trabalhos. Dentre estas, destacamos a análise de itens, que pode ser desenvolvida tanto na perspectiva da abordagem da TCT quanto da abordagem da TRI – que, na prática, são complementares, pois ambas fornecem um conjunto de indicadores a respeito da qualidade das perguntas de um instrumento.

Referências

ABEDI, J. Dimensionality of NAEP Subscale Scores in Mathematics. *CSE Technical Report 428*, University of California, Los Angeles, 1997. Disponível em: <http://www.cse.ucla.edu/products/reports/TECH428.pdf>. (mimeo).

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION – AERA; AMERICAN PSYCHOLOGICAL ASSOCIATION – APA; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION – NCME. *Standars for educational and psychological testing*. Comitê conjunto da AERA, APA e NCME. Washington: AERA Publications Sales, 1999.

APARICIO, A.; ABDOUNUR, O.; BAZÁN, J. L. Uma primeira aproximação aos cursos de graduação e pós-graduação em estatística em universidades de São Paulo. In: SALCEDO, A. *Educación Estadística en América Latina*: tendencias y perspectivas. 1 ed. Caracas: Universidad Central de Venezuela, 2013. p. 257-282. (vol. 1). Disponível em: <https://goo.gl/VTCSgK>. Acesso em: 24 fev. 2017.

ARTILES, I.; MENDOZA, A.; YERA, M. La evaluación del aprendizaje, un indicador para elevar la efectividad del tutor en el contexto de Universalización de la Educación Superior. *Revista Iberoamericana de Educación*, Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura (OEI), v. 46, n. 4, p. 3-14, 2008. Disponível em: <http://www.rieoei.org/deloslectores/2265Olivera.pdf>. Acesso em: 23 fev. 2014.

BAZÁN, J. L.; APARICIO, A. Las actitudes frente a la Matemática-Estadística dentro de un modelo de aprendizaje. *Revista de Educación*, Pontificia Universidad Católica del Perú, Peru, v. 15, n. 28, p. 7-20, 2006. Disponível em: <http://revistas.pucp.edu.pe/index.php/educacion/article/view/2041/1974>. Acesso em: 23 fev. 2017.

BAKER, F.; KIM, S. *Item Response Theory*: parameter estimation techniques. 2 ed. Nova York: Marcel Dekker Inc., 2004.

BOND, T.; FOX, C. *Applying the Rasch model*: fundamental measurement in the human sciences. Mahwah: Erlbaum, 2001.

BORSBOOM, D.; MELLENBERGH, G. J.; VAN HEERDEN, J. The Theoretical Status of Latent Variables. *Psychological Review*, v. 110, n. 2, p. 203-219, 2003.

BURGA, A. La unidimensionalidad de un instrumento de medición: perspectiva factorial. *Reporte técnico*, Oficina de Medición de la Calidad de los Aprendizajes (UMC)/Ministerio de Educación de Perú, p. 1-21, nov. 2005. Disponível em <http://www2.minedu.gob.pe/umc/admin/images/publicaciones/artiumc/2.pdf>. Acesso em: 24 fev. 2017.

DE BOECK, P.; WILSON, M. (Ed.). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach*. Nova York: Springer, 2004.

DE GRUIJTER, D. N. M.; VAN DER KAMP, L. J. T. *Statistical Test Theory for the Behavioral Sciences*. Boca Raton: Chapman Hall/CRC, 2008.

DUCKOR, B.; DRANEY, K.; WILSON, M. Measuring measuring: Toward a theory of Proficiency with the Constructing Measures framework. *Journal of applied measurement*, Berkeley, v. 10, n. 3, p. 296-319, 2009. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.684.6807&rep=rep1&type=pdf>. Acesso em: 23 fev. 2017.

EMBRETSON, S.; REISE, S. *Item response theory for psychologists*. Mahwah: Erlbaum, 2000.

FOX, J. *Bayesian Item Response Modeling*: theory and applications. Nova York: Springer, 2010.

GADERMANN, A. M.; GUHN, M.; ZUMBO, B. D. Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation,* v. 17, n. 3, p. 1-13, 2012.

HAMBLETON, R.; SWAMINATHAN, H.; ROGERS, H. J. *Fundamentals of item response theory*. Newbury Park: Sage, 1991.

KLINE, P. *A Handbook of Test Construction*: Introduction to Psychometric Design. Nova York: Methuen and. Co. Ltd., 1986.

LORD, F. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale: Lawrence Erlbaum, 1980.

LORD, F.; NOVICK, M. *Statistical theories of mental test scores*. Reading: Addison-Welsley Publishing Company, 1968.

MAGNUSSON, D. *Teoría de los tests*. México: Trillas, 1990.

NOVICK, M. The Axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, v. 3, n. 1, p. 1-18, 1966.

NUNNALLY, J. *Teoría Psicométrica*. México: Trillas, 1987.

ORTIZ, P. *El sistema de la personalidad*. Lima: Orion, 1994.

PASQUALI, L. *Psicometria*: teoria dos testes na psicologia e na educação. Petrópolis: Vozes, 2004.

SIJTSMA, K. On the use, the missue, and the very limited usefulness of Cronbach's. *Psychometrika*, v. 74, n. 1, p. 107-120, 2009.

VAN DER LINDEN, W. J.; HAMBLETON, R. K. (Ed.). *Handbook of modern item response theory*. Nova York: Springer, 1997.

WILSON, M. *Constructing Measures*: An Item Response Modeling Approach. Estados Unidos: Lawerence Erlbaum Associates Inc., 2005.

ZUMBO, B. D.; GADERMAN, A. M.; ZEISSER, C. Ordinal Versions of Coefficients Alpha and Theta For Likert Rating Scales. *Journal of Modern Applied Statistical Methods*, v. 6, p. 21-29, 2007.