PMKT
REVISTA BRASILEIRA DE PESQUISAS DE MARKETING, OPINIÃO E MÍDIA

PMKT – Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia
ISSN: 2317-0123 (On-line)
Editor-Chefe: Fauze Najib Mattar
Sistema de avaliação: Triple Blind Review
Idiomas: Português e Inglês
Publicado por: ABEP – Associação Brasileira de Empresas de Pesquisa

Mensuração e Escalas de Verificação: uma Análise Comparativa das Escalas de Likert e *Phrase Completion*

Measurement and Verification Scales: a Comparative Analysis between the Likert and Phrase Completion Scales

Submissão: 1/mar./2014 - Aprovação: 24/jun./2014

Severino Domingos da Silva Júnior
Mestre e Bacharel em Administração de Empresas pela Universidade Federal da Paraíba - UFPB. Professor na União de Ensino Superior de Campina Grande - UNESC. Membro dos Grupos de Pesquisas Consumo e Cibercultura GPCiber e do MEQAD - Métodos Quantitativos em Administração.
E-mail: juniordomingos.sdsj@gmail.com
Endereço profissional: Av. Estrela, nº 499 - Centro – 58306-190 - Cidade de Bayeux/PB – Brasil.

Francisco José Costa
Doutor em Administração de Empresas pela Fundação Getúlio Vargas – FGV-SP. Mestre e Bacharel em Administração de Empresas pela Universidade Estadual do Ceará - UECE. Professor no Programa de Pós-Graduação em Administração da Universidade Federal da Paraíba - UFPB.
E-mail: franzecosta@gmail.com

RESUMO

Este estudo se propôs a analisar a utilização de escalas do tipo Likert e *Phrase Completion* como alternativas de escalas de verificação para mensuração de construtos de múltiplos itens em pesquisas de Marketing e de Administração. Inicialmente, foi realizada uma revisão teórica que contribuiu para a comparação de natureza e de potencialidade das duas alternativas de medição. Foi utilizado o construto religiosidade intrínseca como referência de comparação e, em seguida, foi desenvolvido um questionário contendo itens dos dois tipos de escalas, o qual foi aplicado em uma amostra de 229 respondentes. Foram realizados procedimentos comparativos dos pares de itens correspondentes e das medidas agregadas. Os resultados mostraram que as escalas possuem diferenças entre si na verificação dos pares de itens, entretanto, não houve diferenças significativas na análise dos itens agregados. Isso indica que a escolha da escala é uma decisão dos pesquisadores e estes poderão levar em conta o tipo de pesquisa e as características dos respondentes.

PALAVRAS-CHAVE:
Mensuração, escala Likert, escala *Phrase Completion.*

ABSTRACT

This study analyses the use of Likert and Phrase Completion scales, as scale options for measuring multiple-item constructs in Marketing and Management research. Initially, a literature review was carried, what made possible the comparison of the nature and the potential of the two measurements alternatives. The construct intrinsic religiosity was used as a comparison reference, and, afterwards, a questionnaire was developed containing items from the two types of scales. Data collection generated a sample size of 229 respondents. Comparative proceedings of the paired corresponding items and of the aggregated measures were conducted, and the results demonstrated that the scales have differences between them regarding the verification of paired items; on the other hand, no significant differences were encountered through the analysis of the aggregated items. These findings indicate that the scale alternative is a researcher's decision, which may take into account the kind of research and the sample's profile.

KEYWORDS:
Mensuration, Likert scale, Phrase Completion scale.

1 INTRODUÇÃO

A mensuração é um dos meios pelos quais são acessados e descritos os dados para compreender os fatos e fenômenos de interesse. Por isto, a mensuração é uma questão presente em todas as ciências e são vários os estudos desenvolvidos sobre o tema publicados nos principais periódicos nacionais e internacionais.

O desenvolvimento mais consistente de estudos empíricos quantitativos somente é viável devido aos avanços na teoria e nas práticas de mensuração. Desta forma, a teoria da medição em ciências sociais e comportamentais, como Administração, Psicologia e Sociologia, tem avançado em seus estudos para propor alternativas que possam mensurar mais adequadamente suas variáveis, parte das quais são de conteúdo abstrato (por exemplo, tensão, satisfação, lealdade etc.).

O manuseio de dados e medidas dessas ciências originam informações e conhecimentos que podem ser direcionados tanto para objetivos acadêmicos quanto profissionais, o que sinaliza a necessidade de cuidado especial no processo de definição das medidas geradas; portanto é indispensável um conjunto de procedimentos bem fundamentados para a criação e uso de escalas que atendam adequadamente aos objetivos dos estudos e demandas profissionais. Esta pesquisa visa contribuir nessas análises, por meio da verificação de duas modalidades de escalas de uso corrente em pesquisas de Marketing e de Administração: Likert e *Phrase Completion*.

Conforme levantamento empreendido pelos autores, ao longo das últimas seis décadas uma grande quantidade dos estudos quantitativos desenvolvidos em Marketing utilizou a escala de Likert nos instrumentos de pesquisa que medem construtos como atitudes, percepções, interesses etc. Essa escala é usada para medir concordância de pessoas a determinadas afirmações relacionadas a construtos de interesse. No entanto, como será visto no próximo item, essa escala tem sido bastante criticada de modo que novas alternativas passaram a ser desenvolvidas (COSTA, 2011). Provavelmente, a melhor alternativa já sugerida foi a escala *Phrase Completion*, que mede o construto inserindo sua intensidade no próprio enunciado da escala, facilitando, potencialmente, o entendimento dos respondentes e medindo de forma mais confiável e válida o que está sendo investigado (HODGE; GILLESPIE, 2007).

A seguir são apresentados os fundamentos teóricos sobre escalas de verificação, no terceiro item, os procedimentos metodológicos do trabalho de campo realizado, no quarto estão os resultados do trabalho de campo, no quinto, as considerações finais do estudo e, no sexto e último item, as recomendações para novos estudos.

2 DISCUSSÃO TEÓRICA

A mensuração (ou medição) é definida como atribuição de símbolos, preferencialmente numéricos, à propriedade dos objetos que se deseja medir. Estes símbolos são direcionados a quantificar ou classificar determinadas características. Sendo assim, a medição é um processo de representação, relacionando algum aspecto do mundo real com sistemas simbólicos (MARI, 1996, 1999; FINKELSTEIN, 2003, 2009). Adotando esse conceito, entende-se que a mensuração pode ser realizada para capturar a essência do objeto mensurado, e visa facilitar a manipulação de dados de conjuntos de sujeitos ou simplesmente viabilizar melhor conhecimento do atributo.

Por tais usos se compreende o motivo pelo qual a mensuração é a base de sustentação do

desenvolvimento acadêmico e profissional de várias ciências. Mesmo que boa parte dos estudos desenvolvidos sobre mensuração seja realizada pelas ciências exatas (FINKELSTEIN, 2003), as ciências sociais vêm logrando progressos nessa área (COSTA, 2011).

A despeito de uma série de críticas e eventuais negações da própria possibilidade de medir determinadas variáveis (como desejo ou prazer, por exemplo), pela conceituação proposta, o problema não se encontra no ato de medir, mas no mecanismo de atribuição simbólica adotado pelo pesquisador, inclusive porque nem toda medição é direcionada a quantificar. Mesmo diante de eventuais limitações, a mensuração permanece sendo o mecanismo de viabilidade para o desenvolvimento de pesquisas empíricas associadas a construtos abstratos.

Particularmente nas ciências sociais e comportamentais, a mensuração de variáveis de interesse é realizada por meio de escalas específicas, as quais são construídas de modo a se adaptarem à natureza abstrata de grande parte dos construtos. As ditas escalas de mensuração são parte da instrumentação básica da medição, ganhando formatações variadas como os testes da Psicologia, os exames e provas da Educação ou as escalas diversas em Marketing e Administração.

As escalas no universo das ciências sociais e comportamentais, assim como em todas as outras ciências, são alicerçadas em pressupostos particulares e desenvolvidas por modelos específicos. Conforme Costa (2011), uma escala de mensuração é composta por um conjunto de indicadores, mais uma escala de verificação e um conjunto de regras:

- Os indicadores são os elementos de conteúdo que asseguram a presença do conceito do construto na escala de mensuração. São exemplos de indicadores as afirmações sobre determinado construto, para os quais um sujeito emitirá sua concordância; também são indicadores os pontos extremos de um comportamento (por exemplo, odiar e amar determinada marca).
- A escala de verificação envolve os números que vêm associados aos indicadores para sua medição. Por exemplo, podem-se adotar níveis de concordância de 1 a 5 ou utilizar números de 1 a 7 entre as opções de odiar e amar uma marca.
- As regras são as indicações para uso do instrumento, em termos de sua aplicação e interpretação. Por exemplo, pode-se definir que concordâncias entre 4 e 5 indicam alto nível de avaliação do construto sob análise e que valores entre 1 e 2 indicam baixo nível de avaliação do mesmo construto.

O interesse neste artigo está na análise das escalas de verificação. Estas têm sido motivos de controvérsia ao longo das últimas seis décadas. Por exemplo, em uma escala de concordância, quantos pontos são mais adequados para efeito de medição? Uma medição de 0 a 10 é mais eficiente que uma mensuração de 1 a 5? O uso de concordância para medir um construto é uma boa opção (por exemplo, quer-se medir a satisfação, mas mede-se a concordância com a frase **estou satisfeito**)?

Um dos focos centrais de discussão está no uso ou não, da chamada escala de Likert, a mais adotada em pesquisas. No item seguinte é apresentada esta escala e, no subitem posterior, apresenta-se a alternativa da escala *Phrase Completion*.

2.1 ESCALA DE LIKERT

O modelo mais utilizado e debatido entre os pesquisadores foi desenvolvido por Rensis Likert

(1932) para mensurar atitudes no contexto das ciências comportamentais. A escala de verificação de Likert consiste em tomar um construto e desenvolver um conjunto de afirmações relacionadas à sua definição, para as quais os respondentes emitirão seu grau de concordância. O Quadro 1 mostra um exemplo desta escala para medição de satisfação com um serviço, em 5 pontos.

QUADRO 1
Exemplo de escala de Likert.

ESTOU SATISFEITO COM O SERVIÇO RECEBIDO:				
Discordo totalmente	Discordo parcialmente	Não concordo nem discordo	Concordo parcialmente	Concordo totalmente
1	**2**	**3**	**4**	**5**

Nesta escala os respondentes se posicionam de acordo com uma medida de concordância atribuída ao item e, de acordo com esta afirmação, se infere a medida do construto. Construtos como autoestima, depressão, etnocentrismo, religiosidade e racismo são alguns exemplos recorrentemente mensurados por meio de escalas de Likert. A escala original tinha a proposta de ser aplicada com cinco pontos, variando de discordância total até a concordância total. Entretanto, atualmente existem modelos chamados do tipo Likert com variações na pontuação, a critério do pesquisador.

A grande vantagem da escala de Likert é sua facilidade de manuseio, pois é fácil a um pesquisado emitir um grau de concordância sobre uma afirmação qualquer. Adicionalmente, a confirmação de consistência psicométrica nas métricas que utilizaram esta escala contribuiu positivamente para sua aplicação nas mais diversas pesquisas (COSTA, 2011).

No entanto, mesmo diante de pontos positivos, a escala de Likert possui dificuldades significativas (CUMMINS; GULLONE, 2000; COELHO; ESTEVES, 2007; DAWES, 2008). De acordo com os críticos, perguntas com o modelo Likert solicitam do respondente pelo menos duas dimensões a serem analisados: conteúdo e intensidade. O indivíduo precisa verificar o conteúdo da proposição do item e, em seguida, opinar discordando ou concordando com a afirmação, considerando ainda a intensidade desta concordância. Embora não pareça ser um problema para efeito de uso, os críticos afirmam que esta característica aumenta o nível de complexidade cognitiva da escala, principalmente quando a escala possui muitos pontos (HODGE; GILLESPIE, 2003).

Há também um problema na utilização dessas escalas relacionado à denominação dos pontos (por exemplo, é fácil denominar 3 pontos, com 1 – discordo, 2 – não concordo nem discordo, e 3 – concordo; no entanto, denominar 10 pontos de uma escala de 1 a 10 é muito mais complicado). A própria seleção do número de pontos é complicada.

Estudos empíricos mostram que, em escalas de múltiplos itens com mensuração refletiva em relação ao construto, a confiabilidade é melhor em escalas cujos itens são medidos com mais de 7 pontos, e diminui quando os itens possuem menos de 5 pontos.

Entretanto, menos pontos parecem tornar mais fáceis as respostas, de modo que, ao aumentar o número de pontos ganha-se em consistência psicométrica e perde-se em segurança nas respostas. Obviamente, isso é um ponto fraco dessa escala. Uma opção que, potencialmente, resolve essa dificuldade está em ancorar os níveis extremos de concordância nos limites numéricos e deixar os demais pontos como níveis intermediários de concordância. Isso permite, inclusive, utilizar um número maior de pontos, como 1 a 10 ou 0 a 10, cujos níveis de conhecimento são mais generalizados em países como o Brasil.

Também há dificuldade de decisão sobre o número par ou ímpar de pontos. A indicação corrente é de que uma escala com número ímpar de pontos facilita a resposta por causa do ponto intermediário, que seria um nível neutro entre concordância e discordância. Mas a denominação de neutro a esse ponto central tem um problema do ponto de vista conceitual, pois em uma escala de concordância, quem é neutro não manifesta concordância alguma e aquele número central é um dado ponto de concordância (COSTA, 2011; HODGE; GILLESPIE, 2003).

Outra dificuldade encontrada está na possível perda de informação associada à escala Likert (RUSSELL; BOBKO, 1992), visto que a definição de pontos em uma escala de números inteiros (de 1 a 5 ou 1 a 10, por exemplo), torna as respostas obrigatoriamente discretas, sendo perdidas as referências intermediárias entre esses pontos (por exemplo, uma concordância entre 3 e 4 não tem como ser aferida por determinado sujeito).

Mesmo possuindo procedimentos estatísticos adequados para operacionalização de variáveis discretas esses problemas persistem devido à violação de suposições de abordagens paramétricas convencionalmente aplicadas (como a análise fatorial confirmatória que pressupõe que as variáveis são contínuas e normalmente distribuídas). Possivelmente, a perda de informação resultante das restrições dos dados na escala de Likert seria solucionada de duas formas: a primeira consiste na medição em maior número de pontos; a segunda em aumentar o número de itens de verificação por meio das escalas de múltiplos itens, pois, em sua agregação, a variável final passa a ter, potencialmente, um número muito maior de pontos, aproximando-se da condição de variável contínua.

Por fim, merece destaque ainda a crítica de Rossiter (2002) que, em seu modelo C-OAR-SE, argumenta que o uso de advérbios (discordo totalmente, discordo muito etc.) dificulta ainda mais o posicionamento do respondente. De fato, em uma escala de quatro pontos (por exemplo, com 1 – discordo totalmente; 2 – discordo parcialmente; 3 – concordo parcialmente; 4 – concordo totalmente), não é muito clara a distinção entre uma resposta como a 2 ou a 3, afinal a discordância parcial parece ser equivalente à concordância parcial.

2.2 ESCALA *PHRASE COMPLETION*

Diante das dificuldades associadas à escala de Likert, foram desenvolvidas novas escalas com o objetivo de mensurar os construtos, dentre elas, a escala *Phrase Completion*. Essa escala foi desenvolvida por Hodge e Gillespie (2003) justamente como uma alternativa para resolver as dificuldades da escala de verificação de Likert.

Hodge e Gillespie (2003) propuseram uma escala padrão de 11 pontos, sempre de 0 a 10 na sequência dos números inteiros, em que o 0 tem associação com a ausência de atributo, enquanto o 10 tem relação com a intensidade máxima de sua presença. O Quadro 2 apresenta um exemplo da escala *Phrase Completion*, com a mesma variável do Quadro 1 (satisfação com o serviço).

QUADRO 2
Exemplo da Escala *Phrase Completion.*

MEU NÍVEL DE SATISFAÇÃO COM O SERVIÇO FOI:										
MUITO PEQUENO			**MODERADO**					**MUITO GRANDE**		
0	1	2	3	4	5	6	7	8	9	10

Segundo seus propositores, o fato dessa escala ter 11 pontos (de 0 a 10), facilita a interpretação por parte do pesquisado, visto que, em geral, as pessoas são familiarizadas com esta referência (nas avaliações educacionais, por exemplo). Por isso, a potencial dificuldade de resposta associada ao número de pontos se dissipa. Adicionalmente, o maior número de pontos melhora potencialmente a confiabilidade e a validade da escala, sem possuir os problemas convencionais associados a poucos pontos.

Conforme indicado, na escala de Likert realiza-se uma abordagem indireta para mensurar um construto com a transferência de sua medição direta para uma medida da concordância com uma afirmação associada ao construto. Há, portanto, uma fase intermediária de medição, o que torna possível uma fragilidade nessa prática de medição, uma vez que deve ocorrer perda de informação e de conteúdo entre o que se deseja mensurar e sua transformação em uma afirmação, para depois se avaliar a concordância (COSTA, 2011). Desta forma, a escala *Phrase Completion* procura superar ainda esta dificuldade, buscando medir a intensidade de determinado construto diretamente na própria escala. É dessa possibilidade de aplicação que vem o nome da escala, que seria, em uma tradução para o português, escala de conclusão da frase.

Tomando-se como exemplo o Quadro 2, verifica-se de forma clara essa característica, pois o respondente indica seu nível de satisfação completando a frase dentro da escala de 11 pontos, com três pontos de referência. Naturalmente, esta que é uma vantagem desta escala torna-se uma desvantagem se a pesquisa envolver amostras pequenas, com operacionalização quantitativa baseada em frequências (de modo que se cria uma dispersão sem muito sentido nas opções de resposta e torna possível que pontos fiquem com frequência pequena ou nula).

De acordo com estudos desenvolvidos para testar a estabilidade psicométrica da escala *Phrase Completion* em comparação com a escala Likert em construtos medidos por múltiplos itens refletivos, Hodge e Gillespie (2003, 2007) mostraram que a escala *Phrase Completion* resultou em melhor confiabilidade e melhor consistência fatorial. Isso sugere, portanto, que, além de mais intuitiva e logicamente construída, a escala *Phrase Completion* teria ainda as vantagens do ponto de vista de operacionalização estatística. Naturalmente, apenas dois estudos não são suficientes para uma afirmação dessa natureza, sendo necessários novos esforços para melhor definição de posicionamento.

Entretanto, algumas pesquisas desenvolvidas utilizando a escala *Phrase Completion* possuem uma séria limitação relacionada ao espaço ocupado nos instrumentos de pesquisa. De fato, os estudos que usam escalas de Likert utilizam normalmente uma série de afirmações com a escala de verificação ao lado, ocupando assim, menos espaço nos questionários. Já a escala *Phrase Completion* não tem essa possibilidade, de modo que cada item de verificação demanda ao menos três linhas. Em pesquisas acadêmicas, por exemplo, em que o recurso para motivação das pessoas a responderem questionários é normalmente muito pequeno, um questionário extenso é um potencial gerador de erros de medição.

Por outro lado, em algumas pesquisas de mercado, em que a mensuração pode ser feita com apenas um item e que normalmente avalia um número menor de construtos simultaneamente, isso deixa de ser problema. Desse modo, se for mantida a validade de conteúdo e houver consistência nas medições entre as duas escalas, pesquisadores acadêmicos e de mercado poderão fazer uso da modalidade que for mais adequada em termos de espaço e facilidade de resposta.

Seguindo a mesma ideia de Hodge e Gillespie (2003, 2007) de comparar as medidas com as duas escalas, aqui foi desenvolvido um estudo empírico, cujos detalhes metodológicos são mostrados a seguir.

3 PROCEDIMENTOS DE CAMPO

Conforme indicado na introdução, o presente estudo tem o objetivo de verificar as similaridades e diferenças nos resultados das escalas de verificação de Likert e *Phrase Completion*. Para tanto, foram selecionados cinco itens de mensuração do construto religiosidade intrínseca, da mesma forma em que foram utilizados no estudo por Hodge e Gillespie (2003), com algumas adaptações de significados devido à variação na tradução (embora o interesse seja analisar construtos para pesquisas em Marketing e Administração, optou-se por usar o mesmo construto que os autores citados, tendo em vista a possibilidade de procedimentos comparativos; os resultados não seriam, provavelmente, muito distintos pelo uso de outros construtos; os itens estão expostos no item 4).

Os itens de mensuração do construto de referência (religiosidade intrínseca) mantinham uma perspectiva refletiva em relação ao construto latente. A decisão foi que esses itens seriam apresentados no formato de afirmação para resposta de concordância da escala de Likert (em 11 pontos, de 0 a 10). Em seguida, os itens foram adaptados para o modelo de resposta da escala *Phrase Completion*, com separação por um conjunto de outras questões para minimizar riscos associados ao efeito halo.

A coleta foi realizada com a aplicação de questionários em espaços de aglomeração de duas capitais do Nordeste brasileiro, tais como estabelecimentos comerciais, igrejas, *shopping centers* e eventos culturais realizados. O questionário continha, além das escalas, perguntas sobre o perfil dos respondentes (religião, sexo, estado civil, idade e renda). Assim, foram coletados e transferidos para o *software* SPSS, os dados de 241 respondentes. Destes, 12 questionários foram invalidados, por apresentarem inconsistências como itens sem resposta e informações insuficientes para o estudo. Assim, a análise do resultado foi realizada com 229 questionários válidos.

De acordo com as respostas sobre o perfil dos indivíduos pesquisados, observou-se uma distribuição de gênero relativamente equilibrada (54,6% de mulheres e 45,4% de homens). A maioria alegou ser da religião católica (76,9%), com estado civil solteiro predominante (71,2%), idade entre 21 e 30 anos (37,7%) e renda familiar aproximada a R$ 2.000,00 (40,5%). De acordo com essas informações, verifica-se boa heterogeneidade da amostra, característica que assegura boas condições para a análise de dados.

A análise de dados foi realizada com procedimentos descritivos e diversos testes paramétricos e não paramétricos. Assim, inicialmente foram procedidos os testes de normalidade das variáveis; em seguida foi empreendida a extração das medidas de média, mediana e desvio-padrão de todas as variáveis (nas medidas gerais foram avaliadas ainda a assimetria e a curtose). Foram usados procedimentos de comparação de medidas das variáveis correspondentes de cada uma das duas escalas e, especificamente nas variáveis agregadas, foram aplicados procedimentos de comparação e associação bivariada.

Ao final de cada bloco de procedimentos, foram analisadas divergências e similaridades entre as métricas oriundas de cada tipo de escala (Likert e *Phrase Completion*). Todos os procedimentos foram realizados no *software* SPSS 18 e com base na literatura especializada (CONOVER, 1999;

HAIR et al. 2009; LATTIN; CARROL; GREEN, 2011).

4 RESULTADOS

A seguir são apresentados os resultados obtidos a partir do estudo de campo. No item 4.1 encontram-se os resultados descritivos; no item 4.2 é feita a verificação da estrutura psicométrica; o item 4.3 finaliza com a exposição dos procedimentos adicionais de análise.

4.1 RESULTADOS DESCRITIVOS

Antes de iniciar os procedimentos de extração de medidas, o conjunto dos dez itens foi submetido a uma análise da normalidade para verificar a hipótese de que os dados de cada item poderiam ser oriundos de uma variável com distribuição normal. Usou-se aqui o teste não paramétrico de Kolmogorov-Smirnov (teste KS) e verificou-se que a hipótese nula do teste (que supõe que a amostra foi extraída de uma população normalmente distribuída) foi refutada (α $p < 0{,}001$) em todos os itens, de tal modo que não se deve supor, para efeito de análise e de algumas técnicas aplicadas, que haja normalidade na variável de origem dos dados.

De posse deste resultado, foi procedida a extração das medidas descritivas de cada variável, com as medidas de média, mediana e desvio-padrão. A Tabela 1 apresenta os resultados pelos pares correspondentes de variáveis em cada escala de verificação. As medidas são discrepantes de acordo com os pares, mas em geral, verificou-se que, nos três primeiros pares, as médias da escala de *Phrase Completion* foram mais elevadas que as médias das escalas verificadas por Likert, porém nos dois últimos pares, as médias mensuradas por Likert foram maiores. Em quatro dos cinco pares os valores de mediana seguiram o mesmo comportamento. As dispersões também apresentaram variações semelhantes entre os pares de itens (em alguns pares era maior na escala de Likert e, em outros, era menor na escala *Phrase Completion*).

TABELA 1
Medidas descritivas.

PARES	VARIÁVEIS	MÉDIA	MEDIANA	DESVIO-PADRÃO
Primeiro	Minha crença religiosa afeta minha vida em (de 'nenhum aspecto' até 'todos os aspectos').	6,94	8,00	3,09
	Toda a minha visão da vida é baseada na minha religião.	6,07	7,00	2,89
Segundo	Eu tenho consciência da presença de Deus (de 'nunca' até 'o tempo todo').	9,13	10,00	1,36
	Eu sempre tive um forte sentimento da presença de Deus.	8,81	9,00	1,58
Terceiro	Minha religião em relação à minha vida é (de 'um motivo que não me guia' até 'o maior da minha vida').	7,61	8,00	2,23
	Eu tento dedicadamente viver de acordo com minhas crenças religiosas.	6,76	7,00	2,80
Quarto	Eu leio sobre minha religião (de 'nunca' até 'sempre').	6,15	6,00	2,73
	Eu gosto de ler sobre minha religião.	6,89	7,00	2,74
Quinto	O tempo que eu dedico em momentos religiosos e meditação acontece (de 'nunca' até 'diariamente, sem falta').	6,63	7,00	2,41
	Para mim, é importante ter um tempo dedicado às minhas reflexões e orações.	8,24	9,00	2,04

Fonte: Dados da pesquisa.

Como forma de avaliar a consistência estatística dessas diferenças, foram adotados testes de comparação de médias e medianas. As médias de cada par de itens foram comparadas por meio do teste *t*, que se baseia na estatística *t* de Student e na significância associada à hipótese nula de igualdade de médias na população de origem da amostra pareada; assim, valores de *t* elevados e com significância menor que 0,05 sinalizam que a hipótese nula deve ser refutada. A limitação deste teste vem de seu caráter paramétrico e da suposição de normalidade das variáveis que deram origem às duas amostras pareadas. Por esta razão, o teste não paramétrico de Wilcoxon para dados pareados é uma alternativa adequada, que se aplica particularmente à mediana. Neste teste e, segundo o algoritmo implementado no SPSS, a hipótese nula é de que a mediana das diferenças entre as duas variáveis aleatórias que deram origem à amostra é igual a 0; aqui, com valores elevados da estatística W e significância menor que 0,05, refuta-se a hipótese nula.

Pelos resultados indicados na Tabela 2, é possível observar que, em ambos os testes, não há evidência de igualdade nas variáveis que deram origem aos dados das amostras, ou seja, em todos os pares de variáveis houve diferença significativa entre as extrações com a escala *Phrase Completion* e com a escala de Likert. Isso sugere que, por essa primeira perspectiva (que é exploratória, apenas) as duas escalas não são substituíveis para efeito de averiguação de medidas de posição.

TABELA 2
Testes de médias e medianas.

PARES	TESTE *T*		TESTE WILCOXON	
	ESTATÍSTICA	*P*-VALOR	ESTATÍSTICA	*P*-VALOR
Primeiro	4,064	0,000	-4,428	0,000
Segundo	3,700	0,000	-3,853	0,000
Terceiro	5,856	0,000	-5,789	0,000
Quarto	-5,611	0,000	-5,463	0,000
Quinto	-12,552	0,000	-9,994	0,000

Fonte: Dados da pesquisa.

Por outro lado, conforme indicado na Tabela 1, parece haver compensações entre os itens, sendo possível acreditar que, em uma verificação do conjunto, as diferenças pontuais entre os itens se compensem. Com esse entendimento, os itens foram então agregados da seguinte forma: em cada escala foram extraídos os escores médios de cada respondente (somando os escores dos conjuntos de cada escala e dividindo-os por 5). Seguindo o mesmo procedimento da extração em separado, inicialmente foi verificada a normalidade da variável de origem da amostra, tendo-se observado, pelo teste KS que a hipótese de normalidade é novamente refutada para ambas as variáveis, embora os níveis de significância já se aproximem mais do ponto de corte da normalidade (0,026 na escala *Phrase Completion* e 0,014 na escala de Likert).

Após esse procedimento, foram novamente extraídas as médias, medianas e desvios-padrão de cada variável, além das medidas de assimetria e de curtose (para explorar adicionalmente a normalidade). Os valores que estão indicados na Tabela 3 confirmaram a expectativa indicada de que as médias e medianas seriam muito mais próximas (no caso das medianas estas foram iguais no arredondamento para duas casas decimais). Os valores de desvio-padrão também ficaram próximos, o que indica novamente uma convergência dos resultados agregados. Quanto à assimetria e à curtose, foi observado que, pelos critérios convencionais de extrações do SPSS, os valores sinalizem indícios de normalidade nas variáveis de origem da amostra, embora o teste KS tenha sido contrário a esta evidência.

TABELA 3
Medidas descritivas gerais e testes – Primeira agregação.

MEDIDAS GERAIS	MEDIDAS DESCRITIVAS					TESTES	
	MÉDIA	MEDIANA	DESVIO-PADRÃO	ASSIMETRIA	CURTOSE	*t*	WILCOXON
Medida geral na escala *Phrase Completion*	7,29	7,60	1,77	-0,724	0,225	t = - 0,814; p = 0,417	z = - 0,353; p = 0,724
Medida geral na escala de Likert	7,35	7,60	1,88	-0,581	-0,504		

Fonte: Dados da pesquisa.

Também aqui foram aplicados os dois testes de comparação das variáveis e tanto o teste *t* (agora já mais seguro pela proximidade de normalidade das variáveis), quanto o teste de Wilcoxon mostraram que não há indícios de diferença entre as duas variáveis que deram origem à amostra. A sinalização é, portanto que, em termos de medidas descritivas de posição, as variáveis agregadas convergem em resultados.

Como uma exploração adicional, foi extraída a correlação entre as duas variáveis, tendo-se obtido valores bastante elevados, tanto pelo coeficiente de correlação paramétrico de Pearson (0,813) quando pelo coeficiente não paramétrico de Spearman (0,805). Isto indica que as duas medidas possuem elevado grau de variação conjunta, inclusive variação linear, o que é esperado em caso de substituição de uma medida por outra (conforme indica a correlação pelo coeficiente de Pearson).

Em síntese, estes resultados indicam que, no construto sob análise, quando da avaliação de cada item em particular, não é adequada a substituição de um item mensurado em uma escala pelo seu correspondente em outra escala, pois as medidas posição e dispersão variam. Por outro lado, em uma agregação do conjunto de itens, essas diferenças se anulam e as variáveis agregadas possuem então uma boa sobreposição, sendo possível tomar uma escala ou outra para efeito de descrição da amostra.

4.2 ESTRUTURA PSICOMÉTRICA

Nesta etapa foram avaliadas as características psicométricas da escala pelas ferramentas convencionais de aplicação em conjuntos de itens que são indicadores refletivos em relação ao construto latente. Conforme indica a teoria convencional, os procedimentos são (COSTA, 2011): a verificação da consistência fatorial (com extração pelo método de componentes principais), que se verifica pela variância extraída (desejável que seja maior que 0,5) e pelos escores fatoriais (desejáveis que sejam maiores que 0,4); e a consistência interna, indicativa de confiabilidade, que se verifica pelo coeficiente alpha de Cronbach (desejável que seja maior que 0,6).

A Tabela 4 apresenta os resultados da extração fatorial. Esses resultados permitem visualizar que, independente da escala de verificação, cada conjunto de variáveis manteve-se em um só fator e, em ambos, as variâncias extraídas foram acima do valor de referência de 0,5, inclusive próximos entre si (0,556 nos itens mensurados com a escala *Phrase Completion*, e 0,589 nos itens mensurados com a escala de Likert).

Em relação aos escores fatoriais, foi verificado que, em ambas as extrações, os escores mínimos foram de 0,570 (para a escala Likert) e 0,606 (para a escala *Phrase Completion*), valores acima do ponto de referência. Por estes resultados é possível assegurar, portanto, que os itens de cada escala

de verificação constituem um só fator com boa consistência fatorial.

TABELA 4
Resultados da extração fatorial para as duas escalas utilizadas.

FATOR DA ESCALA DE VERIFICAÇÃO *PHRASE COMPLETION*; VARIÂNCIA EXTRAÍDA = 0,556	
VARIÁVEIS	**ESCORE**
Minha crença religiosa afeta minha vida em (de 'nenhum aspecto' até 'todos os aspectos').	0,608
Eu tenho consciência da presença de Deus (de 'nunca' até 'o tempo todo').	0,570
Minha religião em relação à minha vida é (de 'um motivo que não guia minha vida' até 'o maior da minha vida').	0,843
Eu leio sobre minha religião (de 'nunca' até 'sempre').	0,809
O tempo que eu dedico em momentos religiosos e meditação acontece (de 'nunca' até 'diariamente, sem falta').	0,852
FATOR DA ESCALA DE VERIFICAÇÃO COM LIKERT; VARIÂNCIA EXTRAÍDA = 0,589	
VARIÁVEIS	**ESCORE**
Toda a minha visão da vida é baseada na minha religião.	0,777
Eu sempre tive um forte sentimento da presença de Deus.	0,606
Eu tento dedicadamente viver de acordo com minhas crenças religiosas.	0,838
Eu gosto de ler sobre minha religião.	0,808
Para mim, é importante ter um tempo dedicado às minhas reflexões e orações.	0,787

Fonte: Dados da pesquisa.

Na verificação da consistência interna para indicação de confiabilidade, observou-se que, no caso dos itens que tiveram verificação com a escala *Phrase Completion*, o alpha de Cronbach foi de 0,780. Já no caso dos itens mensurados com escala de Likert o alpha foi de 0,820. Novamente aqui os dois valores foram bem acima do valor de referência (0,6) e foram, inclusive, próximos entre si. A sinalização destes resultados é, portanto, de que a confiabilidade não sofre alteração muito significativa em decorrência da escala de verificação, ou seja, se mantêm bem controlados os erros aleatórios associados ao processo de mensuração nos respectivos conjuntos de itens.

Este resultado é contrário ao que foi alcançado por Hodge e Gillespie (2003), em que os itens do modelo *Phrase Completion* apresentaram melhor ajuste, com maior validade e confiabilidade e menor quantidade de erros de medição. Entretanto, deve-se notar que a pesquisa desses autores foi realizada apenas com estudantes de Pós-Graduação, o que constitui uma amostra com peculiaridades evidentes. Notadamente, com uma amostra mais heterogênea, os resultados deixam evidente que, do ponto de vista da estrutura psicométrica, pelo menos nos métodos de averiguação aqui adotados, não há qualquer problema em utilizar uma ou outra das duas escalas de verificação.

4.3 PROCEDIMENTOS ADICIONAIS

Um primeiro procedimento adicional consistiu em extrair uma nova medida agregada das variáveis, agora pela ponderação dos escores dos respondentes de cada item pelo escore fatorial da variável (segundo a extração efetivada), em cada unidade da amostra. Os resultados, que estão na Tabela 5, geraram valores de medidas muito próximos dos anteriores (indicados na Tabela 3) em termos das cinco medidas utilizadas (média, mediana, desvio-padrão, assimetria e curtose).

Os testes de diferença entre as variáveis (teste *t* e teste de Wilcoxon) também asseguraram que as duas variáveis não são distintas uma da outra, reafirmando o resultado anterior. No teste de normalidade, a verificação foi distinta agora, com uma indicação de normalidade da variável agregada com mensuração por *Phrase Completion*, embora em uma indicação quase no limite de

negação da hipótese nula (p = 0,052). Por fim, na avaliação da correlação, o resultado anterior foi reafirmado, tendo-se observado uma correlação elevada tanto pelo coeficiente de Pearson (0,825) quanto no de Spearman (0,816).

TABELA 5
Medidas descritivas gerais e testes – Segunda agregação.

MEDIDAS GERAIS	MEDIDAS DESCRITIVAS					TESTES	
	MÉDIA	MEDIANA	DESVIO-PADRÃO	ASSIMETRIA	CURTOSE	*T*	WILCOXON
Medida geral na escala *Phrase Completion*	7,19	7,48	1,83	-0,731	0,268	t = - 1,203; p = 0,230	z = - 0,795; p = 0,427
Medida geral na escala de Likert	7,27	7,55	1,94	-0,604	-0,453		

Fonte: Dados da pesquisa.

A sinalização geral destes resultados reafirma o que foi indicado anteriormente, ou seja, em pesquisas empíricas não há maiores variações de resultados pelo uso de um ou outro método de aferição para efeito de extração de medidas descritivas das variáveis agregadas.

Uma segunda verificação adicional relevante foi efetuada pela análise das variações por gênero (destacada aqui por ter sido a variável categórica que dividiu a amostra em duas partes quase iguais). Primeiramente, as variáveis agregadas (pelo último método), foram analisadas por meio de análise de variância e do teste U de Mann-Whitney (não paramétrico), para verificar se o comportamento das variáveis era o mesmo nas duas formas de averiguação. Os resultados estão indicados na Tabela 6 e é possível observar nos dois testes, que há diferença significativa (α p < 0,05) na intensidade da variável por gênero, independente da escala. Isto é mais uma indicação de que o comportamento das variáveis segue paralelo em cada tipo de escala.

TABELA 6
Comparações por gênero.

VARIÁVEIS E TESTES	GÊNERO	NÚMERO	MÉDIA	MEDIANA	DESVIO-PADRÃO
Medida geral na escala *Phrase Completion*	Masculino	104	6,71	6,88	2,08
Anova – F = 13,655, p < 0,001	Feminino	125	7,58	7,81	1,49
U de Mann-Whitney – Z = - 3,018, p < 0,001	Total	229	7,19	7,48	1,83
Medida geral na escala de Likert	Masculino	104	6,92	6,94	2,08
Anova – F = 6,566, p < 0,05	Feminino	125	7,57	7,87	1,77
U de Mann-Whitney – Z = - 2,370, p < 0,05	Total	229	7,28	7,55	1,94

Fonte: Dados da pesquisa.

Depois de procedida uma separação pelo comendo *split-half* do SPSS, os testes *t* e de Wilcoxon de comparação entre as variáveis somente dentro de cada grupo mostraram novamente que não há diferenças entre as duas variáveis na análise restrita do grupo dos homens (teste *t*: t = 1,747, p = 0,084; Wilcoxon: Z = -1,135, p = 0,257) e do grupo de mulheres (teste *t*: t = 0,136, p = 0,892; Wilcoxon: Z = -0,012, p = 0,991). Isto evidencia mais fortemente que o comportamento nas duas aferições é convergente.

5 CONCLUSÕES E CONSIDERAÇÕES FINAIS

Esta pesquisa comparou as alternativas de mensuração de construtos em Marketing e

Administração, a partir de itens medidos com escalas de verificação de Likert e *Phrase Completion*. Como informado, a escala Likert é amplamente utilizada nas ciências sociais, mas possui características que colocam sua eficiência em dúvida e, dentre as alternativas já propostas, a escala *Phrase Completion* parece ser uma das melhores opções.

De acordo com os dados da pesquisa realizada com 229 respondentes, foram verificadas diferenças relativamente pequenas entre as escalas agregadas, apesar de terem sido observadas diferenças estatisticamente significativas nas comparações individuais dos pares de itens. Esse resultado constata que os itens, ao serem analisados separadamente, podem apresentar algumas variações entre as medidas nas duas escalas, o que coloca em dúvida, inicialmente, a possibilidade de uma escala ser utilizada em substituição a outra. Portanto, pela natureza dos resultados e incerteza de mensuração de construtos com múltiplos itens refletivos em relação ao construto, não é possível informar qual seria a melhor das duas escalas, a não ser pela validação de conteúdo e face das etapas preparatórias do procedimento de mensuração (COSTA, 2011).

Por outro lado, quando foram analisados os dados agregados, as medições das duas escalas convergiram em todas as verificações feitas, seja nas comparações de medidas descritivas, nas indicações convencionais de consistência psicométrica, na associação geral ou na segmentada por gênero. Isso mostra que, para avaliações de múltiplos itens com agregação, as duas escalas possuem resultados praticamente iguais.

Deste modo, a escolha de uma escala ou outra é uma decisão do pesquisador, que a fará levando em conta as peculiaridades de público ou interesse de pesquisa. Sendo assim, a escala Likert pode continuar sendo utilizada nas pesquisas acadêmicas de Marketing e Administração, mesmo possuindo uma medição indireta do construto analisado, pois esse tipo de escala é mais adequado para instrumentos longos e tem mais facilidade de adaptação para um número maior de construtos. Em suma, conforme se reafirmou no estudo, essa escala de verificação possui boas propriedades psicométricas, é de fácil organização e tem uma vantagem operacional no tocante à estrutura do instrumento de pesquisa.

Já para as pesquisas de mercado de interesse mais profissional e decisorial, sugere-se que sejam aplicados instrumentos na escala *Phrase Completion* pelo fato do item medir diretamente o construto analisado e assim atender aos objetivos comumente propostos pela pesquisa de mercado. De fato, a escala *Phrase Completion* é mais lógica e intuitiva, embora ocupe maior espaço em instrumentos de aplicação. Nesses termos, e considerando que as pesquisas de mercado normalmente possuem instrumentos mais diretos, é possível entender que esta escala tem maior adequação e não perde em termos de consistência psicométrica de validade e confiabilidade.

6 LIMITAÇÕES DA PESQUISA E SUGESTÕES PARA NOVAS PESQUISAS

Embora os dados tenham sido avaliados com cuidado e rigor estatístico, é necessário reconhecer que houve fragilidades no estudo em relação ao *design* da pesquisa, tanto em termos de acesso quanto em termos de tamanho da amostra. Por isso, é recomendado que outros estudos contribuam com o procedimento amostral para aperfeiçoar as comparações.

Ainda que, no questionário aplicado, tenha havido separação dos itens de mensuração com escala de Likert dos itens de mensuração com escala *Phrase Completion* por meio de outras questões de variáveis categóricas, é possível que a semelhança em conteúdo tenha influenciado as respostas, por

meio do dito efeito halo. Isso possivelmente tenha influenciado na convergência da estrutura psicométrica observada. Por essa razão, recomenda-se a realização de outros estudos que busquem adotar procedimentos de comparação para dirimir dúvidas quanto a este tipo de efeito.

O construto utilizado também tem particularidades que podem explicar os resultados, sendo possível supor variações em outros construtos. Para os objetivos do estudo aqui desenvolvido, não há problemas em termos de resultado, pois o conteúdo do construto não foi objeto de análise, salvo na perspectiva de validação de conteúdo da escala. Ainda assim, recomenda-se que sejam procedidos estudos semelhantes com outros tipos de construtos.

7 REFERÊNCIAS

COELHO, P. S.; ESTEVES, S. P. The choice between a five-point and a ten-point scale in the framework of customer satisfaction measurement. *International Journal of Market Research*, 49 (3), p. 313-339, 2007.

CONOVER, W. J. *Practical nonparametric statistics*. 3. ed. New York: John Wiley, 1999.

COSTA, F. J. *Mensuração e desenvolvimento de escalas:* aplicações em administração. Rio de Janeiro: Ciência Moderna, 2011.

CUMMINS, R. A.; GULLONE, E. Why we should not use 5-point Likert scales: the case for subjective quality of life measurement. International Conference on Quality of Life in Cities, 2. Singapore. *Proceedings*… Singapore: National University of Singapore, 2000.

DAWES, J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, 50 (1), p. 61-77, 2008.

FINKELSTEIN, L. Widely-defined measurement: An analysis of challenges. *Measurement*, 42, p. 1270-1277, 2009.

FINKELSTEIN, L. Widely, strongly and weakly defined measurement. *Measurement*, 34, p. 39-48, 2003.

HAIR JR. J.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. *Análise multivariada de dados*. 6. ed. Porto Alegre: Bookman, 2009.

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

HODGE, D. R.; GILLESPIE, D. F. Phrase completion: an alternative to Likert scales. *Social Work Research*, 27 (1), p. 45-55, 2003.

LATTIN, J.; CARROL, J. D.; GREEN, P. E. *Análise de dados multivariados*. São Paulo: Cengage Learning, 2011.

LIKERT, R. A technique for the measurement of attitudes. *Archives in Psychology*, 140, p. 1-55,

1932.

MARI, L. Notes towards a qualitative analysis of information in measurement results. *Measurement*, 25 (3), p. 183-192, 1999.

MARI, L. The meaning of "quantity" in measurement. *Measurement*, 17 (2), p. 127-138, 1996.

ROSSITER, J. R. The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, 19 (4), p. 305-335, 2002.

RUSSELL, C. J.; BOBKO, P. Moderated regression analysis and Likert scales too coarse for comfort. *Journal of Applied Psychology*, 77 (3), p. 336-342, 1992.