Aplicação do modelo de Rasch

Sobre os autores

Trevor G. Bond
é atualmente professor adjunto da Faculdade de Artes, Sociedade e Educação da Universidade James Cook, na Austrália. Foi recentemente professor adjunto da Faculdade de Educação da Universidade Kebangsaan Malaysia. Antes disso, foi professor e chefe do Departamento de Psicologia Educacional, Aconselhamento e Necessidades de Aprendizagem do Instituto de Educação de Hong Kong e, depois, acadêmico visitante na mesma instituição. Em cursos de graduação e pós-graduação, lecionou disciplinas relacionadas à Psicologia Desenvolvimental e Educacional para alunos de formação de professores e continua a supervisionar alunos de doutorado que usam abordagens de medição de Rasch em sua pesquisa. Em 2005, incentivou os simpósios anuais da Pacific Rim Objective Measurement Society (PROMS) a apoiar o desenvolvimento da capacidade de medição no Sudeste e Leste Asiáticos. Bond ministra regularmente *workshops* sobre a medição de Rasch em países que têm o inglês como segunda língua.

Christine M. Fox
é professora titular de Pesquisa e Medição da Universidade de Toledo, Ohio, onde ministra cursos de pós-graduação em Estatística, Medição e Projeto de Pesquisa. Sua pesquisa é centrada em construção de medidas significativas nas ciências sociais, com interesse específico pelo desenvolvimento de novas metodologias, e medidas para tomadas de decisão de alta importância. Dois de seus principais projetos (em cooperação com a Dra. Beltyukova) envolvem a medida de percepções para decisões de alta importância: desenvolvimento de escalas de experiência de passageiros para a Boeing Commercial Airplanes e atuação como membro da equipe principal de pesquisa para um projeto intitulado "NATO Global Perceptions – Views from Asia-Pacific Region", financiado pelo Programa Ciência para Paz e Segurança da Organização do Tratado do Atlântico Norte (Otan). Em parceria com a Dra. Beltyukova, desenvolve ferramentas de avaliação de risco para subscrição de seguros no setor de atenção à saúde.

Trevor G. Bond
Christine M. Fox

Aplicação do modelo de Rasch

2a. edição

Tradução de Cecília Bartalotti

Revisão técnica de Hudson Golino
Universidade de Virgínia

hogrefe

Título original: *Applying the Rasch model: Fundamental measurement in the human sciences*

Editora: Cristiana Negrão
Tradução: Cecília Bartalotti
Revisão técnica: Hudson Golino
Capa e diagramação: Claudio Braghini Junior
Preparação: Carlos Villarruel
Revisão: Joana Figueiredo

CIP-BRASIL. CATALOGAÇÃO NA PUBLICAÇÃO
SINDICATO NACIONAL DOS EDITORES DE LIVROS, RJ

B694a
Bond, Trevor G., 1944-
Aplicação do modelo de Rasch / Trevor G. Bond, Christine M. Fox; tradução Cecília Bartalotti ; revisão técnica Hudson Golino. - 2. ed. - São Paulo : Hogrefe, 2020.

Tradução de: Applying the Rasch model fundamental measurement in the human sciences
Inclui bibliografia e índice
ISBN 978-65-5072-007-0

1. Psicologia - Métodos estatísticos. 2. Ciências sociais - Métodos estatísticos. 3. Psicologia - Pesquisa - Metodologia. 4. Ciências sociais - Pesquisa - Metodologia. 5. Psicometria. I. Fox, Christine M. II. Bartalotti, Cecília. III. Golino, Hudson. IV. Título.

20-65812 CDD: 150.287
CDU: 159.938

Este livro segue as regras da Nova Ortografia da Língua Portuguesa.

Editora Hogrefe CETEPP
Rua Barão do Triunfo, 73 - conjunto 74
Brooklin, São Paulo – SP, Brasil
CEP: 04602-020
Tel.: +55 11 3900-1670
www.hogrefe.com.br

ISBN: 978-65-5072-007-0
Impresso no Brasil.

Para Mike Linacre,
cooperativo, companheiro e colaborativo.

Sumário

Prefácio à edição brasileira

Este livro dos professores Trevor Bond e Christine Fox apresenta uma introdução detalhada a alguns dos mais básicos modelos psicométricos desenvolvidos por Georg Rasch e posteriormente ampliado por outros pesquisadores da área de métodos quantitativos, como David Andrich e outros. Rasch era um estatístico dinamarquês que trabalhou com psicometria e avaliação de desempenho educacional e habilidades psicológicas. No entanto, podemos sintetizar o seu trabalho com foco principalmente na teoria de medida em psicologia (ver Rasch, 1961, 1977). O seu trabalho teve um profundo impacto nas áreas de psicologia e educação, em todo o mundo, e o seu principal livro (*Probabilistic models for some intelligence and attainment tests*) possui mais de 10.400 citações, de acordo com o Google Acadêmico, das quais aproximadamente 25% foram feitas em trabalhos acadêmicos publicados entre 2015 e 2019. Sua influência, além de profunda, não podia ser mais atual.

A obra de Bond e Fox, por sua vez, também gerou um impacto notável em todo o mundo, tendo sido citada mais de 6.700 vezes desde a publicação da primeira edição em inglês, em 2001. Hoje pessoas de diversos países podem ler a obra em sua língua materna, o que provavelmente contribui para a formação de milhares de cientistas, professores e profissionais de saúde.

Por esse motivo, sinto-me honrado em ter sido convidado a revisar a tradução do livro para o português brasileiro. Tratou-se de uma experiência bastante interessante, uma vez que foi um dos primeiros livros de psicometria que li (e compreendi), influenciado pelo meu orientador (Prof. Cristiano Gomes da Universidade Federal de Minas Gerais – UFMG). Bond e Fox pegaram um ferramental técnico denso de psicometria e o transformaram em um conteúdo acessível e interessante para boa parte dos leitores, sejam eles alunos de graduação, pós-graduação ou pesquisadores formados. E esse é um feito notável, uma vez que explicar conceitos e modelos psicométricos complexos de forma didática requer uma mistura de conhecimento, dedicação e habilidade de comunicação. Essas características, quando combinadas, produzem obras de rara beleza educacional, ainda mais em uma área tão pouco articulada em termos de comunicação com o público geral como a área da psicometria e dos métodos quantitativos.

Espero que o livro de Bond e Fox contribua para a formação técnica de estudantes e pesquisadores brasileiros interessados em psicometria e na medida de traços e habilidades educacionais e psicológicos ou na mensuração de construtos de outras áreas (como saúde). Este livro está alinhado com a obra que publicamos em 2015 (ver Golino, Gomes, Amantes, & Coelho, 2015),

na medida em que ambos trazem conteúdos complementares, uma vez que decidimos nos concentrar mais na teoria matemática da medida, nos modelos de Rasch tradicionais e nos modelos estendidos, enquanto a obra de Bond e Fox apresenta os modelos básicos e suas aplicações em diversos campos do conhecimento, com exemplos ilustrativos que são fundamentais para todos aqueles que estejam aprendendo esse conteúdo pela primeira vez.

Hudson Golino
Universidade de Virginia
Charlottesville, Estados Unidos

Referências

Golino, H. H., Gomes, C. M. A., Amantes, A., & Coelho, G. (2015). *Psicometria contemporânea: Compreendendo os modelos Rasch*. São Paulo: Casa do Psicólogo.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1980, pp. 321-334.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, *14*, 58-93.

Prefácio

Não é mera coincidência que, depois de meu doutorado em Metodologia de Pesquisa Quantitativa na Alemanha, eu tenha me interessado pelo modelo de medição de Rasch e que tenha conhecido Trevor Bond enquanto ele trabalhava em Hong Kong. Ironicamente, para mim, a faísca inicial foi o capítulo de encerramento – "Uma visão geral sintética" – da segunda edição deste livro, que contém todo um conjunto de exemplos ilustrativos e explicações sobre a importância da invariância de medição. Foi, de fato, por causa desse capítulo em particular que decidi me aprofundar mais no tema da medição de Rasch. Sei que alguns de meus colegas torcem o nariz para essa introdução menos formalizada à medição de Rasch e concordo que, caso a intenção seja se aprofundar cada vez mais no entendimento do modelo de Rasch, a formalização matemática é inevitável. No entanto, se os especialistas em Rasch não puderem comunicar nem os aspectos mais fundamentais de suas melhores ideias para aqueles que não receberam o aperto de mão mágico da matemática, suas ideias provavelmente ficarão condenadas a permanecer como o conhecimento secreto de uma espécie de sociedade secreta. Felizmente, a comunicação dos princípios da medição de Rasch pela exposição das ideias conceituais subjacentes e implicações práticas é o fio condutor de todo este livro.

Então, de volta ao Instituto de Educação de Hong Kong, novembro de 2006: depois de ajudar Trevor na montagem de uma oficina sobre medição de Rasch, sentei-me e ouvi sua apresentação introdutória. Eu já era bastante competente em medição de Rasch, em sua base matemática e sobretudo no princípio de objetividade específica de comparações, mas a questão de Trevor feita na apresentação foi uma surpreendente revelação do *significado* da medição de Rasch: "Se todos os itens de um teste pretendem medir a mesma 'coisa', por que deveria fazer diferença quais itens usamos ao compararmos pessoas?". Eu conhecia o princípio matemático de objetividade específica de comparações entre duas pessoas arbitrariamente escolhidas, que pode ser mais formalmente expresso como:

> Uma vez que, de acordo com a equação do modelo de Rasch, a probabilidade de solucionar um item depende apenas da dificuldade do item e da habilidade da pessoa ou da diferença entre o parâmetro da pessoa e o parâmetro do item, é possível calcular a diferença na probabilidade de solução de uma pessoa *j* e *j* como a diferença entre as capacidades da pessoa: $\beta_j' - \alpha_i - (\beta_j - \alpha_i) = \beta_j' - \beta_j$, (com α_i: parâmetro de dificuldade de um item *i*, e β_j: parâmetro de habilidade de uma pessoa *j*). Uma comparação

de pessoas é, portanto, igual à diferença entre suas posições na dimensão latente. Isso é independente do item *i* que foi usado para a comparação.

Embora essa seja uma descrição perfeitamente correta da implicação matemática, a interpretação de Trevor deu sentido aos números ao relacioná-la ao entendimento sólido (espera-se) implícito ou talvez até explícito de todos quanto à necessidade da invariância de medição mesmo nas ciências sociais.

Sempre tive dificuldade para comunicar a propriedade de comparações invariantes e, admito, tinha dificuldade para *compreender* as implicações dessa propriedade. A ilustração de Trevor da ideia conceitual por trás da invariância de medição fez a ponte sobre minha própria lacuna entre teoria e prática – e entre formalização matemática e comunicação.

Contudo, esta nova edição do livro não deve ser vista como um mero guia prático sobre como construir escalas de medição de acordo com o modelo de Rasch. No Capítulo 12 – "Pressupostos e requisitos do modelo de Rasch: ajuste do modelo e unidimensionalidade" –, a análise crítica de práticas comuns para avaliar o ajuste do modelo ou do item não deixa dúvida sobre as desconfianças dos autores em relação a muitas ações adotadas atualmente. Infelizmente, há um número muito grande de artigos sobre o modelo de Rasch que evitam cuidadosamente testes de ajuste do modelo, e isso sugere, em minha opinião, que a medição de Rasch aplicada corre o risco de se condenar a um pragmatismo medíocre. Este livro é, assim, diametralmente oposto ao que Andreski (1972) descreveu como "o costume de evitar críticas mútuas (que) serve meramente como um escudo contra a responsabilidade por negligência". Bond e Fox, além disso, não deixam dúvida de que veem o escalonamento do Rasch como uma hipótese *falsificável* (isto é, uma hipótese que não precisa ser verdadeira para todas as tentativas de construção de escalas).

Ao longo do livro, enfatiza-se a construção adequada de uma teoria substantiva *antes* da aplicação do modelo de Rasch. Assim, se a medição de Rasch não puder ser obtida, não devemos culpar o modelo de Rasch. Em vez disso, devemos ser honestos e perguntar a nós mesmos se a teoria em consideração é suficientemente forte para levar a escalas que satisfaçam os requisitos do modelo de Rasch. Não sei se podemos desenvolver teorias sobre capacidades cognitivas ou personalidade que levem ao modelo de Rasch. Mas não conheço nenhum caso de um método nas ciências naturais –que psicólogos tendem a emular – que tenha sido desenvolvido em um vácuo teórico e depois usado com sucesso para fazer descobertas substanciais ou mesmo revolucionárias. É por isso que fico cada vez mais desconfiado da aplicação cega do modelo de Rasch em décadas recentes.

Muitos devem saber que não sou um "verdadeiro crente" da medição de Rasch (por exemplo, Heene, 2013), e um prefácio a este livro de um crítico de algumas práticas do Rasch pode parecer deslocado. No entanto, a luta por padrões de medição mais rigorosos nas ciências sociais, combinada com integridade e honestidade científicas, é um valor que deve sempre ser apreciado, e eu realmente aprendi muito sobre isso neste livro.

Moritz Heene
Universidade Ludwig Maximilian
Munique, Alemanha

Referências

Andreski, S. (1972). *Social sciences as sorcery*. London: Deutsch.

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology*, *4*, 246. doi:10.3389/fpsyg.2013.00246

Introdução

Sempre tivemos a intenção de produzir um livro, escrito em estilo acessível, que facilitasse um entendimento mais profundo do modelo de Rasch sem exigir que os leitores tivessem uma formação matemática ou estatística sofisticada. Com mais de seis mil citações até o momento, nossos leitores fizeram deste livro um texto clássico. A terceira edição de nosso livro, com frequência chamado simplesmente de "Bond & Fox", examina as propriedades essenciais da medição intervalar do modelo de Rasch e demonstra seu uso com exemplos de várias ciências humanas, entre elas resultados em desempenho educacional, desenvolvimento humano, atitudes e saúde. Um glossário abrangente e inúmeras ilustrações ajudam o leitor a compreender melhor os mecanismos do modelo. Depois de aprenderem a aplicar a análise de Rasch, os leitores poderão fazer suas próprias análises e interpretar os resultados usando os conjuntos de dados que são utilizados como exemplos no livro e os que se encontram no *link* https://www.winsteps.com/BF3/bondfox3.htm..

Estamos sempre interessados em ouvir as opiniões nem sempre elogiosas de nossos leitores, nossos colegas, resenhistas do livro e nossos críticos. Mas, mesmo assim, não seguimos o conselho de alguns colegas e de um resenhista de amenizar nossa crítica à teoria clássica dos testes (TCT) e a outros métodos da teoria de resposta ao item (TRI).

No entanto, nesta edição, refletimos sobre alguns dos aspectos em comum – bem como sobre as diferenças distintivas – entre essas abordagens da psicometria. Não estamos fazendo uma "guerra religiosa" por "verdadeiros crentes" e não depreciamos nem descartamos nossos colegas com interesses e conhecimentos mais amplos na TRI. Apenas reafirmamos os pilares da abordagem de Rasch para construir e monitorar variáveis e contrastamos as consequências de usar abordagens voltadas para dados e teorias. Claramente, recomendamos a abordagem de que os "dados se ajustam ao modelo", e não o contrário. Apesar do melhor que o modelo de Rasch tem a oferecer, a escolha da metodologia está sempre nas mãos do analista.

Novidades desta edição

Na terceira edição, enfatizamos o seguinte:

- Muitas análises estatísticas usadas nas ciências humanas requerem a contribuição de dados intervalares; e a medição de Rasch centra-se em proporcionar medições intervalares.

- A medição de Rasch concentra-se, *a priori*, na construção e monitoração (controle de qualidade) de escalas.
- As técnicas de análises de Rasch também podem ser usadas *post hoc* para revelar a extensão em que a medição intervalar pode ser representada nos conjuntos de dados já existentes.

Para fazer isso, esclarecemos a relação entre medidas Rasch, TCT e TRI (especialmente no final do Capítulo13).

Partindo da premissa de que medições científicas devem permanecer invariantes no tempo e espaço, mostramos como a medição de Rasch proporciona os meios para representar e examinar a invariância de medição.

No novo Capítulo 9 – "Criação de medidas, definição de padrões e regressão de Rasch" –, convidamos três outros estudiosos do Rasch para contribuir com seu conhecimento na aplicação do modelo.

Dividimos os capítulos essenciais (4, 6, 7 e 8) em duas seções:

(a) uma introdução voltada para o entendimento básico, seguida por
(b) uma seção para desenvolver um entendimento estendido.

Condensamos a exposição da segunda edição em cada capítulo e, depois, estendemos essa cobertura com uma investigação e interpretação aprofundadas usando uma ampla variedade de exemplos da literatura. Isso foi feito para que os leitores possam selecionar com facilidade o nível de leitura e atividade mais apropriado para suas necessidades.

O Capítulo 6, baseado no *rating scale model*, foi reescrito com base na análise do *Children's Empathic Attitudes Questionnaire* – CEAQ. (O conjunto de dados e o tutorial do *Computer Anxiety Index* – CAIN – foram movidos para a plataforma *on-line*.)

Incluímos um conjunto de atividades (leituras e análises) em quadros dentro do texto de cada capítulo. Convidamos nossos leitores a aproveitar o tempo e a oportunidade para melhorar suas habilidades na aplicação efetiva do modelo de Rasch, em vez de apenas lerem o que temos a dizer a respeito.

Além disso, incluímos:

- Mais ferramentas de aprendizagem, como introduções aos capítulos, termos-chave em negrito, resumos dos capítulos, atividades e listas de leituras sugeridas.
- Uma variedade maior de exemplos de líderes estabelecidos e emergentes em análise de Rasch.
- Um conjunto de dados parcial para a analogia do percurso a ser adotado, a fim de que os leitores possam trabalhar mais facilmente com esses conceitos.

- Mais detalhes sobre como um conjunto de itens pode ser "construído" e entendido usando uma combinação de diagnóstico do modelo de Rasch e uma teoria substantiva (Capítulo 4).
- A discussão estendida de invariância agora examina DIF, DRF e ancoragem (Capítulo 5).
- Novas interpretações de outros conjuntos de dados sobre medição de facetas (Capítulo 8).
- Um novo capítulo inclui criação de medidas *post hoc*, regressão de Rasch abordando relações de variáveis independentes e dependentes (IV-DV) e definição de padrões objetivos em situações de teste simples e complexas (Capítulo 9).
- Uma variedade mais ampla de exemplos de medições Rasch por toda a extensão das ciências humanas (Capítulo 10).
- Apêndice A estendido "Para começar", que proporciona um acompanhamento passo a passo de uma análise de Rasch de arquivos de dados .txt e SPSS para mostrar aos leitores como conduzir análises de Rasch por conta própria.

Website[1]

No *link* https://www.winsteps.com/BF3/bondfox3.htm, oferecemos:

- *Software* gratuito da análise de Rasch, Bond&FoxSteps e Bond&FoxFacets, pela generosidade de Mike Linacre (Winsteps), que são pré-carregados com os conjuntos de dados introdutórios para os capítulos 4, 6, 7 e 8 e guias tutoriais passo a passo.
- Conjuntos de dados para outras análises descritas no livro.
- A planilha Invariance para o Capítulo 5 na forma de um arquivo Excel.
- Uma pasta de PDFs de instruções passo a passo para cada uma dessas análises. Muitas delas estarão disponíveis em outros idiomas além do inglês, para atender a nossos muitos leitores que têm o inglês como segunda língua.
- *Links* para artigos e *websites* úteis para leituras adicionais sobre temas importantes da medição de Rasch.
- Outros conjuntos de dados com orientação para análises.

Público-alvo

Esta abordagem revela mais claramente nossas estratégias pedagógicas para alcançar toda a faixa de nosso público-alvo.

1 Informamos que não há uma versão em português do *site* que acompanha este livro na edição em inglês, mas o *software* encontra-se disponível para *download* no *link* https://www.winsteps.com/BF3/bondfox3.htm (NRT).

Para aqueles que são *relativamente novos* na medição de Rasch e especialmente para os que estão aprendendo sem a orientação de um colega experiente no Rasch, sugerimos fazer sua leitura seguindo estas etapas:

1. Trabalhar da Introdução até o final do Capítulo 13, completando as leituras adicionais e os exercícios de análises de dados de cada capítulo. Em sua primeira leitura, complete apenas as seções de introdução básica (e atividades correspondentes) dos capítulos 4, 6, 7 e 8.
2. Releia as seções de "Entendimento estendido" desses quatro capítulos e complete as análises restantes de cada seção.
3. Complete a leitura e atividades descritas no Apêndice A, "Para começar". As instruções *on-line* ajudarão você a se tornar mais independente para fazer suas próprias análises de Rasch.

Para aqueles que já estão mais familiarizados com os métodos de Rasch, nosso conselho geral é que completem *integralmente* os capítulos, as leituras e as análises na ordem em que os escrevemos.

No caso dos professores experientes, especialmente os que têm muitos alunos para os quais a matemática é parte integrante de seu entendimento e que se sentirem frustrados com a generalidade e falta de matemática de nossa apresentação, insistimos para que usem seu conhecimento superior e o de seus alunos sobre o modelo de Rasch para planejar suas próprias leituras e sequências de análises relevantes para a classe que melhor atendam às suas necessidades.

Nosso livro é usado com frequência como texto para cursos de pós-graduação em Medição (avançada), Teoria de Resposta ao Item, Métodos de Pesquisa (avançados) ou Análise Quantitativa ministrados nas áreas de psicologia, educação, desenvolvimento humano, saúde, enfermagem, administração de empresas e outras ciências sociais e da vida, por causa de sua introdução acessível da medição de Rasch.

Agradecimentos

A proposta de escrever uma terceira edição deste livro – que se tornou um sucesso de vendas – surgiu quando fomos convidados por nossa editora, Debra Riegert, a apresentar uma proposta para exame da Routledge/Taylor & Francis. Agradecemos a Debra e à equipe da T&F que levaram nossas ideias até a publicação. Parte do processo da proposta envolveu a solicitação pela editora de análises da proposta à luz de seções selecionadas da segunda edição existente, e agradecemos a esses analisadores o tempo, a reflexão e o esforço com que contribuíram para essa tarefa.

Agradecemos ainda a vocês, nossos leitores, que reconheceram o valor do modelo de Rasch como o apresentamos nas duas edições anteriores. Nosso obrigado pelas vendas do livro e, mais particularmente, pelas citações. Acreditamos que vocês encontrarão nesta edição ainda mais material para refletir e apreciar. Obrigado a nossos colegas de medição de Rasch, que nos proporcionaram muitas lições para aprendizado, além de críticas muito produtivas. Somos gratos aos leitores que enviaram perguntas, comentários e correções por *e-mail*. É sempre estimulante manter contato com as pessoas que leem a obra tão detalhadamente. Chan Wai Fong aprendeu medição de Rasch sozinha com a segunda edição; obrigado, Wai Fong, por suas leituras similarmente cuidadosas de cada capítulo desta nova terceira. Agradecemos aos alunos de pós-graduação de George Engelhard, que também ofereceram contribuições sobre as versões preliminares dos capítulos 1 a 4.

Um agradecimento especial aos nossos colegas Svetlana Beltyukova, Gregory Stone e Yan Zi, que escreveram o Capítulo 9:

Dr. Yan Zi é professor assistente do Departamento de Currículo e Instrução do Instituto de Educação de Hong Kong. Yan Zi publica nas áreas de avaliação psicológica e educacional, sendo especializado na aplicação da análise de Rasch. É BSc e MEd pelo Instituto Wuhan de Educação Física (China) e PhD pela Universidade James Cook (Austrália). Preside um grupo de acadêmicos na tradução do Bond & Fox para o chinês.

Dr. Gregory E. Stone é professor de Pesquisa e Medição da Universidade de Toledo (Ohio) e parceiro na empresa de consultoria MetriKs Amérique. É muito publicado na América do Norte, e seu livro *Applications of Rasch measurement in criterion-referenced testing* traz seu conhecimento do Rasch para esse tema importante. Em 2008, Stone recebeu do governador do estado de Veracruz, no México, a homenagem de Visitante Ilustre. É BA pela Faculdade Shimer, MA pela Universidade Loyola de Chicago e PhD pela Universidade de Chicago.

Dra. Svetlana Beltyukova é professora associada de Pesquisa e Medição da Universidade de Toledo (Ohio). Seu trabalho acadêmico e de consultoria centra-se em aplicações criativas do modelo de Rasch em pesquisas da satisfação de estudantes, saúde e pesquisas de serviços humanos, bem como no uso da regressão de Rasch para identificação de fatores-chave na atenção à saúde. É PhD pela Universidade de Linguística de Kiev (Ucrânia) e pela Universidade de Toledo.

Agradecemos a Yan Zi, Jason Fan e à equipe de jovens pós-doutorandos de Hong Kong e da China a tradução do Bond & Fox para o chinês. Parabéns pelo trabalho!

Em resposta parcial aos críticos que tanto questionam por que nós, autores e leitores, insistimos em defender e usar apenas um pequeno nicho da TRI e da teoria estatística, citamos um editorial escrito no *New York Times* sobre uma teoria de finanças corporativas:

> *Esse é o verdadeiro teste de uma teoria brilhante*, diz um membro do comitê do Prêmio Nobel de Economia. *O que antes era considerado errado mais tarde demonstra-se óbvio*. As pessoas veem o mundo como são treinadas para vê-lo e resistem a explicações contrárias. E é isso que faz a inovação ser mal recebida e a descoberta ser quase impossível.
> *Uma inovação científica importante raramente consegue se impor conquistando e convertendo gradualmente seus oponentes. [...] O que acontece é que seus oponentes gradualmente morrem, e a geração seguinte está familiarizada com a (nova) ideia desde o início.* Não é surpresa que as descobertas mais profundas com frequência sejam feitas pelos jovens ou pelos de fora, nenhum dos quais aprendeu ainda a ignorar o óbvio ou a viver com a sabedoria aceita.
>
> "Ortodoxia pura" (17 de outubro de 1985).

Trevor Bond e Christine Fox

1
Por que a medição é fundamental

De acordo com os estudantes de doutorado da Malásia, há um grupo de professores de ciências sociais bastante intransigente que, durante a defesa da tese, insiste em examinar os candidatos quanto à natureza dos dados que eles estão analisando. Em particular, questionam se os dados são de fato intervalares ou meramente ordinais em natureza. Aparentemente, essa disposição um tanto antiquada levou à reprovação de várias defesas de doutorado; os candidatos que não conseguiam comprovar a natureza intervalar de seus dados recebiam a solicitação de refazer suas análises estatísticas, substituir o *r* de Pearson pelo *rho* de Spearman, e assim por diante. No mundo ocidental, a maioria dos professores, pelo menos em educação, psicologia e outras ciências humanas, não tem feito objeções a esses detalhes: o *r* de Pearson parece funcionar igualmente bem com todos os tipos de dado – uma vez que o SPSS não sabe de onde os dados vêm e, aparentemente, muitos de seus usuários também não. O lado positivo dessa dificuldade é que muitos desses professores intransigentes agora entendem que medidas derivadas de análises de Rasch podem ser consideradas intervalares e, portanto, permitem o uso do amplo conjunto de cálculos estatísticos que existem nas ciências sociais. Infelizmente, porém, medição não é um tópico rotineiramente ensinado nos currículos padrão do mundo ocidental, e o recurso alternativo é analisar dados ordinais como se fossem medidas intervalares.

Apesar desses professores antiquados e de um pequeno número de teóricos da medição, há mais de meio século os pesquisadores em ciências sociais conseguiram iludir a si mesmos sobre o que a medição realmente é. Em nossa vida cotidiana, fazemos uso explícito e implícito de sistemas de medição calibrados para comprar gasolina e água, medir e cortar madeira, comprar tecidos, juntar os ingredientes para uma receita de cozinha e administrar doses apropriadas de remédios para parentes doentes. Por que, então, quando vamos para a universidade ou para uma empresa de testes fazer pesquisas em ciências sociais com o objetivo de conduzir uma investigação psicológica ou implantar uma pesquisa padronizada, passamos a tratar e analisar esses dados como se os requisitos para a medição que nos serviram tão bem de manhã não se aplicassem mais à tarde? Por que mudamos nossa definição e padrões de medição quando a condição humana é o foco de nossa atenção?

Os sistemas de medição são ignorados quando expressamos rotineiramente os resultados de nossas intervenções de pesquisa em termos de níveis de probabilidade $p < 0,01$ ou $p < 0,05$ ou – melhor ainda – como tamanhos de efeito. Os níveis de probabilidade só indicam o quanto é provável ou improvável que A seja mais do que B ou que C seja diferente de B, e o tamanho de efeito serve para nos dizer quanto as duas amostras sob análise diferem. Em vez de se concentrarem na construção de medidas da condição humana, psicólogos e pesquisadores de outras ciências humanas focaram a aplicação de procedimentos estatísticos sofisticados em seus dados. Embora a análise estatística seja uma parte necessária e importante do processo científico, e os autores jamais, de maneira nenhuma, desejariam substituir o papel que a estatística desempenha no exame de relações entre **variáveis**, o argumento ao longo deste livro é que os pesquisadores quantitativos nas ciências humanas estão excessivamente concentrados em análise estatística e não se preocupam nem perto do suficiente com a natureza dos dados nos quais usam essas estatísticas. Portanto, não é propósito dos autores substituir a estatística quantitativa pela medição de Rasch, mas voltar o foco de parte do tempo e energia usados em análises de dados para o pré-requisito da construção de medidas científicas de qualidade.

Aqueles professores intransigentes mencionados no início do capítulo recorrem, claro, às diretrizes aprendidas com Stevens (1946). Todo aluno de Psicometria 101 ou Métodos Quantitativos 101 fica com a lição de Stevens entranhada para sempre. Em resumo, Stevens definiu medição como a atribuição de números a objetos ou eventos de acordo com uma regra, e, desse modo, alguma forma de medição existe em cada um de quatro níveis: **nominal**, **ordinal**, **intervalar** e **de razão**. A maioria de nós aceita hoje que medições no nível de razão provavelmente permanecerão além de nossa capacidade nas ciências humanas, no entanto a maioria de nós pressupõe que os dados que coletamos pertencem a escalas de nível intervalar.

Parece intrigante, porém, que aqueles que se propõem a ser cientistas da condição humana, especialmente em pesquisa psicológica, de saúde e educacional, aceitem suas "medidas" de nível ordinal sem nenhuma aparente reflexão crítica, quando elas não são, de fato, medidas. Talvez devamos ler o próprio Stevens (1946, p. 679) com um pouco mais de atenção: "Na verdade, a maioria das escalas usadas ampla e efetivamente por psicólogos são ordinais". Ele especificou, depois, que a única estatística "permissível" para dados ordinais eram medianas e percentis, deixando médias, desvios padrão e correlações apropriadas apenas para dados intervalares ou de razão. E, ainda mais surpreendente, "O coeficiente de correlação Spearman é geralmente considerado adequado para uma escala ordinal, mas, na verdade, essa esta-

tística pressupõe intervalos iguais entre postos sucessivos e, portanto, requer uma escala intervalar" (Stevens, 1946, p. 678). Poderia ser mais claro do que isto: "Com a escala intervalar, chegamos a uma forma que é 'quantitativa' no sentido comum da palavra"? (Stevens, 1946, p. 679). Este é também o nosso argumento: apenas com "intervalos" obtemos o "quantitativo" no sentido comum, o sentido em que usamos medidas científicas em nossa vida cotidiana. Então por que os cientistas sociais ficam tão confusos?

Infelizmente, no mesmo artigo seminal, Stevens (1946, p. 679) depois amenizou essas distinções ordinal/intervalar ao nos permitir recorrer a "uma espécie de sanção pragmática: em inúmeros casos, ela leva a resultados proveitosos". Ele acrescentou um possível exemplo de condição: "Quando apenas os ordenamentos dos dados são conhecidos, devemos proceder com cautela com as estatísticas e especialmente com as conclusões que tiramos delas" (Stevens, 1946, p. 679). Parece que essa "permissão" implícita para tratar dados ordinais como se eles fossem intervalares foi a única conclusão que chegou até os cientistas sociais, que estavam tão obviamente desesperados para usar suas estatísticas sofisticadas em sua profusão de escalas de atitude.

Poderíamos esperar, com razoabilidade, que aqueles que se veem como cientistas sociais aspirassem a ser pesquisadores de mente aberta, reflexivos e, mais importante, críticos. Na ciência empírica, seria de esperar que essa questão da medição fosse um tanto primordial. No entanto, muitas tentativas de levantar essas e outras questões de "será que nossos dados constituem medidas?" resultam no encerramento abrupto das oportunidades de continuar a discussão até mesmo em fóruns especificamente voltados para medição, métodos quantitativos ou psicometria. Será o apego de nossa área à interpretação (errônea?) de Stevens – a ignorância flagrante de que dados ordinais não constituem uma medição – meramente mais um caso da roupa nova do imperador? (Stone, 2002). Vamos examinar os componentes individuais dessa tradição: qual é a prática de rotina, o que implica a definição de medição e qual é a situação de cada um dos quatro níveis onipresentes de medição.

Sob o pretexto de medir, a prática comum tem sido os psicólogos *descreverem* os dados brutos que têm em mãos. Eles informam quantas pessoas responderam ao item corretamente (ou concordaram com o enunciado do item), qual o grau de relação entre uma resposta e outra e qual é a correlação entre cada item e a pontuação total. Essas meras descrições acorrentaram nosso pensamento no nível dos dados brutos, e dados brutos *não* são medidas. Embora os psicólogos geralmente aceitem **contagens** como "medições" nas ciências humanas, esse uso não pode substituir a medição como ela é conhecida nas ciências físicas. Em vez disso, o fluxo de atividade e o peso da importância científica têm sido indevidamente atribuídos a análises es-

tatísticas e não à medição. Essa falta de ênfase, associada à fé ilimitada nas atribuições de números a eventos como suficientes para a medição, cegou os psicólogos, em particular, para a inadequação desses métodos. Michell (1997, p. 374) é bastante contundente quanto a isso em seu artigo intitulado "Quantitative science and the definition of measurement in psychology", em que "a falha continuada [dos psicólogos] em perceber fatos metodológicos relativamente óbvios" é chamada de "distúrbio do pensamento metodológico". A questão permanece: É possível que, como cientistas especializados nas ciências humanas, possamos abrir nossa mente para a possibilidade de que não estejamos medindo absolutamente nada? Ou que, se estivermos, isso se deve tanto a boas intenções e sorte quanto à nossa invocação da metodologia de medição apropriada?

Crianças podem construir medidas

É evidente que, nas ciências sociais, o termo "medição" tem um prestígio não compartilhado pelas palavras "quantitativo" ou "estatística". Talvez pudéssemos aprender algo sobre o que medição realmente envolve se olhássemos para o desenvolvimento de conceitos de medição naqueles em que esse processo ainda está acontecendo: as crianças. Parte do programa de pesquisa de Jean Piaget em Genebra foi estimulada por conversas que ele teve em Davos, na Suíça, com Albert Einstein durante uma reunião em 1928 (Ducret, 1990). Einstein aconselhou Piaget a examinar o desenvolvimento dos conceitos de velocidade, distância e tempo em crianças pequenas para ver quais deles eram logicamente primitivos (isto é, se velocidade = distância/tempo, qual deles se desenvolveria antes dos demais?). Piaget começou a examinar a construção progressiva dos conceitos de comprimento (e medição) em crianças e relatou esses achados em 1948 (Piaget, Inhelder, & Szeminska, 1960).

Os assistentes de Piaget forneceram às crianças conjuntos de materiais para usar em suas investigações e, por meio de uma série de perguntas pouco estruturadas, pediram a elas que desenvolvessem um sistema de medição rudimentar para os materiais (geralmente cavilhas de madeira) fornecidos. Os autores relataram esta sequência temporal e lógica na aquisição de conceitos de medição linear em crianças:

1. As crianças classificaram (agruparam) os objetos fornecidos em uma classe com pelo menos um atributo comum adequado para medição (por exemplo, hastes de madeira) e puseram de lado o resto (por exemplo, copo, bola etc.).

2. Depois seriaram (ordenaram) esses objetos selecionados de acordo com a variação desse atributo (por exemplo, comprimento).
3. Então, identificaram uma unidade de diferença arbitrária entre dois comprimentos sucessivos (por exemplo, uma pequena haste, A, tal que haste C − haste B = haste A). A iteração dessa unidade foi usada para calcular relações de comprimento, tal que haste B = 2 × haste A, haste B + haste A = haste C e assim por diante. As tentativas de medição revelaram que a diferença generalizada entre quaisquer duas hastes adjacentes, X_n e a maior seguinte $X_{(n+1)}$, era a unidade arbitrária, A.
4. Com o tempo, cada criança percebeu que aquela unidade de medição iterada deveria ser padronizada em todos os contextos de medição apropriados, de modo que todos os comprimentos pudessem ser medidos em relação a uma escala de medição linear comum.

Não é preciso ser nenhum Schopenhauer para detectar os paralelos entre os resultados das investigações com crianças pequenas e o que aprendemos sobre níveis de medição com Stevens em nossas aulas introdutórias na faculdade. Para Piaget e as crianças, a sequência de desenvolvimento hierárquica é classificação, seriação, iteração e, depois, padronização; para os níveis de Stevens, nominal, ordinal, intervalar e de razão. A diferença interessante mas crucial – que é bem conhecida por crianças maduras do ensino fundamental e parece ser ignorada por muitos nas ciências humanas – é que, embora classificação e seriação sejam precursores necessários para o desenvolvimento de sistemas de medição, não são, em si e por si mesmas, suficientes para a medição. O atributo distintivo do sistema de medição linear é o requisito de uma unidade arbitrária de diferença que possa ser iterada entre comprimentos sucessivos. Crianças em idade escolar não demoram a perceber que unidades de medição linear convenientes, como largura da mão e comprimento do pé, são inadequadas mesmo para projetos escolares; elas requerem o uso de um pedaço de madeira de comprimento padrão, pelo menos.

É nesse último ponto que os proponentes dos modelos de Rasch para medição focam sua atenção: Como desenvolver unidades de medição que a princípio devem ser arbitrárias, mas possam ser iteradas ao longo de uma escala de interesse de modo que os valores unitários permaneçam os mesmos? Esse é o foco principal da medição de Rasch. A capa de um folheto do Grupo de Interesse Especial em Medição de Rasch da American Educational Research Association trazia o lema "Criando medidas". Cada capa do *Journal of Applied Measurement* afirma o mesmo objetivo de uma maneira diferente: "Construindo variáveis". Pode levar muito tempo até que aqueles de nós que

estão nas ciências humanas consigam adotar de bom grado um ponto de partida genuinamente zero para a medição de desempenho em matemática ou desenvolvimento cognitivo, ou para decidir como é a introversão ou qualidade de vida zero, mas aqueles que trabalham com afinco para criar medidas de modo que as escalas resultantes tenham propriedades de medição intervalar estão fazendo uma importante contribuição para o progresso científico. Para acompanharmos o desenvolvimento de instrumentos nas ciências físicas, precisamos passar mais tempo investigando nossas escalas do que investigando *com* elas. Essas tentativas de construção de medidas vão além de meramente nomear e ordenar indicadores em direção ao talvez inatingível Santo Graal das medidas de razão genuínas.

Em termos dos níveis de Stevens, os autores concluiriam então que os níveis nominal e ordinal NÃO são nenhuma forma de medição em si e por si mesmos. É claro que concordamos que seus níveis intervalar e de razão de fato constituiriam alguma forma de medição genuína. No entanto, com frequência as escalas a que rotineiramente atribuímos esse *status* de medição nas ciências humanas apenas supostamente têm as propriedades de medição de nível intervalar; essas propriedades de medição quase nunca são testadas empiricamente. Não é suficientemente bom alocar números para comportamentos humanos e, depois, simplesmente afirmar que isso é uma medição nas ciências sociais.

Até este ponto do capítulo, os autores ignoraram um aspecto crucial da definição de Stevens (porque queremos dar atenção especial a ele). Stevens nos lembrou de que as alocações numéricas têm que ser feitas "de acordo com uma regra", e aí está a dificuldade. O que essa definição não especifica é que a medição científica requer que as alocações sejam realizadas de acordo com um conjunto de regras que produzirão, no mínimo, uma escala resultante com um valor unitário que manterá seu valor ao longo de toda a escala. Alocações numéricas feitas simplesmente "de acordo com uma regra (qualquer)" produzem muitos dos indicadores úteis da condição humana que usamos habitualmente em nossas pesquisas, mas apenas alguns deles se qualificariam como "medição" assim definida.

Estatística e/ou medição

Uma consequência lastimável da tradição de Stevens, e da posição de outros nessa questão, é que a análise estatística dominou as ciências sociais a quase completa exclusão do conceito de medição. Textos introdutórios e cursos sobre medição nas ciências sociais são, rotineiramente, sobre análise esta-

tística; depois de uma representação simbólica dos quatro níveis de Stevens, o assunto é deixado de lado. Isso não significa que o objetivo dos defensores da medição de Rasch seja substituir nosso uso da estatística convencional. Em vez disso, o objetivo é fornecer aos cientistas sociais os meios para produzir medidas intervalares genuínas e monitorar a aderência dessas escalas aos princípios de medição científica. Dessa forma, a natureza intervalar dos dados – um requisito para muitas de nossas análises estatísticas mais cruciais e úteis –fica explicitada, e não apenas presumida. Por exemplo, muitos inserem seus dados brutos no *software* SPSS ou SEM, pressupondo que eles sejam "medidas" para cada variável, e depois calculam as relações entre eles. A abordagem alternativa é construir cada variável a partir das respostas ao item relevantes, verificar as propriedades de medição de cada uma usando o **modelo** de Rasch e, em seguida, introduzir a estimativa das habilidades das pessoas com controle de qualidade no *software* para a modelagem de equações estruturais ou qualquer outra análise paramétrica.

Por que medição fundamental?

Ben Wright (da Universidade de Chicago) divertia, irritava, provocava ou esclarecia os membros de sua audiência tirando uma régua dobrável de uma jarda do bolso de trás para ilustrar os argumentos que estava expondo sobre o uso da medição de Rasch. Ele denunciava as escalas usadas nas ciências humanas dizendo que eram como uma régua feita de elástico, que tinham segmentos feitos de borracha, tinham lacunas, não eram retas e assim por diante. A analogia da medição física era útil tanto para demonstrar um modelo a se tentar alcançar como para expor as deficiências das técnicas inferiores de construção de escalas. A régua de uma jarda (ou o metro) torna óbvias para nós as propriedades de uma escala de **medição fundamental** baseada no atributo extensivo de comprimento. Como Piaget revelou, crianças do ensino fundamental logo descobrem que podem concatenar comprimentos para mostrar as relações aditivas na escala de medição linear física; elas podem juntar fisicamente "unidades" arbitrárias ou hastes para acrescentar comprimento.

Isso nos dá grande poder quando combinamos as propriedades do sistema de números naturais para refletir as relações iterativas ao somarmos comprimentos unitários (100 cm = 1 metro, 1.000 m = 1 km etc.). Pensemos em como nossos predecessores costumavam alinhar os habitantes locais para obter medidas lineares baseadas em "pés" (*pieds* em francês) e "polegares" (polegadas imperiais ou *pouces* em francês). É interessante conferir

o papel humano no desenvolvimento de medidas físicas (Stone, 1998), as extremas dificuldades políticas, científicas e pessoais para estabelecer a base para o metro (Alder, 2002) e como o processo para fazer uma estimativa da altura de algo tão óbvio e sólido como o Monte Everest foi saudado como "uma das obras mais estupendas de toda a história da ciência", mesmo com a equipe tendo acesso a correntes de agrimensor calibradas com precisão (Keay, 2000).

Há muitas lições implícitas na leitura da história por trás da régua de Ben Wright. A primeira é que muitas medidas locais arbitrárias foram usadas antes de se tornarem padronizadas e intercambiáveis. A segunda é que, mesmo quando as medidas são padronizadas e, ao que tudo indica, suficientemente confiáveis e precisas, iterar as unidades em contexto para fazer estimativas de tamanho físico ainda pode ser uma tarefa desafiadora. Moral: se considerarmos que fazer medidas científicas nas ciências humanas é difícil demais para continuarmos insistindo, talvez não saibamos a história do desenvolvimento de algo tão simples como o metro e como é difícil usar esse dispositivo simples para tarefas de medição aparentemente diretas. Mesmo o modelo padrão do metro– dois pinos de ouro em uma barra de platina em um ambiente cuidadosamente controlado – é agora meramente de interesse histórico; a definição atual de metro não é nada tão rudimentar quanto isso. É claro que a desvantagem da analogia da régua de Ben, como seus críticos regularmente argumentavam, é que não podemos alinhar fisicamente pedaços da psique humana para produzir medidas, como podemos fazer com centímetros para produzir metros.

Medidas derivadas

Aqueles que estão nas ciências físicas já haviam descoberto que, embora a medição fundamental seja possível quando unidades podem ser fisicamente concatenadas (como em peso, ângulos, tempo etc.), essas escalas estavam em minoria, mesmo nas ciências físicas. A natureza aditiva de outras medidas da ciência física – e a densidade é um excelente exemplo – precisa ser descoberta (ou construída) indiretamente, em vez de ser demonstrada em ações físicas repetidas com unidades concretas. Adicionar um litro de água, massa de um quilograma e densidade um, a outra quantidade idêntica de água dará dois litros de água com massa de dois quilogramas, mas a densidade permanece em apenas um. As unidades de volume e peso podem ser fisicamente concatenadas, mas não a unidade de densidade –ainda que possamos estar cientificamente corretos ao nos referirmos a substâncias como tendo duas, três ou

mesmo metade ou um terço da densidade da água. A escala de densidade é derivada da razão constante entre massa e volume para qualquer substância: 1,0 para água pura, 19,3 para ouro, 1,7 para magnésio e assim por diante.

Se, então, a densidade é uma medida derivada, como medíamos a densidade em nossas aulas de ciências na escola? Bem, se o currículo, o livro ou o professor de ciências fossem sensíveis às bases desenvolvimentais do entendimento de ciências pelas crianças, esses experimentos não aconteceriam até muito mais tarde do que os exercícios baseados apenas em medir comprimento, peso e tempo. (Piaget descobriu que a concepção de volume das crianças é construída posteriormente à de comprimento, peso etc. e que a densidade provavelmente não é entendida pelas crianças antes do final do ensino fundamental ou início do ensino médio. Além disso, as concepções de probabilidade, que são centrais para as medições Rasch, só são construídas no fim da adolescência!)

Então, nossa professora de ciências do ensino médio nos deu uma coleção de objetos e nos fez medir o peso e o volume de cada um. Depois disso, tínhamos que inserir os valores nas células de uma tabela na sala de aula para calcular as densidades relativas: Densidade = Massa/Volume. E, se a professora de fato soubesse seu trabalho, teria nos incentivado a "descobrir" por nós mesmos que a densidade de qualquer substância (por exemplo, cobre) acabava sendo sempre consistentemente a mesma, ainda que os tamanhos e as formas dos objetos de cobre variassem consideravelmente de grupo para grupo na sala de aula: Ah! A maravilha de descobrir **invariância** (densidade) não facilmente detectável em meio a essa variação tão óbvia (peso e volume)! Nesse exemplo simples, temos uma das pedras angulares de nossos esforços científicos: nossa tarefa é encontrar medidas, regras e teorias invariantes em meio às variações que são óbvias para todos. Entretanto, precisamos explicar as variações inesperadas que observamos ao aplicarmos essas medidas, regras e teorias invariantes em uma nova situação. Buscamos a constância em face da mudança e mudança onde esperamos constância. Essa parte do empreendimento científico é central para nossa consideração da invariância de medição no Capítulo 5.

A Tabela 1.1 mostra não só os valores crescentes da escala de medição fundamental para peso da esquerda para a direita na linha superior (0,2, 0,4, 0,6, ..., 1,2 etc.) e uma escala aditiva similar para volume (0,5, 1,0, ..., 3,0) indo de cima para baixo na coluna da esquerda, mas também uma escala de medição derivada para densidade aumentando na diagonal (linha pontilhada) do canto inferior esquerdo da tabela para o canto superior direito.

TABELA 1.1 Cálculos da densidade de materiais em exercício de ciências na sala de aula

Massa	0,2 kg	0,4 kg	0,6 kg	0,8 kg	1,0 kg	1,2 kg
Volume						
0,51	,4	,8	1,2	1,6	2,0	2,4
1,01	,2	,4	,6	,8	1,0	1,2
1,51	,13	,27	,4	,53	,67	,8
2,01	,1	,2	,3	,4	,5	,6
2,51	,08	,16	,24	,32	,4	,48
3,01	,07	,13	,2	,27	,33	,4

Assim, os cientistas físicos têm medições fundamentais (que se concatenam fisicamente) e medições derivadas (detectadas ou construídas indiretamente) para cobrir os atributos físicos mensuráveis dos objetos. Em uma decisão pertinente para os cientistas sociais, o Comitê de Ferguson, em 1940, determinou que nada no mundo da quantificação psicológica tinha propriedades das escalas de medição físicas fundamentais ou derivadas (Ferguson, 1940). Nossa dependência atual dos quatro níveis de Stevens é indiretamente atribuível à crítica daqueles presentes no comitê, especialmente do proeminente físico britânico N. R. Campbell, que adotavam as ideias de medição da ciência física. Coincidente e infelizmente, as tentativas do próprio Stevens de medir a percepção de volume sonoro haviam atraído a atenção negativa do Comitê de Ferguson, e o trabalho seminal de Campbell foi fundamental para descartar sem mais considerações a pretensão à medição científica daqueles envolvidos no que chamaríamos frouxamente de psicometria.

A resposta de Stevens (1946, p. 667) foi redefinir medição em prol dos psicólogos: "Parafraseando N. R. Campbell (Final Report, p. 340), podemos dizer que a medição, no sentido mais amplo, é definida como a atribuição de numerais a objetos e eventos de acordo com regras". Stevens recorreu à autoridade de um dos críticos mais severos de sua posição usando o que parecia ser, à primeira vista, reformulações razoáveis de citações do próprio Campbell para justificar os agora muito conhecidos quatro níveis de medição de Stevens na psicologia: nominal, ordinal, intervalar e de razão – cada um dos quais constituía alguma forma de medição.

Medição conjunta

Robert Duncan Luce e colegas mostraram que tanto o físico Campbell como o psicometrista Stevens estavam errados. Luce e Tukey (1964) argumentaram que seu conceito de medição conjunta simultânea era um novo tipo de medição fundamental que incluía as categorias existentes de medição fundamental (por exemplo, peso, volume) e de medição derivada (por exemplo, densidade, temperatura) das ciências físicas e, o que é mais importante para nós, prepararam o caminho para a detecção de estruturas de medição em atributos não físicos, como os **construtos** psicológicos. Voltando à matriz de densidades da Tabela 1.1, a chave para a medição não reside na combinação de duas escalas de medições fundamentais de peso e volume para produzir uma terceira escala de medição derivada de densidade que conserve as propriedades cruciais da medição científica já inerentes em peso e volume. De acordo com Luce e Tukey (1964), o indicador crucial de uma estrutura de medição aditiva nos dados (para densidade e, possivelmente, também para alguns atributos psicológicos) está nas relações observáveis entre as próprias células da matriz, umas com as outras e com o conjunto delas.

Vejamos o potencial se pudéssemos aplicar algumas ideias da medição conjunta simultânea a uma ideia decorrente de testes educacionais ou psicológicos. Se tivermos algum indicador dos dois atributos (aspectos ou facetas) de uma situação de teste que podem ser ordenados de menos para mais– digamos, capacidade dos candidatos de menos capaz a mais capaz e dificuldade dos itens de menos difíceis a mais difíceis –, poderemos tentar verificar se as estruturas de medição de escala de nível intervalar de Luce e Tukey poderiam existir na matriz de dados resultante.

Imaginemos que a Tabela 1.2 seja baseada em um teste relevante de 100 itens aplicado a uma amostra de 100 pessoas apropriadas. Na linha superior, ordenamos alguns dos nossos itens do mais difícil à esquerda (1 pessoa correto/99 incorretos) ao mais fácil à direita (99 corretos/1 incorreto) e, na coluna da esquerda, ordenamos nossas pessoas do mais capaz (99 respostas corretas/1 incorreta) ao menos capaz (apenas 1 resposta correta/99 incorretas). Em essência, o que Luce e Tukey (1964) nos pedem é que procuremos na Tabela 1.2 os padrões de relação entre as células – os mesmos tipos de relações que são evidentes na matriz de resultados de peso/volume/densidade no exemplo da escala de medição derivada na Tabela 1.1.

TABELA 1.2 Pessoas ordenadas por capacidade (linha) e itens ordenados por facilidade (coluna)

Items / *Pessoas*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*	*Capacidade*
P												99/1
Q												90/10
R												80/20
S												70/30
T												60/40
U												50/50
V												40/60
W												30/70
X												20/80
Y												10/90
Z												1/99
Facilidade	1/99	10/90	20/80	30/70	40/60	50/50	60/40	70/30	80/20	90/10	99/1	

Os axiomas matemáticos pelos quais uma matriz de dados pode ser testada para verificar se satisfaz os requisitos das estruturas de medição conjunta simultânea são, como poderíamos imaginar, bastante complexos, mas uma ou duas ilustrações básicas podem ser suficientes para fins de uma introdução conceitual. Dedique um momento a examinar novamente a relação entre as células na matriz de densidade na Tabela 1.1 anterior para tentar encontrar:

(a) Qual é a relação entre qualquer célula e a que está à sua direita em qualquer linha?
(b) Qual é a relação entre qualquer célula e a que está acima dela em qualquer coluna?
(c) Qual é a relação entre qualquer célula e a que está diagonalmente à direita e acima dela?

Cada célula na tabela de densidade (Tabela 1.1 e ver Figura 1.1) tem um valor que é *menor* que

(a) o valor da célula à sua direita na linha (isto é, A < B);
(b) o valor da célula acima dela na coluna (isto é, C < A);
(c) o valor da célula diagonalmente à direita e acima dela (isto é, C < B).

Um pequeno esquema resumindo essas relações entre células pode ser representado como na Figura 1.1.

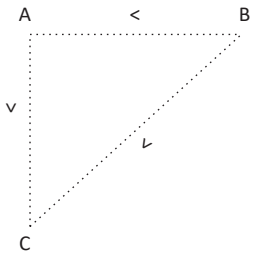

FIGURA 1.1 Relações entre células adjacentes em uma matriz de medidas.

Volte à Tabela 1.1 dos cálculos de densidade para confirmar que essas relações "menor que" se mantêm entre todos os pares adjacentes de células, conforme detalhado no modelo da Figura 1.1. Além disso, todas essas relações "menor que" existem simultaneamente. Mas esteja ciente de que os requisitos (axiomas) de Luce e Tukey para a medição são muito mais abrangentes do que isso.

O modelo de Rasch para medição

A formulação de Georg Rasch (1960) de seus *Probabilistic models for some intelligence and attainment tests* (*Modelos probabilísticos para alguns testes de inteligência e desempenho*), totalmente independente de Luce e Tukey, via a necessidade de um conjunto similar de relações na matriz de dados resultante do uso de um teste bem construído. O princípio que ele enunciou é agradavelmente direto:

> uma pessoa que tem maior habilidade que outra deve ter a maior probabilidade de solucionar qualquer item do tipo em questão, e, similarmente, um item ser mais difícil que outro significa que, para qualquer pessoa, a probabilidade de solucionar o segundo item é a maior (Rasch, 1960, p. 117).

Uma característica central do **modelo de Rasch** é uma tabela de probabilidades de **respostas esperadas** destinada a abordar a questão-chave: Quando uma pessoa com essa habilidade (número de itens de teste corretos) encontra um item dessa dificuldade (número de pessoas que conseguiram responder o item corretamente), qual é a probabilidade de ela dar a resposta correta ao item? Resposta: A probabilidade de sucesso depende da diferença entre a habilidade da pessoa e a dificuldade do item.

E isso é o que faz alguns adeptos da medição de Rasch serem realmente entusiastas quanto às possibilidades de medição científica para o modelo de Rasch: a tabela de probabilidades esperadas (ver Tabela 1.3 para um fragmento) gerada usando a formulação de Rasch tem o mesmo conjunto de relações "maior que/menor que" entre as células que a matriz de densidades na Tabela 1.1.

TABELA 1.3 Tabela de probabilidades de sucesso quando habilidade confronta dificuldade

Itens	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*	
Pessoas									→	fácil		Habilidade
P mais	,500	,866	,924	,949	,963	,973	,980	,986	,991	,995	,999	99/1
Q capaz	,134	,500	,653	,741	,801	,847	,884	,915	,942	,968	,995	90/10
R	,076	,347	,500	,603	,682	,746	,801	,851	,896	,942	,991	80/20
S	,051	,259	,397	,500	,585	,659	,726	,789	,851	,915	,986	70/30
T	,037	,199	,318	,415	,500	,578	,653	,726	,801	,884	,980	60/40
U	,027	,153	,254	,341	,422	,500	,578	,659	,746	,847	,973	50/50
V	,020	,116	,199	,274	,347	,422	,500	,585	,682	,801	,963	40/60
W	,014	,085	,149	,211	,274	,341	,415	,500	,603	,741	,949	30/70
X	,009	,058	,104	,149	,199	,254	,318	,397	,500	,653	,924	20/80
Y	,005	,032	,058	,085	,116	,153	,199	,259	,347	,500	,866	10/90
Z	,001	,005	,009	,014	,020	,027	,037	,051	,076	,134	,500	1/99
Facilidade	1/99	10/90	20/80	30/70	40/60	50/50	60/40	70/30	80/20	90/10	99/1	

O modelo de Rasch incorpora um método para ordenar pessoas (por exemplo, de uma amostra de crianças em idade escolar) de acordo com sua capacidade e ordenar itens (por exemplo, de um teste diagnóstico de cálculos numéricos) de acordo com sua dificuldade. O princípio de Rasch é que a medição de nível intervalar pode ser derivada quando os níveis de algum atributo aumentam junto com os aumentos nos valores de dois outros atributos. Na Tabela 1.3, podemos ver no fragmento que os níveis de um atributo (probabilidade de resposta correta) aumentam com os valores dos dois outros atributos: dificuldade (facilidade) do item e capacidade da pessoa. Assim, as relações puramente ordinais entre os níveis de probabilidades são indicativas de uma estrutura de medição quantitativa de nível intervalar para todos os três atributos. As relações do modelo da Figura 1.1 que se aplicavam para a tabela de densidades na Ta-

bela 1.1 também se aplicam às probabilidades de respostas esperadas de Rasch na Tabela 1.3. Na verdade, a relação requerida não é tão rigorosa para Rasch; é "igual ou menor que" (≤) em vez de apenas "menor que" (<). A exposição definitiva da teoria da medição conjunta simultânea é de Krantz, Luce, Suppes e Tversky (1971). Embora tanto Narens e Luce (1986) como Michell (1990) ofereçam explicações mais simples, a mais acessível é a de Michell (2003).

Uma analogia adequada para medição nas ciências humanas

Podemos entender facilmente por que Ben Wright usou uma régua simples (medição linear) para sua analogia de medição em vez da escala derivada de densidade. No entanto, embora os princípios de medição subjacentes à régua sejam o ideal a que nós, pesquisadores nas ciências humanas, poderíamos aspirar, usar a régua como a principal analogia tem algumas desvantagens que não são tão úteis para muitos no campo.

A régua pode ser usada como uma analogia para medição fundamental somente em situações em que os objetos em si podem ser fisicamente concatenados. Isso não só parece um pouco forçado para aplicar ao estudo de atributos humanos, como também é um princípio que não pode funcionar nem mesmo com medidas derivadas nas ciências físicas, como densidade e temperatura. A história da medição linear física é tão longa, e a medição usando uma régua é tão onipresente, que muitas vezes não vemos de imediato os paralelos entre os problemas com que se depararam os desenvolvedores de escalas e os problemas de medição que nós, nas ciências humanas, atualmente enfrentamos. Por exemplo, nós presumimos que o uso dessas medidas lineares padronizadas não é problemático e é rotineiramente "livre de erros", ainda que apenas um pouco de leitura e alguma reflexão pudessem revelar o contrário. Com a exceção de um único país importante, o uso do sistema de medição linear métrico é tão rotineiro que a progressão em calibrações de milímetros para centímetros e para metros e quilômetros é tida como óbvia – tudo de que precisamos é de uma régua suficientemente longa ou iterações suficientes de uma régua mais curta.

Seguindo e desenvolvendo as ideias de Bruce Choppin (1985), perguntamos: A termometria – a medição da temperatura – poderia ser uma analogia mais útil para aqueles de nós que estão tentando medir atributos humanos? A história do termômetro é muito mais recente; há relatos escritos de testemunhas oculares dos primeiros esforços de Hookes, Galileu e outros. O fascinante relato de 300 páginas de Chang (2004), *Inventing temperature: measurement and scientific progress*, pode ser visto como definitivo. Até o título parece promissor para os cientistas sociais: A temperatura foi inventada em vez de descoberta? Medição e progresso científico andam lado a lado? Então, quais

são os paralelos para nós? Lembre-se de seu projeto de medição em ciências humanas favorito ao refletir sobre o que vem a seguir. A temperatura não é medida diretamente; ela é estimada indiretamente registrando seus efeitos sobre outras substâncias, como mercúrio, álcool colorido e lâminas bimetálicas. Os cientistas têm teorias sobre a maneira como certas substâncias se comportam em resposta a mudanças de temperatura, e essas teorias mudaram com o tempo. Não usamos um único termômetro prototípico em todas as situações de medição de temperatura (ou nem mesmo na maioria). Na verdade, a maioria dos termômetros comuns tem uma faixa de escala bastante limitada e aplicações restritas: um termômetro médico de mercúrio é quase inútil na cozinha, a menos que se ache que o cozinheiro está doente. Sabemos que podemos nos mover facilmente entre graus Celsius (°C) e graus Fahrenheit (°F), e aqueles com um pouco de física no ensino médio podem falar sobre a escala Kelvin (K). Entendemos que 0°C e 100°C foram estabelecidos arbitrariamente por cientistas mais antigos nos pontos de congelamento e ebulição da água pura ao nível do mar para nossa conveniência, não porque essas sejam as extremidades científicas da escala de temperatura. Poderíamos começar a medir temperaturas em 0°C, 0°F ou 0 K, dependendo da nossa finalidade ou vontade. A teoria que postula 0 K (-273°C) como o zero absoluto também sugere que medir de fato a temperatura zero absoluto é, em si, impossível (Choppin, 1985).

Não temos problemas (além de sentirmos muito frio) quando temperaturas negativas são registradas: -10°C ou "sete abaixo de zero" não nos deixam perplexos quanto a ter uma quantidade negativa de temperatura! Nenhum princípio individual de construção de termômetros funciona em toda a escala de temperatura: expansão de líquidos, lâminas bimetálicas, mudanças na condutância elétrica, todos têm aplicações limitadas. Todos nós podemos usar alguns termômetros de maneira eficaz. Mas outros exigem conhecimento especializado e são usados apenas em ambientes especializados. Alguns termômetros bastante funcionais saem praticamente de graça; outros são caros, delicados e raramente sequer imaginados por nós, pessoas comuns. Sabemos que qualquer coleção de termômetros variará em suas leituras, ainda que no mesmo ambiente e ao mesmo tempo. Em situações de alto risco, quando nosso filho está muito doente, checamos a temperatura da criança duas ou três vezes em seguida para nos certificarmos de que medimos certo, mas, voltando do hospital para casa, confiamos que o termostato do carro vai ligar e desligar "mais ou menos" nas temperaturas corretas para manter o motor funcionando sem problemas. Esperamos receber de um termômetro o que pagamos por ele: alguns dólares para o motor do carro ou para temperatura na sala de estar; um pouco mais para controlar a temperatura na adega; e muito mais no centro cirúrgico para

monitorar a temperatura corporal durante grandes cirurgias. E que preço os governos deveriam estar dispostos a pagar pela medição da temperatura do núcleo de um reator nuclear?

Nossos esforços de medição nas ciências humanas talvez combinem melhor com a analogia da termometria. Enquanto a escala inteira (cf. temperaturas de 0 K a + ∞) poderia representar o desenvolvimento desde antes do nascimento (voltando mesmo até a concepção para alguns atributos) até a morte (e mesmo além para conceitos de espiritualidade), cada um de nós poderia trabalhar com apenas uma pequena parte da variável total de cada vez. As tarefas/testes/instrumentos que desenvolvemos provavelmente são tão especializados e tão diferentes quanto um termômetro médico infravermelho e o termostato em um Ford Mustang 1965. Alguns serão bastante baratos e terão consequências pouco importantes (como o termômetro doméstico); outros serão de alto custo, alta importância e alta manutenção (como o necessário para monitorar e controlar a temperatura do núcleo daquele reator nuclear, talvez). Reconheceremos imediatamente que a exatidão e a precisão do teste provavelmente serão dependentes de custo/esforço e que todas as estimativas necessariamente contêm erro, mas que essas qualidades (exatidão/precisão/erro) serão apropriadas para os requisitos das tomadas de decisão. Os nomes e tamanhos de nossas unidades de escala de medição podem variar (como acontece com graus Celsius, Fahrenheit e Kelvin). No entanto, embora muitos de nossos indicadores pareçam notavelmente diferentes e funcionem de maneiras obviamente diferentes em uma variedade de contextos aparentemente não relacionados, o objetivo final será a **calibração** do teste, lista de verificação e assim por diante, em uma única escala de medição de nível intervalar subjacente que tenha aplicabilidade geral em uma variedade de condições humanas (como no caso da temperatura). Para aqueles que precisam de um reforço de otimismo e motivação para enfrentar as tarefas à nossa frente, deem uma olhada em "Thermoscopes, thermometers, and the foundation of measurement" (Sherry, 2011). Sherry (2011) explica como essas lições derivadas da história da termometria podem ser utilizadas para nos orientar em nossos problemas de medição nas ciências humanas. Ele aborda especificamente a afirmação de que escalas de termoscópio ordinais evoluíram para escalas de termômetro intervalares por causa do trabalho experimental de Joseph Black (*c.* 1760).

Parece que os problemas que temos nas ciências humanas para desenvolver, padronizar e converter escalas de medição talvez sejam muito mais manejáveis quando vistos pela perspectiva dos principais avanços em conceitualização e medição da temperatura em meros séculos. E as deficiências de nossas tentativas de construção de escalas têm paralelos óbvios na variedade de termômetros que estão regularmente em uso, em-

bora muitos deles produzam leituras que mal são "boas o suficiente para o trabalho do governo"!

E isso nos leva a outra questão filosófica crucial para aqueles envolvidos em tentar medir as forças e deficiências humanas que se enquadram no título geral de "**traços latentes**". Lembremo-nos do termômetro simples de álcool colorido em um vidro: não medimos a temperatura diretamente, medimos seu efeito em outros objetos. Olhamos o comprimento do tubo de líquido vermelho e lemos a temperatura. Olhamos unidades de comprimento e *inferimos* unidades de temperatura. A afirmação paralela de Ben Wright seria algo assim:

> Se eu quiser medir a capacidade matemática de uma criança, tudo que tenho é o número de itens que ela marcou corretamente. Não é isso que quero. Então preciso ir *do que tenho e não quero* (uma pontuação) *para o que quero e não posso ter* (sua capacidade matemática). ISSO é chamado de "inferência".

Esse é o princípio subjacente às nossas tentativas de medir traços latentes, e o modelo de Rasch é o método que usamos para inferir medidas intervalares de traços latentes a partir de contagens brutas de itens de teste corretos.

Em conclusão

Há muito a ganhar aprendendo e refletindo sobre os problemas que afligem o desenvolvimento e o uso científico de escalas de medição nas ciências físicas. Relatos populares de Alder (2002), Keay (2000) e Sobel (1996) são fáceis de ler e informativos. Muitos dos capítulos de Chang (2004) são acessíveis ao metrologista amador; e Sherry (2011) pode ser exatamente a lição prática de que precisamos. É reconfortante saber que não estamos sozinhos em nossos problemas. E isso também ajuda a entender um pouco como acabamos ficando tão aquém das nossas expectativas razoáveis de medição científica nas ciências humanas. Michell (1999) oferece um relato muito legível dos principais atores, eventos e motivações aparentes. Ele também apresenta uma introdução concisa da relação entre a medição científica e as ideias de Luce e Rasch (Michell, 2003). Embora muitos que defendem a medição de Rasch não concordem com todos os prognósticos de Michell sobre a medição em nosso campo, ele certamente aborda muitas questões importantes que temos em comum, mas que raramente são consideradas em outros fóruns em nossa disciplina (ver Bond, 2001).

Neste capítulo, o termo "fundamental" foi usado de duas maneiras diferentes, mas igualmente importantes. A medição do tipo que usamos em nossa vida diária – escalas com valores unitários iterativos – é *fundamental* para

pesquisas lógicas de base empírica nas ciências humanas. As propriedades da medição científica são mais óbvias no que é chamado de *medição fundamental*, em que atributos como peso e comprimento podem ser concatenados fisicamente ao longo da escala de medição. Muitas escalas de medição nas ciências físicas são derivadas, de forma que, embora as unidades de medição possam ser iteradas, o atributo em si (por exemplo, temperatura e densidade) não pode ser fisicamente somado.

Luce e colegas descreveram os princípios e as propriedades da medição conjunta que trariam às ciências humanas o mesmo tipo de medida rigorosa de que as ciências físicas já desfrutavam havia um tempo considerável. De fato, os sistemas de medição fundamental e derivada das ciências físicas são casos especiais (restritos) de medição conjunta; Luce chamou a medição conjunta de "um novo tipo de medição fundamental" (Luce & Tukey, 1964).

Os modelos de Rasch para medição são atualmente a aproximação amplamente acessível mais próxima, nas ciências humanas, desses princípios de medição fundamental. É claro que a adoção de uma abordagem tão obviamente chauvinista enfrenta o risco de deixar de fora muitos de nossos colegas que usam abordagens quantitativas em pesquisas em ciências humanas. Mesmo aqueles completamente dedicados ao desenvolvimento e uso de escalas calibradas de Rasch às vezes nos imploram para pisar mais leve, ser mais circunspectos, parecer menos assertivos. A defesa que os autores fazem da abordagem do modelo de Rasch para a medição não pretende ser ofensiva; talvez confrontadora, mas nunca ofensiva. É encorajador que colegas que têm padrões tão elevados para medir água, tecidos e farinha possam ter agora as ferramentas para alcançar esses mesmos padrões na medição de desempenho matemático, introversão, desenvolvimento cognitivo ou qualidade de vida relacionada à saúde.

Este livro é um convite (exortação?) ao estabelecimento de padrões impossivelmente altos para a medição na pesquisa em ciências humanas e ao trabalho incremental para atingir esses padrões. Devemos lembrar que aterrissar na Lua já foi, certa vez, uma meta aparentemente inalcançável.

Resumo

Neste capítulo, construímos as bases para entender a diferença crucial entre medição e estatística. Argumentamos que a dependência nas ciências sociais da estrutura de Stevens para o que constitui medidas e suas análises estatísticas correspondentes levou a uma ignorância disseminada da medição e a um excesso de confiança na significância estatística para interpretar pesquisas. Esse descuido, intencional ou não, teve graves consequências para a compreensão dos fenômenos sob investigação.

Leituras sugeridas

Introdutórias (você deve ler estes)

Bond, T. G. (2001). Book review *Measurement in psychology: A critical history of a methodological concept*. *Journal of Applied Measurement*, *2*(1), 96-100.

Choppin, B. H. L. (1985). Lessons for psychometrics from thermometry. *Evaluation in Education*, *9*(1), 9-12.

Michell, J. (2003). Measurement: A beginner's guide. *Journal of Applied Measurement*, *4*(4), 298-308.

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

Para estudantes mais avançados de medição

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. New York: Cambridge University Press.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundation of measurement. *Studies in History and Philosophy of Science*, *42*, 509-524.

ATIVIDADES

Leia Stevens (1946), em particular sua seção sobre escalas ordinais, e os avisos de cautela que ele faz em relação às suas análises estatísticas. "Na verdade, a maioria das escalas usadas ampla e efetivamente por psicólogos são escalas ordinais" (Stevens, 1946, p. 679).

Essa visão das ideias de Stevens combina com a versão que você entendia antes de ter lido Stevens nas próprias palavras dele? Por quê? Por que não?

Encontre um artigo de pesquisa importante em sua área e identifique como ele aborda as ideias de estatística e medição e os alertas de cautela de Stevens.

Leia Choppin (1985) observando os paralelos e as diferenças entre princípios de termometria e medições nas ciências humanas.

Complete uma tabela como a que segue resumindo os paralelos e as diferenças:

Termometria	*Ciências humanas*
Muitos termômetros	Muitos instrumentos
Escalas em sua maior parte intervalares	Escalas em sua maior parte ordinais
Algumas escalas de razão	Algumas escalas intervalares

Referências

Alder, K. (2002). *The measure of all things*. New York: Free Press.

Bond, T. G. (2001). Book review *Measurement in psychology: A critical history of a methodological concept*. *Journal of Applied Measurement*, *2*(1), 96-100.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. New York: Oxford University Press.

Choppin, B. H. L. (1985). Lessons for psychometrics from thermometry. *Evaluation in Education*, *9*(1), 9-12.

Ducret, J. J. (1990). *Jean Piaget: Biographie et parcours.* Neuchatel: Delachaux et Niestle.

Ferguson, A. (1940). Quantitative estimates of sensory events. The advancement of science. *Report of the British Association for the Advancement of Science*, *2*, 331-349.

Keay, J. (2000). *The great arc: The dramatic tale of how India was mapped and Everest was named*. New York: Harper Collins.

Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement, volume 1: Additive and polynomial representations*. New York: Academic Press.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1-27.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355-383.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. New York: Cambridge University Press.

Michell, J. (2003). Measurement: A beginner's guide. *Journal of Applied Measurement*, *4*(4), 298-308.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166-180.

Piaget, J., Inhelder, B., & Szeminska, A. (1960). *The child's conception of geometry*. London: Routledge and Kegan Paul. (Obra original publicada em 1948).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in the History and Philosophy of Science*, *A*, *42*, 509-524.

Sobel, D. (1996). *Longitude: The true story of a lone genius who solved the greatest scientific problem of his time*. Harmondsworth, UK: Penguin.

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

Stone, G. E. (2002, April). The emperor has no clothes: What makes a criterion-referenced standard valid? *International Objective Measurement Workshop*, New Orleans, LA, 5.

Stone, M. H. (1998). Man is the measure... the measurer. *Journal of Outcome Measurement*, *2*(1), 25-32.