REMark – Revista Brasileira de Marketing
e-ISSN: 2177-5184
DOI: 10.5585/remark.v13i2.2685
Data de recebimento: 08/01/2014
Data de Aceite: 17/03/2014
Editor Científico: Otávio Bandeira De Lamônica Freire
Avaliação: Double Blind Review pelo SEER/OJS
Revisão: Gramatical, normativa e de formatação

MENSURAÇÃO EM MARKETING: ESTADO ATUAL, RECOMENDAÇÕES E DESAFIOS

RESUMO

Este artigo tem por finalidade debater o tema de mensuração de construtos em Marketing, sumarizando as principais discussões sobre o assunto. Inicialmente, discutimos a origem das preocupações e os desdobramentos na área desde os anos de 1970. Em seguida, apresentamos os principais modelos consolidados (modelo clássico de Churchill, modelo COARSE e modelo de mensuração formativa). Na sequência, apresentamos preocupações atuais que se somam à teorização clássica, com algumas recomendações relevantes (especialmente sobre mensuração por múltiplos itens, mensuração por um único item, escalas de verificação e aspectos transculturais). Ao final, apresentamos considerações sobre tendências de mensuração em Marketing, com ênfase em Teoria da Resposta ao Item, operadores Bayesianos e estimação por mínimos quadrados parciais. O artigo atualiza o debate sobre o tema e tem a possibilidade de contribuir para estudiosos e pesquisadores de Marketing que demandem uma visão atual sobre mensuração e recomendações para pesquisas.

Palavrachave: Mensuração; Escalas; Validação; Confiabilidade.

MEASUREMENT IN MARKETING: CURRENT SCENARIO, RECOMMENDATIONS AND CHALLENGES

ABSTRACT

The purpose of this article is to discuss about construct measurement in Marketing by summarizing the main considerations about the subject. First, it discusses the origins of the debates about the theme since the 1970s and describes its main consolidated models (the classical Churchill's model, the COARSE model and the formative measurement model). Then it presents current concerns about the classical approach with relevant recommendations (particularly regarding multi-item measurement, single-item measurement, rating scales and cross-cultural aspects). At the end, it presents considerations about measurement trends in Marketing with emphasis on the Item Response Theory (IRT), Bayesian estimators and Partial Least Squares (PLS). The article updates the debate on the theme and contributes to Marketing experts and researchers who demand a current view about measurement and recommendations for research development.

Keywords: Measurement in Marketing; Marketing Scales; Validity; Reliability.

Felipe Zambaldi[1]
Francisco José da Costa[2]
Mateus Canniatti Ponchio[3]

[1] Doutor em Administração de Empresas pela Fundação Getulio Vargas - FGV. Professor da Fundação Getulio Vargas – FGV, Brasil. E-mail: felipe.zambaldi@fgv.br

[2] Doutor em Administração de Empresas pela Fundação Getulio Vargas – FGV. Professor da Universidade Federal da Paraíba, UFPB, Brasil. E-mail: franzecosta@gmail.com

[3] Doutor em Administração de Empresas pela Fundação Getulio Vargas – FGV. Professor da Escola Superior de Propaganda e Marketing de São Paulo (ESPM-SP), Brasil. E-mail: mponchio@espm.br

1 INTRODUÇÃO

O processo de construção do conhecimento científico depende, em grande parte, da capacidade dos pesquisadores de mensurarem adequadamente os conceitos abordados em seus estudos. Diferentemente de algumas áreas da ciência em que a maior parte dos conceitos trabalhados pode ser diretamente observada (tais como altura, peso e idade), nas ciências sociais em geral, e em Marketing em particular, frequentemente nos deparamos com construtos de natureza mais abstrata e que não podem ser diretamente acessados, como é o caso de satisfação, lealdade, felicidade, materialismo e atitude à marca.

Mensurar valores, crenças e atitudes depende, em um primeiro momento, de um grande esforço de definição conceitual e delimitação do construto (por exemplo, o que exatamente estamos querendo dizer quando utilizamos o termo satisfação?). Num segundo momento, supondo superada a barreira da comunicação representada pela clareza conceitual, necessitamos de uma estratégia de mensuração. Nosso objetivo deverá ser o de posicionar unidades de análise (produtos, consumidores, e empresas, por exemplo) num eixo de acordo com a posse de menos ou de mais de determinada característica de interesse que esteja sendo mensurada, ou seja, precisamos definir um sistema de indicação de intensidade (ou quantidade) para o construto que previamente definimos.

Vejamos um exemplo simples, relacionado à medição de inteligência. Mesmo que consigamos chegar a uma definição consensual sobre o que é inteligência (basta uma rápida revisão na literatura para encontrar visões complementares sobre o construto), não é possível observar rótulos nos braços das pessoas contendo indicações da sua quantidade de inteligência. Esse conceito de natureza latente (está presente no objeto, mas não o observamos) não pode ser diretamente medido e, portanto, deve ser acessado por meio de estratégias indiretas de mensuração.

Neste artigo, procuramos fazer um *tour de force* sobre as práticas de mensuração em Marketing, campo tipicamente interessado em atribuir valores a conceitos não observáveis diretamente para posterior operacionalização estatística dos dados gerados para análise de hipóteses envolvendo os construtos. Para tanto, inicialmente posicionamos o problema da mensuração de construtos abstratos e latentes sob os pontos de vista histórico e atual, apresentando as abordagens conhecidas como clássicas e os avanços mais recentes, de forma a introduzir o debate contemporâneo sobre o tema. Em seguida, provemos a descrição de procedimentos e recomendações na construção de escalas, com destaque para a aferição de validade e confiabilidade dos instrumentos e depois para lidar com escalas de verificação (coerência entre escala e conteúdo; número de pontos; estratégia de agregação; e uso de técnicas estatísticas), preocupando-nos em fornecer alternativas para o uso de indicadores formativos, além dos refletivos, mais comuns na literatura. Particularmente, buscamos munir os leitores com conteúdo para prosseguirmos com uma discussão sobre cuidados para a elaboração e uso de escalas em estudos interculturais, com destaque à necessidade de adaptações de escalas quando aplicadas em contextos distintos, e levantamos tendências em mensuração em Marketing motivadas pelos debates atuais e suas respostas frente às fragilidades dos modelos mais comuns, abordando especificamente a Teoria de Resposta ao Item, os estimadores Bayesianos e os modelos de *Partial Least Squares*. Concluímos por meio de considerações e reflexões sobre o material que apresentamos.

2 O PROBLEMA HISTÓRICO E ATUAL DA MENSURAÇÃO DE CONSTRUTOS ABSTRATOS E LATENTES

Em um interessante artigo que se propôs a dar uma visão geral da Estatística, Pereira (1997) realçou que a mensuração é um dos elementos centrais do processo estatístico (que o autor defendeu ser a 'tecnologia da ciência'). Na visão de Pereira, o processo científico convencional, que desenvolve a avaliação empírica para análise de proposições e hipóteses, passa sucessivamente pela decisão de mensuração das variáveis de interesse no campo empírico, em seguida pela captação dos dados a partir da escala de mensuração utilizada, e depois pela análise desses dados, etapa na qual são aplicadas diversas técnicas estatísticas disponíveis.

A estrutura da pesquisa acima relatada, assim também considerada em outros autores (ver Pedhazur & Schmelkin, 1991), coloca para os pesquisadores a necessidade de considerarem esses procedimentos (mensuração, *design* e análise) como um roteiro de referência para construção do conhecimento. Ao que nos parece, a ênfase geral da pesquisa nas ciências sociais e comportamentais privilegiou historicamente a dimensão de análise, com maior atenção nas técnicas de análise estatística.

Internacionalmente, a consideração da mensuração como parte central do processo de pesquisa quantitativa em Marketing alcançou um status diferenciado a partir dos anos 1960. Já no Brasil, essa tendência parece ter se consolidado mais recentemente, a partir dos anos 2000, como natural evolução da pesquisa de orientação mais acadêmica que as escolas de Administração adotaram desde então. A análise de mensuração é atualmente requerida na maior parte dos relatos de pesquisa apresentados nas formas de dissertações, teses e artigos.

Na verdade, o campo da pesquisa em Marketing absorveu uma preocupação há anos recorrente nos campos de Educação e Psicologia, contextos em que a mensuração é objeto de estudo e aprofundamento teórico e operacional há mais de um

século. A razão desta aproximação é simples: trabalhamos em Marketing com construtos abstratos (como satisfação, identidade, apego, lealdade...) para os quais pressupomos uma intensidade mensurável, mas para os quais não temos ainda instrumentos de acesso direto dessa intensidade. O mesmo ocorre na suposição de existência de uma intensidade mensurável em estresse (em Psicologia) ou conhecimento e aprendizado (em Educação), por exemplo; em todos os dados, não temos um instrumento que alcance diretamente estes construtos. Ou seja, pesquisamos em Marketing construtos latentes, que requerem uma estratégia de medição própria e diferenciada daquelas utilizadas, por exemplo, na área de Finanças para medir lucro, ou na área de Produção para medir defeitos de qualidade.

Absorvemos em Marketing a maior parte do conteúdo substantivo da teoria da mensuração da Psicologia e da Educação, para viabilizar, mais recentemente, uma contribuição mais própria do nosso. Para construir um referencial de base do que atualmente já temos consolidado em Marketing, expomos rapidamente algumas informações sobre o tema nestes dois campos.

Em Psicologia, o problema da mensuração vem desde quando profissionais da área optaram por desenvolver testes (métricas) para avaliação de seus construtos e variáveis. O campo da testagem psicológica (ver Urbina, 2004) e a disciplina de Psicometria buscam desenvolver testes e métodos desde o final do século XIX, na tentativa de medir, por exemplo, valores pessoais, tendências profissionais ou predisposição a determinados comportamentos, por meio de instrumentos de lápis e papel (ou equivalentes, como os atuais instrumentos digitalizados e aplicados via *internet*). Neste campo encontra-se um dos principais periódicos acadêmicos de mensuração, que é a revista *Psycometrika*, fundada em 1936 e já acumulando diversas contribuições teóricas que transbordam o próprio campo da Psicologia.

Já em Educação, o problema da mensuração alcança a grande maioria das pessoas escolarizadas, uma vez que as conhecidas provas escolares são na verdade instrumentos de medição de aprendizagem que os professores aplicam durantes suas disciplinas. Nesse campo, a mensuração é relatada como parte central da área especializada de Avaliação educacional, que inclui tanto a avaliação de aprendizagem de conhecimentos transferidos por docentes, quanto a avaliação de competências (como nos concursos públicos) e a avaliação de programas e instituições (como as avaliações institucionais e as avaliações de cursos e programas de pós-graduação). Foi no campo da Educação que se desenvolveram mais recentemente os principais estudos de Teoria da Resposta ao Item (TRI), comentada posteriormente.

Embora não seja seguro fixar uma data de referência, podemos afirmar que, em Marketing, o primeiro grande passo para a definição de uma prioridade da questão de mensuração vem do artigo de Gilbert Churchill, publicado em 1979 no prestigiado *Journal of Marketing Research*, e que trouxe uma crítica bem fundamentada das práticas então vigentes na área, que, segundo o autor, eram extremamente frágeis. O alerta da época era simples, mas até hoje atual: não é possível acreditar no valor de uma operacionalização de números (ou seja, nas técnicas de análise) se não sabemos ao certo o que está por detrás desses números (ou seja, nas decisões de mensuração e de *design* para a sua captação).

Churchill resgatou toda a construção anterior que já se fazia dentro da academia de Marketing[4], Psicologia e Educação, e propôs um passo a passo a ser usado por pesquisadores na construção de métricas. Seu modelo vem sendo recorrentemente citado e utilizado em pesquisas em Marketing (na ocasião de desenvolvimento deste artigo havia mais de 9600 citações no Google Acadêmico), mas não esteve isento de limitações e críticas.

Na realidade, o modelo de Churchill está orientado a desenvolver medidas segundo alguns pressupostos que, se não são considerados válidos, podem ser motivo de proposição de outros modelos de construção de métricas. As críticas centrais vieram do seguinte: sob o pressuposto de mensuração segundo a dita 'teoria da amostra de domínio', são utilizados sempre múltiplos indicadores para medir um construto, e a análise de validação pode ser feita por meio de técnicas como análise fatorial (para identificação ou reafirmação da existência de um fator subjacente – o construto latente – explicando a variação dos itens), e do coeficiente alfa de Cronbach (para atestar a consistência interna do conjunto de itens). Pela negação deste pressuposto (ou de sua aplicação generalizada), vieram os desenvolvimentos de mensuração por um único item e com análise qualitativa da validade (principalmente defendido por John Rossiter em seu modelo COARSE), ou a mensuração formativa, em que não se supõe um fator subjacente explicando a variação de um conjunto de indicadores, mas supondo, inversamente, que é a variação dos itens que implica a variação do construto formado (há diversos defensores dessa controversa tese, merecendo destaque o texto de Diamantopoulos & Winklhofer (2001).

Pela análise de pesquisas e publicações recentes, podemos afirmar que o cenário atual focaliza o debate destas três perspectivas: modelo clássico (com inspiração no modelo de Churchill, 1979); mensuração formativa; e mensuração isenta de maiores elaborações quantitativas e com maior foco na validação qualitativa. Outros desenvolvimentos parecem ser a pauta de pesquisa e aplicação futura, com a expansão do uso da Teoria da Resposta ao Item, de ampla utilização no campo da avaliação educacional, e chegando aos

[4] Já em 1965, Charles Lee debatia a questão da mensuração no contexto mais amplo da pesquisa quantitativa e suas dificuldades e especificidades; cf. Lee (1965).

poucos no universo de Marketing (ver Andrade, Tavares & Valle, 2000; Lucian, 2012).

3 O PROBLEMA DA CONSTRUÇÃO DE ESCALAS: ALTERNATIVAS CLÁSSICAS E AVANÇOS

Para ilustrar particularidades na construção do conhecimento em variados campos científicos, Mari (2005) comparou o emprego de axiomas na ciência formal (citando como exemplo a geometria euclidiana na qual os elementos de construção da teoria estão alicerçados em axiomas) à dependência da mensuração de fenômenos da ciência empírica. O autor argumenta que, nas ciências empíricas, coexistem entre os cientistas diferentes entendimentos epistemológicos em relação à mensuração ou mesmo à possibilidade de atribuição de um número a um fenômeno.

3.1 O modelo clássico de Churchill

Particularmente nos estudos em Marketing, predomina, desde 1979, a proposta realizada por Gilbert Churchill e seus desdobramentos, compondo o que conhecemos como abordagem clássica da mensuração em Marketing. Conforme indicado acima, os procedimentos propostos por Churchill foram motivados por sua percepção de que os esforços de mensuração no campo tendiam a ser carentes em termos de rigor. Nesse contexto, o autor apresentou definições para validade e confiabilidade, seguramente as duas mais fundamentais no processo de validação de instrumentos de medidas. As definições fornecidas de Churchill para validade e confiabilidade ainda são adotadas pela maior parte dos pesquisadores em Marketing.

O autor define como validade a capacidade de uma medida capturar em seus escores o fenômeno sob análise sem ruídos, e como confiabilidade a propriedade de medidas de um mesmo construto serem concordantes entre si. Ou seja, a validade concerne a assegurar que a escala mede o que interessa medir, e a confiabilidade concerne a desenvolver esta medida com o mínimo de erros (que são esperados no processo científico, mas que precisam ser minimizados).

A proposta de Churchill para validação de medidas consiste em passos sequenciais, alguns dos quais podem ser realizados mais de uma vez ao longo do mesmo processo. O primeiro passo se refere a especificar o domínio teórico do construto, ou defini-lo teoricamente, e deve ser realizado com base em revisão de literatura. Em seguida, o autor propõe a geração de um conjunto de itens (questões) que constituirão a primeira versão do instrumento de medida. Essa etapa é dependente da anterior (especificação do domínio de construto) e se realiza com base na revisão da literatura, na consulta a estudos empíricos já publicados, na criação de exemplos e incidentes relevantes ao domínio conceitual e em pesquisas qualitativas com respondentes-chave, realizadas por meio de grupos de foco, por exemplo. Em posse do primeiro conjunto de itens, faz-se uma coleta de dados para um pré-teste. Com seus resultados, procede-se à etapa de purificação do instrumento com o intuito de verificar quais itens devem permanecer e quais itens devem ser excluídos ou adaptados. As ferramentas propostas por Churchill para essa etapa são o cálculo do coeficiente alfa de Cronbach como medida de confiabilidade e também a análise fatorial exploratória, que pode indicar confiabilidade quando as cargas fatoriais dos itens que medem o construto forem altas, além de auxiliar o pesquisador a compreender as diferentes dimensões presentes no instrumento que está desenvolvendo (se houver mais do que uma). A purificação também pode ser feita por meio da análise fatorial confirmatória (que Churchill prefere, por presumir que as etapas anteriores se realizam de forma rigorosa e permitem a formulação prévia sobre a dimensionalidade do instrumento de medida).

A etapa de purificação pode levar os pesquisadores de volta ao passo da geração do conjunto de itens e a alterações no primeiro conjunto proposto. Com um novo conjunto de itens em mãos, procede-se a uma nova coleta de dados e a uma nova depuração, o que pode se repetir até que o pesquisador considere ter uma medida confiável e que bem represente as eventuais dimensões do construto. Esse processo, no entanto, pode ser muito custoso e representar algum desperdício das unidades amostrais, já que muitas coletas de dados não são definitivas. Após o pesquisador obter uma purificação satisfatória, segue nova coleta de dados, essa definitiva, sobre a qual se verifica a confiabilidade novamente por meio do coeficiente alfa ou, alternativamente, por meio da divisão do instrumento em dois conjuntos de itens diferentes e da apuração do grau de associação entre eles, ou ainda da confiabilidade teste-reteste, que consiste em aplicar o instrumento ao mesmo grupo de respondentes em dois momentos distintos e comparar seus resultados. Churchill considera preferível, no entanto, o uso do alfa de Cronbach.

A coleta definitiva também se presta ao teste de validade de construto. Para aferir validade convergente e validade discriminante, a recomendação de Churchill é o uso da Matriz Multitraço Multimétodo, que consiste em verificar associações entre traços (construtos) obtidos por diferentes métodos, ou seja, com aplicação da mensuração por diferentes instrumentos, diferentes formas e momentos de coleta, e até diferentes amostras. A matriz formada por estes procedimentos torna-se um instrumento que provê comparações entre: 1) a variação comum contida dentro de uma escala com diversos itens para um mesmo construto, coletados pelo mesmo método; 2) a associação entre as medidas de um mesmo construto obtidas por diferentes métodos; 3) a associação entre diferentes construtos obtidas por um método comum; e 4) a associação entre diferentes construtos obtidos por métodos distintos. O sentido de fazer essas

comparações é que, quando há uma alta variação comum entre os itens de um mesmo construto, há validade convergente, ou seja, eles convergem para uma medida comum. Essa variação comum deve ser maior do que as associações dessas medidas com diferentes construtos obtidos por diferentes métodos e maior do que as associações entre diferentes construtos obtidos por meio de um mesmo método.

Além disso, é esperado que a associação entre um mesmo traço (construto) coletado por diferentes métodos deve ser maior do que a associação entre traços distintos, sejam eles coletados pelo mesmo método ou não. Quando essas condições são satisfeitas, obtemos evidências de haver validade discriminante, ou seja, de fato temos medidas diferentes para construtos distintos. É comum usarmos o coeficiente de correlação de Pearson para medir as associações propostas. A variação comum entre os itens do construto costuma ser obtida por meio da análise fatorial (embora essas sejam medidas de associação linear, seu uso apresenta resultados satisfatórios, em geral).

Churchill também propõe que se verifique a validade de critério para garantir a validade de construto. De maneira breve (vamos nos aprofundar nesse assunto adiante), a validade de critério se observa quando verificamos uma associação esperada, preferencialmente significante, entre a medida para o construto que estamos validando e outras medidas (em geral de operacionalização mais consolidada) às quais devem se associar do ponto de vista teórico. Se a validade de construto (em seus diversos subtipos) não for alcançada, a proposta de Churchill é recomeçar o processo do início, desde a especificação de domínio do construto.

Quando, finalmente, obtemos uma indicação segura da validade de construto, Churchill propõe que a medida seja apresentada por meio de estatística descritiva da sua distribuição na amostra. Os procedimentos propostos por Churchill e alguns desdobramentos sugeridos em trabalhos posteriores têm sido amplamente adotados pelos pesquisadores da área de Marketing (por exemplo, Netemeyer, Bearden & Sharma, 2003; Costa, 2011).

No entanto, sua aplicação rigorosa é muitas vezes inviável por conta da necessidade de várias coletas de dados, o que pode esbarrar em limitações de tempo e em limitações orçamentárias, e também na dificuldade de se coletarem dados por métodos distintos, o que inibe o uso da Matriz Multitraço Multimétodo.

3.2 Uma alternativa ao modelo clássico: o modelo COARSE

A proposta de Churchill recebeu muitas críticas daqueles que a consideram muito enfática em termos de ajustes estatísticos frente às etapas qualitativas da validação, além de ser dependente dos pressupostos do coeficiente alfa como medida de confiabilidade e da análise fatorial como técnica para verificação de validade. Ademais, os procedimentos se prestam ao desenvolvimento de escalas de múltiplos itens, sob o pressuposto de que estes variam por conta da variação do construto latente (ou seja, têm relação refletiva com o construto). Diante de tais críticas, John Rossiter desenvolveu uma proposta alternativa em 2002, o modelo COARSE, privilegiando os procedimentos qualitativos na validação dos instrumentos de medida.

A sigla COARSE refere-se a seis passos que o pesquisador deve seguir de acordo com o modelo: *Construct definition*; *Object classification*; *Attribute classification*; *Rater Identification*; *Scale formation*; e *Enumeration*. Em português, temos: Definição do construto; Classificação do objeto; Classificação do atributo; Identificação do avaliador; Formação da escala; e Enumeração. O modelo está bem detalhado em Rossiter (2011), e apresentamos a seguir esses passos que, no mínimo, são referência de aprimoramento para eventuais limitações do modelo clássico de Churchill.

O primeiro passo, o de definição do construto, consiste em escrever uma definição em termos de objeto, atributo e entidade avaliadora. O objeto é o foco da medida como, por exemplo, uma propaganda. O atributo é o que será medido no objeto como, por exemplo, as reações afetivas à propaganda; e a entidade avaliadora é quem fará a avaliação do objeto e do atributo como, por exemplo, um grupo de consumidores-alvo.

Partimos então para o segundo passo, a classificação do objeto, que conta com entrevistas abertas com respondentes-chave. O objeto pode ser classificado como concreto simples, abstrato coletivo ou abstrato formado. Um objeto concreto é aquele que qualquer respondente conhece o significado e o reconhece, como, por exemplo, o conceito de controle de qualidade de serviços. Objetos abstratos coletivos são heterogêneos aos olhos dos respondentes-chave, mas compõem uma categoria clara aos olhos do pesquisador, como por exemplo, bebidas com gás (como refrigerantes, águas gaseificadas com sabor, ou água com gás). Os objetos abstratos formados são aqueles cuja interpretação variam perceptivelmente entre pessoas e são vistos como portadores de diferentes componentes como, por exemplo, pode ser o conceito de capitalismo. Se o objeto for classificado como concreto, um único item basta para medi-lo. Para os objetos abstratos, múltiplos itens são necessários. Nessa etapa, começamos a escrever os itens do instrumento de medidas, para que reflitam o objeto.

O terceiro passo é a classificação de atributos, também com base em entrevistas abertas com respondentes-chave. Os atributos se classificam como concretos, formados ou suscitados (*eliciting*). Os concretos são aqueles cuja interpretação é praticamente unânime entre respondentes, como o conceito de intenção de compra, por exemplo. Os formados são abstratos e o que os caracteriza é a soma de uma série

de componentes que, se somados em alguma combinação, os formam (e são por isso chamados de formativos); um exemplo pode ser o conceito de liderança. Os suscitados, por sua vez, também são abstratos, mas são traços internos dos respondentes que podem causar as respostas aos itens do instrumento de medida (que são indicadores da manifestação do atributo, na literatura convencional chamados de refletivos). Um exemplo pode ser o envolvimento pessoal de alguém com algo. Na classificação de atributo, continuamos a escrever os itens do instrumento, usando a estratégia de item único para atributos concretos e de múltiplos itens para os abstratos (formados e suscitados). Após esse passo, é possível voltar à definição do construto e incluir nela os componentes de objeto e de atributo identificados nas fases de classificação.

O quarto passo é identificar a entidade avaliadora, ou o grupo de pessoas que julgará os itens do instrumento de medida. Em outras palavras, esse passo consiste em identificar detalhadamente os respondentes. Para essa etapa, é importante que especialistas tenham avaliado os resultados dos passos anteriores e participado de seu aprimoramento. Nesse passo também definimos se será necessário estimar confiabilidade entre respondentes, e entre itens de atributos suscitados.

O quinto passo é a formação da escala. Aqui, combinamos os textos que contêm os componentes do objeto e os atributos para geração dos itens. Selecionamos os tipos de escala que serão usados, tendo como insumo as entrevistas abertas previamente realizadas com os respondentes-chave, e realizamos um pré-teste com respondentes pertencentes à população de interesse, visando a garantir que as formulações dos itens sejam compreensíveis. Em caso de atributos suscitados, testamos sua unidimensionalidade. Por fim, se o instrumento for de múltiplos itens, embaralhamos a ordem de sua apresentação, mesclando as sequências de componentes distintos dos atributos e do objeto, para evitar reconhecimento por parte dos respondentes e assim evitar que assumam um padrão de resposta induzido pelo instrumento.

O último passo é a enumeração, que consiste em construir os escores da escala (estratégia de agregação) com base em índices ou médias; transformá-los em pontuações com sentido interpretativo, como pontuações de 0 a 10, ou de -5 a 5, no caso de atributos bipolares; e reportar a confiabilidade da escala.

A proposta de Rossiter com o modelo COARSE foi bem recebida por valorizar aspectos qualitativos e conceituais da mensuração, além de expandir o leque de métodos para além da análise fatorial e do uso do alfa de Cronbach, incorporando a possibilidade de adoção de indicadores únicos e formativos. No entanto, embora a proposta tenha trazido luz ao debate ao incorporar elementos não considerados pela abordagem clássica (marcada pelo uso da análise confirmatória, de índices de confiabilidade e, posteriormente, de modelos de equações estruturais), a operacionalização desses elementos é ainda um desafio por conta de limitações relativas ao repertório metodológico dos pesquisadores em marketing, aos recursos computacionais disponíveis e às propriedades das técnicas propostas.

Entendemos que a força dos argumentos de Rossiter está menos no seu conjunto de passos (por vezes confuso), mas na orientação intensiva para a validade de conteúdo, que se faz em detalhes minuciosos, chegando até a um detalhamento dos respondentes, e com acompanhamento continuado de especialistas no construto de interesse.

3.3 Relativizando a refletividade: a mensuração formativa

Os debates atuais sobre mensuração em nossa área permanecem diante de uma série de questões ainda em debate, todas motivadas pelo fato de que as variáveis de interesse em Marketing costumam ser latentes e de mensuração indireta. Tomemos como exemplo o medo. Sabemos que existe, sabemos do que se trata, mas não temos como medir diretamente o medo de uma pessoa; no máximo podemos observar sintomas do medo que alguém sente ou pedir que essa pessoa manifeste de alguma forma, talvez com palavras ou testes, se sente medo, e quanto. Ou seja, podemos observar o medo indiretamente, por meio de indicadores que nos permitem inferir o quanto de medo há em alguém. Em grande parte dos casos, usamos múltiplos indicadores para fazer essa estimativa a respeito do valor de um construto latente. Em outros, acreditamos que um único indicador pode ser suficiente.

Os indicadores empregados para medir construtos latentes costumam ser classificados em duas naturezas: refletiva e formativa. Os refletivos são aqueles que refletem a intensidade do construto quando o acessamos, e os formativos são aqueles que, quando combinados (somados de alguma maneira), formam os construtos. Vamos nos valer de exemplos para melhor esclarecer os dois tipos de indicadores (no primeiro exemplo ilustraremos indicadores refletivos, e no segundo exemplo, formativos).

Imaginemos inicialmente que nosso interesse seja medir a altura de uma pessoa. Sabemos que podemos medir a altura de uma pessoa diretamente, mas, para fins didáticos, vamos assumir que queremos adivinhar a altura das pessoas sem medi-la diretamente, apenas pela observação de sua manifestação em respostas que as pessoas possam dar a duas perguntas. A primeira pergunta pode se referir ao grau de dificuldade que a pessoa tem para pegar um objeto que esteja na prateleira mais alta de um recinto. A segunda, à necessidade de a pessoa esticar ou dobrar as pernas ao dirigir um carro. Presume-se que uma pessoa alta deve alcançar o objeto na prateleira com maior facilidade do que pessoas baixas, e também que deve ter pernas compridas e assim as dobra para dirigir, ao passo que

uma pessoa baixa deve ter pernas curtas e por isso precisa esticá-las. Assim, ocorre que as respostas às perguntas são manifestações (ou sintomas) do construto altura, e refletem sua intensidade. Presumimos também que, por refletirem o mesmo construto, as respostas às perguntas devem ser correlacionadas entre si. Essas características fazem das respostas às duas perguntas indicadores refletivos da altura.

Vamos agora assumir que pretendemos estimar a quantidade de álcool ingerida por pessoas que saíram de uma festa, mas não temos como fazer um exame de sangue nessas pessoas e nem como estimar essa taxa por meio do uso de um bafômetro. Podemos perguntar a essas pessoas quantas doses beberam de uísque, vodca, cerveja e/ou outras bebidas alcoólicas. A combinação das doses nos permite estimar a quantidade de álcool ingerida, se soubermos o teor alcoólico contido em cada dose. Nesse caso, a combinação das doses provê uma soma que nos permite estimar o que não observamos diretamente. Os indicadores em conjunto formarão a taxa de álcool que cada pessoa bebeu. Várias combinações independentes podem levar a quantidades similares de álcool ingerida; por exemplo, uma pessoa pode beber apenas vodca e ter a mesma quantidade de álcool no sangue de uma outra que bebeu uísque e cerveja. Outra pessoa pode ter ingerido muito álcool, tendo bebido apenas uísque. Assim, não é necessário que as respostas às diferentes perguntas (quantidade de doses ingeridas de cada bebida) estejam correlacionadas entre si para que formem a medida de ingestão de álcool. São essas as características que fazem dessas perguntas indicadores formativos da ingestão de álcool.

Embora seja uma estratégia bem fundamentada e lógica, a mensuração formativa encontrou dificuldades operacionais. De fato, mesmo havendo recomendações para avaliação estatística de validade e confiabilidade (ver uma síntese em Costa (2011)), nenhuma delas alcançou a consistência de um coeficiente alfa de Cronbach nem a completude e adequação de uma análise fatorial. Edwards (2011) chega a chamar esta estratégia de mensuração de falaciosa e desaconselha completamente seu uso.

4 PROCEDIMENTOS E RECOMENDAÇÕES

Nesse item, apresentamos os principais procedimentos e provemos recomendações práticas para a desafiadora tarefa de desenvolver e validar escalas em marketing. Particularmente, detalhamos práticas para aferição de validade e confiabilidade na mensuração refletiva de múltiplos itens, práticas essas que configuram o *mainstream* na área. Em seguida, abordamos procedimentos para verificar validade e confiabilidade na mensuração por meio de item único.

4.1 Validade e confiabilidade na mensuração refletiva de múltiplos itens

Talvez como decorrência da ampla repercussão do artigo de Churchill (1979), em que foi proposto um paradigma para mensuração de construtos latentes refletivos em Marketing, e de outros (por exemplo, Gerbing & Anderson, 1988) que também dedicaram atenção a aspectos de mensuração e apontaram falhas nos procedimentos usuais então vigentes, nas últimas décadas tem sido comum encontrarmos, em artigos, o emprego de análises fatoriais exploratórias e confirmatórias para verificar a estrutura dimensional de variáveis, bem como estratégias para analisar validade convergente e discriminante (por exemplo, por meio da Matriz Multitraço Multimétodo), e a modelagem por equações estruturais, entre outros.

No entanto, ainda parece haver necessidade de chamar a atenção dos pesquisadores para a questão da mensuração. Jarvis, Mackenzie e Podsakoff (2003), em substancial esforço de análise do emprego de modelos de mensuração no campo de Marketing, apontaram que ainda havia confusão quanto à distinção entre construtos de natureza formativa e refletiva entre artigos científicos publicados em periódicos de prestígio no campo (*Journal of Marketing Research*, *Journal of Marketing*, *Journal of Consumer Research* e *Marketing Science*). Dos 1.192 construtos utilizados em 178 artigos analisados, extraídos dos quatro periódicos citados, 1.146 (96,1%) foram modelados como refletivos e 46 (3,9%), como formativos. No entanto, na visão dos autores, dos 1.146 construtos refletivos, 336 deveriam ter sido modelados como formativos (o que representa uma taxa de erro de classificação de 29,3%). Entre os 46 modelados como formativos, os autores entenderam que 17 deveriam ter sido classificados como refletivos (taxa de erro de classificação de 37,0%). Simulações conduzidas no mesmo estudo apontaram para a gravidade desse erro de classificação, que, no limite, pode ser a origem de erros nos resultados de testes de hipóteses e, naturalmente, na elaboração de considerações finais de pesquisas.

Conforme indicado no item 3 acima, a natureza do construto influencia as maneiras de avaliar sua confiabilidade e validade. Considerando a medição refletiva de múltiplos itens, comentamos nesta seção sobre estratégias de avaliação desses aspectos. Nossa impressão ao apreciar artigos científicos na área de Marketing, particularmente os produzidos pela comunidade acadêmica brasileira, é que os relatos acerca dos aspectos operacionais das escalas empregadas para mensurar construtos latentes priorizam características associadas à confiabilidade, e pouca atenção é dedicada aos aspectos de validade. Talvez esta realidade esteja associada ao fato de existirem formulações matemáticas amplamente disseminadas em pacotes estatísticos para avaliar confiabilidade, mas o mesmo não se pode dizer da

facilidade de verificação da validade. É fundamental ter clara a concepção de que medidas válidas são necessariamente confiáveis, mas atingir confiabilidade satisfatória não é condição suficiente para assegurar validade. A seguir comentamos os dois conceitos.

4.1.1 Confiabilidade

Em definição da *American Psychological Association* (1985, p. 19, tradução nossa), "confiabilidade refere-se ao grau em que pontuações de testes estão livres de erros de mensuração". Pedhazur e Schmelkin (1991) classificam esses erros em duas categorias: sistemáticos (vieses de mensuração em uma mesma direção em sucessivas rodadas de coleta de dados) e não sistemáticos (aleatórios ao longo de sucessivas rodadas de mensuração). Para uma revisão mais extensa sobre tipos de erros, recomendamos a leitura de Nunnally (1978).

Ao discorrer sobre propriedades de estimadores (em nossa opinião, extensíveis a instrumentos de mensuração), Bussab e Morettin (2007) propõem uma analogia aos tiros dados por quatro rifles. A Figura 1 ilustra o desempenho de cada um deles.

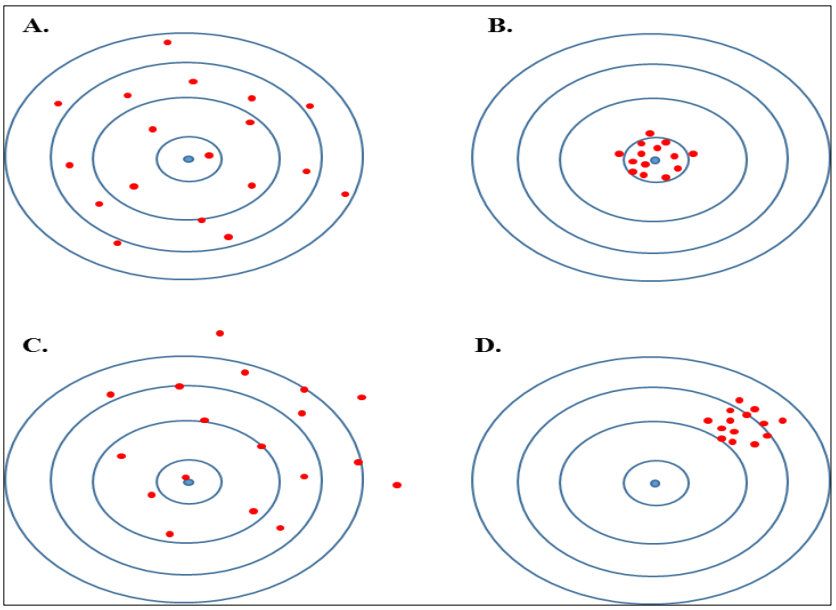

Figura 1 – Exemplos de estimadores (viés e precisão)
Fonte: Bussab e Morettin (2007, p. 291)

Na figura 1.A, temos o exemplo de um estimador não enviesado, porém pouco preciso (o espalhamento dos tiros em torno do alvo é elevado); em 1.B, temos um estimador não enviesado e preciso (ocorrem erros aleatórios pequenos em torno do alvo); em 1.C, o estimador é enviesado e pouco preciso; em 1.D, por fim, o estimador é preciso, porém enviesado. Obviamente, uma escala de mensuração desejável é aquela que retorna a pontuação mais próxima possível da *real*, e com baixa variabilidade quando utilizada repetidas vezes (ou seja, 1.B).

O tipo de erro que controlamos em análises de confiabilidade é o de *precisão* (consistência, espalhamento ao redor de um alvo). Deixemos de lado, por um momento, o componente sistemático do erro de mensuração, presente em 2.C e 2.D (este será abordado adiante). Representando por *R* a pontuação real (a qual queremos *descobrir*), por *O* a pontuação observada (mensurada) e por *E* o componente de erro de mensuração (desvio do observado para o real), temos na mensuração refletiva que: $O = R + E$.

Na situação em que nossa métrica não possui erro sistemático, podemos dizer que, se efetuadas repetidas rodadas de mensuração, o valor esperado (E) do componente de erro será zero (ou seja, em linguagem estatística, $E(E)=0$). Como consequência, o valor esperado da pontuação observada será igual à pontuação real (em termos estatísticos, $E(O)=R$). Este é o princípio central da conhecida teoria clássica de mensuração.

Como, então, acessar a confiabilidade em termos de *precisão*? Operacionalmente, buscamos indicações de que a proporção da variância em uma medida atribuível ao valor real de um construto latente sendo mensurado, seja elevada em comparação com a variância atribuível a componentes de erro (DeVellis, 1991). Alguns exemplos de abordagens são discutidos a seguir.

Poderíamos pensar em medir um mesmo grupo de indivíduos duas (ou mais) vezes, em diferentes momentos; esperaríamos que os valores obtidos, por indivíduo, ficassem próximos, se não idênticos. Ignorando o inconveniente de precisarmos contatar os mesmos indivíduos em dois momentos distintos, existem, ao menos, dois problemas com essa abordagem, conhecida na literatura como teste-reteste e usualmente operacionalizada por meio do cálculo do coeficiente de correlação linear entre os dois vetores de pontuações (Pedhazur & Schmelkin, 1991): a) o *carry-over effect* (participar de um estudo pode influenciar as

respostas do indivíduo em sua participação seguinte); e b) mudanças 'naturais' da pontuação do indivíduo ao longo do tempo (por exemplo, podemos imaginar que o nível de etnocentrismo de um indivíduo aumente ou diminua ao longo de sua vida). Se, por um lado, aumentar o intervalo de tempo entre as duas mensurações pode contribuir com a diminuição do *carry-over effect*, por outro, pode agravar o problema das mudanças 'naturais', e vice-versa. Evidentemente, esses riscos aumentam quando utilizamos múltiplos itens para mensurar um construto, como é o caso da mensuração refletiva.

Em síntese, não é uma tarefa fácil segregarmos confiabilidade de estabilidade temporal ao empregarmos a técnica de teste-reteste, e por isso não encorajamos seu uso para mensuração com múltiplos itens (e caso utilizada, as interpretações devem ser ponderadas à luz dos argumentos apresentados), embora seja possível seu uso em outras estratégias de medição, conforme será apresentado posteriormente.

Especificamente para os construtos de múltiplos itens a literatura especializada já apontou boas soluções. De fato, já há métodos matemáticos eficientes para aferir confiabilidade a partir dos dados de apenas uma rodada de coleta; são exemplos o coeficiente alfa de Cronbach (Cronbach, 1951), o índice de confiabilidade composta (Fornell & Larcker, 1981) e a análise fatorial exploratória (ver Aranha & Zambaldi, 2008). Esses métodos têm como pressuposto a teoria da amostra de domínio, de acordo com a qual existiriam diversos indicadores observáveis cujas variações seriam provocadas por um construto de natureza latente comum.

Retomemos o exemplo do construto latente refletivo *inteligência*. Supondo que cheguemos a um consenso sobre sua definição conceitual, podemos imaginar características de indivíduos a partir das quais inferir sua inteligência. Um exemplo seria o tempo necessário para solucionar problemas. Convencionemos que indivíduos mais inteligentes resolvem problemas mais rapidamente. Se elaborarmos um instrumento de mensuração com dez tipos desses problemas e estes forem resolvidos por, digamos, 300 indivíduos, esperaremos que os tempos de resolução de cada tipo de problema estejam positivamente correlacionados (quanto mais dependente da variação em *inteligência* for a variação nesses tempos, melhor para a nossa medida).

Apesar das conhecidas limitações aplicáveis ao coeficiente alfa de Cronbach (por exemplo, o fato de que mantidos outros aspectos inalterados, quanto mais itens semelhantes e quanto maior o número de itens em uma escala, maior tende a ser seu valor, e, principalmente, o fato de que um valor elevado para a medida não *assegura* unidimensionalidade de construto), seu uso é justificável na avaliação da confiabilidade de uma escala, em particular em estágio inicial de purificação de seus itens. Interpretamos baixos valores de alfa (não há consenso sobre um valor mínimo aceitável; recomendamos ao menos 0,60) como indicativos de baixa consistência interna e consequente necessidade de descarte de indicadores, elaboração de novos ou adaptação de existentes (enfatizamos que, quando nosso construto é de natureza formativa, não faz qualquer sentido esperarmos um valor de alfa de Cronbach elevado, pois a correlação entre os itens não é pressuposta).

Como alternativa ao coeficiente alfa como medida de confiabilidade, podemos empregar o índice de confiabilidade composta proposto por Fornell e Larcker (1981). A confiabilidade composta pode ser obtida por meio de Análise Fatorial e indica a proporção de variância do escore verdadeiro de um construto em relação à variância total do escore calculado. Por não apresentar o inconveniente de se inflar com a inclusão de itens na escala, seu uso tem se popularizado e o consideramos preferível ao uso do alfa de Cronbach. No entanto, o índice de confiabilidade também não é capaz de garantir a unidimensionalidade de um construto. Assim como para o coeficiente alfa, também consideramos desejáveis valores superiores a 0,60.

Quanto à análise fatorial exploratória, deveríamos esperar cargas fatoriais elevadas (no mínimo iguais a 0,40 ou 0,50; ressaltamos que não há um valor mínimo consensual) entre os indicadores e o fator que representa a dimensão a qual deveriam pertencer[5].

É possível, por exemplo, ao incluir diversos itens com redação semelhante em uma escala, inflar seus índices de consistência interna. Isso, no entanto, não torna mais efetivo o instrumento de mensuração, além de tomar espaço em questionários e de deixá-los mais longos sem necessidade. Nesse sentido, devem ser tomados cuidados na fase de geração de itens para que aspectos complementares de um mesmo construto sejam capturados. Recomendamos o artigo de Lee e Hooley (2005) sobre os fundamentos teóricos, aplicações e limitações das técnicas de coeficiente alfa e análise fatorial, e Costa (2011) sobre estágios no desenvolvimento dos itens de uma escala.

4.1.2 Validade

Entendemos por validade de mensuração de um construto o quanto uma proposta de medida

[5] Em nossa percepção, em geral quando a análise fatorial exploratória é reportada em artigos de Marketing no âmbito da academia brasileira, utilizam-se procedimentos de rotação ortogonal (que pressupõem correlação linear nula entre os fatores extraídos). No entanto, parece razoável supor ser comum que dimensões de um mesmo construto refletivo (quando lidamos com construtos multidimensionais) estejam correlacionadas. Por isso, entendemos que o procedimento de rotação apropriado seria o oblíquo (para uma cobertura mais específica sobre o assunto, recomendamos a leitura de Stewart (1981)). Convém ressaltar que, entre indicadores formativos, não necessariamente devemos esperar as altas cargas fatoriais mencionadas.

realmente afere aquilo a que está se propondo medir. A eventual presença de erros sistemáticos (ver item anterior) deverá ser capturada ao empregarmos procedimentos efetivos de validação. É importante destacar, de início, que conseguimos tão somente acumular 'evidências' de que nosso instrumento de mensuração seja válido; não é possível *ter absoluta certeza* que a validade ocorra, já que isto requereria que o construto latente objeto de mensuração pudesse ser observável.

Nosso objetivo, ao buscar evidências de validade para uma escala, é proporcionar condições razoáveis de medição de construtos, para que então hipóteses que o envolvam possam ser testadas. Diferentemente dos métodos para verificar confiabilidade, os métodos disponíveis para avaliar a validade são dependentes da habilidade do pesquisador para desenvolver estratégias mais ou menos eficientes. Estas estratégias podem mirar três tipos de análise de validade[6]:

a) de translação – é um tipo de validação não estatística e qualitativa que envolve o exame sistemático do conteúdo do instrumento de mensuração para avaliar se seus componentes representam adequadamente facetas do construto (situação em que dizemos haver validade de conteúdo) e se há adequação de redação e forma para aplicação dentre a população a que se destina (situação em que dizemos haver validade de face). Em geral, este tipo de validação é conduzido por especialistas (pesquisadores ou participantes); é possível também utilizar potenciais respondentes como juízes;

b) de critério – envolve a análise da associação prevista entre nossa medida e uma variável tomada como critério, representativa do construto. Por exemplo, as medidas de uma escala de propensão ao comportamento doador podem ser comparadas com o comportamento doador, digamos, verificado no ano subsequente. A validação de critério, nesse caso, é qualificada como *preditiva*. É possível empregar validação de critério *simultânea*, por exemplo, ao mensurar materialismo entre religiosos e entre estudantes de negócios, tal como conduzido por Belk (1985);

c) de construto – refere-se a quanto a operacionalização de um construto o mostra aderente ao que a teoria diz, em termos de sua definição e propriedades. Verificam-se sua estrutura dimensional e seu relacionamento com outros construtos. São subtipos as validades: convergente; discriminante; nomológica; e grupo-conhecido. Aqui, as associações encontradas entre o construto e outros são confrontadas com as expectativas teóricas, e técnicas como a Matriz Multitraço Multimétodo, a Análise Fatorial Confirmatória (AFC) e a Modelagem por Equações Estruturais (SEM) são úteis para essas checagens.

Os tipos de estratégias de análise de validade apresentados devem ser vistos como complementares. Raramente encontramos, em artigos na área de Marketing, o uso simultâneo de todos. Para ilustrarmos como estas estratégias se aplicam, tomemos o exemplo de Richins e Dawson (1992). Esses autores, ao desenvolverem e proporem uma escala largamente utilizada para mensurar materialismo, utilizaram estratégias de validação de critério (simultâneo). Em questionários enviados aos respondentes, além de incluírem os indicadores da escala de valores materiais, também apresentaram perguntas como: qual é o nível de renda necessário para satisfazer suas necessidades?; qual é a importância relativa de valores tais como segurança financeira, relacionamento agradável com outros, e auto realização?; o que o respondente faria caso ganhasse, sem esperar, determinada quantia de dinheiro (uso egoísta ou altruísta)?; entre outras. Usaram, então, uma sólida fundamentação teórica para justificar comportamentos esperados de grupos de indivíduos mais materialistas e menos materialistas, e averiguaram se a pontuação de valores materiais indicada pelo instrumento de mensuração proposto servia para predizer o comportamento nas perguntas de verificação apresentadas. Convém reforçar, neste exemplo, o esforço de reflexão acerca das características esperadas para grupos de indivíduos mais e menos materialistas, e de criação de protocolos para buscar validação.

As técnicas de análise de validade em cada uma das estratégias apresentadas são muitas e sua exposição está além do escopo deste artigo. Podemos afirmar que os métodos clássicos de avaliação de validade por estas estratégias estão bem documentados (cf. DeVellis, 1991; Netemeyer, Bearden & Sharma, 2003; Costa 2011). No entanto, em anos recentes, o uso de técnicas estatísticas mais sofisticadas para análise de validade tem se intensificado. Por exemplo, Gonçalves (2013) utiliza um modelo de análise fatorial confirmatória de terceira ordem para verificar confiabilidade e validade convergente da escala de satisfação com atributos. Esse construto foi definido como tendo três dimensões primárias – núcleo do serviço, aspectos periféricos da qualidade do serviço e valor. Por sua vez, a dimensão de aspectos periféricos da qualidade do serviço possui três subdimensões, e a de valor, outras duas.

Já Yi e Gong (2013) propuseram mensurar o comportamento de cocriação de valor do consumidor por meio de uma abordagem hierárquica e multidimensional. Como estratégias de validação (convergente, discriminante e nomológica), os autores

[6] Ressaltamos que, embora o foco desse subitem (4.1) seja na mensuração refletiva de múltiplos itens, essas estratégias de análise de validade são aplicáveis a outras alternativas de medição, como será observado mais adiante. A variação de aplicação está nas técnicas utilizadas.

empregam modelos de análise fatorial confirmatória (de primeira e terceira ordens) e modelo PLS (*partial least squares*).

4.2 Validade e confiabilidade na mensuração por único item

A estratégia de mensuração por múltiplos itens, objeto de aplicação de relevantes técnicas (como modelagem por equações estruturais, por exemplo), pressupõe que um construto bem delimitado tem sua medição a partir do levantamento das pontuações para dois ou mais itens. Nesta perspectiva, e conforme indicado acima, cada item mensura uma faceta do construto, que, pela teoria da amostra de domínio, tem associação direta com o construto por possuir uma parte de sua variação oriunda da variação do fator latente (a outra parte da variação se explica por um erro aleatório). Há, por outro lado, uma alternativa de medição bastante usada nas pesquisas em Marketing, que consiste na mensuração de construtos por um único item em lugar de um conjunto deles.

O pressuposto central da teoria da amostra de domínio facilita sobremaneira a validação estatística de medidas de um construto ou dimensão. De fato, se consideramos que a validade de conteúdo e de face de um conjunto de itens está boa (essa etapa é mais qualitativa), a validade estatística é facilmente verificada pela análise da adequação fatorial e da consistência interna. Por outro lado, em uma averiguação por um único item não há sentido algum em sua submissão a uma extração fatorial ou a extração de um coeficiente de consistência interna, como o alfa de Cronbach ou o índice de confiabilidade composta. Isto faz com que sejam utilizadas técnicas de análise de validade distintas. A seguir apresentamos os principais procedimentos de análise de validade, considerando primeiro a avaliação qualitativa e depois as alternativas de avaliação estatística[7].

4.2.1 Etapa qualitativa de validação

Na avaliação qualitativa, neste tipo de escala os cuidados são os mesmos daqueles aplicados nas escalas de múltiplos itens, e a meta é simples: fazer o enunciado do item refletir plenamente o conteúdo do construto, manifesto em sua definição. Além da clara associação com a definição, ou seja, validade de conteúdo, e para assegurar boa validade de face, o enunciado precisa ser sucinto e compreensível, mesmo que a escala já seja menor em tamanho (em comparação com a mensuração por múltiplos itens). Em outras palavras, o fato de a mensuração ser baseada em um só item não implica que se utilize um item muito extenso ou com vocabulário inapropriado para a compreensão do respondente, mesmo que o construto mensurado seja abstrato. Isto gera um desafio maior para o pesquisador, tendo em vista a necessidade de consolidar em um só enunciado toda a significação de um construto, além de requerer uma forma de apresentação que seja coerente com a escala de verificação a ser utilizada.

Como método para este desafio, dois procedimentos precisam ser cuidadosamente utilizados: primeiro, o item deve ser elaborado e submetido à apreciação de especialistas no tema e/ou pesquisadores experientes; segundo, o item deve ser exposto a futuros potenciais pesquisados, para verificar sua compreensão da associação do conceito ao item. Estes procedimentos ajudam a garantir validade de conteúdo (associação do item com a definição) e face (apresentação e compreensibilidade do item).

John Rossiter (2011), em seu modelo COARSE, é enfático em afirmar que a etapa qualitativa de mensuração por um único item é a principal, senão a única, forma de garantir a validade de uma escala. Ainda assim, entendemos que a indicação reiterada de validade de conteúdo e face por especialistas ou potenciais respondentes da escala não é suficiente, ou ao menos não haveria perdas por sua confrontação com resultados de uma aplicação concreta da escala na mensuração do construto a que se propõe a medir.

4.2.2 Etapa quantitativa de validação e confiabilidade

A análise da consistência de uma escala de um só item se reafirma com dados oriundos de sua aplicação a partir da avaliação da aderência dos resultados amostrais à expectativa de comportamento da variável que deu origem à amostra, da validade de critério, da validade de grupo conhecido e do procedimento de teste-reteste. Vejamos alguns detalhes e recomendações.

Sobre a aderência da escala ao esperado, tomemos por pressuposto que a métrica é direcionada a medir um construto cuja medida segue alguma distribuição de probabilidade em nível populacional. Por exemplo, é possível supor que o 'nível de satisfação de cidadãos com o governo' segue uma distribuição simétrica de comportamento aproximadamente normal, ou que o 'nível de disposição de jovens à participação cívica' é assimétrica à direita, com maior concentração em escores mais baixos de uma escala. Nesses termos, se uma escala é aplicada para medir estes construtos, o comportamento dos escores da amostra deve refletir aproximadamente o modelo de distribuição esperado.

Do ponto de vista operacional, esta verificação pode ser feita de forma exploratória ou por meio de testes, porém recomendamos uma avaliação exploratória e bem fundamentada. Por exemplo, uma avaliação do histograma ou de um gráfico de ramo-e-folha dos valores da amostra já pode ser suficiente para sinalizar se o formato da amostra se aproxima da

[7] Levando em conta a finalidade deste artigo de servir de referencial de consulta a pesquisadores e considerando ainda a menor tradição de uso e de desenvolvimento na literatura de Marketing, optamos por detalhar melhor estes procedimentos e dar mais recomendações, diferente do que fizemos no item 4.1, para o qual o desenvolvimento teórico e de aplicações é muito mais amplo.

expectativa de distribuição pressuposta. Naturalmente, nem sempre é possível supor uma distribuição para a variável de referência, o que dificulta este tipo de análise.

Também é recomendado verificar o comportamento da variável em relação a algumas medidas estatísticas. Por exemplo, é quase sempre esperado que a escala capture a variação real de intensidade do construto existente no universo de interesse da pesquisa. Assim, se em uma população com dispersão sabidamente moderada na intensidade do construto, uma escala gera um desvio padrão muito pequeno ou muito grande, então isto pode ser sinalização de problemas de adequação da métrica para capturar o comportamento esperado dos dados.

Quanto à análise de validade, diferente da mensuração por múltiplos itens, recomendamos somente dois procedimentos em caso de mensuração por um só item: a validade de critério e de grupo conhecido. Do ponto de vista de validade de critério, o procedimento consiste em analisar o comportamento da escala em sua performance de predição ou associação do construto sob medição com relação a outro construto com escala previamente validada (quando esta predição ou associação é esperada). Por exemplo, suponhamos que estamos analisando uma escala de um único item para medir o 'nível de consciência ambiental declarada', assim enunciado 'sou uma pessoa sensível a questões ambientais' (para verificação em uma escala de concordância); se sabemos que a consciência ambiental é preditora da 'predisposição à compra de produtos com selo de sustentabilidade', e se já temos uma escala validada para este construto, então podemos facilmente verificar se nossa escala é válida ou não aplicando as duas métricas simultaneamente, e verificando se a associação esperada emerge, ou seja, se há correlação significativa entre as medidas dos dois construtos, ou se uma análise de regressão consegue níveis adequados de ajustamento (conforme o que se espera em termos de intensidade e direção da previsão).

De forma semelhante à validade de critério, também é possível analisar o comportamento esperado de uma medida em relação a grupos ou variáveis categóricas específicas, na dita validade de grupo conhecido (esta estratégia é pouco usada em mensuração por múltiplos itens). Por exemplo, em uma escala de um único item para medir 'confiança nos governantes municipais' assim enunciada 'em geral, confio nos governantes de minha cidade' (com aferição em uma escala de concordância), e se sabemos que pessoas com vinculação partidária com a liderança possuem avaliação mais positiva que pessoas sem vinculação, então a escala será válida se conseguir refletir esta diferença. Isto pode ser verificado, por exemplo, por técnicas estatísticas como análise de variância, teste t de Student, ou por meio técnicas não paramétricas correspondentes (teste de Kruskal-Wallis ou teste de Wilcoxon-Mann-Whiteney). Assim, caso os dados se comportem como esperado e com a indicação destes testes, é possível assegurar, ou não, a validade de grupo conhecido.

Por fim, e como forma de verificação de confiabilidade, escalas com um só item podem ser avaliadas por seu comportamento segundo diferentes momentos de aplicação no tempo, no procedimento dito de teste e reteste[8]. Isto ocorre mediante a aplicação da escala junto a um grupo de respondentes em um dado momento no tempo, e depois se faz uma segunda aplicação com este mesmo grupo, passado um tempo curto o suficiente para que a intensidade do construto não varie muito, mas distante o suficiente para que os respondentes não se lembrem da resposta dada anteriormente. A confiabilidade é assegurada se a correlação dos dados nas duas aplicações for suficientemente grande para refletir o comportamento esperado de convergência de comportamento (recomendamos ao menos 0,8).

O quadro 1 sumariza os procedimentos indicados e nossas recomendações.

AVALIAÇÃO	RECOMENDAÇÃO
Validade de conteúdo e face	Exposição da escala a especialistas e potenciais respondentes e avaliação qualitativa dos resultados.
Adequação de performance	Análise de medidas e do comportamento (distribuição) dos dados amostrais em comparação com a expectativa de comportamento.
Validade de critério	Análise de associação ou predição da escala em relação a outros construtos com escalas já validadas e comparação com resultados esperados.
Validade de grupo conhecido	Análise de medidas da escala em relação a grupos de sujeitos e comparação com resultados esperados.
Confiabilidade teste-reteste	Avaliação da associação entre as medidas geradas pela escala em dois momentos distintos no tempo e comparação com a expectativa de elevada associação.

Quadro 1 - Procedimentos de validação para escalas de item único

[8] Conforme indicamos acima, não recomendamos este procedimento para mensuração de múltiplos itens, devido ao fato de termos métodos de verificação consistentes para uma só verificação. Este não é o caso da verificação por um único item, razão pela qual o procedimento ganha utilidade.

4.3 Complementos Relevantes: dimensionalidade, organização de instrumentos e variância comum ao método

Uma questão relevante nos debates atuais sobre mensuração em Marketing remete à dimensionalidade de um construto. Um construto não precisa ser necessariamente unidimensional, podendo possuir diversas dimensões (subconstrutos) ou atributos (na elaboração de Rossiter). Tomemos como exemplo a confiança, construto que pode ter, de acordo com a literatura, múltiplas dimensões, como percepção de honestidade, de benevolência e de competência. Nesse caso, entendemos que para medir a confiança seria necessário medir as três dimensões, ou seja, se os respondentes percebem o objeto de análise como honesto, benevolente e competente. As três dimensões, ou atributos, podem, inclusive, ser abstratas e com isso requerem múltiplos itens para suas medidas. A análise fatorial confirmatória é uma técnica útil para análise de dimensionalidade do instrumento (ver Aranha & Zambaldi, 2008), mas está limitada ao ajuste de modelos refletivos. É necessário ressaltar que o teste de dimensionalidade de uma escala não deve se basear no coeficiente alfa de Cronbach, no índice de confiabilidade composta, nem na análise fatorial exploratória, mas em procedimentos mais robustos.

Além das preocupações com os procedimentos qualitativos e quantitativos para a construção e validação de instrumentos de medida, enfrentamos aquelas concernentes aos seus métodos de aplicação. Neste domínio, incluímos a forma de coleta (como por meio de entrevistas ou por autopreenchimento, por exemplo), os momentos de aplicação e as distintas amostras às quais podemos aplicar os instrumentos. Cada variação na aplicação é sujeita a viés e, quando esse viés exerce grande influência nos dados obtidos, enfrentamos um fenômeno indesejado, conhecido como variância comum ao método, que consiste em um padrão comum a todas as respostas (ou à maior parte delas) por parte dos respondentes, seja por apresentarem comportamento socialmente desejável, por tentarem adivinhar o que se quer medir e procurarem direcionar a medida, por tentarem parecer coerentes, ou por sofrerem algum viés oriundo da forma de coleta (como falta de compreensão de um item ou algum tipo de indução por parte do entrevistador).

O uso de múltiplos métodos para coleta dos dados de um construto visando a mitigar a variância comum ao método é dispendioso em termos de tempo e de outros recursos e, por essa razão, os pesquisadores, impossibilitados de empregarem ferramentas como a Matriz Multitraço Multimétodo, lançam mão de técnicas para minimizar o potencial viés decorrente do uso de um método único. Uma das formas de dificultar o reconhecimento do que se quer medir por parte dos respondentes pode ser mesclar a ordem de apresentação dos itens das dimensões presentes no instrumento, como já mencionamos ao apresentar a proposta de Rossiter (2002, 2011). Outra seria o uso de itens inversos (aqueles com relação conceitual negativa com o construto) entre itens com relação positiva com o construto (ver Wong, Rindfleisch & Burroughs, 2003; e Aranha & Zambaldi, 2008). Por exemplo, para medir competência, podemos colocar no instrumento afirmações que remetam a esse atributo, juntamente com um item que remeta à incompetência. A presença de itens inversos tende a obrigar que o respondente se concentre mais em suas respostas, por não poder adotar um padrão automático ao responder (como alta concordância com todos os itens, por exemplo). Evidentemente, os itens inversos devem ter seus valores invertidos para análise e cômputo de escores. Ademais, são de difícil elaboração, pois costumam conter negativas, o que pode confundir os respondentes.

5 CONSIDERAÇÕES SOBRE ESCALAS DE VERIFICAÇÃO

Um importante aspecto da mensuração de construtos em Marketing é o que chamamos de "escala de verificação", que está associada à referência que o respondente tem para apontamento do número que indicará a medida do construto. Com efeito, quando o respondente aponta a intensidade de medição de interesse, normalmente ele o faz pela indicação de um número que escolhe dentre um conjunto de opções (por exemplo, 5 pontos numerados de 1 a 5 em uma escala de concordância). É sempre um bom desafio para os pesquisadores apontarem alternativas de números adequadas aos diferentes propósitos de pesquisa.

Rossiter (2011) chega a afirmar que a validade de uma escala se faz pelo somatório da validade do conteúdo do item (o enunciado de uma afirmação para captação da concordância, por exemplo) com a validade da escala de verificação (ou o número de pontos e o sentido que eles têm para o respondente). É fácil concordar com o posicionamento de Rossiter, o que faz necessária uma atenção especial sobre esta decisão de medição.

Apresentamos aqui quais são as principais decisões a serem tomadas e as alternativas mais apropriadas para cada contexto de decisão. Em geral, as decisões são concernentes à coerência da escala de verificação com a apresentação do item; ao número de pontos da escala; à estratégia de agregação; e às alternativas de operacionalização estatística.

5.1 Coerência entre escala e apresentação do conteúdo

Em relação à coerência entre a escala e o conteúdo, a preocupação é assegurar que a escala de verificação esteja coerente com o enunciado do item. Por exemplo, se o pesquisador decide enunciar um item como afirmação, a escala de verificação tem sentido em ser uma escala de concordância com a afirmação, em diferentes níveis. Este é o caso mais utilizado em mensuração em Marketing, com a utilização da dita 'escala de Likert', proposta por Rensis Likert (1932). O problema recorrentemente observado é aquele em que a escala de verificação vem na forma de concordância, mas sem que o enunciado seja apresentado com a afirmação (para a qual o respondente deve indicar se concorda ou não e em que nível).

Não há sentido, por exemplo, em pedir para um usuário avaliar um serviço (por exemplo, "avalie a qualidade do serviço de transporte público'), e na sequência colocar uma escala de concordância de 5 pontos (por exemplo, de 1 para discordância total até 5 para concordância total). Evidentemente, a forma de evitar este tipo de problema consiste tão somente em analisar a coerência entre as alternativas de resposta e a forma como o item está apresentado, o que se faz por uma análise cuidadosa do pesquisador, além de uma consulta a especialistas e potenciais usuários.

Ainda relativo à coerência entre escala de verificação e enunciado, um aspecto pouco destacado quando se desenvolvem escalas concerne à valência dos itens. Este problema emerge especialmente quando a medida sob análise envolve atitudes. Pela própria conceituação, atitudes estão associadas a avaliações gerais, que, na maioria das vezes, variam de um sentido negativo a um sentido positivo. Ou seja, a indicação de uma medida relativa à atitude pode trazer duas informações ao mesmo tempo: primeiro, se é uma avaliação positiva ou negativa; segundo, qual a magnitude em qualquer das duas opções (ou seja, se negativa, quão negativa, e se positiva, quão positiva).

Rossiter (2011) sugere que, sempre que um construto for de avaliação, ou de forma mais generalizada, se for 'bipolar', a escala de verificação mais coerente é aquela em que há alternativas de valores negativos, nulos e positivos. Retomando o exemplo acima, uma medida de avaliação da qualidade do serviço de transporte, com item enunciado como "avalie a qualidade do serviço de transporte público', as alternativas de resposta mais coerentes seriam (em uma escala de 5 pontos), -2, -1, 0, +1, +2. Isto não impede, por exemplo, que seja dada ao respondente a opção de marcar um ponto em uma escala de 11 pontos de 0 a 10 (0, 1, 2, ..., 10) ou 1 a 100 (deixando um espaço para o respondente indicar um número entre 0 a 100). Entendemos que a decisão não define algo certo ou errado, mas sim algo 'mais adequado' para cada pesquisa e cada ato de medição.

5.2 Número de pontos da escala

Por simples prática, a grande maioria das métricas em Marketing utiliza escalas de mensuração do tipo intervalar (aquele em que se arbitra um ponto de mínimo, ou de máximo (ver Stevens, 1946), com averiguação em um número limitado de alternativas (por exemplo, uma escala de 7 pontos, em que o 1 indica a magnitude mínima e 7 a máxima). A vantagem central desta decisão está relacionada à geração segura de respostas e à facilidade para o respondente. A perda central vem da impossibilidade ou inadequação do uso de determinadas técnicas estatísticas.

Concernente a este último aspecto, sabemos que as principais técnicas estatísticas utilizadas nas pesquisas em marketing pressupõem que algumas distribuições sejam contínuas. Este é o caso, por exemplo, da técnica clássica de regressão normal linear, que, por ter pressupostos para o erro do modelo, requer que a variável resposta seja do tipo contínua. Da forma como costumamos fazer análises, a aceitação do pressuposto de continuidade torna-se complicada por operamos com uma mensuração discreta e limitada a um determinado número de pontos.

Não há na verdade uma regra para definição de número de pontos, mas é possível afirmar que a escala deve ter tantos pontos quantos possíveis. Na verdade, se for possível em uma métrica dar ao respondente a possibilidade de indicar o número, bastaria ao pesquisador apontar os limites da escala; isto inclusive daria à medida um senso de continuidade que viabilizaria aplicações de técnicas estatísticas sem ter que fazer maiores concessões. No entanto, essa alternativa tem restrições operacionais, pois, dado que a maioria das pesquisas é feita com questionários, a indicação de um total de pontos torna mais fácil a coleta de respostas.

Há, por outro lado, delimitadores a serem considerados. Entendemos que o tamanho dos questionários é um primeiro delimitador do número de pontos, sendo necessário considerar que muitos pontos tendem a ocupar mais espaço e isto pode tornar os questionários muito longos e comprometer as respostas. Além disto, é necessário levar em conta a capacidade de os respondentes emitirem uma resposta confiável com determinados números de pontos. Este último aspecto é especialmente relevante para os casos em que os respondentes necessitam de denominações sobre os pontos, ou seja, da indicação de significado de intensidade de cada ponto da escala. Por exemplo, em uma escala de 5 pontos é fácil denominar os pontos como: 1 – discordo totalmente; 2 – discordo parcialmente; 3 - concordo/discordo moderadamente; 4 – concordo parcialmente; 5 – concordo totalmente. Por outro lado, em uma escala de 11 pontos (digamos, de 1 a 11), torna-se bastante complicado dar expressão para cada ponto[9].

A recomendação que damos é a seguinte: se

[9] A denominação de pontos é mais problemática nos casos de números ímpares de pontos, pois há tendência de associar o ponto central à condição de indiferentes ou neutro, o que efetivamente não tem sentido, pois o indiferente ou neutro simplesmente não tem respostas na escala (por exemplo, uma pessoa neutra na concordância com uma determinada

houver espaço, devemos utilizar tantos pontos quanto possíveis, evitando, por outro lado, dar uma denominação específica para cada ponto. Uma estratégia interessante parece ser utilizar escalas de 10 ou 11 pontos (de 1 a 10, ou 1 a 11, ou -5 a +5), denominando somente os extremos e com uma sinalização de significado da região intermediária (ver Hodge e Gillespie, 2007). Aplicações com este tipo de escala têm sido apontadas como consistentes, e em boa medida facilitam a resposta, pois na cultura brasileira estamos habituados em emitir posições de 0 ou 1 a 10 (ver Barboza *et al*. 2013).

A opção pela quantidade de tantos pontos quanto possíveis é, no entanto, controversa, e depende da capacidade de o respondente compreender o funcionamento da escala. De acordo com nossa experiência de campo, especialmente entre respondentes com baixa escolaridade, diminuir o número de opções pode ser interessante, pois deixa a indicação de resposta mais simples. Podemos usar itens de aquecimento, como por exemplo, 'Hoje está frio', ou 'Eu gosto de futebol', para verificar a compreensão de como indicar concordância aos itens que serão lidos. Isto é possível quando a aplicação é feita por um entrevistador, presencialmente.

5.3 Estratégia de agregação

O problema da agregação existe quando utilizamos uma escala de múltiplos itens para mensuração de um dado construto ou dimensão. A demanda vem da necessidade de, eventualmente, analisarmos a medida total do construto (por vezes, esta medida não se faz necessária, como é o caso, por exemplo, de pesquisas que testam modelos por modelagem de equações estruturais). Quando a agregação é necessária, salientamos aqui três opções para o caso de construtos com mensuração refletiva e uma para os demais casos.

Se temos um conjunto de itens que medem refletivamente um construto, e se este conjunto de itens está adequado em termos de estrutura fatorial e de consistência interna, uma primeira recomendação, e a mais comum de todas, consiste na utilização das estratégias de agregação da análise fatorial, que está presente na maioria dos pacotes computacionais. Ou seja, na extração fatorial podemos solicitar que o software gere uma medida geral do fator. O problema desta estratégia é que, nas rotinas atualmente implementadas, a variável que se gera é padronizada de tal modo que sua média é 0 e sua variância é 1, o que normalmente difere das medidas das escalas de origem dos itens (que são entre 1 e 5, ou 1 e 7, entre outras).

Por esta razão, se a estrutura psicométrica estiver adequada, é possível manter a medida agregada na mesma escala das variáveis pela extração das médias aritméticas simples de cada respondente no conjunto de itens (ou seja, extraindo as médias dos escores de cada respondente (Bagozzi & Edwards, 1998) ou, em uma segunda alternativa, pela extração de uma média ponderada dos escores por respondente, utilizando como fator de ponderação os escores fatoriais dos respectivos itens. Esta segunda estratégia tem a vantagem de, além de manter a escala agregada nos limites das escalas originais, dar pesos maiores aos itens mais correlacionados ao construto latente (lembramos que o escore fatorial é uma medida de correlação entre a variável e o fator latente).

Se um construto tem mensuração por múltiplos itens, mas sem supor relação refletiva, a melhor estratégia de agregação é pela extração de uma medida ponderada por respondente. Aqui temos, por outro lado, a necessidade de justificar os fatores de ponderação; caso não haja uma boa justificativa, qualquer agregação é arriscada. É possível a agregação pela média aritmética simples dos escores por respondente para um caso extremo de ausência total de um referencial de ponderação, porém as análises das medidas precisam sempre levar em consideração possíveis problemas oriundos deste procedimento.

5.4 Operacionalização estatística

Fazemos aqui breves considerações sobre a operacionalização estatística de dados oriundos de escalas costumeiramente utilizadas. Este assunto chega a ser polêmico a depender do pesquisador e do seu nível de exigência teórica. Por esta razão, nos limitamos a apontar algumas avaliações e recomendações de prática, passíveis, naturalmente, de contestação.

Conforme já informado, várias técnicas pressupõem continuidade das variáveis para sua aplicação, como é o caso de parte dos modelos lineares convencionais. Por esta razão, se estamos operacionalizando dados oriundos de escalas com mensuração por um número determinado de pontos, os dados dificilmente têm comportamento semelhante a uma variável contínua. Isto faz com que o uso de técnicas de regressão múltipla do modelo normal linear (e mesmo parte das técnicas de modelos lineares generalizados, regressão quantílica e outras), por exemplo, não possa ser realizado quando a variável resposta for medida em uma escala de Likert, por exemplo[10].

Entendemos que a alternativa mais coerente para superar esse tipo de embate consiste em ampliar o leque de técnicas, triangulando tantas quantas possíveis, e analisando convergências, similaridades,

afirmação na verdade não pontua em uma escala que mede justamente o nível de concordância).

[10] Esta afirmação é controversa, pois, por vezes, se confunde continuidade da variável com continuidade da escala, o que, efetivamente, são conceitos distintos. Por um caminho ou outro, não é incomum encontrarmos aplicações de modelos com pressuposto de continuidade com variáveis medidas com escalas de pontos (da mesma forma como encontramos diversas aplicações de técnicas paramétricas sem a total segurança quanto aos pressupostos de distribuição envolvidos).

analogias e discrepâncias, de forma a poder construir um referencial completo sobre a realidade em estudo a partir dos dados disponíveis (Haig, 2005). Isto se faz pela apropriação de técnicas complementares àquelas correntemente utilizadas (que, entendemos, podem sim continuar sendo aplicadas, dando-se o devido desconto na avaliação dos resultados). Ou seja, entendemos ser apropriado aplicar, além das técnicas convencionais, métodos paramétricos de previsão/associação para dados discretos (presente nos modelos lineares generalizados e de análise de dados categorizados, por exemplo; ver Faraway, 2006, Sheather, 2009), além de técnicas não paramétricas ou semiparamétricas (ver Kloke & McKean, 2012; Hao & Naiman, 2007).

O quadro 2 sumariza os procedimentos indicados nessa seção.

AVALIAÇÃO	RECOMENDAÇÃO
Coerência escala-conteúdo	Analisar cuidadosamente a associação entre o conteúdo do item e as alternativas numéricas disponibilizadas aos respondentes e indicar números adequados ao sentido do item.
Número de pontos	Utilizar tantos quantos possíveis, levando em conta, por outro lado, o espaço ocupado no questionário e a facilidade de resposta para os respondentes.
Estratégia de agregação	Para múltiplos itens, se a mensuração for refletiva, confirmar a consistência psicométrica e agregar pela técnica da análise fatorial ou nos escores por respondente, seja por média aritmética simples ou ponderada pelos escores fatoriais.
Técnica estatística	Fazer análises complementares envolvendo técnicas clássicas com outras técnicas paramétricas, além de métodos não paramétricos e semi paramétricos.

Quadro 2 - Procedimentos para escalas de verificação

6 REFLEXÕES SOBRE MENSURAÇÃO EM PESQUISAS INTERCULTURAIS

Pesquisas interculturais têm se tornado comuns nas ciências sociais e servem aos propósitos de testar a generalização de teorias ou prover um tratamento experimental 'natural' para estudar a influência da cultura no comportamento. Alternativamente, podemos pensar que a pesquisa em uma única cultura pode levar a uma visão parcial da realidade ou à generalização (equivocada) dos resultados de uma cultura como se fossem universais (Steenkamp, 2005).

É comum, na Psicologia, encontrarmos esforços para acessar dimensões universais da personalidade, tais como valores, crenças e emoções; no entanto, é possível que sistemas culturais moldem essas características individuais de maneiras distintas. Como aponta Church (2010), a existência de dimensões universais de diferenças individuais, que podem ser acessadas livres de contexto e de maneiras equivalentes entre culturas, é questionada.

Em particular, na comunidade acadêmica brasileira de Marketing, é frequente o emprego de versões (com variados graus de adaptação) de escalas desenvolvidas em outros países. Além da atenção aos aspectos de confiabilidade e validade das medidas, ao aplicá-las a um contexto distinto do qual foram projetadas, e em particular quando existe intenção do pesquisador de realizar comparações interculturais, outros tipos de ruídos devem ser observados. Van de Vijver e Leung (1997) os classificam em três grupos: viés de construto, de método e de item.

O **viés de construto** ocorre quando as definições de um construto sofrem sobreposição apenas parcial entre culturas. Dizemos, nesses casos, que há falta de equivalência conceitual. Church (2010) oferece como exemplo o conceito de motivação para a realização (*achievement motivation*), que pode ser mais socialmente orientado – enfatizando metas de grupos sociais ou familiares – em culturas coletivistas, na comparação com a concepção ocidental, que enfatiza esforço individual para obtenção de metas pessoais. Sobre esse tema, Milfont e Fischer (2010) apresentaram uma revisão da literatura de equivalência de mensuração e um modelo passo a passo de verificação por meio do emprego de análise fatorial confirmatória.

O **viés de método** pode assumir três formas (Church, 2010): (i) viés de amostra; (ii) viés de instrumento; e (iii) viés de administração. Um exemplo de *viés de amostra* poderia ocorrer ao investigarmos indivíduos pertencentes a uma determinada camada socioeconômica. Qual seria o critério de equivalência entre respondentes brasileiros, norte-americanos e japoneses, por exemplo? O uso de um critério de estratificação socioeconômica adequado para a realidade brasileira tal como proposto por Kamakura & Mazzon (2013) (provavelmente) não é diretamente extensível à realidade desses outros dois países. Renda ou poder aquisitivo seriam abordagens superiores para identificar equivalência? Críticas semelhantes podem ser feitas a estudos que buscam mensurar a pobreza das nações (existiria um critério universal de pobreza ou este é um conceito que deve levar em consideração especificidades regionais?).

O *viés de instrumento* refere-se à diferença na

interpretação do instrumento de coleta de dados pelos respondentes, por exemplo, resultante da redação das questões. Wong, Rindfleisch e Burroughs (2003) apontaram problemas com a administração de itens redigidos em ordem direta entre respondentes do leste asiático; argumentam que, em função da maior inclinação para concordar com sentenças proferidas por terceiros, itens redigidos em forma de pergunta poderiam ser mais adequados para capturar valores. Reardon e Miller (2012) sugerem que pode haver benefícios ao usar metáforas em escalas, na comparação com o uso de formatos mais tradicionais, como Likert e diferencial semântico. Já o *viés de administração* refere-se à dificuldade de comunicação entre o pesquisador e o pesquisado.

O **viés de item** ocorre quando indivíduos com a mesma 'quantidade' de uma característica, mas pertencentes a diferentes grupos culturais, exibem diferentes probabilidades de resposta a itens em uma direção esperada. Quanto à equivalência linguística, o procedimento de tradução reversa é provavelmente o mais empregado entre os estudos brasileiros, mas há também outros disponíveis. Pode-se, por exemplo, administrar um instrumento em dois idiomas para pessoas bilíngues e comparar a correlação entre as respostas.

Com a crescente globalização da ciência e das sociedades, pesquisas interculturais provavelmente continuarão a ganhar importância, assim como a necessidade de endereçar com sucesso questões ainda não resolvidas de mensuração nesses contextos. De acordo com Church (2010), medidas válidas entre culturas irão requerer dos pesquisadores desenvolvimentos contínuos em métodos estatísticos para determinar equivalência de mensuração. Por exemplo, modelos lineares hierárquicos e sua habilidade de simultaneamente testar hipóteses tanto no nível individual como cultural de análise provavelmente crescerão em importância.

7 TENDÊNCIAS DE MENSURAÇÃO EM MARKETING

Neste item apresentamos alguns tópicos de tendência de mensuração em marketing. Nossa seleção foi baseada na avaliação da literatura recente em pesquisas e mensuração em Marketing, e os temas de referência foram os seguintes: Teoria da Resposta ao Item, estimadores Bayesianos e modelagem por mínimos quadrados parciais (*partial least squares – PLS*).

7.1 Teoria de Resposta ao Item

Segundo Church (2010), a Teoria de Resposta ao Item (TRI) tem sido empregada para medir uma série de construtos latentes, como inteligência, traços de personalidade, individualismo e coletivismo, por exemplo. Seu uso já ocorre há mais de 60 anos, sendo mais comum nos campos de Educação e Psicologia (Samartini, 2006), mas já encontra aplicações em pesquisas brasileiras na área de Marketing (ver Lucian, 2012).

Embora a nomenclatura TRI represente um conjunto de modelos com especificidades diversas, a maioria deles possui em comum o estabelecimento de dois parâmetros. O primeiro refere a quanto o item (questão) se aproxima do traço a ser medido; e o segundo a quanto do traço está presente no respondente (um terceiro parâmetro associado a aleatoriedade pode ser modelado, a depender do interesse de pesquisa). Por essa razão, os pesquisadores em Educação desenvolveram diversos estudos utilizando TRI, valendo-se do parâmetro do item como uma medida de dificuldade de questões em uma avaliação e do parâmetro do respondente como medida da habilidade (ou conhecimento) dos estudantes. A técnica se popularizou como forma de padronizar resultados de alunos que realizam avaliações com questões distintas, de forma que seus desempenhos sejam comparados.

O campo da Psicologia, tradicionalmente envolvido com a mensuração de traços latentes, também apresenta vasta coleção de aplicações de TRI ao procurar quantificar a aderência dos itens de um instrumento ao construto a ser medido e a presença do construto nos respondentes. Nos campos da Administração, como o de Marketing, por exemplo, ainda são menos comuns as aplicações de TRI, predominando as chamadas abordagens clássicas como a análise fatorial (AF) para aplicação em modelagem por equações estruturais (SEM).

Há, no entanto, uma tendência de aumento do uso de TRI no campo da Administração e, em particular, de Marketing, motivada por propriedades que permitem maior quantidade de informações e resultados mais estáveis do que os da abordagem clássica. Uma das vantagens do uso de TRI é que, ao obtermos uma medida de quanto o item contém do traço, bastam poucas questões para que possamos identificar sua intensidade em um respondente. Isso é possível pelo fato de os modelos de TRI fornecerem uma distribuição de probabilidades para as respostas possíveis para cada questão em função do nível em que o traço está presente no respondente.

A nomenclatura TRI reúne uma família de modelos diferentes, podendo coletar dados binários ou escalares (Scherbaum, Finlinson, Barden & Tamanini, 2006). Os avanços em termos de ferramentas e aplicações têm sido maiores para dados binários e, por essa razão, acreditamos que o uso de TRI em Marketing (campo aculturado ao emprego de escalas de múltiplos pontos), embora crescente, é ainda incipiente e tende a permanecer assim no médio prazo.

Os modelos de TRI dividem-se em modelos cumulativos e de desdobramento (Samartini, 2006; Scherbaum et al., 2006). Modelos cumulativos

presumem que as respostas possíveis para um item implicam ordem e que o avanço nessa ordem representa o aumento do traço em análise. As escalas de concordância, nesse contexto, indicariam que quanto mais um respondente concordar com uma afirmação (não sendo essa inversa), mais possuirá do traço. Os modelos de desdobramento, por sua vez, não presumem acúmulo do traço na ordem das respostas possíveis a um item. Imaginemos, por exemplo, a seguinte afirmação: 'fumar deveria ser permitido apenas em ambientes abertos'. Uma pessoa absolutamente a favor da permissão do fumo em qualquer ambiente discordaria totalmente da afirmação, assim com concordaria plenamente uma pessoa absolutamente contra a permissão de fumar em qualquer ambiente. As pessoas que não estivessem nos extremos de opinião sobre a permissão para fumar distribuiriam suas respostas nos níveis intermediários de concordância. Enquanto não presumem acúmulo do traço de acordo com uma ordem das respostas aos itens, os modelos de desdobramento trazem uma distribuição de probabilidades para cada resposta possível como uma função da presença do traço em cada respondente.

A comparação de resultados entre TRI e análise fatorial confirmatória, sejam obtidos por meio de simulações ou de estudos empíricos, tem demonstrado maior adequação por parte da TRI (ver Salzberger & Koller, 2013; e Buchbinder, Goldszmidt & Parente, 2012) na validação de medidas. Aparentemente, as medidas validadas por TRI funcionam de maneira mais estável em contextos distintos, ao passo que as validadas por meio de AFC requerem maiores adaptações em contextos distintos (por contextos distintos, entendemos variações entre formas de coleta de dados (entrevistas pessoais ou por telefone, e autopreenchimento de questionários, por exemplo), de momentos de coleta, e de amostras que representem populações distintas (Meade & Lautenschlager, 2004).

A explicação para essas diferenças pode ser provida por meio de propriedades teóricas da TRI. Uma delas é a que a obtenção de características dos itens e dos indivíduos por meio das respostas dadas são independentes entre si. Em outras palavras, é possível determinar os parâmetros dos itens (dificuldade da questão ou presença do traço) com base em diferentes conjuntos de respondentes representativos de populações diversas (Salzberger & Koller, 2013; Scherbaum et al., 2006).

Os modelos clássicos (análise fatorial e de consistência interna) são baseados em correlações para a construção de escores e estimação de parâmetros para os itens. Por não contarem com a separação entre os parâmetros dos itens e dos respondentes, seus resultados ficam restritos às características da amostra e, consequentemente, à sua representatividade. Esse é um dos prováveis motivos para a necessidade de adaptações de escalas já validadas pela abordagem clássica em uma cultura quando se conduzem estudos interculturais, uma vez que os escores são construídos de forma limitada à amostra original e os parâmetros dos itens são dela dependentes, o que não ocorre em TRI, ao menos teoricamente. Essa fonte de viés da abordagem clássica é uma de suas desvantagens em termos de estabilidade de resultados. Outra vantagem da TRI é que o erro padrão dos itens varia ao longo de todos os níveis do traço, ou seja, é possível determinar o traço latente para cada um de seus níveis (Scherbaum et al., 2006).

No entanto, ao contrário das abordagens clássicas, os modelos de TRI não medem a confiabilidade de um instrumento completo de medida quando usamos escalas de múltiplos itens, já que a confiabilidade em TRI é apurada por item (Scherbaum et al., 2006), não dispondo de medidas gerais como o índice de confiabilidade composta, por exemplo, o que pode ser considerado uma desvantagem.

Outra desvantagem do uso de TRI é a necessidade de amostras maiores do que as necessárias nas abordagens clássicas (Church, 2010; Scherbaum et al., 2006). Ademais, o uso de TRI é complexo para os usuários que não dominam estatística de forma avançada e, por haver carência de recursos computacionais em termos de software com interfaces amigáveis para TRI, acreditamos que ainda há e haverá inibições de seu uso fora das áreas de educação e psicologia, em que seus desenvolvimentos foram mais proeminentes.

Os modelos de TRI apresentam também duas premissas e, ao menos uma delas pode ser interpretada como uma desvantagem frente à abordagem clássica, tratando-se da unidimensionalidade do traço. Os modelos TRI costumam presumir que um instrumento mede um traço único, embora existam modelos multidimensionais de TRI que são, no entanto, por demais complexos e de difícil implementação (Buchbinder et al., 2012; McDonald, 2010). Os modelos da abordagem clássica ajustam mais facilmente múltiplos traços na validação de instrumentos.

A outra premissa dos modelos TRI é conhecida como independência local (ou condicional) e significa que as respostas dadas para um item dependem exclusivamente do traço latente e não afetam as respostas nos outros itens e nem são por elas afetadas. Essa premissa pode ser uma explicação para o fato dos pesquisadores que preferem a TRI afirmarem que os parâmetros dos itens não dependem da amostra e com isso suas estimativas são estáveis. No entanto, o argumento é questionável, pois se trata de uma premissa, nem sempre observável.

A comparação entre as propriedades de TRI e das abordagens clássicas permite que pensemos em situações em que a escolha por uma ou outra abordagem seja mais ou menos adequada. A abordagem clássica poderia ser indicada para quando imaginarmos haver constância do erro padrão ao longo dos níveis do traço em um item. No entanto, uma situação como essa não é muito plausível, fazendo da escolha por um modelo de TRI preferível. TRI também

deve ser uma melhor escolha quando não dispomos de amostras representativas da população para a qual se pretende desenvolver a medida, por conta da independência entre os parâmetros dos itens e dos respondentes. Essa mesma propriedade, por conferir maior estabilidade dos parâmetros de item independentemente do contexto, também credencia TRI como a primeira opção para a criação de novas medidas ou o refinamento de medidas existentes.

Ao objetivarmos aferir a confiabilidade geral do instrumento, o uso da abordagem clássica seria mais recomendável, ao passo que TRI seria mais apropriada para obter confiabilidade nos diferentes níveis do traço, por item. Outro critério a ser considerado pode ser a parcimônia, também discutível. Se por um lado, a unidimensionalidade em TRI presume um modelo mais parcimonioso, por outro lado, trata-se de uma restrição a modelos multidimensionais, que podem fazer mais sentido em algumas formulações teóricas. Ademais, os modelos de TRI são de aplicação mais complexa e exigem mais recursos computacionais e técnicos do que os modelos clássicos, sobretudo quando ajustamos modelos multidimensionais, o que faz do uso de TRI menos parcimonioso.

Uma aplicação em que se destaca a aderência da TRI é o caso de respostas extremas, ou daquelas situações em que o respondente se posiciona nos extremos das questões (de Jong, Steenkamp, Fox & Baumgartner, 2008). Justamente pela propriedade dos modelos de decomposição apresentarem probabilidades distintas para cada resposta em função do valor do traço de cada indivíduo, sem presumir acúmulo do traço de acordo com a ordem das respostas possíveis, é possível discriminar melhor o traço nos respondentes que escolhem total concordância ou total discordância com um item do que nos modelos cumulativos. A esse atributo, soma-se a possibilidade de haver variação do erro padrão do traço em cada um de seus níveis, permitindo diferentes graus de precisão para indivíduos que estão nos extremos ou em níveis intermediários do traço. A capacidade de bem lidar com o estilo de resposta extrema é também um benefício do uso de TRI para o tratamento da variância comum ao método – a tendência de o respondente se posicionar de forma unívoca (podendo ser muito favorável ou muito desfavorável ao traço ao longo de suas respostas) por todo o instrumento de medida. Ao permitir variação na estimativa dos parâmetros do item para diferentes níveis do traço, o estilo de resposta extrema não contaminará as estimativas dos respondentes que estão em níveis intermediários.

O estilo de resposta extrema permite que os modelos de TRI identifiquem questões que funcionam melhor como dicotômicas, não escalares. Por essa razão, os avanços em TRI têm sido maiores para o desenvolvimento de instrumentos que coletem dados binários e com isso a tradição de escalas de múltiplos itens observada em Marketing pode inibir o aumento do emprego de métodos de TRI no campo.

O Quadro 3 sintetiza os fundamentos da abordagem clássica e de TRI, suas vantagens e desvantagens, e aplicações mais apropriadas em cada caso.

ASPECTOS	ABORDAGENS CLÁSSICAS	TEORIA DE RESPOSTA AO ITEM TRI
Fundamentos	Determinam escores individuais e parâmetros dos itens (como cargas fatoriais, variância média extraída e confiabilidade, por exemplo) com base na estrutura de correlações. Os resultados não são independentes do contexto em que os dados são coletados (formas de coleta, momentos de coleta e amostras distintas).	Calcula parâmetros para o item (aderência ao construto medido) e para o respondente (valor do traço) de forma independente. Costuma presumir unidimensionalidade (nos modelos mais simples) e independência local.
Vantagens	Menor complexidade. Ampla disponibilidade de recursos computacionais. Facilidade de ajuste de modelos multidimensionais. Requerem amostras menores do que os modelos de TRI. Geram índices de confiabilidade globais.	Os modelos não precisam ser cumulativos. Maior estabilidade dos parâmetros de itens em dados coletados em contextos distintos. Variação do erro padrão do item de acordo com o nível do traço no respondente. Permite aferir confiabilidade por item. Com poucas perguntas, é possível estabelecer o valor do traço no indivíduo.
Aplicações recomendadas	Quando consideramos haver constância do erro padrão ao longo dos níveis do traço em um item. Quando desejamos obter um indicador global de confiabilidade do instrumento. Quando ajustamos um modelo de medida multidimensional.	Quando consideramos haver variação do erro padrão ao longo dos níveis do traço em um item. Quando não podemos garantir a representatividade da amostra. Para a criação de novas medidas e/ou para refinar medidas existentes. Quando desejamos obter a confiabilidade por item do instrumento. Quando nos defrontamos com o estilo de respostas extremas.

Quadro 3 - Comparação entre TRI e a abordagem clássica de mensuração (AF e SEM).

7.2 Estimadores Bayesianos

De acordo com Raudenbush e Bryk (2002), a estatística clássica (não estamos nos referindo à abordagem clássica da mensuração em Marketing, mas à abordagem conhecida como clássica no campo da Estatística) assume que os parâmetros populacionais são constantes (fixos) e que os dados utilizados em estudos empíricos representam amostras probabilísticas em um universo de amostras possíveis. Já na abordagem Bayesiana (inspirada no teorema de Bayes), a ideia de probabilidade não é representada por frequência relativa em amostras repetidas, mas por quantificar a incerteza do pesquisador sobre os parâmetros desconhecidos que geram os dados amostrados. Nessa abordagem, os próprios parâmetros possuem uma distribuição de probabilidades que descreve a incerteza do pesquisador sobre os seus valores.

Na visão clássica, a estimativa de um ponto (e também de um intervalo de confiança) representa uma boa inferência para o valor do parâmetro quando obtida por meio de um método confiável, o qual assumimos possuir propriedades teóricas adequadas. O parâmetro populacional não é considerado uma variável aleatória e por essa razão não podemos atribuir a ele uma probabilidade. Na verdade, é o cômputo do intervalo em que o parâmetro deve estar contido que deve capturá-lo com algum grau de confiança.

A Estatística Bayesiana, por sua vez, assume que os parâmetros possuem uma distribuição de probabilidades e assim podemos fazer inferências com base nessa premissa. Uma distribuição *a priori* descreve as crenças do pesquisador a respeito do parâmetro antes da coleta dos dados. Depois que os dados estão disponíveis, revisamos essa distribuição *a priori* com base no que neles observamos, com o intuito de propor uma distribuição *a posteriori*, que combina as evidências que os dados trazem à proposta prévia. A estimativa de um ponto, nesse caso, pode ser a tendência central da distribuição *a posteriori* (como sua média ou sua mediana, por exemplo). A estimação de um intervalo na abordagem Bayesiana pode se basear numa amplitude de valores possíveis para o parâmetro, que serve como base para calcularmos a probabilidade *a posteriori* de os valores do parâmetro estarem dentro do intervalo.

As distribuições *a priori* podem incorporar eventual conhecimento prévio sobre os parâmetros, ou podem trazer pouca informação na construção à *posteriori*, quando comparada à informação trazida pelos dados. Essa segunda situação (quando há *prioris* pouco informativas) trata de *prioris* conhecidas como *prioris* de referência, que não dependem de ajuste fino para serem propostas. Seu benefício é o fato de 'deixarem os dados falarem' por si mesmos. A aplicação de inferência Bayesiana é arriscada quando trabalhamos com amostras pequenas, que em geral requerem *prioris* informativas (Congdon, 2006). O risco está no fato de que as *prioris* em amostras pequenas prevalecem no resultado final (*a posteriori*), que acaba por refletir os julgamentos pessoais prévios do pesquisador. Já quando contamos com amostras muitos grandes, em geral, os valores dos parâmetros estimados por meio de abordagens clássicas e das medidas de tendência central das distribuições *a posteriori* da abordagem Bayesiana tendem a coincidir (ou ser muito semelhantes entre si) (Raudenbush & Bryk, 2002).

A base da inferência Bayesiana contemporânea para estimação de parâmetros é o uso dos métodos conhecidos como *Markov Chain Monte Carlo* (MCMC), que consistem em simulações sequenciais para a distribuição de parâmetros em longas cadeias (Gamerman & Lopes, 2006). O interesse é sumarizar os parâmetros resultantes de um método MCMC sob a forma de esperanças, densidades e probabilidades (Congdon, 2006) obtidas por meio de simulações inspiradas no princípio de Monte Carlo e pouco confiáveis quando não são (aproximadamente) normais ou quando são multimodais.

O método original de Monte Carlo presume um conjunto de simulações independentes entre si. Os métodos MCMC, por sua, vez, geram simulações pseudoaleatórias por meio de cadeias de Markov, em que os parâmetros são considerados sequências de variáveis aleatórias. Uma cadeia somente pode ser considerada de Markov se apenas o passo anterior for relevante para o próximo (Rossi, Allenby & McCulloch, 2006); a simulação a partir de uma cadeia estável de Markov converge para uma distribuição estacionária. Assim se estabelece um esquema de simulação MCMC que converge para a estabilidade.

Há muitas questões a respeito da obtenção de convergência dos métodos de simulação MCMC. Costuma ser necessário estabelecer uma sequência inicial e curta de simulação (*burn in*), que não será aproveitada na distribuição final, por conta de os parâmetros simulados inicialmente poderem ser inadequados; as simulações obtidas por MCMC são autocorrelacionadas e assim muitas são necessárias para prover resultados utilizáveis (Rossi *et al*., 2006). Ademais, pode haver alguma demora para que se encontre a região da densidade *a posteriori* em que a tendência central do parâmetro se encontra, o que dependerá do tamanho da amostra, da complexidade do modelo e do método de simulação. Se as cadeias forem desenvolvidas de forma satisfatória, a autocorrelação tenderá a zero conforme a simulação avançar. Caso contrário, pouca informação acerca da distribuição *a posteriori* será provida em cada iteração e uma simulação de maior tamanho será necessária (Congdon, 2006).

Há diversos esquemas de simulação MCMC; o algoritmo que serve como base a todos é conhecido como Metropolis-Hastings (Congdon, 2006). Outro esquema bastante popular é o amostrador de Gibbs, um caso especial do algoritmo Metropolis-Hastings capaz

de simular distribuições marginais em sequência; embora gere sequências autocorrelacionadas, o amostrador de Gibbs termina por "se livrar" dos valores iniciais da cadeia e converge para uma distribuição estacionária.

Especialmente relevante à modelagem de variáveis latentes é o conceito de aumento de dados (*data augmentation*), usado para moldar a verossimilhança de um modelo de alguma natureza (como a modelagem por equações estruturais, por exemplo); o amostrador de Gibbs pode ser usado para essa finalidade. O conceito de aumento de dados consiste em adicionar informação indisponível (como a estimação de variáveis latentes) ao conjunto de dados por meio de sua modelagem. Rossi, Allenby & McCulloch (2006) demonstram que uma variedade de modelos podem ser construídos por meio de aumento de dados quando não observamos variáveis diretamente. Para aprofundamento a respeito dos algoritmos de simulação MCMC, recomendamos a leitura de Gamerman & Lopes (2006) e de Rossi, Allenby & McCulloch (2006).

Particularmente em relação ao uso da Análise Fatorial Confirmatória na validação de construtos, a abordagem Bayesiana possui algumas vantagens em relação à estatística clássica. Em primeiro lugar, os pesquisadores que preferem o uso da inferência Bayesiana consideram que com ela é possível usar amostras menores do que na abordagem frequentista (clássica) (Rossi, Allenby & McCulloch, 2006); o argumento, no entanto, somente é verdadeiro quando temos *prioris* informativas, o que, como já mencionamos, é arriscado. Para mitigar o risco, sugerimos que pesquisadores realizem extensiva revisão da literatura e de resultados empíricos passados para definir as distribuições *a priori* que utilizarão em seus modelos.

Outra vantagem é que o uso de estimadores Bayesianos não precisa violar premissas acerca de distribuição das variáveis utilizadas. Grande parte dos itens das escalas em Marketing são coletados como variáveis ordinais (graus de concordância, por exemplo), mas tratados por modelos da estatística clássica que presumem que os dados coletados sejam normalmente distribuídos, como é o caso da estimação por máxima verossimilhança, o caso mais frequente em Análise Fatorial Confirmatória e em Equações Estruturais. É muito improvável uma variável ordinal se distribuir normalmente, ou até mesmo impossível, considerando que a distribuição normal é exclusiva para variáveis quantitativas contínuas. Ao não presumir normalidade na distribuição dos dados, a inferência Bayesiana se adéqua melhor à modelagem de variáveis ordinais (Byrne, 2001).

A abordagem clássica também confia por vezes na aproximação assintótica para prover funções de densidade de probabilidade para o conjunto de estimadores amostrais. Mesmo que as aproximações assintóticas não presumam a normalidade dos dados, podem não se manter plausíveis em modelos não lineares, diferentemente do que ocorre com a estimação Bayesiana (ver Zellner & Rossi, 1984). Ademais, a modelagem por aproximação assintótica necessita de amostras muito grandes, uma clara desvantagem em relação à modelagem Bayesiana. Por fim, os modelos Bayesianos são menos sensíveis à presença de *outliers*, pois a distribuição de parâmetros é estimada com base na maior parte da amostra e menos nos casos extremos (Hahn & Doh, 2006).

Por considerar uma distribuição possível para os parâmetros na população, e não a existência de um parâmetro fixo populacional, alguns autores consideram a inferência Bayesiana como a forma mais adequada (senão a única) de ajustar modelos em Marketing (ver Rossi, Allenby & McCulloch, 2006; e Park & Kim, 2013). O argumento está no fato de que conseguimos modelar os comportamentos e atitudes de cada indivíduo em função de suas características individuais em vez de estimar um parâmetro médio para a população inteira (uma limitação dos modelos da estatística clássica). Como em Marketing é relevante compreender os agentes de forma personalizada, essa propriedade dos modelos Bayesianos tem impulsionado o uso desse tipo de inferência na área. Tal benefício dos métodos Bayesianos se aproxima de um dos benefícios dos modelos de TRI que, ao modelarem os parâmetros dos respondentes, também podem ser considerados Bayesianos em sua natureza. No entanto, a estimação Bayesiana em Análise Fatorial e Equações estruturais, ao contrário de TRI, não separa os parâmetros dos itens dos respondentes e se baseia na estrutura de associação entre os dados (assim como na estatística clássica) e, portanto, depende fortemente das características da amostra. Efetivamente, a distribuição de parâmetros *a posteriori*, quando usamos *prioris* de referência, representa bem a amostra utilizada que, portanto, deve ser bastante representativa da população.

Os modelos Bayesianos têm sido crescentemente empregados em diversos campos por conta de sua natureza intuitiva e de suas vantagens frente à inferência clássica. Esse movimento têm sido impulsionado pelo aumento de recursos de *software* de prateleira capazes de prover a estimação Bayesiana por meio de interfaces amigáveis, e também pelo avanço de *hardware* capaz de processar simulações com sequências muito grandes (na casa dos milhares). Um exemplo de aplicação desse tipo é o algoritmo presente no pacote de Equações Estruturais AMOS. No entanto, as ferramentas de prateleira provêm pouca flexibilidade aos pesquisadores em termos de escolha do simulador MCMC ou mesmo de extração de resultados individualizados para cada respondente, o que seria um dos principais benefícios da escolha por um modelo Bayesiano. Existem ferramentas mais flexíveis, como o pacote R, por exemplo, que exigem, no entanto, conhecimento avançado em estatística e habilidades em programação, muitas vezes incomuns entre as habilidades dos pesquisadores em nossa área.

O uso da inferência clássica por máxima verossimilhança na construção de escalas é adequado

quando observamos normalidade nos dados e a ausência de dados discrepantes, mas essa é uma situação improvável e sua ausência fomenta a escolha por modelos Bayesianos. Modelos MCMC também são preferíveis quando temos amostras pequenas, mas para isso dependemos de *prioris* informativas, rigorosamente determinadas por formulações teóricas coerentes e pela consulta a resultados de estudos passados; caso contrário, os parâmetros serão fortemente dependentes de idiossincrasias do pesquisador que propõe a *priori*.

O uso de variáveis contínuas para pontuar os indicadores também permite a estimação por máxima verossimilhança, enquanto o uso de variáveis ordinais e/ou discretas leva à escolha pela estimação Bayesiana. O Quadro 4 sintetiza os fundamentos da inferência clássica e da inferência Bayesiana em Análise Fatorial e em Modelagem por Equações Estruturais, suas vantagens e desvantagens, e aplicações mais apropriadas em cada caso.

ASPECTOS	ESTIMAÇÃO CLÁSSICA	ESTIMAÇÃO BAYESIANA
Fundamentos	Presume a existência de um parâmetro fixo na população e calcula seu intervalo de confiança por meio de uma abordagem frequentista (em geral, por meio de Máxima Verossimilhança).	Presume a existência de uma distribuição de parâmetros na população e a estima com base em uma formulação *a priori*, a ser aprimorado para uma distribuição que leva em conta os dados coletados, chamada de *posterior*. A estimação é feita por simulações *Markov Chain Monte Carlo* (MCMC).
Vantagens	Mais difundida em pacotes estatísticos de prateleira e de execução mais simples.	Não presume normalidade na distribuição dos dados; não se limita à modelagem com variáveis quantitativas contínuas; possui baixa sensibilidade a dados discrepantes. Estima valores individuais para os respondentes, em vez de um parâmetro médio para a população.
Aplicações recomendadas	Amostras grandes, com dados normalmente distribuídos, ausência de *outliers*, variáveis quantitativas contínuas.	Amostras pequenas, com variáveis de diversas naturezas (como qualitativas e discretas, por exemplo), com presença de *outliers*.

Quadro 4 - Comparação entre a estimação clássica e a Bayesiana em Análise Fatorial e em Modelagem por Equações Estruturais.

7.3 Modelagem por mínimos quadrados *parciais (partial least squares* - PLS)

O uso de Análise Fatorial Confirmatória e de Modelagem por Equações Estruturais em Marketing, conforme já mencionamos, tem sido mais comum por meio da estimação por máxima verossimilhança. Uma alternativa que tem se mostrado viável e ganho espaço em pesquisas empíricas na área, com forte influência do campo dos Sistemas de Informação, é o uso dos modelos por mínimos quadrados parciais, em inglês *partial least squares*, conhecidos como PLS. Embora os modelos baseados em covariância (como os estimados por máxima verossimilhança, por exemplo) sejam mais conhecidos pelos pesquisadores em nossa área, os modelos PLS também são modelos de equações estruturais, porém baseados em variância.

A diferença fundamental entre os modelos baseados em covariância e aqueles baseados em variância é que os primeiros obtêm índices de ajuste globais ao comparar as covariâncias (ou correlações, em modelos padronizados) estimadas pelo modelo com aquelas de fato observadas nos dados coletados e procedem a testes de quiquadrado para avaliar as diferenças. O ajuste geral do modelo é acessado nesses casos por indicadores baseados na estatística de quiquadrado e em testes que verificam a significância dos erros oriundos da diferença entre o que é observado e o que é estimado. Já nos modelos baseados em variância, não existem estatísticas globais de ajuste, sendo o modelo avaliado pela significância das relações propostas entre variáveis (também disponíveis nos modelos estimados com base em covariância) e pelo total de variabilidade das variáveis de interesse que o modelo consegue explicar (R^2).

Por não depender de uma estrutura de covariância para ajustar um modelo, o método PLS tem sido empregado em pesquisas que utilizam indicadores formativos, dado que não exige que haja correlação entre os indicadores usados para medir um mesmo construto. Esta tem sido considerada uma boa razão para o uso do método por diversos pesquisadores que

se propõem a fazer estudos com indicadores formativos, mas também tem sido o alvo das discussões sobre suas falhas (ver Diamantopoulos, 2011). Justamente por não contar com indicadores de ajuste globais, os modelos PLS estão limitados a verificar se as relações propostas fazem sentido individualmente, mas não permitem verificar se o modelo como um todo é plausível. Por essa razão, a literatura costuma indicar o uso de PLS em modelos exploratórios, em que há pouca teoria desenvolvida (ver Hair, Ringle & Sarstedt, 2011; Henseler, Ringle & Sinkovics, 2009; Marcoulides & Saunders, 2006; e Ringle, Sarstedt & Straub, 2012). No entanto, é discutível a adequação dessa situação em procedimentos de validação nomológica, em que justamente a relação teórica entre construtos e variáveis deve estar bem amadurecida para o pesquisador, sendo esse em geral o caso para que se usem equações estruturais.

É necessário informar que os modelos baseados em covariância (à exceção dos modelos de análise fatorial, sendo que a análise fatorial confirmatória é um caso específico de modelagem por equações estruturais) não são necessariamente modelos exclusivos para indicadores refletivos. No entanto, em termos práticos, acabam sendo, pois por se basearem na covariância dos itens usados para medir um mesmo construto, costumam não serem 'identificados' (isto é, não convergir para um ajuste) quando não se estipula essa covariância ou quando ela não é suficiente grande para que o modelo 'rode'. Por essa razão, mesmo que os modelos baseados em covariância sejam mais recomendáveis para a validação nomológica com medidas de diversas naturezas (formativas ou refletivas) por contarem com indicadores globais de ajuste, terão maior dificuldade para convergir e exigirão maior complexidade para serem ajustados (Diamantopoulos, 2011). Um problema adicional dos modelos PLS é o fato de não permitirem a estimação de erros para os indicadores formativos, ao passo em que nos modelos baseados em variância os erros dos indicadores formativos estão presentes; entendemos não ser razoável não estimar erros de medida.

Outras indicações para o uso de PLS encontradas na literatura são as situações em que não temos amostras grandes, já que menos parâmetros são estimados em comparação aos modelos baseados em covariância, e assim economizamos graus de liberdade (embora saibamos que, em qualquer estimação frequentista, os intervalos de confiança serão maiores para amostras pequenas, ou seja, menos precisos). Ademais, os modelos baseados em variância também não exigem a normalidade da distribuição dos dados coletados. O uso de PLS tem sido facilitado e crescido pela disponibilidade de ferramentas com interface amigável, como SmartPLS e PLS-Graph.

O Quadro 5 sintetiza os fundamentos da Modelagem por Equações Estruturais baseada em covariância (em particular a estimação por máxima verossimilhança) e baseada em variância (em particular a estimação por mínimos quadrados parciais - PLS), com vantagens e desvantagens e aplicações mais apropriadas para cada caso.

ASPECTOS	MÁXIMA VEROSSIMILHANÇA	PLS
Fundamentos	Obtém índices de ajuste globais ao comparar as covariâncias estimadas pelo modelo com aquelas de fato observadas nos dados coletados e procedem a testes de quiquadrado (ou nele inspirados) para avaliar as diferenças encontradas.	Acessa a significância das relações propostas entre variáveis e a variabilidade das variáveis de interesse que o modelo consegue explicar (R^2).
Vantagens	Calcula índices de ajuste global do modelo, alguns com testes de significância. Estima erro para as medidas formativas, quando há.	Não presume normalidade na distribuição dos dados. Requer amostras menores.
Aplicações recomendadas	Amostras grandes, com dados normalmente distribuídos. Presença apenas de indicadores refletivos.	Amostras menores, com dados sem distribuição normal. Pesquisas que utilizam indicadores formativos.

Quadro 5 - Comparação entre a Modelagem por Equações Estruturais por máxima verossimilhança e por mínimos quadrados parciais (PLS).

8 CONSIDERAÇÕES FINAIS

Parafraseando uma recomendação de Pedhazur & Schmelkin (1991), para ser significante, qualquer atividade de pesquisa, incluindo a leitura de artigos científicos, deve alicerçar-se em reflexões críticas; aos estudantes de iniciação científica, mestrado, doutorado, e pesquisadores, recomendamos

que reflitam criticamente acerca de suas escolhas metodológicas, em particular as que envolvem mensuração.

Com efeito, de pouco ou nada vale o emprego de modelagens estatísticas sofisticadas se o banco de dados a partir do qual essas análises serão feitas contém números que não refletem adequadamente os fenômenos que devem representar. Com este pensamento em mente, desenvolvemos neste artigo uma revisão ampla, da evolução histórica, do estado atual e das tendências futuras do problema da mensuração dos construtos em Marketing.

Em nossa visão, o desenvolvimento acadêmico e profissional de Marketing é uma variável dependente do desenvolvimento de pesquisas para aperfeiçoar conhecimentos em nosso campo. Mas a pesquisa em Marketing, por sua vez, é dependente do nível de desenvolvimento metodológico, que passa pelas questões de mensuração, de *design* e de análise de dados. Sem dúvidas, não há desenvolvimento de pesquisas sólidas em Marketing sem uma concentração cuidadosa com a mensuração que se faz das variáveis e construtos teóricos. Em linha com a percepção de Lee e Hooley (2005), recomendamos que pesquisadores em Marketing dediquem o tempo necessário para que seus modelos de mensuração sejam percebidos como sólidos; apenas posteriormente faz sentido elaborar modelos avançados para testar hipóteses entre construtos.

Nosso artigo fez uma incursão ampla sobre o assunto. Embora tenhamos na restrição do espaço o impedimento de sermos exaustivos, cuidamos de não deixar de fora qualquer dos temas centrais sobre o assunto, o que nos faz acreditar que, do ponto de vista acadêmico, nossa pesquisa alcança utilidade a pesquisadores, iniciantes ou veteranos, quando estes procurarem uma atualização e uma visão global sobre o assunto.

Além disto, entendemos que este artigo traz uma contribuição potencial para o campo da educação em Marketing, especialmente para a área de Pesquisa de Marketing ministrada em cursos de graduação ou para disciplinas de conteúdo metodológico de cursos de pós-graduação. O artigo pode ser utilizado, portanto, como componente de uma disciplina mais geral, assim como um texto introdutório de uma disciplina mais específica sobre mensuração (já temos acumulado experiências de disciplinas dessa natureza na pós-graduação brasileira, como por exemplo na EAESP/FGV (São Paulo) e na ESPM (São Paulo), na FUMEC (Minas Gerais) e na UFPB (Paraíba).

A exposição feita não deixa dúvidas de que já avançamos muito em termos teóricos, inclusive com uma crescente contribuição de pesquisadores de Marketing para o tema mensuração (diferentemente do que já ocorreu no passado, quando o campo de Marketing dependia dos desenvolvimentos oriundos dos campos de Psicologia e Educação). Colocamos como desafio para pesquisadores brasileiros seguirmos nos apropriando do assunto, evoluindo nos avanços de fronteira e buscando expandir nossas análises sobre o tema. O conteúdo exposto também mostra o quanto ainda temos a avançar, e os desafios são bastante motivadores. Nossa demanda agora é seguir adiante nos estudos e nas aplicações para aperfeiçoar ainda mais o conhecimento que produzimos em Marketing.

REFERÊNCIAS

American Psychological Association. (1985). Standards for educational and psychological tests. Washington, DC: Author.

Andrade, D. F., Tavares, H. R., & Valle, R. C (2000). Teoria da resposta ao item: conceitos e aplicações. *14º Simpósio Nacional de Probabilidade e Estatística – SINAPE*. São Paulo: Associação Brasileira de Estatística.

Aranha, F., & Zambaldi, F. (2008). *Análise fatorial em administração*. Sao Paulo: Cengage Learning.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*(1), 45-87.

Barboza, S. I. S., Carvalho, D. L. T., Soares Neto, J. B. & Costa, F. J. (2013). Variações de Mensuração pela Escala de Verificação: uma análise com escalas de 5, 7 e 11 pontos. *Teoria e Prática em Administração*, *3*(2), 99-120.

Belk, R. W. (1985). Materialism: trait aspects of living in the material world. *Journal of Consumer Research*, *12*(3), 265-280.

Buchbinder, F., Goldszmidt, R., & Parente, R. (2012). Item Response Theory and Construct Measurement in Emerging Markets. *Research Methodology in Strategy and Management*, *7*, 73-100.

Bussab, W. O., & Morettin, P. (2007). Estatística Básica. São Paulo: Saraiva.

Byrne, B. (2001). *Structural Equation Modeling with Amos: Basic Concepts. Applications. and Programming*. Mahwha, New Jersey: Lawrence Erlbaum.

Church, A. (2010). Measurement issues in cross-cultural research. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The Sage Handbook of Measurement* (pp. 151-176). London, UK: Sage Publications.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research (JMR)*, *16*(1).

Congdon, P. (2006). *Bayesian Models for Categorical Data*. Chichester, England: John Wiley &Sons, Ltd.

Costa, F. J. (2011). Mensuração e Desenvolvimento de Escalas. Rio de Janeiro: Editora Ciência Moderna.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, *16*(3), 297-334.

de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research*, *45*(1), 104-115.

Devellis, R. F. (1991). *Scale development*: theory and applications. Newbury Park, CA: SAGE Publications.

Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *Mis Quarterly*, *35*(2), 335-358.

Diamantopoulos, A. & Winklhofer, H. M. (2001) Index construction with formative indicators: an alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, *14*(2), 370-388.

Faraway, J. J. (2006). *Extending linear models with R*. Boca Raton, FL: Chapman & Hall/CRC.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of marketing research*, *18*(8), 382-388.

Gamerman, D., & Lopes, H. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* Boca Raton, FL: Chapman & Hall/CRC.

Gerbing, D. W., & Anderson, J. (1988). An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment. *Journal of Marketing Research*, *25*, 186-192.

Gonçalves, H. M. M. (2013). Multi-group invariance in a third-order factorial model: attribute satisfaction measurement. *Journal of Business Research*, *66*, 1292-1297.

Hahn, E. D., & Doh, J. P. (2006). Using Bayesian methods in strategy research: an extension of Hansen et al. *Strategic Management Journal*, *27*(8), 783-798.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371-388.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *The Journal of Marketing Theory and Practice*, *19*(2), 139-152.

Hao, L. & Naiman, D. Q. (2007). *Quantile regression*. Thousand Oaks: Sage Publications.

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in international marketing*, *20*(1), 277-319.

Hodge, D. R. & Gillespie, D. F. (2007). Phrase completion scales: a better measurement approach than Likert scales? *Journal of Social Service Research*, *33*(4), 1-12.

Jarvis, C. B., Mackenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199-218.

Kamakura, W. A., & Mazzon, J. A. (2013). Socioeconomic status and consumption in an emerging economy. *International Journal of Research in Marketing*, *30*(1), 4-18.

Kloke, J. D., & Mckean, J. W. (2012). Rfit : Rank-based estimation for linear models. *The R Journal*, *4*(2), 57–64.

Lee, C. E. (1965). Measurement and the development of science and marketing. *Journal of Marketing Research*, *2*(1), 20-25.

Lee, N., & Hooley, G. (2005). The evolution of "classical mythology" within marketing measure development. *European Journal of Marketing*, *39*(3), 365-385.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives in Psychology*, *140*, 1-55.

Lucian, R. (2012). *Mensuração de atitudes*: a proposição de um protocolo para elaboração de escalas. Tese (Doutorando em Administração). Programa de Pós-Graduação em Administração da Universidade Federal de Pernambuco – PROPAD-UFPE.

Marcoulides, G. A., & Saunders, C. (2006). Editor's comments: PLS: a silver bullet?. *MIS quarterly*, *30*(2), iii-ix.

Mari, L. (2005). The problem of foundations of measurement. *Measurement*, *38*(4), 259-266.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*(2), 99–114.

Meade, A. W., & Lautenschlager, G. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111-121.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures*: issues and applications. Thousand Oaks: Sage.

Nunnaly, J. (1978). *Psychometric Theory*. New York: McGraw-Hill Book Company.

Park, H. J., & Kim, S. H. (2013). A Bayesian network approach to examining key success factors of mobile games. *Journal of Business Research*, *66*(9), 1353-1359.

Pedhazur, E., & Schmelkin, L. P. (1991). *Measurement, design and analysis*: an integrated approach. Hillsdale: Lawrence Erlbaum Associates Inc. Publishers, 1991.

Pereira, B. B. (1997). Estatística: a tecnologia da ciência. *Boletim da Associação Brasileira de Estatística*, ano XIII, n. 37, 2º quadrimestre, 27-35.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*. (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Reardon, J., & Miller, C. (2012). The effect of response scale type on cross-cultural construct measures: an empirical example using Hall's concept of context. *International Marketing Review*, *29*(1), 24-53.

Richins, M. L., & Dawson, S. (1992). A Consumer Values Orientation for Materialism and Its Measurement: Scale Development and Validation. *Journal of Consumer Research*, *19*(3), 303-316.

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: a critical look at the use of PLS-SEM in MIS quarterly. *MIS quarterly*, *36*(1), iii-xiv.

Rossi, P., Allenby, G., & McCulloch, R. (2006). *Bayesian statistics and marketing*. Chichester, England: John Wiley and Sons, Ltd.

Rossiter, J. R. (2002). The COARSE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*(4), 305-335.

Rossiter, J. R. (2011) *Measurement for the Social Sciences*: the COARSE method and why it must replace psychometrics. New York: Springer.

Salzberger, T., & Koller, M. (2013). Towards a new paradigm of measurement in marketing. *Journal of Business Research*, *66*(9), 1307-1317.

Samartini, A. L. S. (2006). *Modelos com variáveis latentes aplicados à mensuração de importância de atributos*. Doctoral thesis, Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas (FGV/EAESP), Sao Paulo, Brazil.

Scherbaum, C., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research. *The Leadership Quarterly*, *17*(4), 366–386.

Sheather, S. J. (2009) *A modern approach to regression with R*. New York: Springer.

Steenkamp, J. -B. E. (2005). Moving out of the US silo: A call to arms for conducting international marketing research. *Journal of Marketing*, *69*(4), 6-8.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, *18*(2), 51-62.

Urbina, S. (2004). *Essentials of psychological testing*. New Jersey: John Wiley & Sons, Inc..

Van de Vijver, F. J., & Leung, K. (1997). Methods and data analysis for cross cultural research. Thousand Oaks, CA: SAGE.

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, 30(1), 72–91.

Yi, Y., & Gong, T. (2013). Customer value co-creation

behavior: scale development and validation. *Journal of Business Research, 66*, 1279-1284.

Zellner, A., & Rossi, P. E. (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics, 25*(3), 365-393.