UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DE
RIBEIRÃO PRETO
DEPARTAMENTO DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO DE ORGANIZAÇÕES

MARCOS ANGELI GHISI

Uma avaliação do desenvolvimento e da validação das escalas de marketing
em periódicos e anais de eventos brasileiros

ORIENTADOR: PROFA. DRA. ADRIANA BACKX NORONHA VIANA

RIBEIRÃO PRETO

2014

Prof. Dr. Marco Antonio Zago
Reitor da Universidade de São Paulo

Prof. Dr. Dante Pinheiro Martinelli
Diretor da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto

Profa. Dra. Sonia Valle Walter Borges de Oliveira
Chefe do Departamento de Administração

MARCOS ANGELI GHISI

Uma avaliação do desenvolvimento e da validação das escalas de marketing
em periódicos e anais de eventos brasileiros

Tese apresentada ao Programa de Pós-Graduação
em Administração de Organizações da Faculdade de
Economia, Administração e Contabilidade de
Ribeirão Preto da Universidade de São Paulo, para
obtenção do título de Doutor em Ciências. Versão
Corrigida. A original encontra-se disponível na
FEA-RP/USP.

Orientadora: PROFA. DRA. ADRIANA BACKX
NORONHA VIANA

RIBEIRÃO PRETO

2014

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

FICHA CATALOGRÁFICA

Ghisi, Marcos Angeli

Uma avaliação do desenvolvimento e da validação das escalas de marketing em periódicos e anais de eventos brasileiros. Ribeirão Preto, 2014.

169 p. : il. ; 30 cm

Tese de Doutorado, apresentada à Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto/USP. Área de concentração: Administração das Organizações.

Orientador: Noronha Viana, Adriana Backx.

1. Escalas de mensuração. 2. Psicometria. 3. Marketing. 4. Validade. 5. Confiabilidade.

DEDICATÓRIA

Dedico este trabalho aos meus pais,
Ignácio (*in memoriam*) e Odete, por
me apoiarem incondicionalmente em
todos os momentos da minha vida

AGRADECIMENTOS

À minha orientadora Profa. Dra. Adriana Backx Noronha, pela contribuição competente e valiosa para a concretização deste trabalho.

Ao Prof. Dr. Eduardo Eugênio Spers, meus sinceros agradecimentos pelo apoio nas fases de estruturação do trabalho e pelos momentos de discussões enriquecedoras que muito contribuíram para a realização deste estudo.

À minha irmã Flávia, por ler e corrigir pacientemente a minha tese, contribuindo de diversas formas para a melhoria do texto.

À minha irmã Renata, pelo incentivo profissional e hospitalidade oferecida durante o período em que cursei aulas em São Paulo.

À Faculdade de Economia, Administração por ter sido um grande diferencial na minha formação acadêmica.

Aos funcionários da Faculdade de Economia, Administração e Contabilidade, em especial à Érika de Lima Veronezi Costa e Matheus Carlos da Costa, pela disponibilidade, paciência e auxílio nos momentos críticos e burocráticos inerentes à execução do curso de doutorado.

A todas as pessoas que contribuíram direta ou indiretamente para a realização deste estudo.

Uma avaliação do desenvolvimento e da validação das escalas de marketing em periódicos e anais de eventos brasileiros

RESUMO

A atividade de mensuração constitui um pré-requisito central para o próprio desenvolvimento científico. Entretanto, ela não é um processo trivial e simples. A complexidade se deve ao fato de que dificilmente um estudo pode ser projetado e controlado para obter uma mensuração precisa e não ambígua, e por conta disso, a ocorrência de erro é praticamente inevitável. Alguns instrumentos podem ser altamente confiáveis, enquanto outros possuem menor grau de precisão. Alguns podem ser considerados altamente precisos, mas por outro lado, podem não medir aquilo que foram projetados para medir. Por essa razão, a qualidade do resultado da mensuração depende muito das medidas que são selecionadas ou desenvolvidas pelo pesquisador e como elas se adaptam às diferentes circunstâncias as quais são ou poderiam ser aplicadas. Nas últimas décadas, a literatura internacional sobre desenvolvimento e validação de escalas adquiriu um elevado grau de desenvolvimento metodológico e científico. Especificamente sobre a área de marketing, há vários livros sobre compilações de escalas, os chamados *handbooks* de escalas de marketing. Em contraste, a literatura nacional sobre escalas nessa área, desenvolvidas e/ou validadas, parece ainda estar em fase embrionária. Pode-se dizer que a preocupação com o tema parece ter evoluído no Brasil mais recentemente, a partir dos anos 2000, como natural evolução da pesquisa de orientação mais acadêmica que as escolas de Administração adotaram desde então. Este trabalho tem como objetivo avaliar a qualidade das escalas de mensuração propostas e publicadas na área de marketing nos anais de eventos da Anpad no Brasil nos últimos 3 anos. Na revisão da literatura primeiramente busca-se levantar a teoria sobre mensuração, entender os problemas e erros relacionados ao processo de medir e os modelos que podem ser utilizados para a construção de instrumentos de medida confiáveis e válidos. A metodologia empregada foi a Revisão Bibliográfica Sistemática da Literatura, com o emprego do método Revisão Integrativa, que possibilitou encontrar e selecionar de forma sistemática as escalas propostas e validadas na área de marketing, dentro do período considerado, na base de dados da Anpad (EnAnpad, EMA, EnEPQ, EnADI, entre outros). A análise das escalas encontradas identificou problemas de qualidade no que se refere à validade de conteúdo e de construto. Verificou-se ainda que a análise fatorial exploratória aparece como a técnica de uso mais frequente, embora um grande número de pesquisadores a utilize erroneamente para fins confirmatórios do que exploratórios, como deveria ser. Já no campo da confiabilidade, atestou-se o grande uso da medida de alpha de Cronbach, mas também foi verificada a utilização da confiabilidade composta como medida alternativa. A falta de detalhamento na redação das decisões tomadas e resultados obtidos inviabiliza maiores análises sobre os resultados. Sugere-se aos autores da área que procurem descrever melhor todos os passos tomados para o desenvolvimento do instrumento, para que se possa auferir a qualidade da escala criada, entre outros pontos.

Palavras-chave: escalas de mensuração; psicometria; marketing; validade; confiabilidade.

An assessment of the development and validation of marketing scales in Brazilian journals and proceedings events

ABSTRACT

The activity of measurement is a key prerequisite for the proper scientific development. However, it is not a trivial and simple process. The complexity is due to the fact that hardly a study can be designed and managed to get a precise measurement and unambiguous, and because of that, the error occurrence is almost inevitable. Some instruments may be highly reliable, while others have lower degree of accuracy. Some may be considered highly precise but on the other hand, can not measure what are designed to measure. For this reason, the quality of the result of the measurement depends heavily on measures that are selected or developed by the researcher and how they adapt to different circumstances which are or could be applied. In recent decades, the international literature on development and validation of measures acquired a high degree of methodological and scientific development. Specifically on marketing, there are several books on scales compilations, known handbooks of marketing scales. In contrast, the literature on national scales in this area, developed and / or validated, seems to be still in the embryonic stage. It can be said that the concern over the issue seems to have evolved in Brazil more recently, from the 2000s, as a natural evolution of the more academically oriented research that business schools have adopted since. This study aims to assess the quality of the measurement scales proposed and published in marketing in the proceedings of Anpad events in Brazil over the past three years. The review of the literature primarily seeks to raise the theory of measurement, understand the problems and errors related to the process of measuring and models that can be used to build reliable and valid measurement instruments. The methodology used was the Bibliographical Systematic Literature Review, with the use of the method integrative review, which made it possible to find and selecting in a systematic way the scales proposed and validated in marketing, within the period considered in the database Anpad (EnANPAD , EMA, EnEPQ, EnADI, among others). The analysis of the scales identified quality problems with regard to the content and construct validity. It was also found that the exploratory factor analysis appears as the most frequently used technique, although a large number of researchers use it wrongly for confirmatory purposes than exploratory, as it should be. In the field of reliability, there is a great use of Cronbach's alpha measure, but was also verified the use of composite reliability as an alternative measure. The lack of detail in the writing of the decisions taken and results achieved prevents further analysis on the results. It is suggested to the authors of the area looking to better describe the steps taken for the development of the instrument, so that one can obtain the scale of quality created, among other things.

Keywords: measurement scales; psychometrics; marketing; validity; reliability.

LISTA DE GRÁFICOS

Gráfico 1: Comparação do uso de escala de quatro pontos versus cinco pontos.	51
Gráfico 2: Exemplo do efeito <i>ceiling</i>	52
Gráfico 3: Número de itens em uma mesma dimensão	130
Gráfico 4: Histograma da razão do nº de respondentes por item da escala.....	139
Gráfico 5: Variância total extraída pelos fatores da AFE.....	142

LISTA DE QUADROS

Quadro 1: Relação de alguns autores que apoiam a classificação de Stevens	33
Quadro 2: Resumo das principais características das escalas primárias de medição.	34
Quadro 3: Enfoque epistemológico em ciências e matemática	42
Quadro 4: Principais técnicas de estimação da confiabilidade.....	54
Quadro 5: Níveis recomendados do coeficiente <i>alpha</i>	60
Quadro 6: Relações entre o coeficiente <i>alpha</i> e características de desenho de pesquisa	62
Quadro 7: Contrastes entre as concepções tradicionais e modernas de validade	70
Quadro 8: Tipos de validade, subtipos, conceito e operacionalização	72
Quadro 9: Validade Preditiva e Validade Concorrente ou Simultânea	80
Quadro 10: Tipos de revisão de literatura.	112
Quadro 11: Resultados da estratégia de coleta de dados empregada – 1º refinamento	119
Quadro 12: Resumo da Metodologia de Pesquisa Empregada.....	123
Quadro 13: Identificação das escalas encontradas	126
Quadro 14: Número de dimensões das escalas.....	128
Quadro 15: Frequência do nº de itens por dimensão	129
Quadro 16: Tipo de escala de avaliação utilizada nas escalas.....	132
Quadro 17: Nº de pontos da escala de avaliação	132
Quadro 18. Síntese dos procedimentos qualitativos realizados pelos autores das escalas	133
Quadro 19: Frequência de realização dos testes preliminares das análises multivariadas	135
Quadro 20. Estatísticas do estudo piloto	136
Quadro 21: Método de levantamento segundo o modo de aplicação (estudo oficial).....	137
Quadro 22. Técnica de amostragem mais utilizada.....	137
Quadro 23: Relação nº de respondentes por item (estudo oficial)	138
Quadro 24. Método de extração de variância utilizado na AFE.....	140
Quadro 25: Principais critérios utilizados para definição do número de fatores na AFE	141
Quadro 26: Método de rotação utilizado na AFE.....	141
Quadro 27: Procedimentos realizados na análise fatorial exploratória – estudo oficial.....	142
Quadro 28: Indicadores de confiabilidade utilizados	143
Quadro 29. Frequência verificada dos tipos de confiabilidade e validade encontradas nas escalas analisadas	151

LISTA DE FIGURAS

Figura 1: Relação entre os conceitos relacionados à mensuração.	23
Figura 2: Tipos de construtos.	24
Figura 3: Tipos de vieses de resposta.	48
Figura 4: Avaliação da confiabilidade de uma escala.	53
Figura 5: Avaliação da validade de uma escala.....	73
Figura 6: Passos para o desenvolvimento de escalas – Modelo Churchill (1979)	92
Figura 7: Passos para o desenvolvimento de escalas – Modelo Vieira (2011)	99
Figura 8: Estrutura do capítulo de metodologia da tese.	107
Figura 9: Tipos de revisão da literatura.	110
Figura 10: Etapas da Revisão Bibliográfica Integrativa.....	114

LISTA DE FÓRMULAS

(1) Transformação permissível para escala ordinal	36
(2) Transformação permissível para escala intervalar	38
(3) Transformação permissível para escala razão.....	39
(4) Coeficiente de confiabilidade <i>split-half</i> ajustado.....	57
(5) Coeficiente alpha de Cronbach	59
(6) Fórmula de Spearman-Brown para ajuste do coeficiente alpha ao tamanho do teste.....	64

SUMÁRIO

1	INTRODUÇÃO	5
1.1	Contextualização do tema	5
1.2	Oportunidade de pesquisa e justificativa	7
1.3	Problema e objetivos de pesquisa	13
1.4	Estrutura do trabalho.....	14
2	REFERENCIAL TEÓRICO – TEORIA DA MEDIDA	17
2.1	Definição de mensuração	17
2.2	Conceitos relacionados	19
2.3	Mensuração e ciência.....	24
2.4	A natureza da medida	28
2.5	A base axiomática da medida	29
2.6	Níveis da medida	32
2.6.1	Escala nominal.....	34
2.6.2	Escala ordinal	36
2.6.3	Escala intervalar	37
2.6.4	Escala razão	39
2.7	Críticas à proposta de classificação de Stevens	40
3	REFERENCIAL TEÓRICO – ERRO DE MENSURAÇÃO	45
3.1	Erro de medida.....	45
3.2	Vieses de resposta em escalas multi-itens	47
4	REFERENCIAL TEÓRICO – CONFIABILIDADE DE ESCALAS.....	53
4.1	Estabilidade – técnica teste-reteste	54
4.2	Equivalência – técnica de formas paralelas ou alternativas.....	56
4.3	Consistência Interna – <i>split-half</i>	57
4.4	Consistência Interna - Coeficiente <i>alpha</i> de Cronbach.....	59

4.4.1	Críticas ao coeficiente <i>alpha</i> de Cronbach	61
4.5	Procedimentos para melhoria da confiabilidade	65
5	REFERENCIAL TEÓRICO – VALIDAÇÃO DE ESCALAS	67
5.1	Validade de translação.....	73
5.1.1	Validade de conteúdo.....	74
5.1.2	Validade de face.....	77
5.2	Validade de critério	79
5.3	Validade de construto.....	81
5.3.1	Validade convergente.....	84
5.3.2	Validade discriminante	85
5.3.3	Validade nomológica	87
5.3.4	Validade de grupo conhecido.....	89
6	MODELOS DE CONSTRUÇÃO DE ESCALAS	91
6.1	Modelo clássico de Churchill (1979)	91
6.2	Críticas ao modelo de Churchill.....	93
6.2.1	A questão da dimensionalidade.....	94
6.2.2	A orientação para construtos refletivos.....	95
6.2.3	A questão das escalas de múltiplos itens	96
6.3	Modelo de Vieira (2011)	98
6.4	Modelo proposto por Hair <i>et al.</i> (2009)	104
7	METODOLOGIA DE PESQUISA.....	107
7.1	Delineamento da pesquisa.....	107
7.2	Método de pesquisa.....	110
7.3	Etapas Metodológicas.....	113
7.3.1	Etapa 1 - Identificação do tema e seleção da questão de pesquisa	114
7.3.2	Etapa 2: Estabelecimento de critérios de inclusão e exclusão	115
7.3.3	Etapa 3: Identificação dos estudos pré-selecionados e selecionados	118

7.3.4	Etapa 4: Categorização dos estudos selecionados	120
7.3.5	Etapa 5 - Análise e interpretação dos resultados	123
7.3.6	Etapa 6 - Apresentação da revisão/ síntese do conhecimento	123
7.4	Resumo da Metodologia de Pesquisa Empregada	123
8	ANÁLISE DOS RESULTADOS	125
8.1	Bloco 1 - Identificação e caracterização das escalas encontradas	126
8.2	Bloco 2 - procedimentos qualitativos para desenvolvimento das escalas.....	133
8.3	Bloco 3 – Análises preliminares em dados multivariados	134
8.4	Bloco 4 – informações sobre a amostra em estudo oficial.....	136
8.5	Bloco 5 – procedimentos de análise fatorial exploratória em estudo oficial	140
8.6	Bloco 6 – procedimentos de verificação de confiabilidade	143
8.7	Bloco 7 – procedimentos de verificação da validade.....	145
9	CONSIDERAÇÕES FINAIS.....	151
9.1	Limitações do estudo	154
9.2	Sugestões de pesquisas futuras	155
	REFERÊNCIAS BIBLIOGRÁFICAS	157

1 INTRODUÇÃO

1.1 Contextualização do tema

Em linhas gerais, a mensuração pode ser definida como o processo padronizado de atribuição de números ou outros símbolos a certas características de objetos de interesse, de acordo com algumas regras pré-especificadas (AAKER; KUMAR; DAY, 2001). Nesse sentido, o processo de mensuração essencialmente está relacionado com os métodos utilizados para fornecer descrições quantitativas do grau em que indivíduos manifestam ou possuem características específicas (GHISELLI; CAMPBELL; ZEDECK, 1981).

Rapidamente a ciência se apercebeu das vantagens consideráveis que ela pode obter ao utilizar a linguagem matemática para descrever o seu objeto próprio de estudo, e que é o uso deste modelo que vem possibilitando distinguir níveis de progresso no conhecimento científico (PASQUALI, 2011). Corroborando esta ideia, Costa (2011, p. 11) afirma que “a atividade de mensuração constitui um pré-requisito central para o próprio desenvolvimento científico”.

Embora a primeira vista esta atividade possa parecer simples, nas ciências sociais e comportamentais as características ou atributos mensurados raramente podem ser auferidos diretamente, mas por outro lado, apresentam manifestações indiretas que então podem ser mensuradas. Neste caso, a medida é produto de uma operação de mensuração baseada em indícios que se supõe estarem relacionados com o atributo do objeto medido. Como se pode perceber, esta atividade é extremamente complexa, com muitos pesquisadores questionando até mesmo a validade do próprio processo de medir nessa área.

O fato é que os esforços de pesquisa em mensuração são próprios de toda e qualquer ciência ou área acadêmica, e não apenas das ciências exatas e das engenharias, ainda que sejam estas áreas as que alcançaram um maior nível de desenvolvimento (COSTA, 2011). Como lembra Erthal (2011), nas ciências exatas, ao contrário das ciências sociais, a medida consegue atingir grande exatidão, o que conduz ao estabelecimento de relações facilmente comunicáveis e precisas, ao passo que nas ciências sociais, a mensuração é algo bem mais complexo, visto que nem sempre se trata de algo observável, como é o caso das variáveis físicas.

Apesar da mensuração de variáveis abstratas ser algo muito mais complexo, chegando até ao nível de contestação da própria possibilidade de mensuração, muitos pesquisadores consideram perfeitamente possível medi-las. Na verdade, os pesquisadores que defendem o

uso da medida no contexto das ciências sociais reconhecem as dificuldades na quantificação dos fenômenos pesquisados e admitem que os instrumentos utilizados para descrever indivíduos quantitativamente em termos de várias características psicológicas ou comportamentais (instrumentos de medida) nunca medem estas características com uma perfeita precisão. Por conta disso, eles consideram necessários modelos apropriados para tentar minimizar os efeitos que geram erros na medida. Como afirma Costa (2011, p. 18), “a questão não é de medir ou não medir; trata-se de buscar minimizar erros de mensuração”. Isto implica em um grande esforço conceitual e metodológico, que orbita em torno de modelos e teorias próprios para estas áreas (ciências sociais e comportamentais), e da apropriação de conceitos, modelos e ferramentas de uso geral nas ciências físicas e tecnológicas (COSTA, 2011). Reforçando este pensamento, Erthal (2009) também afirma que é na aplicação dos métodos científicos e na utilização de instrumental padronizado que se permite medir e classificar o comportamento humano com a menor ambiguidade possível.

De fato, o processo de mensuração não é uma atividade trivial e simples. A complexidade se deve ao fato de que dificilmente um estudo pode ser projetado e controlado para obter uma mensuração precisa e não ambígua, e por conta disso, a ocorrência de erro é praticamente inevitável, isto é, os instrumentos de medida nunca estão livres de erro. Alguns instrumentos podem ser altamente confiáveis, enquanto que outros possuem menor grau de precisão. Outros podem ser considerados altamente precisos, mas por outro lado, podem não estar medindo aquilo que foram projetados para medir. Por conta disso, a qualidade do resultado da mensuração depende muito das medidas que são selecionadas ou desenvolvidas pelo pesquisador e como elas se adaptam às diferentes circunstâncias as quais são ou poderiam ser aplicadas.

Por esta razão, é importante que o pesquisador desenvolva instrumentos adequados para que as medidas efetuadas correspondam efetivamente ao que se deseja medir (validade da medida) e para que o erro seja o menor possível (confiabilidade da medida) diante dos recursos disponíveis. Assim, um bom instrumento de medida é aquele que pode ser considerado confiável (consistente e preciso) e que é válido para medir aquilo que foi projetado para medir.

Dado esse panorama, cabe ao pesquisador, ao utilizar as técnicas de medidas, conhecer suas limitações, sabendo ser impossível abranger a totalidade de cada um dos fenômenos pesquisados e entender que, sendo estes fenômenos demasiadamente complexos, esse objetivo dificilmente será alcançado. O erro sempre irá existir, mas cabe ao pesquisador procurar desenvolver procedimentos metodológicos e sistemáticos que procurem minimizá-los, ou

ainda, estimar o nível de erro que se está incorrendo. Desta forma, o avanço das pesquisas na área das ciências sociais pode depender muito da qualidade com que os instrumentos de medida são construídos, desenvolvidos e validados. Como afirma Costa (2011), a necessidade de se desenvolver instrumentos de mensuração cada vez mais consistentes é a única forma possível para aprimorar a mensuração nas ciências sociais.

1.2 Oportunidade de pesquisa e justificativa

Nas ciências sociais, muitos construtos nada mais são do que “abstrações teóricas”, “estados cognitivos inobserváveis, individuais (por exemplo, atitudes) ou compartilhados (por exemplo, valores culturais), que podem existir mais na mente de cientistas sociais do que na dos sujeitos, foco da pesquisa” (SPECTOR, 1992, p.13). Por essa razão, Figueiredo, Rocha e Silva (2009) argumentam que existe a dificuldade de se avaliar e desenvolver escalas para mensuração de um construto desta natureza, como é o caso da distância psíquica.

De forma complementar, Zabaldi, Costa e Ponchio (2014) discutem que, diferentemente de algumas áreas da ciência, em que a maior parte dos conceitos trabalhados podem ser diretamente observados (tais como altura, peso e idade), nas ciências sociais de maneira geral, e, em Marketing, especificamente, frequentemente é possível deparar-se 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.

Apesar das dificuldades e das discussões entre pesquisadores sobre a mensuração em ciências sociais e comportamentais, ela tem sido um tema constantemente pesquisado. Segundo Costa (2011), de modo geral, o problema da mensuração chegou às ciências sociais e comportamentais de maneira pontual, mas nunca de modo generalizado.

Mesmo na área de Psicologia, onde a mensuração já possui um elevado grau de desenvolvimento através da Psicometria, enquanto partes ou correntes da Psicologia utilizam métodos quantitativos muito avançados, outras partes ficam distantes destes, optando por análises qualitativas ou estudos de casos. O mesmo ocorre com a área da Educação, que se apropria de ferramentas de mensuração de modo mais enfático na subárea de avaliação educacional, mas inclusive nessa subárea, há um amplo desenvolvimento teórico sem suporte de métodos de mensuração ou operacionalizações quantitativas.

Especificamente no campo das ciências sociais, verifica-se um avanço mais enfático na área de economia, em função principalmente do forte desenvolvimento da quantificação da econometria, o que desencadeou, em paralelo, uma maior preocupação com a questão da

mensuração das variáveis econômicas. No campo da administração, por exemplo, é possível verificar variações de intensidade no uso de métodos e ferramentas de mensuração, tendo algumas subáreas mais intensivas, como Finanças e Produção, que possuem variáveis mais objetivas e que são largamente utilizadas em estudos e aplicações práticas; outras com uso parcial e pontual, como Marketing e Recursos Humanos; e outras áreas, com uso muito restrito, como a Teoria das Organizações (COSTA, 2011).

De fato, muitas das variáveis utilizadas em **finanças** podem ser medidas na escala de razão, tais como lucros, receita, despesas, ativos, etc., o que permite até a geração de índices para avaliar o desempenho empresarial, como os índices de estrutura de capital, liquidez e rentabilidade (MATARAZZO, 1997). Na área de **produção**, o uso da medida prevalece principalmente na área de planejamento e controle da capacidade produtiva, dos estoques, da cadeia de suprimentos e principalmente, no controle de qualidade (SLACK *et. al.*, 1997). Variáveis como erros de processo e tempo padrão de produção são medidas comuns nessa área.

Em **marketing**, mais especificamente na subárea do comportamento do consumidor, a maioria das variáveis é mensurável somente por escalas de intervalos, tais como satisfação, identificação, lealdade, confiança, comprometimento, intenções futuras, percepção de valor, percepção de risco, de qualidade, entre outras. Assim como em Psicologia, mensurar variáveis abstratas, como atitudes e percepção, é um ponto central para o desenvolvimento prático e teórico de Administração. Por essa razão, a atividade de mensuração em Administração é fortemente dependente do desenvolvimento de escalas, o que se faz já no nível mais acadêmico. Em razão da preocupação mais enfática dada pela área de Marketing, as questões mais acadêmicas acerca da mensuração acabaram tendo seu principal ambiente de discussão nos periódicos acadêmicos da área, com as teorizações sendo expandidas para outros contextos da Administração (COSTA, 2011).

Considerando a publicação internacional na área de marketing, a consideração da mensuração como parte central do processo de pesquisa quantitativa alcançou um status diferenciado a partir de 1960 (ZABALDI; COSTA; PONCHIO, 2014). Uma parte deste desenvolvimento pode ser atribuída, de certa forma, ao clássico artigo de Churchill publicado em 1979. Nesse artigo, Churchill (1979) criticou a falta de uma metodologia adequada aos pesquisadores da época que trabalhavam com mensurações em marketing, afirmando ainda que esta seria uma das principais razões para a baixa qualidade encontrada em parte da literatura de marketing. Para Jacoby (1978, p. 91) *apud* Churchill (1979, p. 64):

Mais assombroso do que o grande número de nossas medidas é a facilidade com que elas são propostas e da falta de crítica com que elas são aceitas. Na verdade, a maior parte das nossas medidas são medidas apenas porque alguém diz que elas são, e não porque elas têm sido demonstradas para satisfazer critérios de medição padrão (validade, confiabilidade e sensibilidade). Dito de uma forma diferente, a maioria das nossas medidas não são mais sofisticadas do que inicialmente, afirmar que o número de pedras que uma pessoa pode contar em um período de 10 minutos é uma medida de inteligência dessa pessoa; em seguida, conduzir um estudo e descobrir que as pessoas que podem contar muitas pedras em dez minutos também tendem a comer mais; e, finalmente, concluir com isso: as pessoas com alta inteligência tendem a comer mais.

Percebendo então a carência de uma metodologia que pudesse guiar os pesquisadores da área, Churchill (1979) propôs um modelo em que as medidas dos construtos de interesse dos profissionais de marketing poderiam ser desenvolvidas. De fato, esse modelo rapidamente passou a ser adotado em larga escala pela comunidade científica da área, ficando mais tarde conhecido como o paradigma¹ de Churchill. Esse acontecimento pode ser observado no discurso de Costa (2011, p. 23):

Um foco mais concentrado em mensuração foi desencadeado a partir do início dos anos 1980, como consequência da publicação da proposta de Gilbert Churchill Jr., no ano de 1979, que inclusive veio a se constituir em um ‘paradigma’ de desenvolvimento de escalas em Marketing por mais de duas décadas.

Para se ter uma ideia, este clássico artigo já recebeu mais de 3200 citações, segundo pesquisa realizada no Web of Knowledge (2014) e norteou outros desenvolvimentos nessa área, como o trabalho de Gerbing e Anderson (1988), que apresentou uma atualização ao modelo de Churchill, incorporando a questão da unidimensionalidade de escala.

O fato é que nessas últimas décadas, a literatura internacional sobre desenvolvimento e validação de escalas já adquiriu um elevado grau de desenvolvimento metodológico e científico. Os livros sobre o tema na área de pesquisa social aplicada são vastos, tais como as obras de DeVellis (2012), Pedhazur e Schmelkin (1991) e Netemeyer, Bearden e Sharma (2003), entre muitos outros.

Especificamente sobre a área de marketing, há vários livros sobre compilações de escalas, os chamados *handbooks* de escalas de marketing, como as obras de Bearden, Netemeyer e Haws (2011), que consolidaram em um exemplar mais de 150 medidas encontradas na literatura de marketing, e Bruner, Hensel e James (2005), que já lançaram o

¹ Um paradigma pode ser conceituado como um grupo de suposições relacionadas sobre o mundo que é compartilhada por uma comunidade de cientistas que o investigam (KUHN, 1962 *apud* DESHPANDE, 1983). Este grupo de suposições fornece então um modelo conceitual e filosófico que passa a ser uma referência inicial como base de modelo para estudos e pesquisas.

quarto volume do seu livro. Outras coletâneas de escalas em áreas mais específicas incluem as obras de Rubin, Palmgreen e Sypher (1994) na área de comunicação, Hensel, James e Bruner (2008) na área de propaganda, e a de Robinson, Shaver e Wrightsman (1991) na área de atitudes e medidas psicológicas e sociais.

Em contraste, a literatura nacional sobre compilações de escalas em marketing desenvolvidas e/ou validadas no país parece ainda estar em uma fase embrionária. Um primeiro movimento nesta área foi desenvolvido por Vieira (2011), que levantou um repertório de mais de 50 medidas existentes no cenário nacional para mensuração da performance organizacional e de marketing. Entretanto, esta obra possui uma importante limitação, visto que o autor não descreve em detalhes os procedimentos que foram utilizados para a inclusão das escalas, limitando-se apenas a dizer que não foram inseridos artigos que apenas fizeram uma análise fatorial exploratória de itens e que não partiram de uma base teórico-conceitual sólida. Além desta limitação, outro ponto negativo é que o autor não analisou a qualidade técnica dos instrumentos inseridos, mas apenas reuniu artigos publicados sobre escalas que em seu julgamento, poderiam ser considerados como uma medida psicométrica válida. Como afirma o próprio autor:

A obra não objetiva verificar a qualidade do conteúdo do artigo que apresenta, testa e valida uma escala. Este é um procedimento altamente subjetivo e relativo. Portanto, a tarefa de verificar se os procedimentos de tradução reversa, de validação, de testes, de aplicação em amostra, de opção pela rotação de análise fatorial, do cálculo da confiabilidade, dentre outros, não foi preocupação deste livro, mas sim é uma responsabilidade dos autores dos trabalhos citados (VIEIRA, 2011, p. 8).

Pode-se dizer que a preocupação com a mensuração de variáveis parece ter evoluído no Brasil mais recentemente, a partir dos anos 2000, como natural evolução da pesquisa de orientação mais acadêmica que as escolas de Administração adotaram desde então. Atualmente, segundo Zabaldi, Costa e Ponchio (2014), a análise de mensuração é requerida na maior parte dos relatos de pesquisa apresentados nas formas de dissertações, teses e artigos.

Sobre a questão da validação, entretanto, em recente estudo realizado em 2013 por Alves *et al.* (2013), no qual foram analisados 91 artigos publicados entre 2004 e 2012, com o objetivo de entender como se comporta a publicação de artigos sobre validação de instrumentos e as formas de validação mais utilizadas pelos pesquisadores nos eventos e periódicos de administração no Brasil, concluiu-se que, apesar dos avanços sobre o tema no país, os estudos ainda estão concentrados em grupos específicos. Para os autores, os

procedimentos mais utilizados foram em relação à validação de construto “tanto os de Convergente ou de Consistência dos Itens” (ALVES, *et al.* 2013, p.13). Outro ponto importante levantado pelos autores foi de que não existe por parte dos pesquisadores brasileiros do campo de administração o uso de mais de duas formas de validação, o que reforça, sob esse aspecto, a carência de realização dos três métodos de validação mais conhecidos, recomendados e apresentados na literatura (conteúdo, construto e critério), que permitiria uma possível generalização dos resultados gerados pelo instrumento.

Em uma pesquisa complementar, mas agora considerando a área de marketing, Froemming *et al.*, (2000) em uma análise dos artigos de pesquisa *survey* publicados na RAE, na RAUSP e no ENANPAD, no período de janeiro de 1990 a dezembro de 1999, discutem que em linhas gerais há pouca preocupação com aspectos relevantes da qualidade metodológica da investigação em marketing. Para os autores, dentre diversos aspectos está a falta de preocupação de pesquisadores no que tange à validade de instrumentos.

Sampaio *et al.*, (2012), também em estudo da área de marketing, uma década após o estudo de Froemming *et al.* (2000), afirmam que os pesquisadores da área tornaram-se mais criteriosos após o ano 2000, aumentando o registro em suas pesquisas quanto à questão da validade (apesar de ainda ser percentualmente baixa).

Em âmbito internacional, percebe-se claramente a crescente preocupação por parte dos autores estrangeiros em relação a qualidade das pesquisas e sobre o andamento da área de marketing e seu futuro (FIRAT *et al.*, 2010; REIBSTEIN; DAY; WIND, 2009; YADAV, 2010). A edição de julho de 2011 do *Journal of Marketing*, comemorativa ao 75º aniversário do periódico, traz uma sessão especial dedicada a refletir a respeito do futuro da disciplina e das publicações em marketing, em especial, quanto à relevância da pesquisa, do distanciamento entre a teoria e a prática, da necessidade de maior colaboração entre a academia e as empresas, da excessiva ênfase de rigor metodológico e de fragmentação da área e da necessidade de incremento de artigos com contribuições conceituais (MACINNIS, 2011).

Neste contexto, a principal contribuição acadêmica desta tese está em discutir os principais procedimentos metodológicos adotados pelos pesquisadores de marketing no desenvolvimento de escalas de mensuração nesta área, procurando avaliar ainda a qualidade desses instrumentos de medida e, de uma forma conjunta, fornecer um panorama geral e mais recente do nível de desenvolvimento metodológico em que se encontram os artigos publicados nessa área aqui no Brasil. Além disso, com os resultados desse trabalho, será possível elaborar recomendações metodológicas aos pesquisadores brasileiros com o intuito

de superar os problemas que forem verificados e elevar a qualidade das medidas desenvolvidas na área de marketing.

Outra contribuição relevante desta tese se refere à criação de um inventário com métricas nacionais encontradas nos anais de eventos da Anpad nos últimos 3 anos. Por meio de um processo sistemático de busca de artigos realizado nesta tese, as métricas encontradas poderão ficar a disposição para uso de outros pesquisadores ou gestores.

A criação de um inventário de métricas em marketing referida aqui é bastante justificável. Em primeiro lugar, dado que os profissionais de marketing necessitam de instrumentos confiáveis e válidos que mensurem o comportamento do consumidor e outras variáveis de marketing, como as percepções de produto, atendimento, propaganda, entre outras, esta tese tem seu mérito em reunir em apenas um único local diversas medidas de marketing existentes no cenário nacional. Profissionais de marketing poderiam, desta maneira, ter fácil acesso a essas medidas e, por meio da aplicação desses instrumentos, poderiam obter um diagnóstico do objeto investigado, auxiliando o processo de tomada de decisões.

Em segundo lugar, acadêmicos de marketing teriam uma fonte de pesquisa de escalas que poderiam ser utilizadas individualmente ou em conjunto com outras métricas para testar hipóteses de pesquisa. Como lembra Cooper e Schindler (2003), um dos objetivos da mensuração é apoiar ou refutar hipóteses. Desta maneira, trabalhos científicos poderiam ficar ainda mais robustos com a inserção de novas relações entre variáveis e/ou controlando o efeito de outras (por exemplo, inserindo a Escala de Desejabilidade Social de Marlowe-Crowne para controlar os efeitos do viés de desejabilidade social em uma pesquisa).

E finalmente, o desenvolvimento de um inventário de escalas também poderia permitir aos pesquisadores identificar lacunas de pesquisa dentro da área de marketing que representariam oportunidades para novos desenvolvimentos e/ou validações de escalas no contexto brasileiro.

Em relação a esse último ponto, é importante ressaltar que uma escala inicialmente desenvolvida e validada em um certo contexto, para uma determinada população, a qual possui cultura e idiomas específicos, não pode ser simplesmente traduzida e depois utilizada para avaliar individualmente outras populações as quais possuem culturas diferentes daquela primeira. Como lembra Pasquali (2011), independentemente da qualidade dos itens, a resposta aos mesmos pode ser desvirtuada por fatores relativos ao sujeito que a eles reage, e nesse caso, a cultura atua como causa de erros de resposta, estando relacionada ao problema de transferência de instrumentos psicológicos para outras populações para as quais eles não foram especificamente construídos e validados.

Desta forma, escalas que foram desenvolvidas em outros países precisam ser traduzidas e depois validadas no contexto brasileiro para poderem ser efetivamente utilizadas no país. Ter em mãos um inventário das escalas que já foram validadas no cenário nacional pode contribuir de forma positiva para o desenvolvimento e/ou validação de novos instrumentos para a cultura brasileira.

1.3 Problema e objetivos de pesquisa

Com base na discussão precedente, esta pesquisa visa responder a seguinte pergunta: Quais os pontos fortes e fracos das escalas de marketing que foram propostas e publicadas nos principais congressos da Anpad?

Nesse sentido, o objetivo geral desta tese foi avaliar a qualidade das escalas de mensuração propostas e publicadas na área de marketing nos anais de eventos da Anpad nos últimos 3 anos.

Para avaliar as escalas propostas em marketing, é necessário primeiramente conhecer a teoria sobre mensuração, entender os problemas e erros relacionados ao processo de medir e os modelos que podem ser utilizados para a construção de instrumentos de medida confiáveis e válidos. A revisão teórica e a discussão sobre os procedimentos de desenvolvimento de escalas proporcionarão o conhecimento necessário para a definição de um protocolo de pesquisa, um guia que tem como função sistematizar os procedimentos de inclusão ou não de artigos sobre desenvolvimento e/ou validação de escalas na análise de dados. Portanto, são propostos os seguintes objetivos específicos:

- a. revisar a literatura de teoria de mensuração, identificando todos os pré-requisitos necessários para o desenvolvimento de uma escala válida e confiável;
- b. compilar publicações de anais de congressos da Anpad que desenvolveram e/ou validaram escalas de mensuração na área de marketing;

Após o levantamento dos artigos na base de dados, será iniciada a fase analítico-descritiva dos dados, que visa verificar a qualidade dos instrumentos desenvolvidos quanto a sua confiabilidade e validade (conteúdo, critério e construto), bem como verificar quais modelos e procedimentos de desenvolvimento de escalas foram mais utilizados. Assim, são propostos os seguintes objetivos específicos:

- c. avaliar as publicações levantadas quanto aos métodos e procedimentos empregados para verificar a confiabilidade e validade das escalas de marketing desenvolvidas;
- d. gerar um relatório geral sintetizando quais procedimentos de desenvolvimento de escalas foram mais utilizados nos artigos levantados;

Com base no referencial teórico pesquisado, nos dados levantados e nas análises realizadas, por fim, é possível tecer algumas recomendações e limitações sobre o uso das metodologias e procedimentos para o desenvolvimento de escalas, com o intuito de promover um guia para os pesquisadores da área desenvolverem instrumentos de medida mais confiáveis, válidos e que possam ser generalizados. Desta maneira, tem-se o último objetivo específico:

- e. elaborar recomendações e limitações para o desenvolvimento e uso de escalas em marketing;

1.4 Estrutura do trabalho

Este trabalho foi estruturado da seguinte maneira:

- No capítulo 2, é realizada uma revisão da literatura sobre a teoria da medida, apresentando algumas definições de mensuração e sua aplicação nas ciências sociais e comportamentais, debatendo-se a questão epistemológica da utilização da medida nessas ciências. Além disso, também é apresentada a taxonomia padrão de classificação das medidas que normalmente impera nessas ciências, bem como suas principais críticas;
- No capítulo 3, são apresentados os conceitos fundamentais de erro de mensuração, onde são abordados ainda os principais vieses de resposta em escalas multi-itens;
- No capítulo 4, os conceitos fundamentais de confiabilidade de escalas são apresentados, discutindo-se então as principais técnicas que podem ser empregadas para sua verificação;
- No capítulo 5, são descritos os conceitos fundamentais de validade de escalas, apresentando as principais formas e técnicas normalmente empregadas para sua verificação;

- No capítulo 6, são abordados alguns dos modelos de desenvolvimento de escalas, destacando os principais passos utilizados para a criação de medidas confiáveis e válidas;
- No capítulo 7, é apresentada a metodologia que foi utilizada neste trabalho, destacando-se o tipo de pesquisa e as principais etapas de pesquisa que foram executadas;
- No capítulo 8, é apresentada a síntese geral dos principais procedimentos de desenvolvimento de escalas que foram utilizados nos artigos levantados, incluindo os procedimentos de teste de validade e confiabilidade.
- No capítulo 9, são expostas as conclusões do trabalho, bem como suas limitações e as oportunidades de trabalhos futuros decorrentes desta tese.

2 REFERENCIAL TEÓRICO – TEORIA DA MEDIDA

A teoria da medida é considerada uma ramificação da estatística aplicada que procura descrever, categorizar e avaliar a qualidade de medições; melhorar a sua usabilidade, a precisão e o significado de medidas; e por fim, propor métodos para desenvolver novos e melhores instrumentos de medidas (ALLEN; YEN, 1979).

Embora seja classificada como um ramo da estatística, em cada área da ciência o estudo da teoria da medida recebe um nome diferente e específico, composto por um prefixo ao termo “metria”, identificando assim a área em que a medida é aplicada. Por exemplo, em Psicologia, o estudo da teoria da medida aplicada aos comportamentos psicológicos é chamada de Psicometria. Na área de Sociologia, recebe o nome de Sociometria, e em Economia, Econometria. Outros exemplos ainda incluem a Antropometria (Antropologia) e a Biometria (Biologia, no estudo das características físicas e comportamentais dos seres vivos), entre outras mais.

Neste capítulo 2 é realizada a discussão da teoria da medida aplicada às ciências sociais e comportamentais. Inicialmente algumas definições fundamentais serão esclarecidas, como a definição de mensuração e conceitos relacionados a ela. Em seguida, discute-se o uso da mensuração em ciências e as questões epistemológicas que envolvem seu emprego nas ciências sociais e comportamentais. Na sequência, são discutidos aspectos importantes da natureza da medida e sua base axiomática, bem como os níveis em que a medida pode ser classificada (escalas nominais, ordinais, intervalar e de razão). Por fim, são discutidos os principais pontos de críticas quanto à classificação citada anteriormente, do ponto de vista de diversos pesquisadores da área.

2.1 Definição de mensuração

A origem etimológica da palavra mensuração está relacionada à palavra latina *mensura*, que vem de *mensus*, a qual é a forma de particípio do verbo *metiri*, que significa, em latim, medir, avaliar ou estimar medidas. Em seu sentido próprio, *mensura* significa medida, quantidade, dimensão, ou mesmo instrumento usado para averiguação destas propriedades. Em seu sentido figurado, mensuração significa ainda proporção, grandeza, capacidade, alcance (FARIAS, 1985).

Mensuração está, assim, relacionada ao processo de atribuir números às coisas, mas deve-se reconhecer que existem diferentes formas em que as coisas podem ser quantificadas.

Nesta linha de raciocínio, é interessante citar a definição de Stevens (1946), que de uma forma mais genérica, definiu mensuração como a “atribuição de numerais a objetos ou eventos de acordo com regras” (STEVENS, 1946, p. 677). Esta definição tem sua importância quando agrega ao conceito de medir a necessidade de se definir regras para padronizar a forma como os números podem ser atribuídos às coisas.

Entretanto, é importante destacar que quando se lançam medições a respeito de um objeto, não se está medindo de fato o objeto, mas sim, as características desse objeto. Nesse sentido, é destaca-se a definição de Allen e Yen (1978), que definem mensuração como a “designação de números a indivíduos em uma forma sistemática com o objetivo de representar as propriedades deste indivíduo” (ALLEN; YEN, 1978, p. 2).

Das definições apresentadas anteriormente, surgiram definições mais completas e/ou mais técnicas. Entre essas definições, uma que merece destaque é aquela proposta por Erthal (2009, p.20), em que medir significa “atribuir magnitudes a certa propriedade de um objeto ou classe de objetos, de acordo com certas regras preestabelecidas e com a ajuda do sistema numérico, de forma a que sua validade possa ser provada empiricamente”. De acordo com a autora, o processo de medir implica em algumas características importantes. Primeiramente, o processo de medir é sempre quantitativo, visto que ele implica sempre em um resultado numérico e não frases descritivas. Em segundo lugar, apresenta-se em unidades relativamente constantes, desde que as condições de mensuração o sejam. Em terceiro lugar, e especificamente no caso da psicologia, a medida é considerada relativa, por não dispor de um ponto zero absoluto, como certas variáveis da física.

No uso cotidiano, de acordo com Costa (2011), mensuração e medição são utilizadas como palavras sinônimas, embora a primeira remeta eventualmente a um contexto mais técnico. Segundo o autor, a palavra medição e as palavras medir e medida são mais largamente utilizadas, variando sentidos e significados de acordo com a situação empregada. Por isso, a atividade de medir pode estar associada tanto a um procedimento especializado e mais técnico, quanto a um sentido de avaliação mais subjetiva e de referência mais cotidiana.

Uma definição de mensuração mais técnica é dada por Ferris (2004), em que após uma longa revisão conceitual sobre o tema, propôs a seguinte definição:

Mensuração é um processo empírico, que se utiliza de instrumentos para efetuar uma verificação rigorosa e objetiva de algo observável dentro de uma categoria de um modelo deste elemento observável, que distingue uma manifestação de significado de outras manifestações de significados possíveis e distinguíveis. (FERRIS, 2004, p. 107).

A definição de Ferris (2004), ainda que seja mais direcionada ao contexto técnico, tem sua relevância ao incluir na definição a necessidade de se utilizar um instrumento para realizar a mensuração.

Entretanto, alguns autores reconhecem que o processo de mensuração deve ser entendido como um conceito mais amplo, que deve englobar um leque maior de possibilidades. Neste sentido, Hoffman (1983) sugere dois significados para a mensuração. Segundo o autor, medir em um sentido restrito (quantificação), é a comparação experimental de uma quantidade mensurada com um padrão metrológico definido. Medir em um sentido amplo (classificação) é a comparação experimental do objeto medido com um padrão particular. Com base nessa definição, a mensuração pode ser entendida como o ato de quantificar, por meio de padrões metrológicos, ou pelo ato de classificar, quando são fixados padrões numéricos para classificar objetos, como por exemplo, indicando com o número 1 o sexo masculino e com o número 2 o sexo feminino.

Seguindo a linha de Stevens, Hoffman e Ferris, Costa (2011) propôs uma definição de mensuração que inclui as características apontadas por esses autores. Segundo o autor, mensuração pode ser definida como o “procedimento de verificação empírica, a partir de instrumentos adequados, de uma característica bem definida e diferenciada de um objeto, com a finalidade de atribuição de números, seja para definir quantidades, seja para definir classificações” (COSTA, 2011, p. 16).

Avaliando-se as diversas definições de mensuração elaboradas, Costa (2011) afirma que elas evidenciam certas características do processo de medir, que seriam:

- (i) Não se mensura o objeto, mas sim uma característica bem definida deste;
- (ii) A característica deve ser claramente diferenciável de outras características do objeto;
- (iii) Deve possuir uma variação que indique o sentido da regra de atribuição definida.

2.2 Conceitos relacionados

Além da definição de mensuração, é interessante também caracterizar outros termos que estão relacionados a ela. De acordo com Erthal (2009), se a mensuração pode ser considerada como o processo de atribuir símbolos a objetos seguindo regras, os números atribuídos a esses objetos ou eventos devem representar quantidades de atributos. Neste sentido, é importante deixar claro alguns termos utilizados nessa definição:

- **Símbolo:** é o que representa o atributo medido;
- **Objeto:** elemento para o qual a mensuração se dirige;
- **Atributo:** característica do objeto aferida pela mensuração;
- **Instrumento:** meio utilizado para medir o atributo do objeto;
- **Regras:** formulações, previamente estabelecidas, que indicam os procedimentos para a atribuição de símbolos aos atributos dos objetos que determinam as relações entre o objeto e o símbolo;
- **Situação-padrão:** diz respeito ao controle de variáveis que podem interferir no resultado da mensuração, chamado medida.

Segundo Erthal (2009), a medida pode ser classificada de duas formas: a fundamental e a derivada. A **medida fundamental** é aquela que é obtida como resultado da mensuração direta, ou seja, “o atributo de um objeto é alcançado através de uma comparação simples e direta com outro objeto que apresenta uma quantidade padronizada dessa propriedade” (HAYS, 1970 *apud* ERTHAL, 2009). Segundo Pasquali (2011), trata-se de dimensões (atributos mensuráveis diretamente) que permitem a concatenação, isto é, dois objetos podem ser associados, concatenados, formando um terceiro objeto de mesma natureza, que é o caso, por exemplo, com os atributos de massa, comprimento e duração temporal. Já a **medida derivada** é definida como o produto de uma operação de mensuração baseada em indícios que se supõem estarem relacionados com o atributo do objeto medido (ERTHAL, 2009). Essas definições estão relacionadas à definição de construto latente e observável, que será discutida mais adiante.

Para Cooper e Schindler (2003), as variáveis em estudo em uma pesquisa podem ser classificadas como **objetos** ou **propriedades**. Os objetos incluem coisas da experiência comum, como tabelas, pessoas, livros e automóveis, como também, coisas que não são concretas, como genes, atitudes, nêutrons e pressões de grupos. Já as propriedades representam as características dos objetos, como por exemplo, as propriedades físicas de uma pessoa (peso, altura, postura, etc.), as propriedades psicológicas (atitude, inteligência, etc.), as propriedades sociais (capacidade de liderança, afiliação de classe, posição social, etc.), entre outras. Entretanto, os autores indicam que os objetos e suas propriedades não podem ser medidos diretamente, mas sim, por meio de indicadores de propriedade ou indicadores das propriedades dos objetos. Por exemplo, a mensuração da “motivação no trabalho” não poderia

ser medida diretamente, mas por meio de indicadores, como por exemplo, a satisfação do funcionário com a sua remuneração, os benefícios que recebe, a sua relação com o chefe, a infraestrutura que está a sua disposição para desempenhar o seu trabalho, etc. Desta forma, como cada propriedade não pode ser mensurada diretamente, é necessário inferir sua presença ou ausência observando-se algum indicador ou apontador de mensuração.

Outro termo frequentemente utilizado pelos pesquisadores em mensuração e que nem sempre é definido de forma clara é a palavra **escala**. De acordo com Urbina (2007), o termo escala é empregado de muitas formas diferentes, podendo se referir a:

- um teste composto de várias partes, como por exemplo, a Escala de Inteligência Stanford-Binet;
- um subteste, ou conjunto de itens dentro de um teste, que mede uma característica distinta e específica, como por exemplo, a Escala de Depressão do Inventário Multifásico Minnesota de Personalidade (MMPI);
- um conjunto de subtestes que compartilham certas características, como, por exemplo, as Escalas Verbais dos testes de inteligência Wechsler;
- um instrumento separado formado por itens que avaliam uma única característica, como por exemplo, a Escala de Locus de Controle Interno-Externo;
- sistema usado para classificar ou atribuir valor a alguma dimensão mensurável, como, por exemplo, uma escala de 1 a 5 na qual 1 significa discordo totalmente, e 5 significa concordo totalmente.

Em administração, é muito frequente o emprego do termo escala para designar estes dois últimos significados citados por Urbina (2007), além de outro significado que se refere à classificação do nível da medida. Primeiramente, o termo escala pode representar um grupo de itens que procura medir uma (ou mais) características de um objeto, como por exemplo, o arrependimento no processo de tomada de decisão do consumidor, consumo de status, avaliação da imagem de países, consciência ecológica, etc. Itens, neste caso, se referem às questões (ou indicadores) que farão parte do questionário. Em segundo lugar, o termo pode se referir à técnica de escalonamento utilizada pelo pesquisador, como por exemplo as técnicas comparativas (escalas de comparação pareada, escalas de ordem de posto, escalas de soma constante, etc.) ou não comparativas (escalas de Likert, escalas de diferencial semântico de Osgood, escalas de Stapel, escalas de Thurstone, etc.). Finalmente, o termo ainda é

empregado como uma forma mais genérica de classificar o nível de medida das variáveis que um pesquisador possui em mãos, que segundo Stevens (1946), podem ser as escalas nominais, ordinais, intervalares e de razão.

Além desses termos, outro termo importante normalmente utilizado junto com mensuração é o termo **construto**. De acordo com Costa (2011), a expressão ‘construto’ é utilizada para se referir a uma característica de um dado objeto que pode ser mensurada, isto é, que apresenta características ou conceitos que viabilizem sua quantificação ou classificação, sendo bem delimitado em relação a outras características do objeto. Como exemplos na área de administração, é possível citar os construtos de lealdade, intensidade de reclamação, qualidade percebida do serviço, clima organizacional, intensidade ética, satisfação no trabalho, dentre outros. Esses construtos podem ser mensurados diretamente (**construto observável**) ou podem gerar manifestações observáveis (**construto latente**). Para medi-los, é necessário utilizar uma **escala de mensuração**, que inclui um **instrumento** de medição, as **escalas de verificação** (maneira como se atribui números às variações e quantidades da característica mensurada no objeto) e as **regras e orientações** de sua aplicação. Para efeitos de desambiguação, o termo “escalas de verificação” utilizado aqui por Costa (2011) é empregado com o mesmo sentido de técnica de escalonamento, não sendo, portanto, um conceito novo, mas apenas uma outra forma de alcunhar a técnica de escalonamento utilizada.

A forma como esses conceitos estão relacionados é apresentada na Figura 1. De acordo com Costa (2011), a mensuração parte de um objeto, que possui uma característica específica, que por sua vez, gera um construto, seja ele observável ou latente. A avaliação desta característica é realizada por meio de uma escala de mensuração, a qual é composta por um instrumento físico (ou sua descrição), que é determinado por uma escala de verificação específica e por um conjunto de regras e orientações de aplicação e análise.

Figura 1: Relação entre os conceitos relacionados à mensuração.

Fonte: Costa (2011)

Como nas ciências sociais e comportamentais há uma quantidade significativamente grande de construtos de natureza latente, é interessante caracterizá-lo de uma maneira mais específica. Para Costa (2011, p. 33), os construtos latentes representam “a característica de um objeto de interesse que apresenta variações que viabilizam quantificações, e que são muito bem delimitados em relação a outras características do mesmo objeto”. O construto latente ainda pode ser caracterizado como sendo do tipo **reflexivo**, em que são eles quem influenciam outras variáveis, ou **formativos**, os quais são influenciados por outras variáveis. Em outras palavras, essa definição pode ser reduzida a uma definição de causalidade, de modo que, no primeiro caso (reflexivo) o construto causa a variável, e no segundo caso (formativo), é a variável que causa o construto.

Finalmente, merece ainda esclarecer que os construtos podem apresentar apenas uma única dimensão, chamados então de **construto unidimensional**, ou ainda, podem ser multifacetados, possuindo várias dimensões, de modo que não haja como alcançar uma medida adequada sem que ele seja dividido em dois ou mais componentes. Neste último caso,

tem-se um **construto multidimensional** (COSTA, 2011). A Figura 2 ilustra em síntese os tipos de construtos.

Figura 2: Tipos de construtos.

Fonte: Costa (2011)

2.3 Mensuração e ciência

No contexto técnico, em que a palavra mensuração é normalmente mais utilizada, os primeiros esforços na quantificação das características de objetos se deram por meio de **símbolos numéricos**, e podem ser creditados aos primeiros geômetras que precisaram antes acessar as medidas das formas, para então seguir em um desenvolvimento mais abstrato. Segundo Costa (2011), é desse esforço que se deriva a postulação da medição por meio de

comparações com uma medida padrão, a qual Euclides chamava de magnitude padrão em sua obra **elementos**.

Para atender as necessidades dos estudiosos e a complexidade da vida prática, vários conjuntos numéricos foram surgindo em épocas históricas diferentes. Conforme Pasquali (2011), inicialmente só havia os números inteiros, que se mostravam suficientes para a contagem de objetos discretos (razão pela qual são também chamados de números naturais), em que era possível realizar as operações de soma e multiplicação. Entretanto, esse conjunto numérico apresentava problemas quando se queria subtrair um número maior de um número menor. Esta limitação do sistema de inteiros fez com que o sistema fosse estendido para incluir números negativos e o zero. Limitação mais séria ainda ocorria com a divisão, o que trouxe a adoção dos números fracionários. Este conjunto de números (inteiros positivos, negativos, zero e frações) constitui o sistema de números racionais, dado que qualquer número deste sistema pode ser expresso em termos de razão entre dois números inteiros.

Entretanto, de acordo com Costa (2011), algumas medidas não podiam ser comparadas com a magnitude padrão, e que, portanto, não geravam um número racional como medida, mas sim um número não racional, ou irracional (daí vem o nome do conjunto dos números irracionais, isto é, que são impossíveis de serem expressos como uma razão). Esta problemática da operacionalização dos números irracionais levou então à mensuração por meio de outra categoria numérica, que é a dos números complexos.

Como afirma Costa (2011), mesmo com as limitações dos números irracionais, o fato é que o entendimento da mensuração como a atribuição de números a partir da comparação com uma medida de base mostrou-se altamente proveitoso para o avanço do conhecimento e do bem-estar da humanidade e que “a atividade de mensuração constitui um pré-requisito central para o próprio desenvolvimento científico” (COSTA, 2011, p. 11).

Compartilhando este mesmo pensamento, Pasquali (2011) argumenta que a ciência se apercebeu das vantagens consideráveis que ela pode obter ao utilizar a linguagem matemática para descrever o seu objeto próprio de estudo, e que é o uso deste modelo que vem possibilitando distinguir níveis de progresso no conhecimento científico.

De fato, são inúmeros os defensores do uso dos números nas ciências. Mari (2003) cita algumas frases clássicas de notáveis, que exaltaram o valor da medição:

- “O progresso da civilização tem estreita relação com o desenvolvimento de medições” – A. Pérard, engenheiro.

- “Quando você pode mensurar o objeto sobre o qual você está falando, e expressá-los em números, você sabe então alguma coisa sobre este objeto; se não for assim, seu conhecimento é restrito e insatisfatório” – atribuída a Lord Kelvin, físico.
- “Conte o que for contável; meça o que for mensurável, e o que não for mensurável, faça mensurável” – atribuída a Galileu Galilei, físico.

Mari (2005) relata que a mensuração ganhou, nas ciências, um protocolo de verdade, em especial porque viabiliza o desenvolvimento de análises matemáticas. Esta ideia é corroborada por Costa (2011), quando diz que a visão da mensuração, associada à quantificação, está no fato de que ela gera segurança nos procedimentos comparativos, e que com o uso da medida, em particular quando são expressas em quantidades, é possível “proceder a manipulações matemáticas que deverão conduzir a resultados ‘incontestáveis’, uma vez que são resultados de uma operacionalização lógica, perfeita e desprovida de maiores erros” (COSTA, 2011, p. 8). O autor ainda enaltece o valor que os esforços de matematização geraram, por exemplo, para a área de Economia a partir do século XX, e que, em boa medida, a mensuração e a matematização parecem ser formas adequadas de avanços nas práticas e nas pesquisas das ciências sociais.

O fato é que os esforços de pesquisa em mensuração são próprios de toda e qualquer ciência ou área acadêmica, e não apenas das ciências exatas e das engenharias, ainda que sejam estas áreas as que alcançaram um maior nível de desenvolvimento (COSTA, 2011).

Por esta razão, a mensuração calcada no protocolo da verdade tornou-se um meio que viabilizou o desenvolvimento da pesquisa em diversas áreas do conhecimento e abriu caminho para a criação de uma teorização e sistemática próprias que deram origem à Ciência da Mensuração, conforme afirma Costa (2011, p. 11):

O status da mensuração como tendo um protocolo da verdade ensejou o desenvolvimento de uma teorização própria, e de um esforço sistemático de desenvolvimento teórico e instrumental no que comporia uma ‘Ciência de mensuração’ (*Measurement Science*).

Entretanto, segundo Erthal (2009), nas ciências exatas, ao contrário das ciências sociais, a medida consegue atingir grande exatidão, o que conduz ao estabelecimento de relações facilmente comunicáveis e precisas. Já a mensuração em psicologia, por exemplo, é algo bem mais complexo, visto que nem sempre se trata de algo observável, como é o caso das variáveis físicas: “o psicólogo lida quase sempre com construtos hipotéticos. Quando

planeja seus experimentos, recorre, portanto, a diferentes espécies de medição que o habilitem a estimar essas variáveis, o que requer, evidentemente, alto grau de abstração” (ERTHAL, 2009, p. 22).

De forma semelhante, Costa (2011) argumenta que o conceito de mensuração a partir da medida padrão encontra dificuldades na proporção em que estes padrões não são tão seguros e bem determinados quanto nas variáveis físicas.

Pasquali (2011) ainda é enfático em dizer que a teoria da medida está razoavelmente axiomatizada somente nas ciências físicas, aparecendo ainda lacunar nas ciências psicossociais, onde segundo o autor, ainda se discute a viabilidade epistemológica da própria medida.

Na área do comportamento, em especial na psicologia, dada a sua complexidade, a medição se dá através de indícios que se supõe estarem ligados às medidas: “[...] os testes fornecem apenas uma situação padronizada que permite elucidar alguns comportamentos manifestos que se supõe representar a variável psicológica em questão” (ERTHAL, 2009, p. 23). Aliado a isso, outro problema frequente está nas manifestações do comportamento, que são amplamente variáveis ao longo do tempo, significando que medem reações sempre passíveis de mudança. Este também é o caso das atitudes em marketing, por exemplo.

Muitos dos pesquisadores que defendem o uso da medida no contexto das ciências sociais reconhecem as dificuldades na quantificação dos fenômenos pesquisados. Admitem que os instrumentos utilizados para descrever indivíduos quantitativamente em termos de várias características psicológicas ou comportamentais nunca medem estas características com uma perfeita precisão. Entretanto, mesmo um modelo criado para representar um determinado fenômeno que se manifeste incorreto pode, de certa forma, ser útil na descrição do fenômeno e útil para fazer previsões sobre ele. Além disso, um modelo pode ainda ser adotado para simplificar uma situação complexa, para torná-la compreensível e consequentemente “suportável”, mesmo que o modelo seja contrário aos fatos (GHISELLI; CAMPBELL; ZEDECK, 1981). Para minimizar esses problemas, Erthal (2009) afirma que é na aplicação dos métodos científicos e na utilização de instrumental padronizado que se permite medir e classificar o comportamento humano com a menor ambiguidade possível.

Para Costa (2011), apesar da mensuração de variáveis abstratas ser algo muito mais complexo, chegando até ao nível de contestação da própria possibilidade de mensuração, é perfeitamente possível medi-las, mas são necessários modelos apropriados para tentar minimizar os efeitos que geram erros. Parafraseando o autor, “a questão não é de medir ou não medir; trata-se de buscar minimizar erros de mensuração” (COSTA, 2011, p. 18). Isto

implica em um grande esforço conceitual e metodológico, que orbita em torno de modelos e teorias próprios para estas áreas (ciências sociais e comportamentais), e da apropriação de conceitos, modelos e ferramentas de uso geral nas ciências físicas e tecnológicas.

De fato, pelo que se pode perceber, o debate sobre o protocolo da verdade é bem desenvolvido no contexto das ciências físicas e tecnológicas, mas o seu alcance para as ciências sociais e comportamentais ainda é algo muito contestado. Isso abre precedentes para muitas discussões entre pesquisadores, e até mesmo, a condenação dos esforços quantitativos nesta área. Como revela Pasquali (2011, p. 52), “a teoria da medida em ciências não constitui campo pacífico entre pesquisadores, sobretudo em ciências psicossociais” e ainda hoje, a discussão do que exatamente significa “medida” nessa área não está de todo resolvida. Mesmo assim, muitos dos elementos deste debate são de extremo valor para reflexões sobre mensuração em qualquer área do conhecimento.

Dado esse panorama, cabe ao pesquisador, ao utilizar as técnicas de medidas, conhecer suas limitações, sabendo ser impossível abranger a totalidade de cada um dos fenômenos pesquisados e entender que, sendo estes fenômenos demasiadamente complexos, esse objetivo dificilmente será alcançado. O erro sempre irá existir, mas cabe ao pesquisador procurar desenvolver procedimentos metodológicos e sistemáticos que procurem minimizar os erros, ou ainda, estimar o nível de erro que se está incorrendo.

2.4 A natureza da medida

O uso do número na descrição dos fenômenos naturais constitui o objeto da teoria da medida. Entretanto, de acordo com Pasquali (1997, 2011), só há legitimidade do uso no número na descrição dos fenômenos naturais se, e somente se, as propriedades estruturais do mesmo forem salvaguardadas neste procedimento. Segundo o autor, a natureza da medida implica em alguns problemas básicos, tais como o problema de representação ou isomorfismo, a unicidade da representação e o erro.

O **problema da representação ou isomorfismo** está na legalidade de se passar de procedimentos e operações empíricos (a observação) para uma representação numérica desses procedimentos. A questão relevante aqui, então, é se seria justificável designar ou expressar objetos ou fenômenos naturais por meio de números. Se nesta designação se salvarem tanto as propriedades estruturais do número quanto as características próprias dos atributos dos fenômenos empíricos, a resposta para esta questão é positiva.

O **problema da unicidade da representação** envolve a discussão se o número seria de fato a única viável e possível representação das propriedades dos objetos naturais, ou se a tal representação é a mais adequada que se possa imaginar para descrever as propriedades em questão. Embora as discussões sobre a unicidade da representação do número entre cientistas da área psicossocial sejam enormes, mesmo entre aqueles que defendem a medida em ciência alertam que esta representação, ainda que sendo a melhor, apresenta níveis diferentes de qualidade ou precisão, dependendo do tipo de objeto que se está analisando. Assim, a problemática da unicidade da representação é o que irá definir o nível de escala de medida (nominal, ordinal, intervalar ou razão) que o pesquisador tem em mãos.

O **problema do erro** se refere às imperfeições de mensuração. Embora o número matemático seja um conceito unívoco, isto é, que é susceptível a uma só interpretação (sem a mínima variabilidade de interpretação), o número na medida dos fenômenos naturais se adultera um pouco, perdendo sua identidade pontual e absoluta, para se tornar um intervalo em vez de ser um ponto sem dimensões. Assim, na medida, o fato do número se tornar um intervalo, já diz que ele tem variabilidade (variância), e isso é o erro. Esses erros são devidos tanto ao instrumental de observação (os sentidos e suas extensões por meio de instrumentos tecnológicos) quanto às diferenças individuais do observador, além dos erros aleatórios, sem causas identificáveis. Assim, toda e qualquer medida vem acompanhada de erros. A consequência disto é que qualquer número que descreve um fenômeno empírico deve vir acompanhado de algum indicador do erro provável, o qual deve ser analisado dentro de teorias estatísticas para determinar se o valor encontrado e que descreve o atributo empírico está dentro dos limites de aceitabilidade de medida.

2.5 A base axiomática da medida

A legitimidade do uso do número como descritor dos fenômenos naturais só se justifica se realmente for possível utilizar o número para descrever os fenômenos da ciência (isomorfismo) e se as propriedades estruturais do número, bem como dos fenômenos naturais, forem preservadas neste procedimento (PASQUALI, 1997, 2011).

A matemática é um saber que está baseado em convenções: tanto seu objeto (o número), quanto suas regras são convencionadas. Segundo Pasquali (2011), no início do século XX, os filósofos e matemáticos Whitehead e Russell enumeraram 27 axiomas ou regras da matemática, que foram então apresentadas em seu livro *Principia Mathematica*. Um axioma, de acordo com o Dicionário Priberam da Língua Portuguesa (2012), pode ser

entendido como uma “proposição tão evidente que não precisa ser demonstrada”. Conforme Pasquali (2011), destes axiomas, três grandes conjuntos são importantes para o caso da medida e se referem aos axiomas que definem as propriedades numéricas de identidade, ordem e aditividade.

A propriedade da **identidade** define o conceito de igualdade, isto é, que um número é idêntico a si mesmo e somente a si mesmo, tendo 3 axiomas que expressam a relação de “igual a” (=):

- Axioma de reflexividade: $a = a$ ou $a \neq b$. Números são idênticos ou são diferentes;
- Axioma de simetria: se $a = b$, então $b = a$;
- Axioma de transitividade: se $a = b$ e $b = c$, então $a = c$. Duas coisas iguais a uma terceira são iguais entre si.

A propriedade da **ordem** define que todo número é diferente de outro, sendo que as diferenças não são apenas de qualidade, mas também de magnitude, isto é, um número não é só diferente do outro, mas um é maior que o outro. Assim, excetuando o caso da igualdade, os números podem ser colocados numa sequência invariável ao longo de uma escala linear (sequência monotônica crescente). A propriedade de ordem possui 3 axiomas, que expressam a relação de “não igual a” ou “maior que” ($>$):

- Axioma de assimetria: se $a > b$, então $b \neq a$. A ordem dos termos não pode ser invertida;
- Axioma de transitividade: se $a > b$ e $b > c$, então $a > c$;
- Axioma de conectividade: ou $a > b$ ou $b > a$;
- Axioma ordem-denso: números racionais são tais que entre dois números inteiros quaisquer há sempre um número racional; o intervalo entre dois inteiros não é vazio.

Na demonstração empírica dos axiomas de ordem, Guilford (1954) *apud* Pasquali (2011) afirma que a ordem dada pelos números atribuídos aos objetos (transitividade e conectividade) deve ser a mesma obtida pela ordenação empírica destes mesmos objetos e as inversões que ocorrem são consideradas “erros de medida” ou de observação, que devem ser

tratados dentro da teoria da consistência, a qual visa mostrar que, apesar desses erros, há consistência na medida.

A propriedade da **aditividade** define que os números podem ser somados, isto é, podem ser concatenados de modo que a soma de dois números, excetuando o zero, produz um outro número diferente deles próprios, permitindo assim a aplicação das quatro operações de soma, subtração, multiplicação e divisão (sendo estas três últimas redutíveis à primeira). Esta propriedade possui 2 axiomas importantes:

- Axioma da comutatividade: $a + b = b + a$. A ordem dos termos não altera o resultado da adição.
- Axioma da associatividade: $(a + b) + c$ é igual a $a + (b + c)$. A ordem de associação ou de combinação dos termos não afeta o resultado.

A demonstração empírica dos axiomas de aditividade, segundo Pasquali (2011), parece ser possível somente no caso dos atributos extensivos, como massa, comprimento e duração temporal, como também no caso da probabilidade. Dado que a aditividade se baseia na ideia de concatenação, a combinação de dois objetos ou eventos deve produzir um terceiro objeto ou evento com as mesmas propriedades dos dois, mas em grau maior. Isso fica mais fácil de enxergar em uma medição de comprimento, por exemplo. Se fosse adicionado a um objeto de comprimento “x”, medido em uma unidade de comprimento qualquer, outro objeto de comprimento “y”, a união desses dois objetos formaria um objeto “z”, que teria comprimento maior do que os dois objetos individuais.

Logo, parece claro que a utilização da medida, isto é, a utilização do número para descrever os fenômenos naturais, é legítima e adequada. Entretanto, como afirma Pasquali (2011), nem todas as medidas parecem ser iguais no que se refere ao seu nível de qualidade, visto que isso depende fundamentalmente do grau ou da quantidade de isomorfismo que existe entre as propriedades do número e as propriedades naturais, isto é, até que grau seria legítimo utilizar o número para descrever os fenômenos da ciência. Esta afirmação indica a existência de diferentes níveis de correspondência entre o número e os fenômenos naturais, o que implica em diferentes níveis de medida, que serão discutidos em detalhes no tópico a seguir.

2.6 Níveis da medida

As escalas de mensuração representam uma forma de classificação que procura descrever a natureza da informação contida em objetos e fenômenos e, portanto, em variáveis (FÁVERO *et al.*, 2009).

Como a medida consiste na atribuição de números às propriedades das coisas segundo certas regras, conforme comentando anteriormente, ela deve garantir que as operações empíricas salvem os axiomas dos números (PASQUALI, 2011).

Stevens (1946) afirma que as escalas são possíveis, em primeiro lugar, apenas porque existe um certo isomorfismo entre o que se pode fazer com os aspectos de objetos e as propriedades das séries dos números. Lidando com os aspectos dos objetos, invoca-se operações empíricas para a determinação de identidade (classificação), de ordem e para determinar quando as diferenças e quando as razões entre os aspectos dos objetos são equivalentes (aditividade).

De forma semelhante, Cooper e Schindler (2003) afirmam que cada tipo de dado tem seu próprio conjunto de suposições implícitas sobre como os símbolos numéricos correspondem às observações do mundo real. Assim, o processo de mensuração implica no desenvolvimento de regras de mapeamento que devem ser utilizadas para traduzir a observação de indicadores de propriedade. Conforme os autores, as regras de mapeamento possuem quatro características:

1. **classificação**, em que são utilizados números para agrupar ou designar respostas, não existindo uma ordem;
2. **ordem**, em que os números possuem uma ordenação, sendo que um número pode ser maior do que outro, menor do que outro ou igual a outro;
3. **distância**, em que as diferenças entre os números são ordenadas, sendo que a diferença entre qualquer par de números é maior que, menor que ou igual à diferença entre outro par de números;
4. **origem**, em que a série de números tem uma origem única indicada pelo número zero.

Partindo da conceituação de que mensuração, no seu sentido mais amplo, pode ser definida como a atribuição de números a objetos ou eventos de acordo com regras, e que a existência de diferentes regras para atribuir números leva a diferentes formas de escalas e

diferentes formas de mensuração, Stevens (1946) propôs uma forma de classificação padronizada das escalas de mensuração em torno de quatro possibilidades possíveis. O autor foi quem primeiro cunhou os termos nominal, ordinal, intervalar e razão para descrever a hierarquia de escalas de medidas usadas na psicofísica, e classificou os procedimentos estatísticos que seriam permissíveis de acordo com o tipo de escala. Segundo Stevens:

A medição existe em uma variedade de formas e que as escalas de mensuração recaem em um certo grupo definido de classes. Estas classes são determinadas pelas operações empíricas invocadas no processo de 'medição' e pelas propriedades formais (matemáticas) das escalas. (STEVENS, 1946, p. 677)

De acordo com Velleman e Wilkinson (1993), a taxonomia proposta por Stevens foi posteriormente adotada por diversos autores importantes de livros de estatística e, assim, influenciou o raciocínio estatístico de uma geração. De fato, a classificação de Stevens é notadamente a classificação mais aceita nas ciências sociais (CHURCHILL, 1999; FÁVERO *et al.*, 2009) e do comportamento (FÁVERO *et al.*, 2009). Dada a sua simplicidade, é sem dúvida, uma classificação bastante difundida e utilizada, e sua nomenclatura vem sendo empregada em muitos *softwares* estatísticos (FÁVERO *et al.*, 2009; VELLEMAN; WILKINSON, 1993). Programas de computador desenhados para auxiliar na seleção de métodos de análise de dados estão baseados nas suas recomendações e mesmo alguns programas de fins gerais também utilizam essa taxonomia para estruturar sua relação com o usuário (VELLEMAN; WILKINSON, 1993).

Quadro 1: Relação de alguns autores que apoiam a classificação de Stevens

Área	Autores
Metodologia de pesquisa em administração	Cooper e Schindler (2003)
Desenvolvimento de escalas em Administração	Vieira (2011); Costa (2011)
Análise de dados qualitativos	Pereira (2001)
Análise de dados multivariados	Hair <i>et al.</i> (1998); Fávero <i>et al.</i> (2009); Lattin, Carroll e Green (2011);
Pesquisa de marketing	Churchill (1999); McDaniel e Gates (2003); Malhotra (2001); Mattar (1994)
Teoria da mensuração	Allen e Yen (1979)
Psicometria	Pasquali (1997, 2011), Urbina (2007), Erthal (2009)

Fonte: elaborado pelo autor

Muitos autores de livros na área de administração e psicologia utilizam a taxonomia proposta por Stevens como forma de classificação padrão para a categorização de dados. O Quadro 1 apresenta apenas uma pequena relação desses autores.

Como propõe Stevens (1946), dependendo da escala de mensuração em que uma variável é avaliada, diferentes operações matemáticas são possíveis: “as manipulações estatísticas que podem ser legitimamente aplicadas aos dados empíricos dependem do tipo de escala contra a qual os dados estão ordenados” (STEVENS, 1946, p. 677). O Quadro 2 resume as principais características e estatísticas admissíveis das escalas propostas por Stevens (1946), que serão discutidas a seguir.

Quadro 2: Resumo das principais características das escalas primárias de medição.

Escala	Características básicas	Uso em Marketing	Estatísticas admissíveis
Nominal	Os números identificam e classificam os objetos	Marcas, sexo, raças, cores, tipo de lojas, regiões, uso/não uso, gosta/não gosta	Moda, percentagens, teste qui-quadrado, teste binomial, teste de McNemar, teste de Cochran Q
Ordinal	Os números indicam as posições relativas dos objetos mas não a magnitude das diferenças entre eles	Atitudes, preferências, opiniões, classes sociais, ocupações	Medianas, quartis, decis, percentis, teste de Mann-Whitney, teste U, teste de Kruskal Wallis, correlação de postos, ANOVA de Friedman
Intervalar	É possível comparar as diferenças entre objetos; o ponto zero é arbitrário	Atitudes, opiniões, conscientização, preferências, números-índices	Média, intervalo, amplitude total, amplitude média, desvio médio, variância, desvio padrão, teste z , teste t , ANOVA, correlação momento produto, regressão, análise fatorial
Razão	O ponto zero é fixo; podem ser comparadas as razões dos valores da escala	Idade, preço, número de consumidores, volume de vendas, renda, patrimônio	Todos do item anterior mais a média geométrica, média harmônica e coeficiente de variação.

Fonte: Adaptado de Stevens (1946), Malhotra (2001) e Mattar (1994).

2.6.1 Escala nominal

Segundo Stevens (1946), a escala nominal representa a atribuição de números na sua forma mais irrestrita, em que os números são usados apenas como rótulos para identificar e

classificar objetos. Possuem a propriedade da igualdade, em que qualquer permutação de números entre grupos é possível.

Na verdade, esse tipo de dado não requer nem a atribuição de valores numéricos, mas apenas identificadores exclusivos, como números, letras ou cores. Eles são invariáveis sob qualquer transformação que preserve a relação entre indivíduos e seus identificadores. Assim, é permitido fazer qualquer operação nos seus valores desde que não se combinem ou gerem confusão entre as identidades (VELLEMAN; WILKINSON, 1993).

Quando usados para fins de classificação, os números escalonados nominalmente servem como rótulos para classes ou categorias, sendo que as classes devem ser mutualmente excludentes e coletivamente exaustivas. Os objetos em cada classe são considerados equivalentes em relação à característica representada pelo número da escala nominal, e os números não refletem o grau da característica possuída pelos objetos (MALHOTRA, 2001).

Conforme Stevens (1946), a única operação aritmética possível para as escalas nominais é a contagem de membros, sendo desta forma a escala menos poderosa dos quatro tipos de dados: não sugere qualquer ordem ou relação de distância e não têm origem aritmética, desperdiçando qualquer informação que um elemento da amostra possa compartilhar sobre graus de variação da propriedade que está sendo mensurada. Entretanto, uma vez que as classes contenham alguns indivíduos, é possível determinar a classe mais numerosa (moda), e em algumas circunstâncias, é possível testar, por métodos de contingência, hipóteses relacionadas à distribuição de casos entre as classes. Mattar (1994) cita ainda que os testes binomial, qui-quadrado, McNemar e Cochran Q seriam outros testes que poderiam ser utilizados para analisar dados nominais.

Stevens (1946) discute ainda que há pesquisadores que consideram um absurdo conferir a esse processo de atribuição de numerais a nomes a dignidade implícita ao termo mensuração. A esses, o autor reforça a lembrança de que a escala nominal é uma forma primitiva e o uso de numerais como nomes para classes é um exemplo de fato da atribuição de números de acordo com uma regra: a de não atribuir o mesmo número para diferentes classes ou diferentes números para a mesma classe.

Cooper e Schindler (2003) relatam que na pesquisa empresarial e social, os dados nominais são coletados de forma mais ampla do que em qualquer outro tipo de pesquisa. Na área de pesquisa de marketing, segundo Mattar (1994), os dados nominais são intensivamente utilizados para nomear, identificar e classificar variáveis como marcas, cores, regiões, tipo de loja, informações demográficas, como o sexo, raça, ocupação e atitudes, como gosta/não gosta, uso/não uso, entre outros.

2.6.2 Escala ordinal

De acordo com Stevens (1946), a escala ordinal surge da operação de ordenação, indicando a posição relativa dos objetos. Malhotra (2001, p. 239) define a escala ordinal como uma “escala de graduação em que se atribuem números a objetos para indicar até que ponto possuem determinada característica, e se a possuem em maior ou menor grau do que outro objeto, mas sem indicar a magnitude das diferenças”. Assim, um objeto classificado em primeiro lugar tem a característica medida em maior grau do que outro objeto classificado em segundo lugar, mas não é possível determinar a distância que os separam. Objetos classificados como equivalentes recebem o mesmo posto e é possível utilizar qualquer série de números que preserve as relações ordenadas entre os objetos.

De acordo com Pasquali (2011, p. 36), “[...] uma escala numérica pode ser transformada numa outra equivalente se forem respeitados os elementos de invariância nesta transformação”. Assim, uma escala de maior nível pode utilizar as operações estatísticas de uma escala inferior, mas o fará com perda de informação, dado que as estatísticas próprias de uma escala inferior são menos robustas.

Segundo Stevens (1946), as escalas ordinais podem ser transformadas arbitrariamente, desde que a ordenação primária dos objetos seja preservada e a escala transformada compartilhe as mesmas relações de comparações um-a-um entre os objetos e seus valores. Velleman e Wilkinson (1993) afirmam que qualquer transformação monotônica dos valores são permissíveis para as escalas ordinais, podendo aplicar logaritmos ou tirar a raiz quadrada (desde que os valores não sejam negativos), ou ainda, executar uma transformação linear, adicionando uma constante e multiplicando por outra constante positiva. Segundo os autores, assumindo que s seja uma escala utilizada para atribuir números reais em $\Re$ para os elementos de um grupo, uma transformação f é permissível para uma escala ordinal se:

$$s(i) > s(j) \rightarrow f[s(i)] > f[s(j)] \quad (1)$$

Por conta das suas propriedades, além da operação de contagem permitida para dados em escala nominal, as escalas ordinais possibilitam a utilização de estatísticas baseadas em centis, como os percentis, quartis e a mediana. Além dessas estatísticas, também podem ser utilizadas as estatísticas de correlação por postos, a ANOVA de Friedman (MALHOTRA, 2001) e outras estatísticas resumo para dados ordinais, como as estatísticas não paramétricas.

A utilização das escalas ordinais em psicologia é algo frequente e muitas vezes mal empregada. Conforme Stevens (1946) afirma, a maioria das escalas efetivamente utilizadas por psicólogos são do tipo ordinal, e dadas as suas propriedades, as estatísticas envolvendo médias e desvios padrão não deveriam ser utilizadas, visto que estas estatísticas implicam em um conhecimento além da ordenação relativa dos dados.

Na área de pesquisa de marketing, Mattar (1994) comenta que as escalas ordinais têm sido muito utilizadas para obter o ordenamento das preferências, opiniões, atitudes e percepções dos consumidores. Entretanto, segundo Malhotra (2001), dados relativos a atitudes obtidos de escalas de classificação costumam ser normalmente tratados como dados intervalares em pesquisa de marketing, permitindo assim ao pesquisador a utilização de métodos estatísticos mais robustos, como as estatísticas paramétricas.

2.6.3 Escala intervalar

Stevens (1946) classificou este tipo de escala como uma escala quantitativa no sentido real da palavra. É um tipo de escala em que não apenas o sentido de ordem, mas também a distância numérica entre os objetos pode ser aferida. As distâncias entre os valores da escala são fixas e assim, diferenças nos valores numéricos da escala representam as diferenças na característica que se está sendo avaliada. O autor ainda acrescenta que o ponto zero nas escalas intervalares é uma matéria de convenção ou conveniência, tornando a escala invariável quando uma constante é adicionada.

Uma boa definição das escalas intervalares é dada por Malhotra (2001, p. 240). Segundo o autor, a escala intervalar é aquela “em que se utilizam números para classificar objetos, de tal modo que distâncias numericamente iguais na escala representam distâncias iguais na característica que está sendo medida”. Assim, há um intervalo constante ou igual entre os valores da escala, mas por outro lado, a localização do ponto zero não é fixa e tanto este último quanto as unidades de medida são arbitrários. Como o ponto zero não é fixo, não tem sentido calcular razões de valores da escala.

De forma semelhante, Cooper e Schindler (2003) destacam que os dados intervalares têm o poder dos dados nominais e ordinais, além de incorporar o conceito de equidade de intervalo, sendo possível dizer, por exemplo, que a distância entre 1 e 2 é igual à distância entre 2 e 3 e entre 6 e 7. Entretanto, como a escala intervalar não possui um ponto zero absoluto, sendo este definido de forma arbitrária, ela possui as propriedades de classificação, ordem e distância, mas não a propriedade de origem.

Em relação às possíveis transformações das escalas intervalares, Velleman e Wilkinson (1993) afirmam que as transformações permissíveis este tipo de escala são aquelas que preservam as diferenças relativas entre os objetos. Especificamente, a transformação f é permissível para escalas intervalares se houver uma constante c :

$$s(i) - s(j) \rightarrow c\{f[s(i)] - f[s(j)]\} \quad (2)$$

Por conta disso, transformações lineares em que a mesma constante é adicionada para cada valor e/ou cada valor é multiplicado por uma constante são permitidos para as escalas intervalares, mas não se pode, por exemplo, utilizar transformações com base em logaritmos. Segundo Velleman e Wikinson (1993), este tipo de escala permite um menor número de transformações do que as escalas ordinais, sugerindo de certa forma que os valores de dados intervalares carregam consigo mais informação do que esta última.

As técnicas estatísticas que podem ser utilizadas para dados em escala intervalar incluem todas as que podem ser aplicadas a dados nominais e ordinais, além da média aritmética, o desvio padrão, a correlação momento produto, entre outras estatísticas comumente utilizadas. Outras estatísticas, como os testes t , testes F e outros testes paramétricos também podem ser utilizados para este tipo de escala (MALHOTRA, 2001; COOPER; SCHINDLER, 2003). Técnicas estatísticas como a ANOVA, regressão e análise fatorial são igualmente permissíveis (MALHOTRA, 2001).

Na área de psicologia, Stevens (1946) afirma que a maioria das medidas psicológicas aspiram ser escalas intervalares, e em algumas vezes até conseguem, mas o problema em questão normalmente está em criar operações para equalizar as unidades da escala – um problema que não é de fácil solução e que ainda abre diversas brechas para discussões. Outro ponto importante que o autor discute é que ocasionalmente existe a preocupação na localização de um ponto zero verdadeiro, porque os atributos humanos medidos por psicólogos normalmente existem em graus positivos que são extensos quando comparados com a abrangência de sua variação. Como exemplo, o autor cita a inteligência, que é melhor avaliada em escalas ordinais que procuram aproximar as escalas intervalares e que não precisam necessariamente definir o que significaria o nível zero de inteligência.

Em pesquisa de marketing, segundo Mattar (1994), as escalas intervalares são muito utilizadas para medir atitudes, opiniões conscientização e preferências, sendo também frequentemente utilizadas na construção de números-índices, principalmente na economia.

2.6.4 Escala razão

De acordo com Stevens (1946), as escalas do tipo razão são escalas frequentemente encontradas na física e ocorrem quando as operações de equivalência, ordenação, equivalência de intervalos e equivalência de razões podem ser determinadas. Cooper e Schindler (2003) acrescentam que a esse tipo de escala possui uma origem absoluta, representando quantidades reais de uma variável.

Como as escalas do tipo razão preservam as razões relativas, Velleman e Wilkinson (1993) afirmam que as transformações permissíveis para esta escala são aquelas que satisfazem a seguinte relação:

$$s(i) / s(j) \rightarrow cf[s(i)] / f[s(j)] \quad (3)$$

Os dados do tipo razão são considerados os mais poderosos entre os quatro tipos de dados existentes. A grande vantagem deste tipo de escala em relação às demais é que a escala permite as operações de multiplicação e divisão, visto que possui uma origem absoluta. De fato, todas as estatísticas mencionadas até este ponto podem ser utilizadas com escalas do tipo razão. A moda, mediana e média aritmética além das médias geométricas e harmônicas são as estatísticas de tendência central que podem ser calculadas, enquanto que o desvio padrão e o coeficiente de variação são as estatísticas de dispersão permitidas. Testes mais poderosos como os testes *t*, testes F e os testes paramétricos são também permissíveis (COOPER e SCHINDLER, 2003).

De acordo com Stevens (1946), escalas de razão em psicologia são raras, mas não inteiramente desconhecidas. O autor cita a escala Sone como exemplo de uma tentativa deliberada em que observadores humanos julgam razões de sonoridade de pares de tones. Como em todo processo de mensuração, existem limites impostos pelos erros e a variabilidade, mas dentro desses limites, a escala Sone deve corretamente ser classificada como uma escala de proporção. Em marketing, Malhotra (2001) cita o valor das vendas, dos custos, da fatia de mercado e número de clientes como escalas do tipo razão. Idade, preços, número de consumidores, número de lojas, quantidade de produtos consumidos, número de vezes que o produto é comprado ao mês, renda dos consumidores e lucros são outros exemplos citados por Mattar (1994) como escalas deste tipo encontradas em pesquisa de marketing.

2.7 Críticas à proposta de classificação de Stevens

A taxonomia proposta por Stevens procura enquadrar os dados dentro de classes pré-definidas que possuem propriedades matemáticas específicas. O autor classifica não apenas as operações simples que poderiam ser realizadas para cada escala, mas também procedimentos estatísticos que seriam “permissíveis”. Uma escala que preserve seu significado dentro de uma classe de transformações deveria estar restrita a certas estatísticas as quais o seu significado não mudaria com as transformações realizadas na escala. Assim, Stevens (1946) sugeriu limitações à aplicação das estatísticas de acordo com o tipo de escala dos dados.

De fato, as ideias propostas por Stevens foram rapidamente adotadas por muitos pesquisadores e autores de livros. Segundo Velleman e Wilkinson (1993), um dos motivos para isso se deve, talvez, pelo fato que esta classificação fornece uma orientação simples aos pesquisadores e de certa forma, por limitar as estatísticas que podem ser utilizadas para cada tipo de dado, fornece um guia para analistas de dados inexperientes, protegendo-os dos erros na aplicação das estatísticas.

Entretanto, de acordo com Fávero *et al.* (2009), por mais que esta classificação possa parecer simples e seja bastante utilizada e difundida, ela já recebeu inúmeras críticas de diversos autores que argumentam, entre outros pontos, que “a escolha de determinado teste estatístico a ser aplicado a um vetor de dados não depende da escala de mensuração” (FÁVERO *et. al.*, 2009, p. 20).

De uma forma geral, críticas ao trabalho de Stevens tem focado basicamente em 3 pontos fundamentais. Primeiramente, Stevens afirma que a escolha dos métodos estatísticos depende invariavelmente do tipo de escala que o pesquisador possui, o que é contestado por muitos pesquisadores (LORD, 1953; TUKEY, 1961; GUTTMAN, 1977; VELLEMAN; WILKINSON, 1993).

De uma forma cômica e, até certo ponto irônica, Lord (1953) atacou os argumentos de Stevens mostrando que a escolha dos testes estatísticos permissíveis para um certo grupo de dados não depende dos problemas de representação ou unicidade, mas ao invés disso, está relacionado ao significado, que depende da questão que é desenhada para responder. Em seu texto, o autor cria uma história em que seu personagem - um professor que possui a concessão de números nas camisas dos times de futebol da universidade, precisa analisar uma questão levantada por um dos times da universidade, de que seus números são muito baixos. Neste caso, os números das camisas são utilizados apenas para identificar cada jogador, sendo então considerados como dados em uma escala nominal. Entretanto, para analisar a questão

levantada, o professor recorre a um estatístico que realiza uma série de operações nos dados, inclusive calculando a média dos números, para obter uma conclusão ao problema em questão. Questionado pelo professor quanto à legitimidade das operações realizadas, o estatístico profere a célebre frase de que “os números não sabem de onde eles vieram”. Em síntese, o autor utiliza esse exemplo para mostrar que, em outras circunstâncias, esses mesmos números podem assumir uma escala do tipo razão em que médias e outras operações podem ser utilizadas.

Guttman (1977) argumenta de uma forma mais geral que a interpretação estatística dos dados depende da questão que é perguntada aos dados e do tipo de evidência que pode ser aceita sobre essa questão. Nesta mesma linha, Tukey (1961) *apud* Velleman e Wilkinson (1993) também critica as restrições ao uso de estatísticas impostas por Stevens notando a importância do significado dos dados para então se determinar a escala e a análise apropriada que deveria ser utilizada.

Para Velleman e Wilkinson (1993), o uso das categorias propostas por Stevens para selecionar ou recomendar métodos de análises estatísticas é inapropriado e podem estar frequentemente errados, sendo uma prática perigosa para a análise de dados. As categorias não descrevem os atributos dos dados reais que são essenciais para uma boa análise estatística, como também não fornecem um esquema de classificação apropriado para os métodos de análise modernos.

De fato, essa é uma discussão frequente e nem sempre cordial entre os pesquisadores. Citam-se apenas como alguns exemplos, os acalorados embates entre Labovitz (1967, 1968), Champion (1968) e Morris (1968), e entre Gaito (1980) e Townsend e Ashby (1984).

Entretanto, muitas das diferenças de ponto de vista entre pesquisadores da área podem estar em parte relacionadas às diferenças fundamentais entre matemática e ciência. Pasquali (2011) discutiu as diferenças entre ciência e matemática, e se concentrou em mostrar as nuances entre os dois, tanto do ponto de vista dos métodos quanto do objeto de estudo. Segundo o autor, a matemática e a ciência empírica são sistemas teóricos (ou de conhecimento) muito distintos, possuem objetos e metodologias próprias, distintas e irreversíveis entre si. No que se refere ao objeto de estudo, a ciência tem como referente os fenômenos da realidade, enquanto que a matemática tem como referente o símbolo numérico, que de acordo com Frege (1884) *apud* Pasquali (2011), é “um conceito e não uma realidade empírica e nem uma propriedade desta realidade”. No que tange à metodologia, o método das ciências empíricas é a observação sistemáticas dos fenômenos, ao passo que o da matemática é a dedução. Por esta razão, o critério da verdade para a ciência é o teste empírico, que não

está livre de erros, ao passo que para a matemática é a consistência interna do argumento, que está baseado na lógica. Feitas estas considerações, no contexto da teoria da medida e em termos das estruturas epistemológicas, é possível dizer que o sistema científico do conhecimento não tem nada a ver com a matemática e vice-versa. O Quadro 3 apresenta um resumo do enfoque epistemológico em ciências e matemática.

Quadro 3: Enfoque epistemológico em ciências e matemática

	Ciência (empírica)	Matemática
Objeto	Fenômenos naturais	Símbolo numérico
Atitude	Empírica	Transcendental
Metodologia	Observação e controle	Dedução
Verdade	Fato	Teorema
Certeza	Relativa	Absoluta
Crítério de verdade	Teste empírico	Consistência interna do argumento

Fonte: Pasquali (2011)

Tukey (1962 *apud* VELLEMAN e WILKINSON, 1993) argumenta que existem diversas visões do que se constitui uma ciência, mas três componentes são considerados essenciais pela maioria dos pesquisadores, a saber: (i) conteúdo intelectual; (ii) organização em uma forma compreensível; e (iii) confiança no teste de experiência como o padrão final de validade. Com base nestes componentes, Tukey diz que a matemática não é uma ciência, visto que seu padrão final de validade decorre de um tipo estipulado de consistência lógica e provação.

Nessa linha de raciocínio, Velleman e Wilkinson (1993) argumentam que a teoria de mensuração axiomática é de fato matemática, em vez de ciência, pois a condenação de certos métodos estatísticos falha no terceiro teste de Tukey. Segundo os autores, a experiência tem mostrado em uma ampla gama de situações que a aplicação de estatísticas reprovadas a certos tipos de dados podem produzir resultados que são cientificamente significativos, úteis na tomada de decisões e valiosos como base para futuras pesquisas. Esta conclusão também é corroborada por Tukey (1961, p. 245-246) *apud* Velleman e Wilkinson (1993):

Resultados baseados em fundamentos aproximados devem ser utilizados com a subjacente aproximação em mente. Aqueles que buscam a certeza, em vez da verdade, irão tentar evitar este fato. Mas qual conhecimento não está baseado em alguma aproximação? E qual progresso tem sido feito, exceto com o uso deste conhecimento?

Até mesmo Stevens (1946, p. 679) admitiu que bons resultados tinham sido obtidos por psicólogos com o uso de estatísticas ilegais em dados de escalas ordinais. Segundo ele:

De fato, a maioria das escalas utilizadas amplamente e efetivamente pelos psicólogos são escalas ordinais. Na sua propriedade mais estrita as estatísticas comuns que envolvem médias e desvios-padrão não devem ser usadas com essas escalas, visto que estas estatísticas implicam um conhecimento de algo mais do que a relativa classificação de ordem dos dados. Por outro lado, para esta operação estatística 'ilegal' pode ser invocada uma espécie de sanção pragmática: em numerosos casos ela conduz a bons resultados.

Um segundo ponto de discórdia se refere propriamente à taxonomia proposta por Stevens, que é considerada por alguns pesquisadores como sendo muito restrita para ser aplicada para os dados do mundo real, não extinguindo todas as possibilidades de classificação. Neste sentido, diversos autores sugeriram taxonomias alternativas para o enquadramento de dados, sem, entretanto, fazer associação ou recomendar o uso de métodos estatísticos para cada tipo de dados.

Mosteller e Tukey (1977) questionaram, por exemplo, onde deveria ser classificada as frações, como as percentagens, que possuem limites dos dois lados (0% e 100%) e que não podem tolerar aumentos arbitrários na escala. Por esta razão, os autores apresentaram uma proposta de classificação de dados em 7 tipos diferentes de escalas de mensuração:

- Nomes;
- Notas (ordenadas por rótulos, tais como júnior, pleno e sênior);
- Ranking (iniciando por 1, que representa o maior ou o menor);
- Frações (entre 0 e 1, incluindo percentuais);
- Dados de contagem (números inteiros não negativos);
- Montantes (números reais não negativos);
- Balanços (números negativos e positivos);

Mosteller e Tukey (1977) utilizaram estas categorias de escalas para sugerir formas de transformação de dados, incluindo transformações que alterariam valores de um tipo para outro. Entretanto, em nenhum momento os autores sugeriram que essas categorias de dados poderiam restringir a análise (fazendo associações das categoriais com testes estatísticos) ou mesmo a transformação de dados.

Por fim, um terceiro ponto de crítica se refere à degradação dos dados que ocorre com as restrições impostas por Stevens quanto ao uso das escalas. Segundo seus críticos, essas proibições frequentemente levam a uma degradação e rebaixamento dos dados para uma escala de ordenação, e por conta disso, acabam por abdicar o uso dos testes paramétricos e sua eficiência (BAKER; HARDYCK; PETRINOVICH, 1966; BORGATTA; BOHRNSTEDT, 1980).

Para Tukey (1961), os tipos de escalas de Stevens são absolutos e os dados que não são inteiramente intervalares devem ser definidos como dados da escala ordinal. Seria então um mau uso da estatística pensar que os métodos estatísticos são similarmente absolutos.

Baseado no Teorema do Limite Central e em simulações de Monte Carlo, Borgatta e Bohrnstedt (1980) mostraram que para dados típicos, não importa se as escalas são ordinais ou intervalares. Conclusão similar chegou Norman (2010) quanto às escalas do tipo Likert. Apoiando-se nas pesquisas realizadas nos anos 30, o autor afirma que os testes paramétricos são suficientemente robustos para lidar com as violações de seus pressupostos e que desta forma, poderiam ser utilizadas em escalas do tipo Likert. Entretanto, esta discussão envolvendo a utilização de estatísticas paramétricas em dados ordinais, como Norman (2010) e Knapp (1990) e muitos outros autores reconhecem, está longe de terminar e até hoje ainda é motivo de discussão entre os pesquisadores na área de estatística.

3 REFERENCIAL TEÓRICO – ERRO DE MENSURAÇÃO

Os instrumentos usados para descrever quantitativamente indivíduos em termos de várias características psicológicas nunca medem estas características com precisão perfeita. O erro está presente em qualquer medida e sua presença constitui uma ameaça séria à tomada de decisões. Por esta razão, é interessante abordar algumas questões pertinentes relacionadas especificamente aos erros de mensuração, suas principais fontes e os vieses de resposta, que representam um problema comum em escalas multi-itens. Essa discussão será tratada nos tópicos 1 e 2 deste capítulo.

3.1 Erro de medida

Conforme discutido no capítulo anterior, a medição é um número que reflete uma característica ou uma faceta de um objeto, não sendo exatamente o verdadeiro valor da característica de interesse, mas apenas uma observação daquele valor.

Para Pasquali (2011), a medida é um procedimento empírico e por esta razão, não está isenta da ocorrência de erro. Mesmo nas medidas fundamentais, o erro existe. Se for considerado que o processo de medição consiste na determinação da coincidência de pontos, isto é, um sinal no objeto a ser medido e um sinal no instrumento de medida, a medida sempre acompanha algum nível de erro, visto que não há uma coincidência exata de que os dois pontos se fundem num ponto único. O que ocorre é apenas uma justaposição dos dois pontos, sendo que a coincidência se faz dentro de um intervalo: o ponto do corpo medido cai dentro de um intervalo de pontos no instrumento (chamado pelo autor de “extremos de condensação”). Quanto menor este intervalo, maior a precisão da medida.

Desta discussão, é possível concluir que o número utilizado na medida dos fenômenos naturais não é exatamente o número que os matemáticos estudam, embora ele mantenha importantes características em comum, tais como ordem e até aditividade. Pasquali (2011) pontua que o número estudado pelos matemáticos é um conceito absolutamente claro e distinto, enquanto que o número utilizado na medida já não é mais um ponto, mas na verdade, um intervalo. Isto significa que ele pode ser mais ou menos este ponto, admitindo assim variabilidade, o que é, em outras palavras, uma maneira mais elegante de dizer que ele admite erro.

Diferentes fatores podem causar erros de mensuração, que têm como resultado o fato de que a medição, ou escore observado, ser diferente do verdadeiro escore da característica

que está sendo medida. Para Cooper e Schindler (2003), as principais fontes de erro que podem contaminar os resultados são as seguintes: (1) o respondente; (2) a situação; (3) o entrevistador; (4) o instrumento de coleta de dados.

Em relação ao **respondente**, características relativamente estáveis do respondente, como status, classe social, inteligência, conveniência social e educação, entre outras, podem influenciar o escore do teste. Um outro problema que pode ocorrer é quando o respondente sente-se relutante em expressar fortes sentimentos negativos ou possui pouco conhecimento sobre o assunto pesquisado, recusando-se a admitir sua ignorância. Neste caso, essa relutância pode levar ao que Cooper e Schindler (2003) chamam de entrevista de “suposições”, o que notadamente afeta a confiabilidade dos dados. Finalmente, fatores pessoais de curto prazo ou transitórios, como saúde, emoções, fadiga, fome, impaciência, aborrecimento, ansiedade, variações gerais de humor e outras distrações, podem ter um impacto negativo na mensuração.

Os **fatores situacionais** como, por exemplo, a presença de outras pessoas, ruídos e distrações podem afetar negativamente a medição. A presença de uma outra pessoa no ambiente em que está sendo realizada uma pesquisa pode distorcer as respostas, ao participar, distrair ou simplesmente estar presente. Em certos tipos de pesquisas, caso o anonimato não seja garantido, os respondentes podem ficar relutantes em expressar certas opiniões.

O **entrevistador** é outro foco potencial de erro na medição. Ele pode distorcer as respostas ao mudar a redação, parafrasear ou reordenar as questões. A inflexão de voz e atos conscientes ou inconscientes, como sorriso, inclinação da cabeça, etc., podem encorajar ou desencorajar certas respostas. Estereótipos na aparência e na ação também podem gerar vieses. Codificação incorreta, tabulação descuidada e cálculos estatísticos incorretos podem levar também a erros adicionais.

Finalmente, um **instrumento** de medida mal construído pode causar distorções de duas formas. Primeiro, ele pode ser muito confuso e ambíguo. Por exemplo, o uso de palavras e sintaxes complexas, que vão além da compreensão dos respondentes, podem gerar vieses de respostas. Outros problemas incluem as questões indutivas, significados ambíguos, defeitos mecânicos que não permitem colocar a resposta de forma adequada, a omissão de opções de resposta e a impressão de má qualidade. Uma segunda deficiência é a má seleção do universo de conteúdo, quando o instrumento não explora todas as questões potencialmente importantes. Omissões de variáveis importantes numa pesquisa representam um sério problema de validade. Mesmo se os assuntos gerais são tratados na pesquisa, as questões podem não cobrir aspectos suficientes de cada área de preocupação.

Além dessas fontes de erros, outros autores ainda citam os **erros de amostragem** como outra fonte potencial de erro de medida. Para Pasquali (2011), como a pesquisa empírica normalmente não pode ser feita sobre todos os membros de uma população de eventos ou objetos, normalmente é selecionada uma amostra destes eventos ou objetos. Acontece que a escolha de indivíduos no meio de uma população é sujeita a desvios, vieses, isto é, erros, e se o interesse do pesquisador é tirar conclusões ou fazer inferências sobre toda a população da qual a amostra foi retirada, a presença de vieses na amostra com respeito a esta população (falta de representatividade) podem ser desastrosos, levando a inferências errôneas. A utilização de procedimentos de seleção de amostra segundo a teoria de amostragem estatística é a forma de minimizar este tipo de erro.

Entretanto, mesmo quando o pesquisador utiliza procedimentos de amostragem científicas, o **erro amostral** ainda existe, porém, pode ser calculado e controlado *a priori*. Trata-se do erro da amostragem aleatória, que segundo Costa (2011, p. 75), “está ancorado no conceito de distribuição amostral de uma estatística sob análise, e decorre do nível de desvio definido ou calculado em relação à esta estatística, juntamente com sua significância na distribuição de probabilidades que rege a distribuição da estatística”. O pesquisador pode definir o nível de erro que estaria disposto a assumir com base na determinação do tamanho da amostra, ou vice versa.

3.2 Vieses de resposta em escalas multi-itens

Segundo Pasquali (2011), independentemente da qualidade dos itens, a resposta aos mesmos pode ser desvirtuada por fatores relativos ao sujeito que a eles reage. Esses vieses na resposta falseiam os dados, introduzindo correlações espúrias, mesmo em se tratando de bons instrumentos. Segundo o autor, esses erros podem ser classificados em três categorias em termos de suas causas, conforme mostra a Figura 3: (i) cultura/nível socioeconômico; (ii) resposta aleatória; (iii) resposta estereotipada.

Para Pasquali (2011), a **cultura** como causa de erros de resposta está relacionada ao problema de transferência de instrumentos psicológicos para outras populações para as quais eles não foram especificamente construídos e validados. Neste contexto, incluem-se os casos da utilização destes instrumentos para minorias e a adaptação dos mesmos a outras culturas (tradução de testes). Este problema ocorre quando os itens avaliados não possuem o mesmo significado, a mesma interpretação ou não representam uma característica do universo daquela cultura ou minoria.

Figura 3: Tipos de vieses de resposta.

Fonte: Pasquali (2011)

Este problema pode ser observado, sobretudo, na dificuldade relativa de resposta a certos itens para grupos de indivíduos com tradições culturais e de experiência diferentes dos grupos para os quais os testes foram elaborados. Um exemplo desse tipo de problema pode acontecer quando um teste é aplicado a duas amostras diferentes de sujeitos e os escores médios forem diferentes. Essa diferença de média pode ser devida ao fato de que realmente as duas amostras são diferentes na magnitude do traço que o teste mede, ou ainda, pode ter sido provocada por artefatos do teste, tais como o tipo ou formato dos itens, que expressam modos de falar e representar as coisas de forma diferente nas duas amostras. Pasquali (2011) apresenta um bom exemplo desse tipo de problema:

[...] falar de avião e carro para um pigmeu das florestas africanas para se referir a voar e correr provavelmente não diz o que significa para um ocidental, o qual seria mais adequado falar de aves e outras criaturas ou objetos familiares ao pigmeu. Aplicar a um pigmeu testes do tipo mencionado faz com que eles tenham escores inferiores, simplesmente porque a representação comportamental do traço latente que o teste pretende medir não é adequada.

A **resposta ao acaso** é outro viés de resposta comum em escalas de múltipla escolha. De acordo com Pasquali (2011), os fatores que determinam a resposta ao acaso não são determináveis, pois são, por definição, aleatórios e não sistemáticos. Entre os fatores que podem levar à resposta ao acaso estão a má disposição do sujeito em responder ao teste, a incompreensão das instruções, a gozação, a sua falta de atitude de colaboração entre outros. Dentro da Psicometria Clássica e no que se refere aos testes de habilidade, em que o teste possui uma resposta certa ou errada, este tipo de problema é tratado através de uma fórmula

estatística, a qual supõe que os sujeitos que conhecem a resposta certa assim o respondem, enquanto que os que a desconhecem, ou erram ou acertam por acaso. Além disso, supõe que todas as alternativas de resposta sejam equiprováveis, isto é, que têm a mesma chance de serem selecionadas, suposição esta que é evidentemente difícil de ser aceita.

A **resposta estereotipada**, isto é, respostas tendenciosas devido a peculiaridades do sujeito que responde, ocorre sobretudo em testes de personalidade e de atitude. Pasquali (1997) destaca duas causas possíveis da ocorrência desse tipo de erro: a “desejabilidade social”, representando um traço de personalidade que afeta negativamente a objetividade nas respostas de autorrelato e a resposta sistemática, que ao contrário da anterior, é constituída por erros de julgamento.

Edwards (1957, p. vi *apud* Pasquali, 2011) define a **desejabilidade social** como “a tendência dos sujeitos em atribuir a si mesmos, em caso de autodescrição, afirmações de personalidade com valores socialmente desejáveis e em rejeitar aquelas com valores socialmente indesejáveis”. Segundo Pasquali (2011), o sujeito não procura intencionalmente mentir sobre si mesmo, mas o faz sem dar-se conta disso. Assim, a desejabilidade social é considerada uma atitude inconsciente de se apresentar bem diante dos outros e está, sobretudo, quase sempre presente em inventários de personalidade.

A maneira de controlar esta tendência, segundo Pasquali (2011), tem sido a elaboração de uma escala de desejabilidade e incluí-la no inventário, sendo um traço a mais que deve ser mensurado. De acordo com o autor, mesmo adicionando esta escala, fica difícil saber o que fazer com tal dado que alerta sobre o fato de que o sujeito pode ter utilizado a mesma tática de resposta aos demais traços medidos pelo inventário. Entretanto, fica ao menos o alerta da interpretação dos resultados do inventário principal que está sendo medido, quando da presença de índices elevados de desejabilidade social.

A resposta sistemática representa falhas de julgamento, chamados correntemente de erros de resposta. Segundo Pasquali (2011), de uma forma geral, o controle dos erros de resposta sistemática tem se mostrado ainda bastante falho na utilização de escalas de avaliação. Segundo o autor, são exemplos do erro de resposta o efeito halo, o erro de leniência, a tendência central, o erro de contraste, entre outras.

Segundo Pasquali (2011), o termo “**efeito halo**” foi inicialmente cunhado por Thorndike e ocorre quando “um avaliador tende a avaliar um indivíduo de modo semelhante sobre todas as dimensões” (GUILFORD, 1959, p. 146 *apud* PASQUALI, 2011, p. 155). Esse erro é inversamente proporcional à variância nas respostas, acarretando altas correlações entre diferentes fatores e reduzidos desvios padrões.

Em particular, Wirtz (2001) afirma que a existência potencial de efeitos halo é um aspecto preocupante e perturbador de estudos que empregam modelos multi-atributos. Diversas pesquisas em diferentes áreas do conhecimento tem apontado as mesmas observações (MURPHY; JAKO; ANHALT, 1993). De acordo com o autor, os efeitos halo são distorções das percepções do consumidor a respeito das propriedades específicas de atributos. Em termos técnicos, o efeito halo é o excesso de correlação, tanto acima quanto abaixo, da verdadeira correlação entre os atributos.

Wirtz (2001) descreve duas formas de efeito halo que são discutidas na literatura de marketing. Primeiramente, a resposta para um atributo em particular pode ser influenciada pela impressão geral do objeto como um todo ou seus traços afetivos. Como exemplo, o gosto (desgosto) por uma marca pode ter uma influência positiva (negativa) na avaliação de todos os outros atributos deste produto. Em segundo lugar, a resposta para outros atributos pode ser influenciada pela avaliação de um atributo dominante. Exemplificando, a avaliação da qualidade do café servido a bordo por uma companhia aérea pode estar influenciada pela avaliação do serviço de bordo como um todo.

Os efeitos halo em geral assimilam a avaliação de diferentes atributos, nivelam o perfil global das avaliações e reduzem as diferenças entre avaliações da performance de diferentes atributos (MURPHY; JAKO; ANHALT, 1993). De acordo com Wirtz (2001), estes efeitos têm sido amplamente estudados no contexto da psicologia social e na administração de pessoal, e maneiras potenciais de controlar ou reduzir seus efeitos tem sido propostas e testadas, tais como as estratégias listadas abaixo:

- Usar respondentes que são capazes (experiência e conhecimento) para avaliar o produto;
- Aumentar o desejo do respondente em avaliar os diversos atributos, enfatizando um propósito de pesquisa de desenvolvimento (por exemplo, melhorar o produto) do que apenas para fins de avaliação (por exemplo, pesquisa acadêmica);
- Aumentar o desejo do respondente em avaliar os diversos atributos, evitando respondentes com baixo nível de envolvimento (controle do envolvimento medindo-o separadamente);
- Controlar traços afetivos ou reações afetivas em relação ao produto (necessidade de medição do afeto);
- Tornar aleatório a ordem dos atributos dentro do mesmo estudo.

O **erro de leniência** consiste em dizer “apenas coisas boas a respeito de todo mundo” (DUNNETTE, 1983 *apud* PASQUALI, 2011). Estatisticamente, essa tendência é definida como “uma mudança significativa na média das avaliações na direção favorável, de uma condição de avaliação para outra” (SHARON e BARTLETT, 1969 *apud* PASQUALI, 1997).

A **tendência central** ocorre quando um avaliador tende a colocar todos os sujeitos no centro da escala, evitando proferir julgamentos extremos, ocorrendo normalmente, quando os avaliadores não conhecem suficientemente bem os avaliados ou aquilo que se está avaliando. Um caso interessante de tendência central é citado por Churchill (1999). Quando duas escalas semelhantes, uma de quatro pontos sem ponto neutro e outra de cinco pontos contendo um ponto central, foram comparadas para medir a intenção de compra de certos produtos dentro de um período de tempo pré-determinado, os resultados indicaram haver algumas semelhanças e diferenças. Comparativamente, a frequência de respostas nos valores extremos das escalas foram basicamente os mesmos entre as escalas de quatro e cinco pontos, mas a frequência de ocorrência do valor intermediário negativo da escala de quatro pontos (sem ponto central) foi proporcionalmente muito superior ao valor intermediário positivo. Nesse contexto particular, a presença de um ponto central fez com que muitas pessoas que talvez tivessem uma intenção negativa de compra assumissem uma posição neutra. O Gráfico 1 apresenta um exemplo de comparação do uso de uma escala de quatro pontos e cinco pontos para mensurar a intenção de compra de um bilhete de cinema dentro de um prazo de 7 dias.

Gráfico 1: Comparação do uso de escala de quatro pontos versus cinco pontos.

Fonte: Churchill (1999)

O **erro de contraste** consiste na tendência de as pessoas avaliarem os outros de modo oposto do que se avaliam a si mesmas. Por exemplo, a pessoa que é organizada acha que as outras pessoas são desleixadas.

Por fim, Pasquali (1997) ressalta que os pesquisadores têm procurado maneiras de contornar os erros sistemáticos, como por exemplo, eliminar o ponto neutro (central) da escala para inviabilizar a tendência central ou eliminar a parte inferior da escala para descaracterizar a leniência. Contudo, tais investidas não têm surtido efeitos suficientes e claros e, assim, essas tendências ainda continuam sendo um problema substancial na medida da personalidade e das atitudes.

Outro erro de resposta sistemática foi verificado por Oliver (1996), chamado de *ceiling effect* (“efeito teto”), que ocorre quando respondentes são instados a atribuir notas a atributos segundo sua importância em escalas não comparativas. Em estudo realizado por Ghisi, Merlo e Nagano (2004), os autores verificaram a presença deste tipo de erro na utilização de uma escala modificada de diferencial semântico, do tipo *rating*. Conforme pode ser visualizado no Gráfico 2 (gráfico Boxplot), o efeito teto diagnosticado neste caso provocou alta concentração e assimetria de notas no extremo superior da escala, e baixa variância para a maioria dos nove atributos analisados, fazendo com que as reais diferenças de importância entre os atributos fosse, de certa forma, mascarada.

Gráfico 2: Exemplo do efeito *ceiling*.

Fonte: Ghisi, Merlo e Nagano (2004)

4 REFERENCIAL TEÓRICO – CONFIABILIDADE DE ESCALAS

Considerando que nenhuma medida é livre de erro, uma boa ferramenta de mensuração é aquela que possui um elevado grau de acuidade e precisão do procedimento de mensuração, e que mede efetivamente aquilo que foi projetada para medir. O primeiro conceito se refere à característica da confiabilidade da medida e o segundo, à sua validade (COOPER; SCHINDLER, 2003). O conceito de confiabilidade é discutido aqui neste capítulo, enquanto que o conceito da validade é abordado no capítulo 5.

Para Cooper e Schindler (2003), a confiabilidade possui diferentes significados para diferentes pessoas, mas na maioria dos contextos, está relacionada à noção de consistência, isto é, o quanto uma medida fornece resultados consistentes. Para Malhotra (2001), a confiabilidade representa o grau em que uma escala produz resultados consistentes se as medidas são tomadas repetidamente.

Segundo Malhotra (2001, p. 263-264), “as fontes sistemáticas de erro não têm impacto desfavorável sobre a confiabilidade, porque afetam a medida de uma forma constante e não conduzem a inconsistências”. Por outro lado, o erro aleatório causa inconsistências, levando à baixa confiabilidade. Por esta razão, a confiabilidade pode ainda ser definida como o âmbito em que as medições estão livres do erro aleatório ou instável.

Figura 4: Avaliação da confiabilidade de uma escala.

Fonte: adaptado de Cooper e Schindler (2008); Costa (2011).

De forma semelhante, Cooper e Schindler (2003) afirmam que instrumentos confiáveis podem ser usados com a segurança de que fatores transitórios e situacionais não estão interferindo na medida, podendo ser considerados fortes, funcionando bem em diferentes épocas, sob diferentes condições. De acordo com os autores, essa diferença de tempo e condições é a base para as perspectivas comumente utilizadas em confiabilidade –

estabilidade, equivalência e consistência interna. Com base nestas perspectivas, as diversas técnicas de verificação da confiabilidade podem ser agrupadas, conforme mostra a Figura 4. O Quadro 4 abaixo apresenta um breve resumo das principais características dessas técnicas, que serão discutidas em maiores detalhes a seguir.

Quadro 4: Principais técnicas de estimação da confiabilidade

Técnica	Procedimento empírico	Coefficiente	Vantagem	Desvantagem
Teste-reteste	Aplicar mesmo teste T aos mesmos sujeitos em duas ocasiões diferentes, O_1 e O_2 , e calcular correlação $r_{O_1O_2}$	Estabilidade ou constância	Garantia da equivalência (paralelismo), pois se trata do mesmo teste T	Difícil definir intervalo ideal de tempo entre O_1 e O_2 Difícil controlar eventos que ocorrem entre O_1 e O_2
Formas paralelas	Aplicar 2 formas paralelas de um teste, T_1 e T_2 , a amostra representativa de sujeitos e calcular a correlação $r_{T_1T_2}$	Equivalência	Bate diretamente com o conceito de fidedignidade Testes aplicados numa só ocasião	Difícil conseguir formas perfeitamente paralelas, isto é, medir mesmo traço latente com itens diferentes
Duas metades	Aplicar um teste T a amostra representativa; dividir T em duas partes equivalentes (itens pares vs. Ímpares; 1ª ou 2ª metades; ou outra divisão); calcular correlação entre as duas metades; aplicar a correção de Spearman-Brown	Consistência Interna do teste	Exige apenas uma aplicação (evita eventos temporais)	Difícil garantir equivalência das 2 metades Técnica de 1ª e 2ª metades não controla fadiga dos testandos ao tomar 2ª metade e normalmente esta metade contém os itens mais difíceis

Fonte: Pasquali (2011, p. 198)

4.1 Estabilidade – técnica teste-reteste

Para Cooper e Schindler (2003), a perspectiva da estabilidade teste-reteste se refere a capacidade da ferramenta de mensuração apresentar resultados consistentes, com mensurações repetidas da mesma pessoa usando o mesmo instrumento. Segundo Malhotra (2001), os entrevistados recebem conjuntos idênticos de itens da escala, em duas ocasiões diferentes e sob condições tão próximas da equivalência quanto possíveis. O grau de semelhança entre as duas medidas é determinado calculando-se o coeficiente de correlação entre elas, sendo que quanto mais alto o coeficiente de correlação, maior a confiabilidade.

Normalmente, o intervalo de tempo entre testes ou aplicações é, tipicamente, de duas a quatro semanas (MALHOTRA, 2001; COOPER; SCHINDLER, 2003). Entretanto, quando há muito tempo entre as mensurações, há chance de mudança nos fatores situacionais, afetando assim as observações, onde uma mudança nos escores poderia ser incorretamente entendida como uma queda na confiabilidade do processo de mensuração.

Segundo Cooper e Schindler (2003, p. 187), algumas das dificuldades que podem ocorrer na metodologia de teste-reteste e causar um viés descendente na estabilidade incluem:

- Demora entre as mensurações – leva a mudanças nos fatores situacionais;
- Tempo insuficiente entre as mensurações – permite que o respondente lembre as respostas anteriores e repita-as, resultando em indicadores enviesados de confiabilidades;
- Discernimento do respondente de um propósito disfarçado – pode gerar viés se o respondente tiver opiniões em relação ao propósito, mas não associadas com as questões de mensuração atuais;
- Sensibilidade do tópico – ocorre quando o respondente procura saber mais sobre o tópico ou forma opiniões novas e diferentes antes do reteste;
- Introdução de variáveis de moderação estranhas entre as mensurações – pode resultar em mudança na opinião dos respondentes a partir de fatos não relacionados à pesquisa.

Além desses problemas, Malhotra (2001) ainda lembra que pode ser impossível fazer repetidas medições, como por exemplo, quando o tópico da pesquisa é a reação inicial do entrevistado a um novo produto. Outro problema do procedimento teste-reteste, segundo o autor, é que o coeficiente pode ser inflacionado pela correlação de cada item com ele próprio, dado que essas correlações tendem a ser mais altas do que a correlação entre diferentes itens da escala, através de administrações. Como consequência disso, é possível obter altas correlações teste-reteste em razão das elevadas correlações entre os mesmos itens da escala medidos em momentos diferentes, mesmo que as correlações entre diferentes itens da escala sejam extremamente baixas.

4.2 Equivalência – técnica de formas paralelas ou alternativas

Segundo Cooper e Schindler (2003), a equivalência considera quanto erro pode ser introduzido por diferentes investigadores (na observação) ou por diferentes amostras de itens sendo estudados (no questionamento ou na escala) e se refere às variações em um ponto do tempo entre os observadores e as amostras de itens.

Neste caso, o principal interesse normalmente não é saber como os respondentes diferem item por item, mas como um determinado conjunto de itens representa bem as pessoas. A ideia é que, mesmo havendo muitas diferenças nas respostas entre duas amostras de itens, se uma pessoa é classificada da mesma maneira em todos os testes, então os testes têm boa equivalência (COOPER; SCHINDLER, 2003).

A forma usual de testar a equivalência da amostra de itens normalmente ocorre por meio de formas alternativas ou paralelas do mesmo teste administrado às mesmas pessoas simultaneamente. Os resultados dos dois testes são então correlacionados (COOPER; SCHINDLER, 2003).

Um primeiro problema com esta abordagem é a fadiga. Nas circunstâncias em que o teste de equivalência é realizado, a duração do processo de teste pode afetar as respostas das pessoas através da fadiga e a confiabilidade inferida da forma paralela será reduzida proporcionalmente. Para compensar esse problema, alguns pesquisadores recomendam a utilização de um intervalo entre as duas aplicações do teste, em uma abordagem chamada de forma equivalente retardada, podendo ser considerada como uma forma composta de teste-reteste e método de equivalência. Nessas circunstâncias, deveria ser administrado o formulário X seguido do formulário Y para a metade dos pesquisados e o formulário Y seguido do formulário X para a outra metade, a fim de reduzir os efeitos da ordem de apresentação (COOPER; SCHINDLER, 2003).

Cooper e Schindler (2003) citam ainda um segundo problema com este tipo de procedimento, que se refere à dificuldade de criação de um segundo grupo de itens que sejam efetivamente equivalentes. Em muitos casos, o pesquisador pode incluir apenas um número limitado de itens em um instrumento, retirado de um domínio de conteúdo. Para realizar o teste de equivalência, outra amostragem de itens, de número similar, terá que ser retirada do mesmo domínio, o que nem sempre é um procedimento fácil de ser feito. Segundo Cooper e Schindler (2003, p. 188), “mesmo com os procedimentos mais sofisticados, usados por editores de testes padronizados, é raro encontrar questões totalmente equivalentes e intercambiáveis”.

Para Malhotra (2001), as duas formas devem ser equivalentes em relação ao conteúdo, o que, em um sentido estrito, exige que os conjuntos alternativos de itens da escala tenham médias, variâncias e intercorrelações idênticas. Mesmo que essas condições sejam satisfeitas, as duas formas podem não ser equivalentes em conteúdo, em que uma correlação baixa pode refletir uma escala não-confiável como também formas não-equivalentes.

Malhotra (2001) ainda lembra outro problema com os procedimentos de equivalência. Para o autor, este tipo de teste consome muito tempo e a construção de uma forma equivalente da escala é muito dispendiosa.

4.3 Consistência Interna – *split-half*

A terceira abordagem para avaliar a confiabilidade usa a administração de apenas um instrumento ou teste em um único momento para avaliar a consistência interna ou homogeneidade entre os itens (COOPER; SCHINDLER, 2003). Para Malhotra (2001), a confiabilidade da consistência interna serve para avaliar a confiabilidade de uma escala somada (*summated scale*), em que vários itens são somados para formar um escore total. Neste tipo de escala, cada item mede algum aspecto ou faceta do construto medido por toda a escala, e por esta razão, os itens devem ser consistentes na indicação dessa característica.

De acordo com Malhotra (2001), a medida mais simples de consistência interna é a confiabilidade meio-a-meio, também chamado por Costa (2011) e Churchill (1979) de confiabilidade *split-half*, em que os itens da escala são divididos em duas metades e os meio escores resultantes são correlacionados. Correlações altas entre as duas metades indicam alta consistência interna, isto é, há similaridade ou homogeneidade entre os itens. Este procedimento também é chamado de técnica de dividir ao meio (COOPER; SCHINDLER, 2003) e pode ser utilizado quando a mensuração tiver muitas questões ou declarações similares que possam ser respondidas pelas pessoas.

Conforme aponta Costa (2011), o coeficiente de correlação na verificação da confiabilidade *split-half* costuma ser ajustado, em um esforço de melhoramento de sua capacidade de estimação de consistência interna do conjunto de itens, sendo utilizada normalmente a Fórmula 4 mostrada abaixo:

$$C = \frac{2 \cdot r_{s1s2}}{1 + r_{s1s2}} \quad (4)$$

Onde:

C = confiabilidade

r_{S1S2} = correlação entre as metades

As principais formas de divisão da amostra de itens em duas metades são as seguintes (COSTA, 2011; PASQUALI, 2011; COOPER; SCHINDLER, 2003; MALHOTRA, 2001):

- Separar as metades pela ordem de apresentação dos itens – parte inicial e parte final (*first-half last-half*)
- Separar as metades considerando itens ímpares e itens pares (*odd-even*)
- Separar as metades selecionando-se itens de forma aleatória (*random halves*)

Observando-se as diversas possibilidades de divisão do conjunto de itens em duas metades, é fácil perceber que a principal dificuldade com esta abordagem está exatamente na metodologia de divisão das metades. O problema neste caso é que a forma de dividir os itens da escala em duas metades pode influenciar o coeficiente de consistência interna, isto é, os resultados dependem da subjetividade de como os itens da escala foram divididos, conforme aborda Costa (2011, p. 87):

É esperado que as diferentes metades, por diferentes métodos, gerem diferentes medidas de correlação. Assim, é possível que um dado critério de definição de metades gere correlações elevadas, sugerindo uma elevada confiabilidade da escala, e outro critério estabeleça outras metades, do mesmo conjunto de itens, e gere uma correlação menor, e, portanto, sugerindo uma confiabilidade baixa.

Para solucionar este problema, foram propostas outras abordagens para assegurar estimativas de confiabilidade sem necessariamente dividir os itens da escala. Por esta razão, o procedimento *split-half* é muito pouco utilizado nas pesquisas em administração, sendo preterido por procedimentos mais consistentes de análise (COSTA, 2011). Entre essas abordagens, segundo Cooper e Schindler (2003), encontram-se a abordagem Kuder-Richardson 20 (KR20) e o coeficiente *alpha* de Cronbach. O KR20 é um método a partir do qual o *alpha* é generalizado, sendo utilizado para estimar a confiabilidade para itens dicotômicos. O coeficiente *alpha* de Cronbach é considerado o coeficiente mais utilizado para estimar confiabilidade em escalas multi-itens (PETERSON, 1994) e por esta razão, será discutido em maiores detalhes a seguir.

4.4 Consistência Interna - Coeficiente *alpha* de Cronbach

O coeficiente *alpha* de Cronbach foi desenvolvido por Cronbach (1951) como uma medida geral da consistência interna de escalas multi-item, e segundo Malhotra (2001), representa a média de todos os coeficientes meio-a-meio que poderiam resultar das diferentes maneiras de dividir ao meio os itens de uma escala. Conforme Cooper e Schindler (2003), este coeficiente é mais útil para escalas multi-itens no nível de intervalo de mensuração.

Conforme proposto por Cronbach (1951), o coeficiente *alpha* pode ser assim formulado (Fórmula 5):

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \sum_{i=1}^k \sigma_i^2 / \sigma_s^2 \right) \quad (5)$$

Onde:

k = número de itens da escala

σ_i^2 = variância do item i

σ_s^2 = variância da escala

Segundo Peterson (1994), o coeficiente *alpha* de Cronbach é amplamente o coeficiente de confiabilidade mais utilizado pelos pesquisadores e tem se tornado a medida de referência para estimar a confiabilidade de escalas multi-itens. Para Costa (2011), sua rápida adesão é devida à sua consistência e elegância operacional na aferição da confiabilidade.

De fato, o coeficiente *alpha* de Cronbach se tornou uma das bases da teoria da medida, onde seu artigo clássico publicado na revista *Psychometrika* no ano de 1951 já foi citado em mais de 2200 artigos nos últimos 20 anos, segundo o *Social Science Citation Index* (PETERSON, 1994). Outro ponto ainda a ser destacado é que, além do coeficiente *alpha* ser o estimador de confiabilidade mais largamente utilizado, é ainda o objeto de considerável atenção metodológica e analítica (CORTINA, 1993).

Um aspecto importante sobre o coeficiente *alpha* se refere ao nível apropriado de *alpha* que deve ser aceito para que uma escala possa ser considerada confiável. Segundo Peterson (1994), embora esta seja uma questão de grande importância, há pouca orientação na literatura sobre o que poderia ser considerado como um coeficiente aceitável ou suficiente para propósitos de pesquisa. Embora as recomendações de muitos autores difiram em alguns pontos (ver Quadro 5), elas compartilham duas especificidades. A primeira indica que o nível

recomendado de confiabilidade é uma função do tipo de pesquisa empreendida, isto é, se é uma pesquisa exploratória, aplicada e assim por diante. Em segundo lugar, nenhuma das recomendações possui uma base empírica, uma justificativa teórica ou ainda uma lógica analítica por trás, mas sim, parecem estar baseadas na experiência ou intuição desses pesquisadores.

Quadro 5: Níveis recomendados do coeficiente *alpha*

Autor	Situação	Nível Recomendado
Davis (1964, p. 24)	Predição para um indivíduo	Acima de 0.75
	Predição para grupo de 25-50	0.5
	Predição para grupo acima de 50	Abaixo de 0.5
Kaplan e Saccuzzo (1982, p. 106)	Pesquisa básica	0.7 – 0.8
	Pesquisa aplicada	0.95
Murphy e Davidshofer (1988, p. 89)	Nível inaceitável	Abaixo de 0.6
	Nível baixo	0.7
	Nível moderado a alto	0.8 – 0.9
	Nível alto	0.9
Nunnally (1967, p. 226)	Pesquisa preliminar	0.5 – 0.6
	Pesquisa básica	0.8
	Pesquisa aplicada	0.9 – 0.95
Nunnally (1978, pp. 245-246)	Pesquisa preliminar	0.7
	Pesquisa básica	0.8
	Pesquisa aplicada	0.9 – 0.95

Fonte: Peterson (1994, p. 382)

Entre as recomendações apresentadas no Quadro 5, de acordo com Peterson (1994), a proposta de Nunnally (1967, 1978) apresenta-se como a recomendação mais referenciada. Nos últimos 12 anos, segundo o autor, as recomendações de Nunnally já haviam sido citadas mais de 50 vezes no *Journal of Marketing Research* e mais de 36 vezes no *Journal of Consumer Research*.

Pesquisas empíricas tem revelado que os parâmetros mostrados no Quadro 5 parecem ser consistentes, em especial para pesquisas preliminares e elementares. Em estudo realizado por Peterson (1994) que envolveu o levantamento de 4.286 valores de coeficiente *alpha* em diversos periódicos importantes de marketing e psicologia, a média geral dos coeficientes encontrados foi de 0,77, sendo que cerca de 75% dos coeficientes *alpha* observados foram

iguais ou maiores que 0,70. Esses dados corroboram os valores recomendados por Nunnally para pesquisas preliminares e básicas, que foram citadas anteriormente no Quadro 5.

Outro dado relevante encontrado por Peterson (1994) é que apenas 14% dos coeficientes *alpha* observados atingiram ou superaram o valor de 0,90, que é o nível sugerido por especialistas para pesquisas aplicadas, como Nunnally. Em relação a esse achado, Peterson (1994) argumenta que os níveis recomendados na literatura para pesquisas aplicadas são irrealisticamente altas para pesquisas de comportamento do consumidor e marketing. O autor ainda afirma que, como muitos pesquisadores caracterizam suas pesquisas como básicas, a ausência de se atingir um nível de 0,90 possui poucas consequências, e que, de fato, a própria ausência desse nível pode refletir uma boa prática de pesquisa.

4.4.1 Críticas ao coeficiente *alpha* de Cronbach

Por ser o coeficiente mais utilizado pelos pesquisadores como indicação da consistência interna, e de uma forma mais abstrata, da confiabilidade de escalas, o coeficiente *alpha* de Cronbach já recebeu muita atenção (e críticas) de pesquisadores. Muitos estudos foram empreendidos com o intuito de melhor compreender não apenas as potencialidades do coeficiente *alpha*, mas também suas limitações. Entre as principais críticas feitas por pesquisadores ao coeficiente *alpha*, 3 merecem especial destaque: (1) o número de pontos da escala pode influenciar o valor de *alpha*; (2) o aumento do número de itens da escala pode aumentar o valor de *alpha*; e (3) a possibilidade de geração de medidas de *alpha* elevadas somente pela inclusão de variáveis redundantes.

Estudos têm revelado resultados inconsistentes a respeito do efeito do número de pontos da escala sobre o coeficiente *alpha*. Alguns estudos concluíram que o *alpha* de Cronbach não era afetado (ou quase não era afetado) pelo número de categoriais de respostas (AIKEN, 1983; MATELL; JACOBY, 1971; WONG; CHUEN; FUNG, 1993). Outros, por sua vez, revelaram que as estimativas de confiabilidade aumentavam com o acréscimo do número de categorias de resposta comparada à escala de resposta de 2 pontos (GUILFORD, 1954; LISSITZ; GREEN; 1975; NUNNALLY, 1978). Lissitz e Green (1975) realizaram um estudo de simulação e verificaram que o coeficiente *alpha* de Cronbach aumentou quando o número de categorias de respostas aumentou de 2 para 5, mas se manteve constante depois disso. Resultados semelhantes foram também obtidos por Jenkins e Taber (1977) e Bandalos e Enders (1996).

Talvez um dos estudos mais amplos realizados nessa área seja o estudo de Peterson (1994), que procurou verificar, baseado em um estudo de meta-análise, os efeitos de algumas características do desenho de pesquisa sobre o coeficiente *alpha*. Este estudo investigou as publicações de diversos periódicos importantes das áreas de marketing e psicologia entre o período de 1960 a 1992, procurando localizar artigos que publicaram dados empíricos de coeficiente *alpha*. Esta busca resultou em uma coleção de 832 artigos e 4.286 valores de coeficiente *alpha*, que foram então utilizados para testar a influência de 12 variações de desenho de pesquisa. Os resultados desse estudo são mostrados no Quadro 6 em comparação a outro estudo similar, mas de menor abrangência, que foi realizado por Churchill e Peter (1984).

Quadro 6: Relações entre o coeficiente *alpha* e características de desenho de pesquisa

Característica de desenho de pesquisa	Variações analisadas	Relação com o coeficiente <i>Alpha</i>	
		Churchill e Peter (1984)	Peterson (1994)
Tamanho da amostra	<100; 100-199; 200-299; 300 ou mais; não informada	Tamanho da amostra negativamente relacionada ao <i>alpha</i>	Nenhuma relação substancial
Tipo de amostra	Estudantes de faculdade; consumidores; pessoas de negócio; misto; não é possível dizer	Nenhuma relação	Nenhuma relação substancial
Número de categorias (pontos) da escala	Não informada; 2; 3; 4; 5; 6; 7; 8 ou mais	Número de categorias positivamente relacionada à magnitude de <i>alpha</i>	Escala com duas categorias exibiram <i>alphas</i> menores do que aquelas com 3 ou mais categorias
Número de itens da escala	Não informada; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11 ou mais	Relação positiva entre número de itens e tamanho de <i>alpha</i>	Escala com dois e três itens exibiram menores <i>alphas</i> do que aquelas com 4 ou mais itens
Tipo de escala	Likert; Diferencial Semântico	Nenhuma relação	Nenhuma relação substancial
Formato da escala	Apenas as pontas nomeadas; valores numéricos nas categorias internas; valores verbais nas categorias internas; não é possível dizer	Nenhuma relação	Nenhuma relação

Natureza da escala	Número ímpar de categorias da escala; número par de categorias na escala; não é possível dizer	Nenhuma relação	Nenhuma relação substancial
Modo de administração	Auto preenchimento; entrevistador; não informada	Nenhuma relação	Administração por meio de entrevistador produziu menores <i>alphas</i> do que a administração por auto preenchimento
Orientação da escala	Centrado no respondente; centrado no estímulo; ambas	Não foi estudado	Nenhuma relação substancial
Natureza do construto	Dependente; independente; não é possível dizer/ambos	Não foi estudado	Nenhuma relação substancial
Tipo de pesquisa	Desenvolvimento de escala; aplicação de escala; não é possível dizer	Não foi estudado diretamente	Nenhuma relação
Número de itens eliminados durante o desenvolvimento da escala	Número de itens deletados	Não foi estudado	Relação positiva entre o número de itens eliminados e a magnitude de <i>alpha</i>

Fonte: adaptado de Peterson (1994)

Em particular sobre a influência do número de pontos da escala no valor de *alpha*, os resultados da pesquisa de Peterson (1994) indicaram que as escalas com dois pontos apresentaram significativamente menores níveis de coeficiente *alpha* (0,70 de média) do que as escalas com 3 ou mais pontos (médias variando entre 0,76 a 0,78). Além disso, em um cenário com uma escala contendo apenas 2 ou 3 itens, as médias foram ainda menores (0,62) comparadas às escalas com os mesmos números de itens, mas com 3 ou mais pontos (0,74 de média). Este achado corrobora os resultados encontrados por Guilford (1954), Lissitz e Green (1975) e Nunnally (1978), que foram citados anteriormente.

Um segundo tema de crítica pontua que o acréscimo do número de itens em uma escala aumenta o coeficiente *alpha*. Para Cooper e Schindler (2003), existe potencial para inferências incorretas sobre a alta consistência interna quando o teste contém muitos itens, o que inflaciona o índice de correlação. Segundo Pasquali (2011, p. 222), “quanto maior o número de itens tiver o teste, maior será seu índice de precisão, pois o erro tende a zero quando o número se aproxima do infinito, segundo o famoso teorema de Bernoulli”.

Para Costa (2011), esse problema advém da própria estrutura matemática do coeficiente *alpha*, onde em sua fórmula geral, o numerador é multiplicado por *k*, ao passo que o denominador é multiplicado por (*k*-1), isto é, o numerador é sempre elevado por um fator maior do que o denominador. Como consequência, o *alpha* sempre tenderá a ser maior.

Pesquisas empíricas, entretanto, demonstraram que essa relação é demasiadamente pequena. A pesquisa de Peterson (1994), por exemplo, indicou que a variação do número de itens explicou somente 1% da variação do *alpha*. Entretanto, a pesquisa também identificou que as escalas com até 3 itens apresentaram menores coeficientes *alpha* do que escalas com 4 ou mais itens.

Costa (2011) acredita que os pesquisadores não possuem interesse em utilizar um número demasiadamente grande de itens para influenciar o valor de *alpha* artificialmente, pois quanto menor a extensão da escala, mais fácil seria a coleta de dados na medida em que reduz o tempo de resposta e a resistência dos respondentes. Como consequência, teriam-se respostas mais consistentes e melhores resultados na avaliação de confiabilidade. Por isso, segundo o autor, “mesmo que os pesquisadores possam inflacionar o valor de *alpha* pelo aumento do número de itens, normalmente não o fazem, obtendo resultados provenientes da própria adequação dos itens e de sua homogeneidade” (COSTA, 2011, p. 100).

Por conta desse problema, a fórmula de correção de Spearman-Brown é utilizada como ajuste para o efeito do tamanho do teste e para estimar a confiabilidade do teste completo (COOPER; SCHINDLER, 2003). Ela ainda pode ser utilizada para estimar o aumento da confiabilidade da escala associada ao aumento do número de itens (PASQUALI, 2011). De acordo com Costa (2011), a fórmula de Spearman-Brown é a seguinte (Fórmula 6):

$$\alpha = \frac{k \cdot \bar{r}}{1 + (k - 1) \cdot \bar{r}} \quad (6)$$

Onde:

k = número de itens da escala

r = média da correlação entre todos os pares de itens

Uma terceira crítica ao coeficiente *alpha* refere-se a possibilidade de geração de medidas de *alpha* elevadas somente pela inclusão de variáveis redundantes. Existe até o entendimento de que níveis muito elevados de *alpha* não estariam necessariamente medindo a consistência interna de uma escala, mas sim, a redundância entre os itens (BOYLE, 1991).

Neste sentido, quanto mais semelhantes entre si forem os itens, mais elevada tenderá a ser a sua correlação interitem e, por conseguinte, maior tenderá a ser o coeficiente *alpha*. Por esta razão, de acordo com Boyle (1991), escalas com coeficientes *alpha* superiores a 0,90 deveriam ser evitadas, pois estariam indicando um alto nível de redundância, e não de confiabilidade.

Na mesma linha de discussão, Costa (2011) lembra que os itens selecionados para compor uma escala de múltiplos itens são baseados no modelo da amostra de domínio, e por conta disso, os itens deveriam ser distintos entre si, correspondendo a diferentes facetas do construto em análise. Entretanto, se vários itens redundantes são inseridos, sendo eventualmente meras paráfrases de outros itens, a consequência disso será um alto valor de coeficiente *alpha*, que de fato, estaria medindo então a redundância entre os itens. Costa (2011) ainda minimiza este problema afirmando que é esperado que os pesquisadores sejam cuidadosos o suficiente, procurando não incrementar o coeficiente *alpha* simplesmente pela paráfrase de itens, mas sim, buscando abranger as facetas que variam junto com as variações do construto.

Finalmente, também é importante ressaltar que no caso de uma escala multidimensional, isto é, uma escala que possui diversos conjuntos de itens planejados para medir diferentes aspectos (dimensões) de um construto multidimensional, o coeficiente *alpha* não deve ser calculado para a escala como um todo. Segundo Malhotra (2001), como as dimensões de uma escala multidimensional são estruturas independentes, não seria apropriado uma medida de consistência interna calculada através de dimensões. No entanto, se vários itens são utilizados para medir cada dimensão, então é possível calcular a confiabilidade da consistência interna para cada uma das dimensões separadamente.

4.5 Procedimentos para melhoria da confiabilidade

Como último ponto a ser abordado neste capítulo, é interessante ressaltar formas que poderiam ser empregadas para aumentar a confiabilidade da medida. Entre os procedimentos que poderiam ser empregados, Cooper e Schindler (2003, p. 188) apresentam uma lista mais abrangente, que inclui os seguintes procedimentos:

- Minimizar as fontes externas de variação;
- Padronizar as condições sob as quais as mensurações normalmente ocorrem;

- Melhorar os procedimentos de coleta de dados, em especial o investigador, usando pessoas bem treinadas, supervisionadas e motivadas para fazer a pesquisa;
- Estender a amostra de questões de mensuração utilizada adicionando questões similares ao instrumento de coleta de dados ou acrescentando mais observadores ou ocasiões em um estudo de observação.
- Melhorar a consistência interna de um instrumento ao excluir dados de análises de questões que evocam respostas extremas. Esse enfoque supõe que um alto escore total na escala reflete alto desempenho na característica medida, enquanto que uma baixa pontuação significa um baixo desempenho. Selecionando-se os indivíduos com as pontuações extremas, por exemplo, os 20% mais altos e os 20% mais baixos para análises individuais, é possível distinguir através desse processo os itens que mais diferenciam os respondentes que tiveram escores altos e baixos. De acordo com essa abordagem, os itens que têm pouco poder discriminatório podem ser retirados do teste.

5 REFERENCIAL TEÓRICO – VALIDAÇÃO DE ESCALAS

Em muitos trabalhos de cunho científico, a produção do conhecimento está diretamente relacionada ao seu rigor, enquanto na literatura de metodologia, especificamente dos estudos quantitativos, a rigidez é caracterizada pela validade e confiabilidade do estudo, que busca reduzir o erro dos valores observados frente aos valores ditos “verdadeiros” (MONDADORI; LADEIRA, 2007). Para medir, avaliar e quantificar informações, os pesquisadores precisam atentar para os critérios de significância e precisão dos instrumentos de medidas que irão utilizar, o que envolve a análise da validade e confiabilidade dos mesmos (MARTINS, 2006). A confiabilidade já foi discutida no tópico anterior. Aqui, busca-se discorrer sobre o conceito de “validade”.

Validade refere-se ao grau com que um instrumento representa o objeto a ser medido, o que significa compreender a relação entre o instrumento de medida e o que se pretende medir. Para Block (1982) e Fumagalli (2007), para uma medição ter validade, é necessário ter nível zero de erros sistemáticos. De forma similar, Malhotra (2005) discorre que a validade de uma escala ou instrumento pode ser definida como o âmbito em que as diferenças em escores observados da escala refletem as verdadeiras diferenças entre objetos quanto à característica que está sendo medida, e não devido a um erro aleatório ou sistemático. Por essa razão, a validade perfeita exige que não haja qualquer erro na medida.

Tradicionalmente, a validade tem sido definida como uma característica ou qualidade de um teste, um critério para sua aceitabilidade, geralmente utilizada para se referir à reprodutibilidade de uma medida, ou seja, o grau de concordância entre múltiplas medidas de um mesmo objeto (ARMSTRONG; WHITE; SARACCI, 1994).

Acadêmicos a definem como um processo de coletar e avaliar as proeminências da validade de um instrumento, que tem como objetivo examinar com precisão uma determinada medida através dos escores de um teste (STRAUB; CARLSON, 1989; DANTAS, 2007). Selltiz, Wrightsman e Cook (1987), descrevem a validade como o nível por meio do qual uma medida empírica reflete o significado real do conceito ou construto que se quer abranger.

Henning (1987), Hughes (1989/1994), Alderson, Clapham e Wall (1995) e Pasquali (2009) compartilham a ideia de que um teste é válido se mede precisamente aquilo que deve medir. Pasquali (2009, p. 995) ainda reforça,

ao se medirem os comportamentos (itens), que são a representação física do traço latente, está-se medindo o próprio traço latente. Tal suposição é justificada se a representação comportamental for legítima. Esta legitimação somente é possível se

existir uma teoria prévia do traço que fundamente que a tal representação comportamental constitui uma hipótese dedutível desta teoria. A validade do teste (este constituindo a hipótese), então, será estabelecida pela testagem empírica da verificação da hipótese. Pelo menos, esta é a metodologia científica.

Historicamente, a validade constituiu um parâmetro de medida tipicamente discutido no contexto das ciências psicossociais. Não causa surpresa, nesse sentido, que o problema de validade tenha tido, na história da Psicologia, uma posição central na teoria da medida, constituindo-se, na verdade, no seu parâmetro fundamental e indispensável (PASQUALI, 2009).

Nos anos 80 e 90, diversos autores de diferentes campos de atuação passaram a dar maior ênfase à questão da validade, à medida que aumentaram as preocupações com o rigor metodológico e mensurações mais científicas (OLIVEIRA; WALTER; BACH, 2011). Segundo a *American Educational Research Association* (1999), nos últimos 15 anos nota-se uma preocupação dos pesquisadores em caracterizar o processo de avaliação da validade como uma investigação de caráter científico que serve de base para uma argumentação sobre o grau de validade das interpretações dos resultados observados.

Scaramucci (2011) aponta dois acontecimentos importantes ocorridos após 1985 como sendo decisivos para o surgimento do que poderia ser chamada de “uma nova teoria da validade” (SCARAMUCCI, 2011, p.104):

- A criação do código oficial de prática profissional nos Estados Unidos (AERA/APA/NCME – *American Educational Research Association, American Psychological Association* e *The National Council on Measurement in Education – Standards for Educational and Psychological Testing*); e
- A publicação dos trabalhos de Cherryholmes (1988), Moss (1992) e Wiggins (1993) (*apud* SCARAMUCCI, 2011), que questionaram as bases filosóficas da visão tradicional, retomadas no artigo seminal de Messick (1989) intitulado “Validade”, publicado na terceira edição do *Handbook of Educational Measurement*.

Bessa (2007) e Sampaio *et al.* (2012) discutem que a conceituação de validade e a concepção dos processos incluídos na validação refletem facetas importantes tanto do conhecimento científico como de questões sociais de diversas épocas:

- Nos anos 50, a influência do behaviorismo e de certas correntes da filosofia da ciência;
- A partir dos anos 60, a preocupação com diferenças entre grupos populacionais (segundo o gênero, os níveis socioeconômicos, ou os conceitos pré-definidos de etnia);
- A partir dos anos 70 e 80, a influência dos avanços da psicologia cognitiva;
- A partir dos anos 80 e 90, o processo de validação sofre, também, o impacto das ciências da computação;
- Após o ano 2000, os pesquisadores tornaram-se mais criteriosos, aumentando o registro em suas pesquisas quanto às validades externa e nomológica.

Kaaks *et al.* (1994) discorrem que os estudos de validação têm papel fundamental de estimar parâmetros estatísticos não conhecidos dentro de um modelo que especifica relações entre medida. Validar uma escala exige um processo aprofundado de investigação, já que a “validação” não termina em um único momento, mas, pressupõe continuidade e precisa ser repetida inúmeras vezes para o mesmo instrumento. Ela endossa credibilidade, sendo algumas vezes mencionada como o processo que fornece uma evidência documentada de que o método realiza aquilo para o qual é indicado para fazer (RIBANI, *et al.*; 2004).

A questão fundamental para se medir a validade de um instrumento de medidas é dada pela resposta à seguinte pergunta: será que se está medindo aquilo que crê que deve ser medido? Se a resposta for positiva, sua medida é válida, caso contrário, não se pode afirmar com rigor (GRESSLER, 1989 *apud* MARTINS, 2006).

Martins (2006) exemplifica a validação com um exemplo de estudantes brasileiros fazendo um teste de QI (Quociente de Inteligência) em inglês. Para o autor, tal teste é muito mais uma medida da proficiência na língua inglesa do que uma medida (válida) de inteligência, já que pode se ter estudantes brasileiros inteligentes, mas que foram mal no teste por não compreenderem a língua inglesa.

Straub e Carlson (1989) discutem que os critérios de validação de um instrumento de pesquisa científica consistem um elemento fundamental que compõe o rigor metodológico e permite desenvolver uma pesquisa com resultados confiáveis (STRAUB; CARLSON, 1989). De forma complementar, Dantas (2007) argumenta que tanto o pesquisador quanto o sujeito/respondente desempenham papéis importantes nesse processo, cabendo ao primeiro coletar as evidências que confirmam ou não a validade do instrumento.

Discussões e estudos nas últimas décadas apontaram para o crescimento da preocupação em relação à carência de validação nos instrumentos de pesquisas científicas. Oliveira, Walter e Back (2011) apontam que a validação tem sido muitas vezes empregada de forma inadequada, e que poucos pesquisadores dedicam séria atenção ao assunto. Para Boudreau, Gefen e Straub (2001) *apud* Oliveira, Walter e Back (2011, p.2), os estudiosos que desenvolvem pesquisas quantitativas aplicam em maior quantidade critérios de validade em seus trabalhos; contudo, "o campo ainda não atingiu o ponto no qual a validação é a regra, e não a exceção".

Para Scaramucci (2011), a crítica principal ao conceito tradicional de validade se deu ao fato de ser fragmentado e incompleto, uma vez que era elaborado exclusivamente por especialistas em medidas, seguindo uma visão essencialmente psicométrica. Esses pesquisadores, muitas vezes, não levavam em conta, como base para ação, as implicações de valor do significado dos resultados ou escores e nem as consequências sociais do uso desses resultados. Segundo o autor, "o novo conceito, embora contemplando múltiplas facetas, unifica-se em torno da validade de construto, passando a considerar não apenas bases evidenciais, mas também as consequenciais.

Para Messick (1989) *apud* Scaramucci (2011, p.110-111),

validade, portanto, não é uma propriedade do teste ou da avaliação, mas do significado dos seus resultados. Importante nesse caso é o argumento de validade, que tem como objetivo coletar informações a favor ou contra uma determinada interpretação dos escores do teste. O que é validado, portanto, são as inferências derivadas dos resultados ou outros indicadores, assim como as implicações para ação determinadas pela interpretação.

O Quadro 7 traz um resumo dos contrastes entre as concepções tradicionais e modernas de validade.

Quadro 7: Contrastes entre as concepções tradicionais e modernas de validade

DÈCADA DE 90	DÈCADA DE 2010
Validade era considerada uma característica de um teste: a medida que "mede aquilo que se quer medir"	Validade é considerada um argumento relativo à interpretação e uso: a medida em que as interpretações e uso de um teste podem ser justificados
Confiabilidade era vista como distinta e uma condição necessária para validade	Confiabilidade pode ser vista como um tipo de evidência de validade
A validade era frequentemente estabelecida através de correlações de um teste com outros	Validade é argumentada com base em um número de tipos de justificativas e evidências, incluindo as consequências da avaliação

Validade de construto era vista como um dos três tipos de validade (conteúdo, critério e construto)	Validade é um conceito unitário, em que a validade construto ocupa uma posição central (validade de conteúdo e de critério podem ser usadas como evidência da validade de construto)
O estabelecimento da validade era uma tarefa de responsabilidade de pesquisadores da avaliação, responsáveis pelo desenvolvimento de teste de grande escala e alta relevância	A justificativa de validade de um teste é de responsabilidade de todos os usuários de um teste

Fonte: Adaptado de Chapelle (1999, p.258) *apud* Scaramucci (2011)

Para Pasquali (2007), professor emérito da Universidade de Brasília da área de psicologia, com ênfase em psicometria, um dos principais autores de trabalhos sobre “validação”, e referência na pesquisa acadêmica brasileira, em seu artigo “*Validade dos Testes Psicológicos: Será Possível Reencontrar o Caminho?*” na *Revista Psicologia: Teoria e Pesquisa* expõe que existem tantas definições sobre a validade dos testes psicológicos, que acaba gerando uma “confusão” do real significado desse conceito. No artigo o autor destaca que esse problema se agravou após a publicação do trabalho de Samuel Messick (1989, p.13 *apud* Pasquali, 2007, p.99), sobre a validade como sendo “um julgamento avaliativo integrado do grau em que evidência empírica e racionalizações teóricas apoiam a adequação e propriedade de inferências e ações baseadas em escores de teste ou outros modos de avaliação”. Para Pasquali (2007, p.99),

este modo de entender e definir a validade dos testes, que é considerado o modo moderno de entender validade, inclusive assumido pela *American Psychological Association* (APA, 1985), tornou o tema um verdadeiro zoológico. Na verdade, encontra-se na literatura tradicional e atual sobre o tema, especialmente a norte-americana, um elenco sem fim de expressões ou tipos de validades que um teste pode ter (PASQUALI, 2007, p.99).

Na mesma linha, Costa (2011) sinaliza que assim como os estudos sobre confiabilidade, a literatura publicada sobre validade de escala e mensuração é diversa e controversa. O autor ainda discorre que, apesar do esforço de muitos pesquisadores, as conclusões são muitas vezes parciais, o que faz do tema, que já foi amplamente debatido nos campos da Psicologia e Educação, um dos mais interessantes e ricos em ideias e perspectivas de análise quando se pensa em mensuração em ciências sociais e comportamentais.

Cooper e Schindler (2003) no livro *Métodos de Pesquisa em Administração* apresentam duas principais formas de validade de escala: externa e interna. A validade externa dos resultados de pesquisa refere-se à capacidade dos dados de serem generalizados

entre pessoas, ambientes e épocas. Relaciona-se aos estudos que comparam as notas obtidas com medidas externas da competência/capacidade avaliada e, portanto, são correlacionais (ALDERSON; CLAPHAM; WALL, 1995). Já a validade interna limita-se à capacidade de um instrumento de pesquisa medir o que deve ser medido; e relaciona-se com os estudos que analisam o conteúdo do teste e seu efeito percebido (ALDERSON; CLAPHAM; WALL, 1995; COOPER; SCHINDLER, 2003).

Existem diversas possibilidades de averiguação da validade de uma proposta de mensuração (COOPER; SCHINDLER, 2003; MARTINS, 2006; ALEXANDRE E COLUCI, 2011; COSTA, 2011). Por essa razão, a análise de alternativas de verificação constitui a parte mais extensa nos estudos sobre desenvolvimento de uma escala. Para essa tese, buscou-se seguir o parâmetro defendido por Costa (2011), que procurou desenvolver uma proposição de análise aliando praticidade e efetividade de utilização, e ao mesmo tempo, estando de acordo com a tradição teórica da literatura especializada sobre mensuração e validade. O Quadro 8 e a Figura 5 trazem os três principais tipos de validação (Validade de Translação, Validade de Critério e Validade de Conteúdo), com seus respectivos subtipos.

Quadro 8: Tipos de validade, subtipos, conceito e operacionalização

Tipo de Validade	Subtipos	Conceito	Operacionalização
Validade de Translação	<ul style="list-style-type: none"> - Validade de Conteúdo - Validade de Face 	Grau em que o conteúdo dos itens representa adequadamente o universo de todos os itens relevantes sob estudo. Consiste em analisar exhaustivamente a estrutura da escala completa.	Qualitativa
Validade de Critério	<ul style="list-style-type: none"> - Validade Preditiva - Validade Simultânea 	Grau em que a predição é adequada para captar os aspectos relevantes do critério. A validade de critério estabelece a validade de um instrumento de medição comparando-o com algum critério externo.	Quantitativa
Validade de Construto	<ul style="list-style-type: none"> - Validade Convergente - Validade Discriminante - Validade Nomológica - Validade Grupo Conhecido 	Grau em que um instrumento de medidas se relaciona consistentemente com outras medições assemelhadas derivadas da mesma teoria e conceitos que estão sendo medidos. Tenta identificar os constructos implícitos que são mensurados e determinar como o teste representa esses constructos.	Quantitativa

Fonte: Elaborado pelo autor a partir de Cooper e Schindler (2003); Vieira e Slongo (2007); Alexandre e Coluci (2011) e Costa (2011)

Figura 5: Avaliação da validade de uma escala.

Fonte: Elaborado pelo autor a partir de Cooper e Schindler (2003); Vieira e Slongo (2007); Alexandre e Coluci (2011) e Costa (2011)

Os tipos de validade serão descritos nos tópicos a seguir.

5.1 Validade de translação

A **Validade de Translação** (*Translation Validity*) relaciona-se à aferição qualitativa do conteúdo e da forma de apresentação dos itens de escala. Implica em quanto o instrumento fornece cobertura adequada das questões investigativas que orientam o estudo, ou seja, se o instrumento contém uma amostra representativa do universo do assunto de interesse, então a validade de conteúdo é boa (COOPER; SCHINDLER, 2003).

Na literatura há controvérsias sobre a terminologia e o conceito da validade de conteúdo (ALEXANDRE; COLUCI, 2011). Para alguns autores, consiste em julgar em que proporção os itens selecionados para medir uma construção teórica representam bem todas as facetas importantes do conceito a ser medido. Outros estudiosos tentam explicar esse conceito por meio do seguinte questionamento: as perguntas do instrumento são representativas dentro do universo de todas as questões que poderiam ser elaboradas sobre esse tópico? (ALEXANDRE; COLUCI, 2011).

A Validade de Translação é sempre a primeira a ser verificada e possui dois momentos distintos: primeiro faz-se a análise do conteúdo do instrumento e do que se pretende medir (validade de conteúdo); em seguida, analisa-se a forma como esse instrumento se apresentará para medição (validade de face) (COSTA, 2011).

5.1.1 Validade de conteúdo

A validade de conteúdo (*content validity*) relaciona-se ao grau em que os itens/elementos selecionados para verificação são relevantes e representativos do conteúdo ou do domínio do construto sob análise (COSTA, 2011). Para construir a variedade de itens, o pesquisador deve inicialmente definir o constructo de interesse e suas dimensões por meio de pesquisa bibliográfica e consulta a investigadores da área (ALEXANDRE; COLUCI, 2011).

Para Martins (2006), a validade de conteúdo pode ser exemplificada por uma prova de operações aritméticas. Ela não terá essa validade se incluir somente problemas de adição e subtração, e excluir problemas de multiplicação e divisão.

Segundo Alexandre e Coluci (2011), os elementos selecionados são todos os aspectos do processo de medida que podem afetar a coleta de dados. No caso de um questionário, consideram-se como itens as instruções, o formato das respostas e os itens de forma individual. Escolher e utilizar itens em uma escala que se refiram efetivamente ao constructo é a demanda mais básica do processo de construção de uma escala; entretanto, essa é tida como a mais essencial na construção da mesma. Essa etapa de validade, segundo Costa (2011), é a única realmente indispensável para o desenvolvimento de uma escala, e, dependendo do tipo de objetivo e atributo, esse procedimento já seria satisfatório para gerar evidências de validade de uma escala. No processo de validade de conteúdo, ainda de acordo com o autor, é necessário que se verifique, de maneira sistemática, os seguintes aspectos: definição do constructo, a dimensionalidade utilizada na mensuração, o propósito da mensuração e a associação entre estes recortes e os itens propostos.

Considerando um nível mais operacional em relação à realização dos procedimentos de verificação, deve-se buscar garantir os seguintes aspectos (COSTA, 2011):

- Os itens devem refletir as facetas de um construto, e não aos aspectos associados a outros construtos;

- Os itens devem cobrir todas as facetas desse construto, ou ao menos uma amostra significativa e representativa do domínio do mesmo (ou seja, deve-se verificar se não faltam itens indicativos de determinadas facetas relevantes do construto);
- Os itens devem corresponder à condição proposta de mensuração em termos de construtos formativos ou refletivos (ou seja, se a proposta é mensurar como um construto refletivo, deve-se ter segurança de que os itens são efetivamente refletidos pelo construto. Já se a proposta for por uma mensuração na perspectiva do construto formativo, deve-se ter segurança de que os itens são realmente os causadores do construto);
- Deve-se buscar uma proporção equilibrada entre as facetas do escopo e o número dos itens, de modo a evitar que uma dada faceta seja medida com um dado número de itens, e outra faceta igualmente relevante seja mensurada com um número muito maior ou muito menor.

Costa (2011) ainda reforça a importância de se considerar o tempo de uma escala, atentando para o fato de que a variação temporal de um construto poderá ter implicações na validade de conteúdo, e ainda no sentido e no uso de uma escala consolidada. Para o autor, os itens de uma escala não são estáveis em longo prazo, especialmente em razão da variação e do aprimoramento conceitual dos construtos, ou de natureza ética, legal ou moral de uma escala. Nesse sentido, a validade de conteúdo deve ser reavaliada periodicamente, por meio de uma revisão completa da associação entre os itens e o conteúdo de uma escala.

A determinação da validade de conteúdo depende de julgamento e pode ser encarada de várias formas. Uma vez que não existe formalmente teste estatístico para a validade de conteúdo, o julgamento do pesquisador, e discernimento devem ser aplicados (GARVER; MENTZER, 1999). Inicialmente, o planejador pode determiná-la através de uma definição cuidadosa do tópico de preocupação, dos itens a serem mensurados e das medidas a serem usadas. Tal processo é frequentemente intuitivo e único para cada pesquisador. Ou o pesquisador pode também usar um painel de pessoas para julgar como o instrumento atende aos padrões, já que ele julga cada item como essencial, útil, mas não essencial ou desnecessário para avaliar o desempenho de um comportamento importante (COOPER; SCHINDLER, 2003).

De forma complementar, Dantas (2007) argumenta que a validade de conteúdo tende a ser julgada de forma mais qualitativa do que quantitativa baseando-se, fundamentalmente, em resultados empíricos. Esse fato ocorre pois não existem métodos totalmente objetivos para

garantir que o instrumento avalie aquilo que se propõe. A validade de conteúdo para ser verdadeira, precisa, nesse sentido, que o instrumento seja submetido a pelo menos dois especialistas, sendo o mais comum a avaliação por um painel de juízes e leigos que irão avaliar se o instrumento apresenta clareza, compreensão e redundância nos itens (DANTAS, 2007).

De acordo com Alexandre e Coluci (2011), durante o desenvolvimento do instrumento, um dos pontos discutidos nessa avaliação é o número e a qualificação desses juízes. Não há consenso na literatura sobre esses aspectos. Alguns pesquisadores recomendam no mínimo cinco e no máximo dez pessoas necessárias para participar desse processo, enquanto outros sugerem de seis a vinte sujeitos, sendo composto por três indivíduos, no mínimo, em cada grupo de profissionais selecionados para participar. Esses devem receber instruções específicas sobre como avaliar cada item, o instrumento como um todo e como preencher o questionário que orienta a avaliação. Os autores ainda reforçam que quando for um processo de adaptação cultural, é importante formar um comitê multidisciplinar. Neste caso, recomenda-se que deve ser composto por pessoas bilíngues e por especialistas na área de conhecimento do instrumento, que precisam ser informados sobre as medidas e conceitos envolvidos.

Para Cooper e Schindler (2003, pg. 185),

As respostas “essencial” para cada item de cada painelistas são avaliadas por um índice de validade de conteúdo, e aquelas que atingirem um valor estatístico importante, são mantidas. Tanto nos julgamentos informais como nesse processo sistemático, a validade de conteúdo está primariamente relacionada às inferências sobre construção de teste, e não às inferenciais sobre resultados de testes (COOPER; SCHINDLER, 2003, p.185).

Um instrumento de medição deve conter todos os itens de domínio do conteúdo das variáveis que pretende medir. Assim, pode aparecer que uma simples verificação do conteúdo do teste é suficiente para estabelecer a validade com relação a esse objetivo; no entanto, a solução não é tão simples. A amostragem do conteúdo, por exemplo, é uma dificuldade que precisa ser superada. A área de conteúdo a ser testada precisa ser sistematicamente analisada a fim de assegurar que todos os aspectos essenciais sejam, adequadamente, e em proporções corretas, abrangidos pelos itens do teste. Para uma maior garantia da validade de conteúdo de um instrumento de medidas, a área de abrangência deve ser inteiramente descrita anteriormente à construção de um teste ou de outro instrumento de coleta de dados (MARTINS, 2006).

Para Rubio *et al.* (2003), a validade de conteúdo é fundamental no processo de desenvolvimento e adaptação de instrumento de medidas, no entanto, apresenta limitações por ser um processo subjetivo. Dessa forma, sua utilização não elimina a necessidade de aplicação de outras medidas adicionais.

5.1.2 Validade de face

A *validade de face* (*face validity*), também denominada de validade aparente, refere-se à praticidade, à permanência e à representatividade de um conjunto de itens em relação ao construto a ser medido, isto é, o quanto os respondentes e os especialistas consideram os itens apropriados para o construto e para os propósitos da mensuração. Ela deve ser realizada de forma paralela à validade de conteúdo, e, em termos concretos, as duas são indissociáveis, pois o conteúdo está expresso nos enunciados, e não pode haver validade de conteúdo sem validade de face e vice-versa. No entanto, Costa (2011, p.111-112) chama a atenção da importância de se distinguir os dois conceitos:

(...) A validade de conteúdo remete à amostra do domínio do construto, ao passo que a validade de face trata de aspectos complementares, e com foco no enunciado e na praticidade. Assim, quando temos um conjunto de itens propostos para uma escala, a validade de conteúdo procura averiguar se este conjunto representa uma amostra adequada e bem selecionada e representativa do domínio do construto, enquanto a validade de face procura assegurar se os enunciados refletem de fato o que se pretende medir (COSTA, 2011, p.111-112).

Costa (2011) ainda discute que em termos de cuidados e procedimentos, muito do que se aplica à validade de conteúdo pode ser utilizada para a validade de face, recomendando algumas ações específicas:

- Análise da adequação entre os enunciados e a forma de mensuração selecionada (escala de Likert, diferencial semântico, escala de Stapel, entre outras). Para cada alternativa há uma forma diferente de enunciado;
- Usar os futuros respondentes para avaliação dos itens (e não somente os especialistas);
- Evitar aparência de repetições e as redundâncias de itens;
- Fixar critérios de avaliação de qualidade, recomendando-se as correções gramaticais, de ordem de apresentação das palavras na frase, clareza nas expressões e enunciados, extensão dos enunciados, aleatoriedade da disposição, etc;

- Usar a validade face não só para os itens da escala, mas para todas as partes do instrumento, no caso de instrumentos completos.

Para exemplificar a validade de face, tem-se um exame de aritmética para o terceiro ano de graduação em administração. Constituído de problemas de finanças, esse exame terá validade de face se, aparentemente, medir o conhecimento de finanças que se ensina no terceiro ano da graduação. A validação do instrumento deve ser feita por professores dessa matéria naquela turma e pelos alunos, que se submeterão ao exame.

Costa (2011) expõe que a verificação da validade de translação, nas suas duas formas (validade de conteúdo ou validade de face), seguem as seguintes etapas:

1. Definição clara do domínio e as facetas do construto;
2. Análise dos fatores iniciais de influência, tais como dimensionalidade, condição temporal, e finalidade da mensuração;
3. Avaliação da proporção entre o número de itens e as dimensões e facetas do construto, verificando se todas as facetas do construto estão bem cobertas;
4. Definição dos critérios de adequação de conteúdo, tais como: aderência à definição, número de itens por dimensão ou faceta, característica de mensuração formativa e reflexiva;
5. Definição de critérios de adequação de face, tais como: linguagem, expressividade, extensão, disposição dos itens no questionário, tipo de escala de verificação;
6. Uso da própria população (leigos), e um conjunto de especialistas, para análise de itens;
7. Apresentação de procedimentos e os resultados da avaliação de conteúdo nos relatórios da escala.

Uma outra proposta interessante foi desenvolvida por Chester Schriesheime e seus parceiros de trabalho (COSTA, 2011), que sugeriram uma verificação empírica e de base quantitativa da validade de conteúdo, a partir de um protocolo de atividades, que envolviam desde análises descritivas de adequação, até procedimentos mais sofisticados de análise fatorial, contendo as seguintes etapas:

1. Definição dos construtos e dimensões
2. Definição de critérios de adequação, considerando aspectos como clareza, relevância e representatividade, e especificidade do item, em relação à definição enunciada;
3. Construção de um questionário;
4. Submissão do questionário a um conjunto de especialistas e leigos pré-definidos, que serão respondentes potenciais da escala finalizada;
5. Fixação de uma regra de decisão sobre o item;
6. Seleção dos itens para as etapas de validação posteriores.

5.2 Validade de critério

A segunda forma de validade - **Validade de Critério** (*Criterion-Related Validity*) - segundo Pasquali (2009) e Costa (2011), consiste no grau de eficácia que um teste tem em prever um desempenho específico de um sujeito, ou seja, o quanto uma medida co(varia) com outra previamente validada, inclusive na determinação de previsão, quando tal (co)variação é justificada. O desempenho do sujeito torna-se, assim, o critério contra o qual a medida obtida pelo teste é avaliada. Nesse caso, o desempenho do sujeito deve ser medido/avaliado por meio de técnicas que são independentes do próprio teste que se quer validar. Costa (2011) também discute que nessa etapa o importante é saber o comportamento dos itens quando aplicados à mensuração do constructo. Ainda discute que a aderência conceitual é esperada, ou tida como certa, uma vez que a validade de translação vem antes da validade de critério.

Em alguns casos, pode haver vários critérios possíveis, mas nenhum completamente satisfatório. Por exemplo, o total de vendas para o vendedor pode não refletir adequadamente o potencial de mercado, as condições de concorrência ou os diferentes índices de lucratividade dos vários produtos. Pode-se basear na avaliação geral do gerente de vendas; mas o quanto essas impressões são acuradas e não enviesadas? (COOPER; SCHINDLER, 2003).

Se a escala comportar-se como esperado, tem-se “evidências” de validade de critério. Pelo fato dos pesquisadores lidarem com amostra, dificilmente pode-se ter certeza absoluta da validade de critério da escala, pois a correção, ainda que não nula, pode ser próxima da correlação real do universo. O esforço deve então ser direcionado para a geração de

evidências cada vez mais consistentes, mesmo que nunca seja possível gerar certeza de validade (COSTA, 2011).

Cooper e Schindler (2003) reforçam que qualquer medida de critério deve ser julgada em termos de quatro qualidades:

- Relevância – um critério é relevante se for definido e classificado em termos do que se julga ser medidas apropriadas;
- Inserção de viés – obtida quando o critério dá a cada possibilidade igual oportunidade de obter classificação;
- Confiabilidade – um critério confiável é estável ou reproduzível;
- Disponibilidade – a quantidade de dinheiro e de esforço a ser gasta no desenvolvimento do critério depende da importância do problema para o qual o teste está sendo usado.

Quadro 9: Validade Preditiva e Validade Concorrente ou Simultânea

Validade Preditiva	Validade Concorrente ou Simultânea
<ul style="list-style-type: none"> • Refere-se à capacidade da variável de efetivamente prever os resultados no futuro. • Envolve a tomada de evidência em diferentes momentos, e a comparação de resultados com a medida tomada na escala. • Dada a dificuldade de verificação, que requer mais de uma tomada de informações, a validade preditiva costuma ser negligenciada no desenvolvimento de escalas em administração 	<ul style="list-style-type: none"> • Refere-se à capacidade da escala de efetivamente se comportar como esperado em uma coleta simultânea das medidas. • Envolve a tomada de medidas em um mesmo momento, da escala e de um dado critério, a partir das quais se verifica uma relação esperada. A confirmação da relação como esperado dá evidências de validade simultânea. • Essa alternativa de verificação é a mais utilizada nos esforços de produção de escalas em administração.

Fonte: Elaborado pelo autor, a partir de Costa (2011)

Na literatura pode-se distinguir dois tipos de validade de critério, conforme apresentado no Quadro 9: (1) *validade preditiva* (com verificação futura) e (2) *validade concorrente ou simultânea* (com verificação ao mesmo tempo). Segundo Pasquali (2009), a diferença fundamental entre esses dois tipos é em relação ao tempo que ocorre entre a coleta da informação pelo teste a ser validado e a coleta da informação sobre o critério. Se estas coletas forem (mais ou menos) simultâneas, a validação será do tipo concorrente; caso os dados sobre o critério sejam coletados após a coleta da informação sobre o teste, tem-se a validade preditiva. Na mesma linha, Martins (2006) expõe que se o critério se fixa no

presente, tem-se a validade simultânea, já que os resultados do instrumento se correlacionam com o critério no mesmo momento ou ponto no tempo. Já se o critério se fixa no futuro, tem-se a validade preditiva, que se refere à extensão a qual o instrumento (geralmente teste) prediz futuros desempenhos de indivíduos.

O fato de a informação ser obtida simultaneamente ou posteriormente à do próprio teste não é um fator tecnicamente relevante à validade do teste. O fator primordial nesse processo é a determinação de um critério válido. Nesse sentido, podem-se considerar duas etapas para a definição da validade de critério: (1) definição de um critério adequado e (2) medição e validação desse critério, independentemente do próprio teste (PASQUALI, 2009).

Costa (2011) destaca que não há clareza na literatura especializada se a relação entre o construto e o critério adotado deve ou não ser justificada teoricamente. Isso significa que para aferir a validade de critério, primeiramente deve-se definir um determinado critério para julgar o valor da escala, possivelmente o seu efeito futuro ou simultâneo. Este, por sua vez, deve ser mensurável, para viabilizar a comparação estatística (bi ou multivariada) com a escala em teste, além de ser consistente com a prática ou a teoria.

5.3 Validade de construto

A **Validade de Construto** (*Construct Validity*) indica que construto ou característica a escala está medindo, isto é, possibilita determinar qual a característica que explica a variância do teste ou, então, qual o significado do teste. Procura identificar as diferenças individuais dos escores do instrumento de medida, permitindo o pesquisador centralizar-se na característica ou traço que está sendo medido, mais do que no próprio teste (VIANNA, 1983). Sua verificação pode ser procedida pela verificação da hipótese estatística da relação entre os dados e a suposição de paralelismo, ou de divergência teórica, associada especificamente à definição (COSTA, 2011).

Vianna (1983, p.36) ainda destaca que

a constatação da validade de construto resulta do acúmulo, por diferentes meios, de várias provas, que precisam ser analisadas em todos os seus detalhes, a fim de constatar, entre outros aspectos, quais as variáveis com os quais os escores do teste se correlacionam, quais os tipos de itens que integram o teste, o grau de estabilidade dos escores sob condições as mais variadas e o grau de homogeneidade do teste, com vistas a ter elementos que possam esclarecer o significado do instrumento.

Dantas (2007) define a validade de construto como sendo o processo de examinar a ligação teórica dos itens juntamente com a escala hipotética. Para o autor, obter a confirmação

dessa validade envolve o exame lógico das relações que deveriam existir com outras medidas e/ou padrões de valores para grupos que supostamente deveriam divergir nos valores relacionados ao construto.

Uma medida apresenta validade de construto no grau em que estiver de acordo com as correlações previstas em outras proposições teóricas. Se busca-se medir o resultado de um evento, o método em que é definido operacionalmente esse evento precisa corresponder a uma base teórica empírica. Uma vez estabelecido que o construto é importante no sentido teórico, faz-se a adequação do instrumento. Caso haja uma medida do evento conhecida, correlacionam-se os resultados obtidos usando essa medida com aquelas derivadas dos instrumentos (COOPER; SCHINDLER, 2003).

Para Vianna (1983, p.37),

A validação de construtos não se limita apenas a validar um teste, o seu alcance é bem mais amplo, centrando-se seu objetivo na validação da teoria em que se apoiou a construção do instrumento, desse modo, o trabalho de validação de um construto é uma pesquisa científica empírica, porque definidos os construtos que seriam responsáveis pelo desempenho no teste, o avaliador passa a formular hipóteses sobre a teoria dos construtos e, a seguir, testa empiricamente essas hipóteses. A testagem de hipóteses sobre construtos exige a verificação da convergência e discriminatividade (VIANNA, 1983, p. 37).

Para Gressler (1989), a validade de construto será dada pela resposta à seguinte questão: em que medida a definição operacional (construto) de um conceito de fato reflete seu verdadeiro significado teórico? Na mesma linha, Costa (2011) complementa com a seguinte questão: O que responde pela variância da mensuração?

Para avaliar a validade de um construto, de acordo com Malhotra *et al.* (2005), os pesquisadores precisam verificar alguns pontos importantes: as questões de escalas refletem a dimensão que se pretende testar, aquilo que se propõe? Há algum julgamento teórico, por peritos, na estrutura da escala?

A validade de construto refere-se à demonstração de que o instrumento realmente mede aquilo que ele se propõe medir. Procura responder questões teóricas quanto aos motivos pelos quais uma escala funciona e que deduções podem ser feitas em relação à teoria subjacente à escala, e, por essa razão, esse tipo de validade exige uma sólida teoria da natureza do construto que está sendo medido e como ele se relaciona com outros construtos (KERLINGER, 1973; ZIMMER; FERREIRA; HOPPEN, 2007; MENEZES, 2014).

Para Costa (2011), o termo “construto” refere-se a um construto psicológico, uma conceituação teórica sobre um aspecto do comportamento humano que não pode ser medido ou observado diretamente. Sousa e Silva (2005) defendem que o processo de validação de um

construto busca unir dois domínios – um teórico e outro da observação. Como exemplos de construtos, Costa (2011) cita inteligência, motivação para o rendimento, ansiedade, rendimento, atitude, dominância e compreensão de leitura.

A validade de construto é de grande relevância na área acadêmica, considerando que a avaliação se vale, frequentemente, de construtos, que após sua definição operacional, são medidos por intermédio de testes. Pode-se dizer que tanto a validade de conteúdo quanto a validade de critério, ambos importantes de serem avaliados em bons testes, não têm o foco, entretanto, na compreensão dos construtos que os testes medem, conseqüentemente, impõem-se uma nova abordagem para análise dos instrumentos de medida (VIANA, 1983). Costa (2011) fecha essa ideia ao afirmar que o foco na dimensão conceitual é justamente o aspecto que distingue a validade de construto dos demais tipos de validade.

Os primeiros autores que estudaram a validade de construto, segundo Alves *et al.* (2013), foram Cronbach e Meehl em 1955, na área de psicologia, e o foco do trabalho desses autores foi o de especificar a qualidade para investigação antes da publicação de testes. Para Vianna (1983), eles estudaram amplamente a validade de construto e produziram documentos importantes. Também cita o ensaio de Campbell e Fiske (1959 *apud* Vianna, 1983) sobre validade convergente e discriminante como um dos trabalhos marcantes no meio acadêmico.

De acordo com Sampieri (1996) *apud* Martins (2006, p.7),

Difícilmente a validade de constructo será estabelecida em um único estudo. Ela é construída por vários estudos que investigam a teoria do constructo particular que está sendo medido. Medidas variáveis do campo das ciências sociais aplicadas, as medidas de variáveis têm “vida limitada”. Com o acúmulo de resultados de pesquisa, os investigadores descobrem limitações e criam novas medidas para corrigir possíveis problemas. Esse processo leva ao aprimoramento das medidas e a uma compreensão mais completa das variáveis subjacentes que estão sendo estudadas.

O processo de validação de um construto deve, segundo Martins (2006) necessariamente, estar vinculado a uma teoria. Para o autor, não é possível considerar uma validação de constructo na sua totalidade de implementação, a menos que exista um marco teórico que suporte o construto em relação a outras definições.

No domínio da teoria, das ideias, das intuições do pesquisador, o constructo pode originar-se das percepções que o pesquisador tem acerca do fenômeno a ser estudado enquanto no domínio da observação, estão incluídos a observação direta, as anotações informais e, especialmente, as medidas objetivas, formais, baseadas em questionários e/ou escalas (DANTAS, 2007).

As evidências essenciais para esse tipo de validação são obtidas fazendo-se estudos inter-relacionados, visando a verificação empírica por meio de testes estatísticos, das construções teóricas sobre a relação entre as variáveis a serem medidas (ZIMMER; FERREIRA; HOPPEN, 2007; MENEZES, 2014).

A validade de construto é tida como a mais valiosa maneira de avaliar o instrumento. Para Kerlinger (2003), ela também é a mais difícil de ser alcançada. Costa (2011) expõe que ela é uma das mais importantes e gera uma evidência necessária de consistência de medida. Ainda o autor destaca que esse tipo de validade, juntamente com a de translação, está sempre presente nos estudos sobre o desenvolvimento de escalas, o que não ocorre com a validade de critério. Uma explicação parcial para tal ocorrência associa-se à disposição atual de métodos de verificação, que, quando inseridas em propostas de publicações e relatórios, trazem mais consistência e um maior senso de rigor na análise (COSTA, 2011, p.123).

A validade de construto inclui três subtipos: validade convergente, discriminante e nomológica (COSTA, 2011), descritas a seguir.

5.3.1 Validade convergente

A Validade Convergente (*Convergent Validity*) indica a extensão da relação positiva da escala com outras medidas de um mesmo construto (MALHOTRA, 2005). De forma semelhante, Pasquali (2003) discute que ela pode ser entendida como o relacionamento significativo entre duas ou mais medidas de um mesmo construto ou de construtos relacionados em sua teoria, utilizando métodos ou instrumentos de avaliação variados.

Os estudos de validade convergente são caracterizados por variáveis externas utilizadas na validação do teste, que consistem em outros instrumentos medindo os mesmos construtos. Sua finalidade principal é verificar, empiricamente, se o teste em estudo de validação apresenta fortes correlações, estatisticamente significativas, com outros instrumentos que avaliam o construto em questão (NUNES; PRIMI, 2010).

Da mesma maneira que ocorre nos demais tipos de validade, tem-se mais de uma interpretação para a validade convergente. Costa (2011) apresenta, dentre outras possíveis, três alternativas de definição. Primeiramente, tem-se a associação da validade convergente ao grau em que os indicadores designados para medir um mesmo construto são relacionados e convergentes (os indicadores em si). Também pode-se entender que há validade convergente, em uma segunda definição, quando duas medidas distintas de um mesmo construto confirmam a expectativa de serem fortemente relacionadas, o que significa dizer, que existe

um elevado coeficiente de correlação entre as duas medidas. O autor reforça que esse tipo de validade implica em verificar por um mesmo método, por exemplo, um questionário auto-respondido, as duas alternativas de verificação, e em seguida verificar a correlação existente. A validade convergente, considerando uma terceira definição, pode ser entendida como associada à convergência entre diferentes métodos usados para medir um mesmo construto, isto é métodos distintos de aferição de uma mesma medida. Como exemplo, Costa (2011) cita um questionário auto-respondido e um procedimento de observação.

O autor conclui que essas distintas definições encaminham a procedimentos de verificações diferentes e, por essa razão, pode-se observar evidências de validade convergente em trabalhos variados, sem considerar um mesmo sentido de convergência.

Para o primeiro caso (convergência de indicadores do construto), o pesquisador poderia, por exemplo, avaliar os pares de correção entre os indicadores, ou a realização de procedimentos mais sofisticados, como a análise fatorial, sendo essa a mais recorrente, já que existem muitos softwares atualmente que viabilizam a realização de análise fatorial confirmatória (AFC). Na análise de escalas de múltiplos itens (para escalas com somente um item não há opção além da verificação conjunta de outras medidas), pode-se observar se as cargas fatoriais são consistentemente vinculadas ao fator associado à escala, permitindo conferir a aderência do item ao fato. Nesse processo de verificação, analisa-se se as saídas do processo de extração de fatores nos procedimentos de AFC, analisando inicialmente as cargas fatoriais dos múltiplos itens, seguido das estatísticas associadas ao teste de hipótese de nulidade do escore. Duas bases possíveis de análise são as seguintes (COSTA, 2011, p.125):

- Cargas fatoriais devem ser maiores que 0,4 (de preferência 0,6);
- O teste de significância de nulidade do escore (t de *Student*, no software LISREL; CR no software AMOS) deve ser significativo a $p < 0,05$ ou a $p < 0,01$ (COSTA, 2011, p.125).

Para o autor, com estas verificações confirmadas, pode-se dizer que se alcançou a evidência de validade convergente no conjunto de itens.

5.3.2 Validade discriminante

A Validade Discriminante (*Discriminant Validity*) avalia em que grau a medida em questão não está relacionada com outras das quais deveria se diferenciar (MALHOTRA, 2005). É testada por meio da comparação entre medidas que, supostamente, não possuem conexão (BREVER; HUNTER, 2006).

Segundo Hair *et al.* (2013) esse tipo de validade representa quanto um único construto difere de todos os outros e representa um construto especial. Logo, a validade discriminante elevada oferece evidência de que um construto é único e captura alguns fenômenos que outras medidas não conseguem (HAIR *et al.*, 2009).

Para avaliar essa modalidade de validade, a melhor alternativa (se não a única), de acordo com Costa (2011), consiste em tomar medições dos construtos com diferentes escalas e verificar o grau de correlação. Se este for baixo ou nulo, tem-se evidências de validade discriminante.

Para a verificação da validade discriminante, entretanto, há certas dificuldades, conforme as discutidas por Costa (2011). Uma base de referência para validade discriminante, segundo o autor, pode ser infinitamente grande. Além disso, essa base pode ser problemática pela possibilidade de existência de uma elevada associação entre os dois construtos sob análise, e que justifique uma correlação elevada (ao contrário da expectativa, a priori, de correlação baixa entre construtos distintos). Costa (2011) exemplifica citando o caso de uma tentativa de medir a avaliação pelo cliente, de sua experiência no ponto de venda varejista, considerando uma base de referência de duas dimensões – “atendimento e layout”. As escalas que medem cada dimensão devem ser diferentes, entretanto, a correlação pode ser elevada. O autor finaliza dizendo que talvez a maior dificuldade da validade discriminante seja selecionar os construtos dos quais uma determinada medida efetivamente deve diferir. Usualmente tem-se como referência somente os construtos presentes em um mesmo estudo, como ocorre frequentemente em estudos com modelo estruturais sob teste, e em escalas de construtos multidimensionais. Entretanto, saliente que mesmo nessas circunstâncias, ainda há a dificuldade de fixação de critérios para se aceitar uma correlação como baixa ou alta, já que algumas correlações altas são teoricamente justificadas e até mesmo esperadas.

Costa (2011, p. 127) argumenta que,

os autores da área de modelagem de equações estruturais sugerem que se tome como referência, para escalas de mensuração de múltiplos itens, a medida de variância extraída de um dado construto, e a variância compartilhada com os demais. A referência é a seguinte: se a variância extraída for maior que a variância compartilhada entre dois construtos, diz-se que os dois têm medidas efetivamente diferentes, o que evidencia a validade discriminante (COSTA, 2011, p.127)

Silva, Macedo e Silva (2013) discutem que no processo de validação torna-se essencial especificar as hipóteses previstas entre as variáveis envolvidas apontando: 1. O sentido que se espera da relação, seja positivo, negativo, ou ausência de relação; e 2. A

magnitude relativa esperada da relação, na qual se pode comprovar a existência de relações maiores e mais claras.

Morales Vallejo *et al.* (2003) apontam que deve ser realizado um criterioso planejamento do processo da validação discriminante durante a preparação do instrumento para coletar simultaneamente os dados necessários. As correlações entre as medidas envolvidas neste processo podem ser apresentadas através de uma matriz denominada matriz multitraço-multimétodo. Para Costa (2011), essa matriz, também conhecida como MTMM, desenvolvida pelos pesquisadores Donald Campbell e Donald Fiske, consiste em analisar conjuntos de construtos por meio de diferentes métodos de aferição. Os requisitos básicos para sua aplicação envolvem:

- Uso de mais de um construto, e mais de um método de verificação, que pode ou não estar baseado em um mesmo conjunto de itens. Para cada construto anota-se a confiabilidade, e entre os construtos a correlação bivariada;
- Níveis de correlação elevados em um mesmo construto, nos diferentes métodos. Caso esta expectativa se confirme, tem-se a evidência de validade convergente;
- Níveis de correlação reduzidos nos diferentes construtos, dentro de cada método, e entre os diferentes métodos. Caso isso se confirme, tem-se evidências de validade discriminante.

Para Costa (2011, p.130),

É fácil perceber a atração que a matriz MTMM pode exercer sobre os pesquisadores, se considerarmos sua força na indicação de resultados e sua praticidade. No entanto, a utilização desta matriz em estudos de desenvolvimento de escalas tem sido muito reduzida a partir dos anos de 1990, especialmente porque se trata de um método mais demorado, pois requer mais de uma verificação, e ainda porque há muitas alternativas de verificação que dispensam o esforço de mais de uma verificação (COSTA, 2011, p. 130).

5.3.3 Validade nomológica

A Validade Nomológica indica a relação entre os construtos teóricos, de forma a confirmar as correlações significativas entre eles, originadas por uma teoria (MALHOTRA, 2005). A validade nomológica se refere ao grau em que uma ou várias previsões a partir de uma rede teórica a respeito de um construto são confirmadas pela mensuração utilizada (BAGOZZI, 1981; CAMPBELL, 1960). Em última instância, esse tipo de validade procura evidências favoráveis ao conjunto de hipóteses que compreendem o modelo hipotético de

pesquisa. Trata-se de uma tentativa de verificar se as hipóteses do modelo encontram suporte nos dados, sendo, portanto, o último e o mais importante critério de validade, priorizando suportar a qualidade da teoria subjacente (NETEMEYER *et al.*, 2003).

Essa modalidade de validade deve iniciar de relações consagradas na literatura especializada, que geram modelos estruturais nos quais o construto sob mensuração se insere. “Caso a verificação estatística confirme as relações (teoricamente) previstas, no conjunto, tem-se evidência de validade nomológica de escala” (COSTA, 2011, p.133).

Para Vianna (1983), as teorias que procuram explicar um determinado fenômeno compreendem um conjunto interrelacionado de conceitos, proposições e leis, e que, a esse sistema interligado, Cronbach e Meehl (1995) deram o nome de rede nomológica. Segundo o autor,

as leis estabelecem relações entre diferentes características, entre características e construtos, ou entre um construto e outro, os quais ressaltam ainda que as leis e os conceitos devem estar ligados a comportamentos observáveis. Às vezes, entretanto, isso não ocorre. Uma determinada lei ou conceito não se relaciona diretamente com as características, mas com outras leis e conceitos, que, por sua vez, são diretamente ligados a características observáveis. É condição necessária, na pesquisa de validade de construto, que a definição de conceitos ou formulação de leis estejam apoiados, direta ou indiretamente, em dados observáveis (VIANNA, 1983, p.37).

Para Hair *et al.* (2009), deve-se considerar a utilização de três ou quatro índices de ajuste, sendo pelo menos um índice incremental (o qual compara o modelo testado com um modelo alternativo, em geral, um modelo nulo no qual todas as variáveis não são correlacionadas), um absoluto (o qual verifica se o modelo utilizado reproduz os dados observados) e um índice de má qualidade do ajuste (índices que verificam se o ajuste é ruim com base no valor dos resíduos ou diferenças gerados, em que valores mais altos indicam um ajuste ruim).

Costa (2011) discorre que para verificar a validade nomológica, uma alternativa é pesquisar de forma conjunta os construtos que se supõe relacionados, de preferência com antecedentes e consequências do construto mensurado. Pela amostra, procede-se à verificação da consistência estatística das relações teóricas, podendo-se utilizar as seguintes alternativas:

- Procede-se à verificação das correlações entre as várias medidas. Este é o procedimento menos recomendado;
- Procedem-se as análises de regressão sucessivas (uma para antecedentes e outra para consequências, incluindo as mediações). Este procedimento é recomendado, especialmente em casos de amostras reduzidas;
- Procede-se a uma modelagem de equações estruturais, com verificação simultânea das relações. Este é o procedimento mais recomendado quando houver amostras de pelo menos 200 respondentes (COSTA, 2011, p.133).

O procedimento da validade nomológica, apesar de ser recomendado na literatura especializada há décadas, segundo Costa (2011) é especialmente delicado na medida em que requer o suporte seguro na literatura, e, também, que se verifique em um mesmo estudo mais de um construto, com mais de uma escala, por hipótese, válidas. O procedimento, muito mais do que simplesmente validar uma escala, serve ainda para testar, e em boa medida, descobrir relações teóricas entre construtos.

5.3.4 Validade de grupo conhecido

A *Validade de Grupo Conhecido* (*Known-group validity*), segundo Costa (2011) diz respeito ao grau em que uma medida deve variar em grupos distintos, atendendo a uma expectativa bem justificada de variação. Refere-se ao grau em que uma medida que deve variar por grupo realmente varia. Como exemplo o autor cita o caso da mensuração de fanatismo por futebol, que deve ter variações entre homens e mulheres, uma vez que é sabido, por meio de verificações empíricas diversas, que pessoas do sexo masculino são mais apreciadoras de futebol que pessoas do sexo feminino.

A sua mensuração consiste em extrair a medida de uma escala nos dois grupos, e, em seguida, proceder à comparação entre os mesmos. A diferença consistente entre os grupos explica que há evidência de validade de grupo conhecido. Entretanto, a dificuldade está em assegurar qual o comportamento esperado, ou seja, se deve haver diferença, deve-se saber se não há momentos ou circunstâncias específicos para manifestação da diferença. Se houver, precisam antes ser especificados (COSTA, 2011).

Assim como no caso da validade de critério, no qual se espera que haja correção, mas quase nunca há segurança sobre qual deve ser o nível de correlação, também na validade de grupo conhecido deve-se ter uma indicação bem clara da diferença, mas não se sabe ao certo qual deve ser o tamanho da diferença (COSTA, 2011).

A Validade de Grupo Conhecido, segundo Zambaldi, Costa e Ponchio (2014), permite analisar o comportamento esperado de uma medida em relação a grupos ou variáveis categóricas específicas. Essa estratégia é pouco utilizada em mensuração por múltiplos itens. Para explicar seu significado, os autores exemplificam:

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 de técnicas não paramétricas correspondentes (teste Kruskal-Wallis ou teste de Wilcoxon-Mann-Whitney). Assim, caso os dados se comportem como esperado e com a indicação destes testes, é possível assegurar, ou não, a validade de grupo conhecido (ZAMBALDI; COSTA; PONCHIO, 2014).

6 MODELOS DE CONSTRUÇÃO DE ESCALAS

Há mais de três décadas atrás, Gilbert Churchill desenvolveu um paradigma operacional para a construção de escalas de mensuração, que é considerado um dos principais marcos da pesquisa em marketing (LADEIRA, 2010). Este capítulo tem como objetivo apresentar este clássico modelo de Churchill, explorando as etapas propostas de desenvolvimento de escalas de mensuração, assim como discutir algumas de suas críticas, encontradas na literatura. Por fim, entre as diversas propostas de desenvolvimento de escalas atuais, será apresentado o passo a passo do modelo proposto por Vieira (2011), que na opinião deste autor, pode ser considerado um modelo bastante completo no que se refere à discussão dos procedimentos e técnicas estatísticas que podem ser empregados em cada etapa para a avaliação da confiabilidade, dimensionalidade e validade estatística. Por fim, ainda é apresentado uma breve discussão do modelo proposto por Hair *et al.* (2009), que é um dos modelos mais citados em artigos de escala no Brasil.

6.1 Modelo clássico de Churchill (1979)

Fazendo uma crítica sobre a baixa qualidade das medidas utilizadas pelos pesquisadores de marketing da época para calcular suas variáveis de interesse, Churchill (1979) propôs uma série de procedimentos para o desenvolvimento de melhores medidas dos construtos desse campo do conhecimento. De acordo com o autor, esses procedimentos deveriam passar pelas seguintes etapas: (1) especificar o domínio do construto; (2) gerar amostra de itens; (3) coletar dados; (4) purificar a medida; (5) coletar dados; (6) calcular confiabilidade; (7) calcular a validade; (8) desenvolver normas. Essa sequência de procedimentos, ainda segundo o autor, funcionou muito bem em diversas instâncias para produzir medidas com propriedades psicométricas desejáveis. Estas etapas podem ser melhor visualizadas na figura 6, e serão discutidas a seguir.

Esse processo se inicia com a precisa definição do domínio do construto. O pesquisador deve ser exato no delineamento do que está incluído na definição e do que está excluído. É imprescindível que os pesquisadores consultem a literatura quando conceituam construtos e definam domínios.

O passo seguinte é a geração de uma amostragem de itens. Para tanto, o pesquisador deve (1) utilizar a literatura para fornecer uma visão do número de dimensões e componentes que o construto possui; (2) desenvolver pesquisas com pessoas que possuem experiência no

assunto estudado e que podem dar ideias sobre o tema; (3) comparar produtos concorrentes e/ou examinar cartas de clientes; (4) utilizar a técnica de incidente crítico e grupos focais, no qual podem ser criados diferentes cenários descrevendo situações específicas. Então, uma amostra de clientes com experiência no assunto pode participar de um grupo focal, com o objetivo de promover a discussão sobre essas experiências.

Figura 6: Passos para o desenvolvimento de escalas – Modelo Churchill (1979)

Fonte: Churchill (1979)

A seguir, é realizada a primeira coleta de dados (etapa 3), visando promover a purificação dos itens selecionados (etapa 4). Na purificação (etapa 4), deve-se avaliar quais itens possuem baixa correlação inter-item e que talvez não estejam apropriadamente descrevendo o domínio pesquisado. Para tal fim, Churchill (1979) sugere a utilização do coeficiente alpha e a análise fatorial.

Após a escala ser purificada, uma segunda coleta de dados (etapa 5) é realizada com o intuito de finalmente averiguar e confirmar sua confiabilidade (etapa 6) e sua validade (etapa 7). De acordo com Hair *et al.* (1998), a confiabilidade mede o grau de consistência entre

múltiplas medidas de uma variável. É o grau em que as medidas observadas medem o valor “verdadeiro” e é livre de erro. Já a validade é a extensão em que uma medida ou um grupo de medidas representa corretamente o conceito do estudo – o grau em que ela está livre de qualquer erro sistemático ou não-aleatório. Está relacionada com quão bem um conceito está definido pela(s) medida(s). Em outras palavras, validade é o grau em que uma medida corretamente representa aquilo que deve supostamente medir.

Com a escala devidamente validada e sua confiabilidade atestada, deve-se então desenvolver normas (etapa 8) para permitir que um escore individual possa ser interpretado dentro da escala criada. Isso implica em definir médias amostrais e outras estatísticas que sumarizam a distribuição dos escores e forneçam parâmetros para comparações entre escores na escala.

Ressalta-se ainda que esse processo sugerido de oito etapas para o desenvolvimento de escalas multi-item pode requerer reavaliações de etapas anteriores, conforme os fluxos da figura 1 indicam. Por exemplo, se após a purificação da escala o pesquisador ainda não obtiver itens confiáveis, ele deve voltar à etapa 2 para rever a amostragem de itens ou ainda voltar à etapa 1 para reconsiderar a definição do construto pesquisado.

6.2 Críticas ao modelo de Churchill

Com o passar dos anos, o artigo de Churchill (1979) publicado no *Journal of Marketing Research* se tornou a base de apoio dos pesquisadores de marketing para o desenvolvimento de escalas, se tornando um “paradigma” nesse campo da ciência. Para se ter uma ideia desse fenômeno, este clássico artigo já recebeu mais de 3200 citações, segundo pesquisa realizada no *Web of Knowledge* (2014).

Na perspectiva da teoria do conhecimento, contestações, exceções e particularizações são elementos naturais da evolução dos paradigmas, e o modelo de Churchill (1979) não fugiu a essa regra (COSTA, 2011). Embora Churchill (1979) tenha se baseado em um conjunto bem delimitado e fundamentado de pressupostos, estes foram perdendo sentido ao longo dos anos e gerado contestações diversas de vários autores. Entre as críticas, é possível citar (i) a questão da dimensionalidade; (ii) a orientação para construtos refletivos; e (iii) a questão das escalas de múltiplos itens, descritas nos subtópicos 6.2.1, 6.2.2 e 6.2.3.

6.2.1 A questão da dimensionalidade

Um primeiro tópico de discussão que surgiu a respeito do paradigma de Churchill (1979) foi em relação à dimensionalidade do construto. Churchill (1979) partiu de um entendimento de que procedimentos de revisão da literatura, juntamente com as técnicas de análise fatorial exploratória e a análise de confiabilidade, seriam o suficiente para se avaliar a dimensionalidade de construtos. Esse construto poderia ser então estimado a partir de uma medida geral agregada, que poderia ser proveniente de uma combinação dos escores das variáveis, para cada respondente, seja pela soma ponderada dos escores ou por sua média, dependendo da decisão do pesquisador.

No entendimento de Gerbing e Anderson (1988), tal procedimento demandaria uma análise mais profunda em termos da certificação da unidimensionalidade do construto. De acordo com os autores, os processos sugeridos pelo modelo de Churchill eram falhos em certificar-se da condição da unidimensionalidade do construto, uma vez que o modelo utilizava procedimentos de natureza exploratória, como a verificação da confiabilidade (alpha de Cronbach) e a extração de resultados por meio da análise fatorial exploratória, que não poderiam por si só atestar essa condição.

Em relação a essa discussão, Costa (2011) lembra que a análise fatorial é uma técnica que produz resultados indicativos de uma estrutura fatorial, mas que não assegura existir apenas uma dimensão subjacente aos dados, a qual só é de fato atestada pelo uso da técnica de análise fatorial confirmatória:

De fato, os procedimentos de análise fatorial são apenas indicações de uma estrutura fatorial, mas não asseguram que apenas uma dimensão é subjacente a um conjunto de variáveis. Tal condição somente é assegurada por meios da técnica de ‘análise fatorial confirmatória’, método que passou a ser mais amplamente usado a partir dos anos de 1980, com o avanço teórico e a disponibilidade de *softwares* para realizarem as complicadas operações associadas à técnica, havendo também a contribuição da popularização da técnica complementar de ‘modelagem de equações estruturais’ (COSTA, 2011, p.50).

Como uma proposta de atualização do modelo de Churchill, Gerbing e Anderson (1988) propõem a análise da “consistência externa” da escala, um conceito criado pelos autores para se referir à adesão de uma dada variável a um dado construto ou dimensão de um construto, e não a outro diferente. De acordo com os autores, apenas depois da verificação da consistência externa, em complementariedade aos procedimentos de consistência interna sugeridos pelo modelo de Churchill, é que se poderiam então proceder à extração de medidas

agregadas, como também, testar as relações entre diferentes construtos em modelos estruturais.

6.2.2 A orientação para construtos refletivos

De acordo com Costa (2011), Churchill usa o pressuposto de que os construtos de seu modelo de construção de escalas devem ser do tipo refletivos, não contemplando métodos e procedimentos que especificamente viabilizem o desenvolvimento de escalas baseadas em construtos refletivos, os quais nos dias atuais têm ganhado relevância:

[...] os construtos do tipo formativos são amplamente presentes no dia a dia da pesquisa e da prática gerencial, até mais do que se pensou durante mais de 20 anos após a proposta de Churchill. Esta foi, provavelmente, a discussão mais relevante sobre desenvolvimento de escalas nos anos 2000, em especial na dimensão de seleção e análise de itens de mensuração (COSTA, 2011, p. 51).

O fato é que a maioria dos pesquisadores em Ciências Sociais utiliza construtos refletivos em seus estudos, em detrimento aos construtos formativos, muito embora este último possa ser adequado em muitos casos (BOLLEN, 1989). Para Diamantopoulos e Winklhofer (2001), tal negligência é certamente o caso encontrado em marketing, porque praticamente todas as medidas disponíveis são baseadas implícita ou explicitamente em indicadores refletivos, citando, por exemplo, as compilações de Bearden, Netemeyer e Mobley (1993) e Bruner e Hensel (1992, 1996).

A discussão sobre construtos refletivos ou formativos é de fato uma discussão bastante relevante, pois remete aos procedimentos utilizados para o desenvolvimento da escala. Como revela Mackenzie, Podsakoff e Jarvis (2005), esta é uma distinção extremamente importante, porque muitos dos procedimentos de desenvolvimento de escala recomendadas na literatura só se aplicam a construtos com medidas refletivas, e se eles forem aplicados a construtos com medidas formativas, podem minar a validade de construto.

Por exemplo, a maioria dos textos sobre os processos de desenvolvimento de escala recomenda que os itens que possuem baixas correlações com a escala como um todo (*item-to-total correlation*) deveriam ser retirados da mesma para melhorar a confiabilidade da consistência interna. Embora esta recomendação é adequada no caso de indicadores refletivos, pois os itens são amostrados a partir do mesmo domínio de conteúdo, se esta recomendação for seguida por construtos com indicadores formativos, pode resultar na eliminação

precisamente daqueles itens que são mais propensos a alterar o sentido empírico e conceitual do construto (MACKENZIE; PODSAKOFF; JARVIS, 2005).

Esse problema também é apontado por Costa (2011), que discute que no esforço de purificação da escala através do uso de análise fatorial e da análise de confiabilidade, alguns itens são normalmente eliminados e outros mantidos, com base em suas características psicométricas. Entretanto, “excluir itens pode representar perda de representação de facetas relevantes do construto, e a manutenção de alguns itens pode ser um artifício de elevação do alpha de Cronbach ou do escore fatorial por meio de itens redundantes” (COSTA, 2011, p.53).

Como forma de complementar as orientações existentes em desenvolvimento de escalas com algumas sugestões sobre a construção de índices, Diamantopoulos e Winklhofer (2001) propuseram um modelo específico para tratar os construtos formativos, permitindo, assim, que os pesquisadores possam chegar a uma escolha mais embasada quanto ao modelo de medição adequado para as suas necessidades. Pela abordagem detalhada que fizeram do assunto, o estudo desses autores tornou-se um dos mais citados do *Journal of Marketing Research* (COSTA, 2011; WEB OF KNOWLEDGE, 2014).

De acordo com Costa (2011, p.51), “os problemas de uma abordagem de mensuração com itens formativos ou refletivos estão ainda longe de ser superados, e, não sem razão, o reputado *Journal of Business Research* dedicou um número especial ao tema no ano de 2008”.

6.2.3 A questão das escalas de múltiplos itens

De acordo com Churchill (1979), medidas individuais de item único possuem uma série de dificuldades as quais tornam as medidas de múltiplos itens mais vantajosas para uso em mensuração. Conforme o autor, essas limitações são:

1. Itens individuais frequentemente possuem considerável singularidade ou especificidade em que cada item tende a ter uma baixa correlação com o atributo que está sendo medido, como também tende a estar correlacionado com outros atributos;
2. Itens isolados tendem a classificar as pessoas em um número relativamente pequeno de grupos, não as diferenciando bem (por exemplo, numa escala de 5 pontos, é possível diferenciar os respondentes em apenas 5 grupos);

3. Itens individuais tipicamente possuem considerável erro de mensuração, produzindo respostas não confiáveis no sentido de que é improvável que a mesma posição da escala seja marcada em administrações sucessivas de um instrumento.

A utilização de uma escala com múltiplos itens supera essas dificuldades na medida em que um maior número de itens gera uma maior cobertura das diferentes facetas de um construto, além de viabilizar uma maior diferenciação entre os respondentes e elevar a confiabilidade da escala (COSTA, 2011).

Entretanto, estas constatações só poderiam ser verdadeiras se o construto possuísse efetivamente múltiplas facetas e que houvesse a necessidade de diferenciar os escores dos respondentes por parte dos pesquisadores. Porém, um problema que surge é quando se criam muitos itens redundantes, que elevam artificialmente os coeficientes alpha, não incorporando de fato as diferentes facetas do construto. Este problema é levantado por Costa (2011, p.52): “o risco, por outro lado, é que na agregação de itens apareçam alguns deles redundantes ou meras paráfrases de itens já existentes”.

Nos dias atuais, o uso de escalas de múltiplos itens é a prática padrão na pesquisa de marketing acadêmica, que se reflete tanto em monografias de desenvolvimento de escala relevantes como em vários manuais contendo compilações de medidas de marketing (DIAMANTOPOULOS *et al.*, 2012).

Entretanto, recentemente Bergkvist e Rossiter (2007, p. 183) desafiaram esta “convenção” em ambos os fundamentos teóricos e empíricos ao concluírem que os “testes teóricos e resultados empíricos seriam inalterados se boas medidas de item único fossem substituídos no lugar das comumente utilizadas medidas de múltiplos itens”. Do ponto de vista teórico, os autores se embasaram no procedimento de desenvolvimento de escalas desenvolvido por Rossiter (2002) chamado de C-OAR-SE, segundo o qual se o objeto do construto pode ser conceituado como concreto e singular e se o atributo do construto pode ser designado como concreto, não haveria a necessidade de se utilizar uma escala de múltiplos itens para operacionalizar o construto. Do ponto de vista empírico, Bergkvist e Rossiter (2007, 2009) relataram resultados indicando que medidas de um único item apresentaram elevada validade preditiva como as escalas de múltiplos itens. Bergkvist e Rossiter (2009) então concluem que se as medidas de um único item forem cuidadosamente elaboradas, de construtos duplamente concretos, são pelo menos tão válidas quanto as medidas de múltiplos itens do mesmo construto, e que o uso de vários itens para medi-los seria desnecessário.

Contrastando aos resultados de Bergkvist e Rossiter (2007, 2009), a série de estudos empíricos realizados por Diamantopoulos *et al.* (2012) revelou que a validade preditiva de escalas de itens únicos varia consideravelmente entre diferentes construtos (concretos) e estímulos de objetos. Na tentativa de explicar a instabilidade observada, os autores realizaram um estudo de simulação abrangente com o objetivo de identificar a influência de diferentes fatores sobre a validade preditiva de medidas únicas versus multi-item. Os resultados apontaram que, na maioria das condições normalmente encontradas em aplicações práticas, as escalas multi-item claramente superam as escalas de item únicos em termos de validade preditiva, e, apenas em algumas condições muito específicas, os itens únicos conseguem ter uma performance igual às escalas multi-item. Os autores então concluem que as medidas de um único item na pesquisa empírica devem ser empregadas com cautela, e sua utilização se limitar às circunstâncias especiais.

6.3 Modelo de Vieira (2011)

Em seu livro *Escalas de Marketing*, Vieira (2011) discute oito passos para a construção de novas escalas de medidas, conforme apresentado na Figura 7.

O primeiro passo para a construção da escala é a definição conceitual do construto a ser medido. Nesse passo, o pesquisador precisa entender exatamente o que é, conceitual e teoricamente, a variável medida no campo de estudo. Em outras palavras, representa a definição do construto.

O segundo passo do processo é a elaboração de indicadores para mensurar o construto, que pode ser realizado por meio de dois procedimentos. O primeiro é a revisão da literatura, procurando coletar indicadores que supostamente correspondem ao construto via dados secundários. Neste caso, a ideia é identificar itens prontos ou já publicados oriundos de outros trabalhos já realizados. O segundo procedimento sugerido é a realização de um painel incluindo especialistas acadêmicos e/ou profissionais e/ou consumidores para criação da lista geral de indicadores que correspondem ao fenômeno de interesse. Neste processo, muitas vezes é empregada uma pesquisa qualitativa para a geração dos indicadores, tendo como objetivo a criação do máximo número de itens.

Após a criação dos indicadores, o próximo passo (terceiro) é a validação do conteúdo dos indicadores criados. Esta etapa visa avaliar se os itens elaborados na etapa anterior representam adequadamente o domínio ou construto específico de interesse. Neste sentido, o objetivo da validação de conteúdo é consultar *experts* na área para avaliar se os indicadores

realmente representam o construto em questão. É sugerido consultar entre 4 até 10 profissionais, sendo que deve haver uma concordância de aproximadamente 70% entre eles para se considerar um item representativo do conceito em questão. Para tanto, são sugeridas as seguintes etapas: (1) definição do domínio de interesse (conceituação dos comportamentos que definem operacionalmente o construto em questão); (2) seleção de especialistas no domínio do conteúdo de interesse, sendo estes profissionais, acadêmicos ou pessoas que dominam o assunto o qual se está trabalhando; (3) preparação *a priori* de um quadro explicativo para o processo de emparelhamento ou equiparação dos itens com o domínio específico; e (4) a conclusão e resumo dos resultados do processo de emparelhamento ou equiparação dos itens com o domínio de interesse (CROCKER; ALGINA, 1986 *apud* VIEIRA, 2011).

Figura 7: Passos para o desenvolvimento de escalas – Modelo Vieira (2011)

Fonte: Vieira (2011)

No quarto passo (pré-teste), busca-se realizar coletas de dados referentes à escala gerada no passo anterior, com amostragem piloto (ou de teste) do tipo não probabilística. Nestes termos, são realizadas duas fases de análises exploratórias, com o objetivo de purificar os indicadores gerados no passo anterior. Esse refinamento da escala nada mais é do que analisar os dados obtidos por meio de um pré-teste, utilizando amostras de tamanho 130 a 200 (primeira amostragem) e amostras maiores (segunda amostragem). As análises clássicas do processo de refinamento são executadas por meio de análise fatorial exploratória e da análise de confiabilidade, em que se sugere a utilização dos seguintes testes e procedimentos:

- **Teste de adequação da amostra para análises fatoriais KMO** (Kaiser-Meyer-Olkin), sendo esperados valores KMO acima de 0,5;
- **Teste de esfericidade de Bartlett**, sendo esperados valores do p-valor < 0,05;
- **Análise das cargas fatoriais**: espera-se que as cargas fatoriais dos itens estejam dentro da dimensão hipotetizada. Alguns autores sugerem um mínimo de valor para a carga fatorial igual a 0,32 (TABACKNICK; FIDELL, 1996 *apud* VIEIRA, 2011) e 0,40 (COSTELLO; OSBORNE, 2005). Outro ponto que também deve ser observado são as diferenças das cargas fatoriais de um mesmo item nas dimensões, que precisam ser superiores a 0,1, isto é, espera-se que um item tenha uma carga fatorial maior em uma única dimensão e que as cargas fatoriais nas outras dimensões sejam menores, pelo menos, em 0,1.
- **Comunalidade**: representa a quantia total de variância que uma variável original compartilha com todas as outras variáveis incluídas na análise (HAIR *et al.*, 1998). O valor ideal da comunalidade para cada uma das variáveis (itens) é de 0,50.
- **Quantidade de fatores**: Vieira (2011) faz algumas críticas ao uso do critério “autovalor maior do que 1” para extração do número de fatores de uma análise fatorial exploratória. Baseado no estudo de Zwick e Velicer (1986), o autor sugere a utilização do método de “Análise Paralela”, que é um método que visa comparar os autovalores reais obtidos pela base de dados com autovalores randômicos artificiais gerados pela syntax. Este método se mostrou mais eficiente na estimação do número de fatores em simulações de Monte Carlo realizadas no estudo de Zwick e Velicer (1986).

- **Análise do Coeficiente Alfa de Cronbach:** é uma análise realizada com o objetivo de excluir indicadores cuja retirada acarretaria um aumento do Coeficiente Alpha de Cronbach da dimensão correspondente. Sugere-se a verificação da matriz de correlação entre as variáveis e a opção “se o item for eliminado”. Conforme o autor, deve-se evitar correlações muito altas entre as variáveis, o que poderia indicar uma possível colinearidade, o que aumentaria o valor do alfa artificialmente. Caso a escala medida seja considerada métrica, o valor de corte do Coeficiente Alpha sugerido pelo autor como ideal é de 0,70. Porém, o autor ainda lembra que este valor pode variar de acordo com o objetivo do estudo: para estudos exploratórios, um valor de 0,60 pode ser considerado satisfatório (HAIR *et. al.*, 1998); para estudos conclusivos, um valor acima de 0,80 é sugerido (NUNNALLY; BERSTEIN, 1994 *apud* VIEIRA, 2011).

No quinto passo (estudo oficial), procura-se obter uma amostra grande e expressiva da população-alvo para exame do instrumento, com o intuito de verificar se os resultados do pré-teste se repetem no estudo oficial. As análises são as mesmas sugeridas no passo anterior, mas com a inclusão das análises de validação cruzada e análise fatorial confirmatória:

- **Validação cruzada:** sugere-se dividir a amostra aleatoriamente e realizar a análise fatorial exploratória novamente nas amostras, pressupondo que o resultado será idêntico ao feito com a amostra total. Esta técnica permite examinar se a estrutura identificada se repete quando investigada em uma segunda amostra criada artificialmente (LAROS; PUENTE-PALACIOS, 2004 *apud* VIEIRA, 2011). Conforme Kline (1998) *apud* Vieira (2011), a utilização da validação cruzada é importante tanto para soluções fatoriais exploratórias como para soluções fatoriais confirmatórias, sendo que neste último é empregada a análise de invariância em multigrupos.
- **Análise fatorial confirmatória:** nesta análise é interessante avaliar as cargas fatoriais confirmatórias dos indicadores para com o construto latente de interesse. De acordo com Bagozzi e Philips (1982) *apud* Vieira (2011), a análise fatorial confirmatória pode ser utilizada para avaliar a validade dos indicadores dos construtos, devendo ser verificada a significância das cargas fatoriais dos construtos ao nível de 5% (t crítico de 1,65) ou 1% (t crítico de 2,236).

O passo seis no processo de desenvolvimento de escalas é a realização do estudo de validade nomológica, “[...] que objetiva verificar como o instrumento proposto se comporta quando associado à outras escalas” (VIEIRA, 2011, p.20). Procura-se verificar se existe correlação (forte, média ou fraca) ou não entre o construto e outras medidas concorrentes. Para tanto, o pesquisador precisa rever os conceitos das variáveis a serem relacionadas e elaborar uma hipótese de associação: caso a teoria sugerir que elas devem estar associadas positivamente e os testes empíricos indicarem que estas medidas estejam correlacionadas entre si, então pressupõem-se que existe uma **validade convergente**; se a teoria sugerir que elas não devem estar associadas, pois são variáveis distintas, e os testes na amostra empírica indicarem que não estejam correlacionadas, pressupõem-se então que existe uma **validade discriminante**.

Em relação à avaliação da validade convergente, Vieira (2011) atenta ao fato de que o ponto-chave aqui é o de analisar o construto com outros construtos e não com seus próprios indicadores, como fazem alguns pesquisadores, que utilizam o procedimento proposto por Bagozzi e Phillips (1982) de análise fatorial confirmatória como mecanismo que atesta a validade convergente. Desta maneira, o ideal para se testar a validade convergente, segundo o autor, é aplicar a medida e associa-la com outras concorrentes, acompanhando o grau de correlação. O resultado esperado é uma baixa associação dentre elas, sendo correlações menores do que $\pm 0,75$, mas maiores do que $\pm 0,30$.

Para a avaliação da validade discriminante, Vieira (2011) cita 4 procedimentos que podem ser utilizados:

- **Análise Fatorial Exploratória:** para testar a validade discriminante, o pesquisador pode aplicar a escala em estudo em conjunto com outras medidas e realizar uma análise fatorial oblíqua dos resultados, isto é, forçando a associação entre os construtos, verificando se foram formadas estruturas distintas, sendo todas diferentes e com baixo grau de associação (menor do que 0,80). Neste caso, é possível dizer que há validade discriminante.
- **Método de Fornell e Larcker (1981):** para o desenvolvimento deste método, dois cálculos são necessários. Primeiramente, calcula-se a matriz de correlação dos construtos ao quadrado (r^2), que representa a variância compartilhada; depois, realiza-se o cálculo da AVE (*Average of Variance Extracted*) – variância média extraída, a qual mensura a quantidade de erro no construto. Comparando-se a

variância média extraída (AVE) com o quadrado da correlação dos construtos (r^2), se (r^2) for menor do que AVE, afirma-se que existe validade discriminante.

- **Matriz de multitração-multimétodo (MTMM):** o objetivo deste procedimento é testar a convergência entre diferentes medidas do mesmo conceito e, em um segundo momento, examinar diferenças entre medidas de conceitos distintos relacionados, mas conceitualmente diferentes (COOK; CAMPBELL, 1979 *apud* VIEIRA, 2011).
- **Método de Bagozzi e Philips (1982):** por este método, calcula-se a diferença de qui-quadrado entre dois modelos, o primeiro em que a correlação entre os dois modelos estruturais (via estruturas de covariância) é fixada em 1 e outro no qual ela é deixada livre. Considerando o valor obtido para o qui-quadrado do primeiro modelo sendo x e do segundo y , se a diferença entre esses dois modelos $x-y$ for significativa ($p\text{-valor} < 0,05$), então há validade discriminante.

O sétimo passo no desenvolvimento da escala é a realização de um estudo para testar a validade de grupos. Este passo tem como objetivo “entender a extensão na qual altos e baixos escores na medida diferem grupos quanto ao quesito aferido” (VIEIRA, 2011, p.23). A ideia por trás da validade de grupos é testar as diferenças de escore da escala entre grupos que o pesquisador possui conhecimento a priori que sejam efetivamente diferentes. De acordo com Vieira (2011, p.23), “a validade de grupo é extremamente relevante, pois mostra o poder do instrumento em diferenciar grupos que são supostamente diferentes”.

O oitavo e último passo na criação da escala é a realização de um estudo de desejo de aceitação social. De acordo com Vieira (2011), os pesquisadores da área de marketing estão constantemente utilizando técnicas de coleta de dados que dependem de autorrelato, que por sua vez, pode ser enviesado pelo efeito do desejo de aceitação social. Conforme revela Ballard, Crino e Rubenfeld (1988) *apud* Vieira (2011, 24), “a validade dessas percepções é essencial no estudo do marketing e comportamento do consumidor, onde o efeito do desejo de aceitação social (*social desirability*) nas respostas é uma fonte de preocupação”. Uma escala de aceitação social, como a de Marlowe-Crowne, pode ser utilizada para medir a aceitação social e depois correlacionada com a nova escala criada. Neste sentido, espera-se que a escala proposta não tenha correlação com a aceitação social ($p\text{-valor}$ seja não significante). Vieira (2011) cita uma versão traduzida da escala de Marlowe-Crowne, que pode ser encontrada no trabalho de Scagliusi *et al.* (2004).

6.4 Modelo proposto por Hair *et al.* (2009)

Para Hair *et al.* (2009), o desenvolvimento de uma escala de múltiplos itens envolve 6 passos. A seguir, será tecida uma breve síntese deste modelo.

No **primeiro passo**, deve-se definir teoricamente o construto. O pesquisador deve descrever de forma clara uma sentença concisa, que defina o que será medido. O **passo 2** envolve desenvolver uma lista de potenciais itens de escala que correspondam à definição no passo 1. Os itens devem ser suficientes em termos de quantidade para conseguir capturar completamente o domínio do assunto. Além disso, o pesquisador deve determinar qual o tipo de escala será usado para a medição do item, ou seja, se será em formato Likert, diferencial semântico, alguma combinação ou em um formato alternativo.

No **passo 3**, deve-se julgar os itens quanto a conteúdo. Esse trabalho deve ser feito por vários juízes (peritos) e envolve a avaliação de quão precisa está a definição do construto e o quanto os itens correspondem a ele. Deve-se solicitar aos peritos que classifiquem cada item de escala dentre as seguintes opções: a) não corresponde ao construto; b) corresponde de alguma forma ao construto, c) corresponde muito bem ao construto. Alternativamente, nos casos em que estejam sendo julgados múltiplos construtos ao mesmo tempo, podem ser solicitados aos peritos que ordenem o conjunto de itens, relacionando-os às definições apropriadas. De um modo geral, “independente do número de juízes utilizados, itens que recebem menos do que 50% de concordância são candidatos para exclusão da análise” (HAIR *et al.*, 2009, p. 594).

Um outro aspecto importante de verificação nessa etapa diz respeito a análise dos itens quanto à redundância. Se, por exemplo, dois itens são tão semelhantes a ponto de dizerem a mesma coisa, um item deve ser eliminado, o que ajuda a evitar problemas posteriores com os resultados empíricos.

O **passo 4** implica em conduzir um pré-teste para avaliar os itens. O pré-teste deve ser regido para uma amostra que deve responder analogamente às amostras sobre as quais a escala eventualmente será aplicada.

- Após a coleta de dados, deve-se analisar estatísticas descritivas quanto à curtose ou à assimetria significantes. Para Hair (2009, p. 594), “escalas com médias próximas de seus valores mínimo ou máximo podem estar sofrendo efeitos de “pisso” ou “teto”, e são candidatas para modificação ou eliminação”.

- Podem ser criadas escalas múltiplas e, em seguida, podem ser analisadas as correções entre itens e total e entre item e item. “Itens com correlações entre itens e total e entre itens e itens inferiores a 0,5 e 0,3, respectivamente, são candidatos à eliminação” (HAIR *et al.*, 2009, p.594).
- A análise fatorial exploratória pode ser realizada para fornecer uma verificação preliminar sobre o número de fatores e o padrão de cargas. Se é esperado que os itens de carga indiquem uma só dimensão, evidências escorando uma solução de um fator devem ser examinadas, por exemplo, por meio de um critério de raiz latente (autovalor >1) ou uma abordagem semelhante. Se mais fatores são esperados, então a análise fatorial deve exibir alguma evidência de que este número de fatores pode ser sustentado. Torna-se essencial examinar as cargas fatoriais individuais. “Os itens com cargas baixas são candidatos à eliminação” (HAIR *et al.*, 2009, p.594).

No passo 5, são realizadas modificações na escala com base nos resultados anteriores. O pesquisador deve considerar se existe um número adequado de itens para cada fator, antes de dar continuidade. Um construto deve ser refletivo por um mínimo de três itens, preferencialmente quatro. De forma recíproca, nenhum máximo teórico põe limite no número de itens por fator; no entanto, escalas com mais de uma dúzia de itens tornam a análise mais complexa e, frequentemente, conduzem a problemas no fornecimento de evidências de unidimensionalidade.

Finalmente no passo 6, prossegue-se com um teste confirmatório da teoria da mensuração. Neste caso, é verificado o ajustamento do modelo de mensuração, avaliadas a significância das cargas fatoriais, do percentual médio de variância extraída (AVE), dos valores de confiabilidade (alpha, confiabilidade composta) e verificada a validade discriminante. Também pode ser testada a relação do construto com outros dentro de uma teoria de mensuração, a fim de garantir a validade nomológica da medida.

7 METODOLOGIA DE PESQUISA

Para maior compreensão do desenvolvimento desse trabalho, foi elaborada a Figura 8, que apresenta a sequência da estrutura metodológica empregada e que será discutida nos próximos subtópicos.

Figura 8: Estrutura do capítulo de metodologia da tese.

Fonte: Autor (2014)

No tópico 7.1, define-se o tipo de pesquisa empregada. Na sequência, no tópico 7.2, discute-se o método utilizado e, no 7.3, aborda-se as etapas metodológicas adotadas. Finalmente, no tópico 7.4 tem-se a síntese da metodologia do estudo.

7.1 Delineamento da pesquisa

De acordo com Dane (1990), a revisão bibliográfica é importante para definir a linha limítrofe da pesquisa que se deseja desenvolver, e pode ser narrativa ou sistemática (COOK; MULROW; HAYNES, 1997). O primeiro tipo é baseado em uma descrição simplificada de estudos e informações sobre um determinado assunto; já o segundo tipo, apesar de também ter caráter narrativo, é fundamentado na aplicação de métodos com maior rigor científico,

podendo alcançar resultados mais precisos e reduzir erros e o viés do pesquisador responsável pela investigação.

Em recente estudo, Botelho, Cunha e Macedo (2011) discutem as diferenças desses dois tipos de revisão. Para os autores, a revisão narrativa é empregada para descrever o “estado da arte” de um assunto específico, sob o ponto de vista teórico ou contextual, e constitui-se basicamente da análise da literatura, da interpretação e análise crítica pessoal do pesquisador. Esse tipo de revisão não fornece a metodologia para a busca das referências, nem as fontes de informação utilizadas, ou os critérios usados na avaliação e seleção dos estudos em análise. Por outro lado, a revisão bibliográfica sistemática é planejada para responder a uma pergunta específica, utilizando-se de métodos explícitos e sistemáticos para identificar, selecionar e avaliar criticamente os estudos, e para coletar e analisar dados dos artigos incluídos na revisão.

No caso específico de pesquisas avançadas, onde se exige certo ineditismo e originalidade, Webster e Watson (2002) argumentam que a revisão bibliográfica desempenha um papel essencial, e, por essa razão, conduzi-la de forma sistemática e rigorosa, contribui para o desenvolvimento de uma base robusta de conhecimento, facilitando o desenvolvimento da teoria em áreas onde já existem estudos, ou identificando outras com oportunidades para novas pesquisas (WEBSTER; WATSON, 2002).

Rother (2007) chama a atenção ao fato dos trabalhos de revisão bibliográfica sistemática serem considerados “originais”, pois, além de utilizarem como fonte dados da literatura sobre determinado tema, são elaborados com rigor metodológico.

O estudo em questão se trata de uma **Revisão Bibliográfica Sistemática da Literatura**, ou seja, um estudo descritivo, bibliográfico e documental, que segundo Conforto e Amaral (2011), é um dentre os diversos tipos de delineamento de pesquisa e fundamenta-se na busca e análise de artigos de uma determinada área da ciência.

Reconhecida por ser metódica, transparente e replicável, a revisão bibliográfica sistemática é amplamente utilizada na área de saúde (COOK; MULROW; HAYNES, 1997), como forma de obter evidências para dar suporte ao aumento das intervenções e informações científicas (EVANS; PEARSONS, 2001), e mais recentemente começa a ganhar adeptos nas pesquisas em gestão (TRANFIELD; DENYER, 2003; CONFORTO; AMARAL, 2011).

Assim como outros tipos de estudo de revisão, a revisão bibliográfica sistemática é uma forma de pesquisa que utiliza como fonte de dados a literatura sobre determinado tema (SAMPAIO; MANCINI, 2007), constituindo em um método moderno para avaliação de um conjunto de dados simultaneamente (ATALLAH; CASTRO, 1998). Sampaio e Mancini

(2007) ainda destacam que a revisão bibliográfica sistemática permite o estudo do retrospectivo e secundário, no qual a revisão é usualmente desenhada e conduzida após a publicação de muitos estudos experimentais sobre um tema particular, sendo dependente da qualidade da fonte primária.

No caso das pesquisas na área das ciências sociais, Mendes e Fracoli (2008) salientam que a utilização da revisão bibliográfica sistemática justifica-se pela necessidade do emprego de métodos que possibilitem trazer à tona as contradições, transformando-as em conhecimento.

Em seu estudo, Berwanger *et al.* (2007, p.481) argumentam que as revisões narrativas, são simples revisões da literatura, formatadas semelhantemente a capítulos de livro e que geralmente abordam temas de forma ampla e sem metodologia de seleção pré-estabelecida. Já as revisões bibliográficas sistemáticas são investigações científicas, que “possuem metodologia definida ‘a priori’ por um protocolo, em que é definido o passo a passo das etapas metodológicas”.

Para Sampaio e Mancini (2007), as revisões bibliográficas sistemáticas são particularmente úteis para integrar as informações de um conjunto de estudos realizados separadamente sobre determinado assunto/tema, que podem apresentar resultados conflitantes e/ou coincidentes, bem como identificar temas que necessitam de evidência, auxiliando na orientação para investigações posteriores.

Conforto e Amaral (2011, p.2) argumentam que se a pesquisa bibliográfica for conduzida com rigor e de forma sistemática, permitirá que outros pesquisadores possam fazer uso desses resultados com maior confiabilidade, possibilitando complementar estudos já finalizados, focando apenas no tópico em que se deseja pesquisar. Além da economia de tempo e recursos, os resultados de uma revisão bibliográfica sistemática podem identificar, segundo os autores, “lacunas a serem exploradas por outros pesquisadores, mas que não foram identificadas em estudos semelhantes devido à superficialidade e falta de rigor na revisão bibliográfica”. Complementando esse ponto de vista, Berwanger, Buehler e Oliveira (2007) reforçam que as revisões bibliográficas sistemáticas têm grande importância como ferramenta de investigação científica para a tomada de decisão, com custos bem menores do que os exigidos para a realização de estudos de larga escala.

7.2 Método de pesquisa

Para Botelho, Cunha e Macedo (2011), os artigos de revisão utilizam fontes de informações bibliográficas ou eletrônicas para obtenção de resultados de pesquisa de outros autores, buscando fundamentar teoricamente um determinado tema.

Conforme abordado no tópico 7.1, duas categorias de artigos de revisão são encontradas na literatura: as revisões narrativas e as sistemáticas. Esta última pode ser subdividida em quatro métodos (WHITEMORE; KNAFL, 2005 *apud* BOTELHO; CUNHA; MACEDO, 2011; ZANDONAI, 2010; ROTHER, 2007), conforme observado na Figura 9:

Figura 9: Tipos de revisão da literatura.

Fonte: Botelho, Cunha e Macedo (2011, p. 125)

Os quatro métodos utilizam-se de mecanismos explícitos e sistemáticos para identificar, selecionar e avaliar criticamente os estudos (WHITEMORE; KNAFL, 2005 *apud* BOTELHO; CUNHA; MACEDO, 2011; ZANDONAI, 2010; ROTHER, 2007):

- **Meta-análise** – combina os resultados de vários estudos primários, e caracteriza-se pela aplicação de uma ou mais técnicas estatísticas (LOVATTO *et al.*, 2012). O delineamento e as hipóteses devem ser similares, se não idênticos, de modo a facilitar o processo de condução da pesquisa. Na meta-análise, cada estudo é sintetizado, codificado e inserido em um banco de dados quantitativo. Posteriormente, os dados encontrados são transformados em uma medida comum, que servirá de subsídio para a formulação de um cálculo que proporcionará a dimensão geral do efeito ou da intervenção mensurada. Nesse tipo de método as conclusões podem ser generalizadas

para um conjunto de estudos (FAGARD; STAESSEN; THIJS, 1996). Os resultados de uma meta-análise são apresentados de forma quantitativa, com análises estatísticas e intervalos de confiança. Estudos que não foram publicados ou que não apresentam tratamento estatístico tendem a ser excluídos de uma meta-análise, por correrem o risco de apresentar limitação metodológica (DOLLAGHAN, 2007).

- **Revisão sistemática** – tem como princípios a exaustão na busca dos estudos analisados, a seleção justificada dos trabalhos por critérios de inclusão e exclusão explícitos e a avaliação da qualidade metodológica, bem como a quantificação do efeito dos tratamentos por meio de técnicas estatísticas (GALVÃO; SAWADA; TREVIZAN, 2004). Esse tipo de revisão combina a evidência de múltiplos estudos, podendo incorporar resultados de uma pesquisa de meta-análise, como por exemplo, os dados estatísticos. A amostra é composta por estudos quantitativos, de metodologia similar. Com base em sua síntese rigorosa, busca evitar e superar possíveis vieses que o pesquisador possa ter no momento da análise da literatura sobre um tema. Esse método permite limitar o viés de seleção de artigos, e avaliá-los com espírito crítico, sintetizando todos os estudos relevantes em um tópico específico.
- **Revisão qualitativa** – bastante empregada em pesquisas nas últimas décadas, sintetiza exclusivamente os estudos primários qualitativos, podendo diferir em abordagens e níveis de interpretação. Metassínteses, metaestudos, *grounded theory* e a metaetnografia são métodos que sintetizam os achados de estudos qualitativos individuais, transformando-os em ferramentas para a construção de novas teorias.
- **Revisão integrativa** – sintetiza resultados de pesquisas anteriores, ou seja, já realizadas, e mostra, sobretudo, as conclusões do corpus da literatura sobre um fenômeno específico (CROSSETTI, 2012). Tem sido descrita como a mais ampla das categorias de pesquisa bibliográfica sistemática, que tem como objetivo resumir o passado da literatura para fornecer uma compreensão mais abrangente de um fenômeno particular. Trata-se de uma revisão que abrange tanto pesquisas empíricas ou de literatura teórica ou ambas, dependendo do foco do estudo, assim como uma diversidade de fontes de dados, permitindo uma compreensão holística do tópico de interesse. Tem como foco traçar uma análise sobre o conhecimento já construído em pesquisas anteriores (qualitativas e/ou quantitativas) sobre um determinado tema. A integração desses resultados permite a elaboração de uma conclusão geral, e os achados poderão ser generalizados a outras populações com maior confiança.

O Quadro 10 faz um resumo dos tipos de revisão da literatura constantes na revisão bibliográfica sistemática.

Quadro 10: Tipos de revisão de literatura.

Tipo de Revisão e Exemplar	Definição	Propósito	Escopo	Amostra	Análise
Revisão Integrativa (REDEKER, 2000)	Um sumário da literatura, num conceito específico ou numa área de conteúdo, em que a pesquisa é resumida, analisada, e as conclusões totais são extraídas	Revisar métodos, teorias, e/ou estudos empíricos sobre um tópico particular.	Limitada ou ampla	Pesquisa quantitativa ou qualitativa; literatura teórica; literatura metodológica	Narrativa
Meta-Análise (CLEMMENS, 2001)	Um sumário de pesquisas passadas, que usa técnicas estatísticas para transformar descobertas de estudos com hipóteses idênticas ou relativas em uma medida comum e que calcula o efeito total, a magnitude do efeito, e efeitos de subamostras.	Avaliar e estimar o efeito de relacionamentos entre variáveis	Limitada	Pesquisa quantitativa de metodologia similar	Estatística
Revisão Sistemática (FORBES, 1998)	Um sumário de pesquisas passadas, que usa um objetivo e uma abordagem rigorosa de estudos com hipóteses idênticas ou relativas.	Sumarizar evidência concernente a um problema ou tema específico	Limitada	Pesquisa quantitativa de metodologia similar	Narrativa ou estatística
Meta-sumário Meta-síntese Teoria constituída formalmente Meta-estudo (BECK, 2002)	Um sumário de pesquisas passadas, que combina as descobertas de múltiplos estudos qualitativos.	Informar pesquisas ou práticas pela sumarização (resumo) de processos ou experiências	Limitada ou ampla	Pesquisa qualitativa	Narrativa

Fonte: adaptado de Botelho, Cunha e Macedo (2011, p. 128)

O método de pesquisa empregado nesse estudo é a **Revisão Integrativa**, pois tem como propósito sumarizar as práticas metodológicas utilizadas nos artigos publicados de desenvolvimento e/ou validação de escalas na área de marketing nos eventos da ANPAD nos últimos 3 anos. Este procedimento foi escolhido por possibilitar a síntese e análise do conhecimento científico já produzido sobre o tema investigado.

7.3 Etapas Metodológicas

O planejamento de uma pesquisa fundamentada na revisão bibliográfica sistemática deve ser cuidadoso para assegurar a validade de seus resultados (DE-LA-TORRE-UGARTE-GUANILO; TAKAHASHI; BERTOLOZZI, 2011).

Marconi e Mancini (2007) destacam que uma revisão sistemática envolve a avaliação da qualidade metodológica de cada artigo selecionado. Ainda, reforçam que é importante que os pesquisadores elaborem um protocolo de pesquisa, nele contido os seguintes itens: definição de como os estudos serão encontrados; critérios de inclusão e exclusão dos artigos; definição dos desfechos de interesse; verificação da acurácia dos resultados; determinação da qualidade dos estudos; e análise da estatística utilizada.

Em se tratando do método de revisão integrativa, Mendes e Fracolli (2008) reforçam ainda a necessidade de que todas as etapas a serem seguidas sejam descritas passo a passo.

O processo de elaboração da revisão integrativa encontra-se bem definido na literatura, entretanto, diferentes pesquisadores adotam formas específicas de subdivisão das etapas do processo de revisão sistemática (WHITEMORE; KNAFL, 2005 *apud* BOTELHO; CUNHA; MACEDO, 2011; BEYEA; NICOLL, 1998; BROOME, 2006; MARCONI E MANCINI, 2007; FONSECA, 2008; ZANDONAI, 2010; BOTELHO; CUNHA; MACEDO, 2011; CONFORTO; AMARAL, 2011). Este estudo seguiu o modelo desenvolvido por Botelho, Cunha e Macedo (2011) (Figura 10). Cada etapa nele descrita será discutida nos próximos subtópicos.

Figura 10: Etapas da Revisão Bibliográfica Integrativa

Fonte: Botelho, Cunha e Macedo, 2011

7.3.1 Etapa 1 - Identificação do tema e seleção da questão de pesquisa

A primeira etapa serve como norte para a construção de uma revisão bibliográfica sistemática integrativa e pode ser compreendida por quatro passos: definição do problema e formulação da pergunta de pesquisa, definição dos descritores ou palavras-chave, definição da estratégia de busca e definição das bases de dados.

Primeiramente torna-se essencial a definição do problema e a formulação de uma pergunta de pesquisa. Ambos foram discutidos e definidos no capítulo 1 dessa tese.

Posteriormente, Botelho, Cunha e Macedo (2011) apontam a necessidade da definição das palavras-chave, da estratégia de busca, bem como dos bancos de dados a serem utilizados, e reforçam que cada passo dado nessa etapa deve estar relacionado e ser sistemático.

Na presente pesquisa, o processo de procura de artigos utilizou como estratégia a busca por meio de palavras-chave nos títulos dos artigos da base de dados da Anpad, realizada diretamente no site desta instituição na internet. As palavras-chave selecionadas foram aquelas usualmente empregadas nos textos de validação de escalas, a saber: “escala”; “mensuração”; “validação”, “instrumento”, “medida”, “construto”, “medir”, “validando”.

A base de dados pesquisada inclui artigos que foram publicados nos congressos do EnAnpad – Encontro anual da Associação Brasileira dos Programas de Pós-Graduação em Administração (1997 a 2014), do EMA – Encontro da Divisão de Marketing da ANPAD (2004 a 2014), do EnAPG – Encontro de Administração Pública da ANPAD (2004 a 2014), do EnEPQ - Encontro de Ensino e Pesquisa em Administração e Contabilidade (2007 a 2013), do EnADI - Encontro de Administração da Informação (2007 a 2013) e do 3Es – Encontro de estudos em estratégia (2003 a 2013), todos conceituados no estrato E1 da avaliação QUALIS da CAPES (2009). Ressalta-se que a atual avaliação QUALIS de 2012 não mais inclui a classificação de congressos, sendo então utilizada a avaliação do triênio 2007-2009 para este tipo de publicação, sendo esta a última a conter a classificação de eventos.

Considerando todas as palavras chave pesquisados, o resultado da busca sistemática retornou um conjunto de 602 artigos, sendo as palavras instrumento (com 179 artigos) e escala (com 128 artigos) as palavras-chave que retornaram o maior número de artigos.

7.3.2 Etapa 2: Estabelecimento de critérios de inclusão e exclusão

Essa etapa pode ser compreendida pela busca nas bases de dados, visando identificar os estudos que serão incluídos na revisão sistemática integrada. Essa etapa depende muito da boa condução da etapa anterior, pois um problema muito amplo tenderá a conduzir a uma amostra demasiadamente diversificada. Frequentemente, a seleção de artigos inicia-se de forma mais ampla e afunila-se, já que o movimento de busca na literatura nem sempre é linear. É importante destacar ainda que os critérios de inclusão e exclusão podem sofrer uma reorganização durante o processo de busca dos artigos e durante a elaboração da revisão integrativa (BOTELHO; CUNHA; MACEDO, 2011).

Para esse estudo, a definição de inclusão ou não dos artigos na base de dados da pesquisa e sua posterior análise teve como ponto de partida os seguintes critérios:

1. A escala foi desenvolvida dentro da literatura de marketing ou comportamento do consumidor, ou que ainda seja relevante para esta área.
2. O objetivo principal, ou um dos objetivos principais dos autores do artigo foi o desenvolvimento de uma nova escala ou a validação de uma escala proposta inicialmente em outro idioma.
3. A escala desenvolvida no artigo deve ser caracterizada como do tipo multi-item, isto é, composta de no mínimo dois ou mais itens ou questões;
4. O artigo foi publicado entre os anos de 2012 e 2014;

O primeiro critério utilizado para a seleção de artigos para fazer parte da base de dados desta tese visou delimitar o escopo da presente pesquisa, isto é, a de pesquisar as escalas desenvolvidas e/ou validadas na área de marketing. Dado o grande número de escalas encontradas na área psicológica e em outras áreas afins, definiu-se como critério de inclusão apenas as escalas que foram desenvolvidas dentro da literatura de marketing ou comportamento do consumidor, ou ainda aquelas que fossem relevantes para estas áreas. Vale lembrar que um critério semelhante foi utilizado por Vieira (2011), quando o autor delimitou o seu processo de busca excluindo trabalhos de eventos nas áreas de turismo, comunicação, economia, psicologia e engenharia de produção.

O segundo critério abarcou o objetivo principal do artigo em avaliação. Dada a complexidade dos procedimentos necessários para se desenvolver uma escala de mensuração com qualidade, é razoável pensar que o desenvolvimento da escala em si seja o objetivo principal do artigo ou seja um dos objetivos principais do pesquisador. Logo, artigos que apenas utilizaram escalas emprestadas de outros pesquisadores e que não tiveram etapas exploratórias de desenvolvimento do instrumento não foram foco desta tese. Reforça-se assim, o foco e o interesse expresso por este pesquisador em avaliar todo o processo metodológico envolvido na criação e/ou validação de uma medida.

Como terceiro critério, definiu-se pela inclusão de artigos de desenvolvimento de escalas que são compostas por vários itens (*multi-item measures*), isto é, por dois ou mais itens. De acordo com Hair *et al.* (1998), uma escala multi-itens ou escala somada representa um método em que várias variáveis são combinadas juntamente para compor uma medida composta para representar um conceito. Na maioria dos casos, as variáveis separadas são

somadas e então seu escore total ou escore médio é utilizado como uma medida individual nas análises.

O argumento por trás da adoção deste terceiro critério advém da própria natureza dos construtos encontrados nas ciências sociais e comportamentais, onde a área de marketing está inserida. Dado sua natureza latente, estes construtos representam abstrações que só podem ser mensurados indiretamente, por meio de vários indicadores (facetas ou itens). Para Costa (2011, p. 43), “mesmo havendo algumas proposições de que muitos construtos deveriam ser mensurados com apenas um item, a grande maioria das escalas desenvolvidas em Administração para medir construtos latentes são direcionadas a procedimentos com múltiplos itens”. Este pensamento também é compartilhado por Jacoby (1978), que fez sérias críticas ao uso de medidas individuais. Dada a complexidade dos fenômenos pesquisados em marketing, o autor considerou impróprio o uso de medidas individuais como medidas de certos conceitos. Além disso, considerou inválido relacioná-las a outras variáveis para chegar a conclusões baseadas neste tipo de investigação.

Para Gardner (1995), o uso de escalas multi-item, ao invés de uma escala de item único, tem como principal vantagem a melhora da confiabilidade da medida. Churchill (1979) relata que isso ocorre porque o erro de mensuração diminui com o aumento do número de itens quando estes são combinados conjuntamente. Segundo Hair *et al.* (1998), o uso de vários indicadores permite aos pesquisadores especificarem de forma mais precisa as respostas desejadas, não depositando toda a confiança em apenas uma única resposta, mas ao invés disso, em um grupo de respostas que estão relacionadas entre si.

Além desse benefício, conforme pontua Churchill (1979), as escalas multi-itens possuem mais outras duas vantagens. Com a combinação de itens, é possível melhor distinguir os sujeitos avaliados do que escalas de item individuais. Por exemplo, uma escala de um único item medida por uma escala de cinco pontos poderia, no máximo, distinguir a característica medida entre 5 níveis. Já uma escala multi-item contendo 5 itens medidos em uma escala Likert de 5 pontos teria uma capacidade de distinção muito maior, entre 21 níveis (de um mínimo de 5 pontos ao máximo de 25 pontos). Outra vantagem é que as escalas multi-item podem atenuar, com a combinação dos itens, a singularidade e especificidade que caracterizam as escalas individuais, em que o item tende a ter apenas uma baixa correlação com o atributo que está sendo medido como também tende a estar correlacionado com outros atributos.

Finalmente, também deve ser enfatizado que as escalas multi-itens se mostraram superiores às escalas de item individuais na questão da validade preditiva. Baseado em

estudos que revelaram que a validade preditiva das escalas de itens individuais variavam consideravelmente entre diferentes construtos e objetos de estímulo, Diamantopoulos *et al.* (2012) compararam a performance das escalas multi-itens com as escalas individuais utilizando simulação de Monte Carlo. A conclusão dos autores revelou que, sob a maioria das condições tipicamente encontradas em aplicações práticas, as escalas multi-itens claramente sobrepujaram as escalas de item individuais em termos de validade preditiva. Apenas sob condições bastante específicas que as escalas de item individuais conseguiram ter um desempenho semelhante às escalas multi-itens. Por todas estas razões expostas aqui, escalas criadas contendo apenas um único item (*single-item measure*) não representaram o foco desta tese e não foram incluídas na base de dados da pesquisa.

E como quarto e último critério, como o objetivo desta tese é avaliar a qualidade dos instrumentos de mensuração propostos e/ou validados na área de marketing na atualidade, foram selecionados apenas os artigos publicados em um período mais recente, definidos entre os anos de 2012 e 2014.

7.3.3 Etapa 3: Identificação dos estudos pré-selecionados e selecionados

Essa etapa compreende a identificação dos estudos que foram incluídos. Para tanto, realiza-se a leitura criteriosa dos títulos, resumos e palavras-chave de todas as publicações localizadas pela estratégia de busca, para posteriormente verificar sua adequação aos critérios de inclusão na amostra do estudo. Quando o título, ou o resumo ou as palavras-chave não forem suficientes para definir sua seleção, busca-se a publicação do artigo na íntegra. A partir da conclusão desse procedimento, elabora-se uma tabela com os estudos pré-selecionados para o início da revisão integrativa (BOTELHO; CUNHA; MACEDO, 2011).

Desta forma, após a etapa de busca sistemática de artigos por palavras-chave, passou-se para a leitura do título (1º refinamento) e, em um segundo momento, a leitura do resumo e partes do corpo de cada uma das publicações (2º refinamento), tendo como intuito o de verificar primeiramente se o artigo estava relacionado à linha de desenvolvimento ou validação de escalas e se este tinha como tema a área de marketing.

O primeiro refinamento reduziu a base dos 602 artigos da etapa anterior para 289 artigos. Excluindo-se os artigos repetidos, isto é, aqueles que haviam sido selecionados mais de uma vez por diferentes palavra-chave, chegou-se ao final desta etapa de refinamento com um amostra de 210 artigos. Dentre as palavras-chave utilizadas, “escala”, “mensuração”, “validação” e “instrumento” foram aquelas que retornaram um maior número de artigos para

revisão final. A palavra-chave “instrumento”, apesar de retornar muitos artigos na primeira etapa (179), teve apenas 43 artigos selecionadas para a etapa posterior de refinamento, indicando talvez um uso mais genérico e/ou amplo desta palavra nos títulos de artigos. O mesmo pode ser verificado para as palavras-chave “medida” e “construto”. O Quadro 11 apresenta os resultados do primeiro processo de refinamento.

Quadro 11: Resultados da estratégia de coleta de dados empregada – 1º refinamento

Palavra-chave	Artigos retornados c/ palavra chave (Final Etapa 1)	Nº de Artigos selecionados para revisão (Etapa 3 – pós 1º refinamento)	Nº de Artigos selecionados para revisão (Etapa 3 – pós 1º refinamento) em %
Escala	128	91	71%
Mensuração	89	60	67%
Validação	84	54	64%
Instrumento	179	43	24%
Medida	64	19	30%
Construto	49	17	35%
Medir	5	3	60%
Validando	4	2	50%
Total	602	289	48%

Fonte: elaborado pelo autor (2014)

Numa segunda etapa de refinamento, cada um dos artigos selecionados na etapa anterior foi reavaliado, procurando verificar não apenas o título e resumo, mas também o corpo do artigo quando necessário, em especial os procedimentos metodológicos empregados, com vistas a examinar suas características com os critérios pré-definidos no protocolo de pesquisa. Após cuidadosa análise, 121 artigos se qualificaram para compor a base de artigos de escala na área de marketing.

Finalmente, numa terceira e última fase de refinamento, considerou-se apenas os artigos publicados dentro do período definido entre os anos 2012 e 2014, retornando um total de 28 artigos. A leitura completa dos artigos ainda permitiu excluir outros 4 artigos que não se encaixaram dentro dos critérios apresentados anteriormente (um caso em que o artigo tinha objetivo de realizar uma pesquisa bibliométrica das escalas existentes em uma determinada área; dois casos em que o artigo continha apenas 4 páginas e não era possível avaliar os procedimentos utilizados; e um último caso em que não havia objetivo de desenvolver uma

escala). Concluindo, a base final de dados foi composta por 24 artigos, contendo um total de 28 escalas propostas e/ou validadas (em dois artigos, foram propostas 2 escalas e em um outro, três).

7.3.4 Etapa 4: Categorização dos estudos selecionados

A etapa 4 tem como foco a sumarização e documentação, de forma concisa, das informações extraídas dos artigos científicos encontrados nas fases anteriores. Essas informações devem incluir, por exemplo, o tamanho da amostra e a quantidade dos sujeitos, metodologia empregada, mensuração de variáveis, métodos de análise, a teoria ou conceitos embaixadores utilizados.

Um dos instrumentos usados para extrair essas informações é a matriz de síntese, ou matriz de análise - largamente utilizada devido à sua capacidade de resumir aspectos complexos do conhecimento. Segundo Botelho, Cunha e Macedo (2011, p.131):

A matriz pode conter informações verbais, conotações, resumos de texto, extratos de notas, memorandos, respostas padronizadas, e, em geral, dispor de dados integrados em torno de um ponto ou temas de pesquisa. Em suma, a matriz deve conter informações sobre aspectos da investigação e permitir que o pesquisador tenha uma visão geral de dados relacionados a um desempenho de certos pontos. A matriz serve de ferramenta de interpretação e construção da redação da revisão integrativa para os pesquisadores.

No seu estudo, Botelho, Cunha e Macedo (2011) afirmam que a construção da matriz pode ser flexível, uma vez que não existe matriz de síntese correta, apenas funcionais ou não, sendo sua construção influenciada pela interpretação do pesquisador e da forma como ele organiza seus dados.

Para a análise da matriz de síntese o pesquisador deve criar categorias analíticas que facilitem a ordenação e sumarização dos trabalhos. Broome (2006) *apud* Botelho, Cunha e Macedo (2011) reforça que para a análise das informações coletadas nos estudos, é necessário que o pesquisador crie categorias analíticas que facilitem a ordenação e a sumarização de cada estudo. Para o autor, essa categorização pode ser realizada de forma descritiva. Para categorizar e analisar as informações, o pesquisador pode utilizar diferentes métodos, tais como: análises estatísticas; listagens de fatores que mostram o efeito das variáveis ao longo do tempo de estudo; escolha ou exclusão de estudos, entre outros.

Para Ganong (1987), o pesquisador pode proceder de diferentes maneiras para criar categorias. O importante é que ele deixe clara a maneira como ele analisa os dados extraídos

dos artigos científicos. Em seu estudo o autor cita o exemplo da listagem de variáveis feita por meio da escolha de estudos válidos que se aproximem da questão de pesquisa do estudo.

Para a categorização dos dados dos artigos levantados na etapa anterior, foi definido um procedimento sistemático de tabulação considerando um grupo de variáveis que pudessem levantar as principais práticas de desenvolvimento de escalas sugeridas por diversos autores da área (CHURCHILL, 1979; HAIR *et al.*, 2009; VIEIRA, 2011; COSTA, 2011; ZAMBALDI; COSTA; PONCHIO, 2014). Essas variáveis foram agrupadas em 7 blocos diferentes, discutidos a seguir.

O **primeiro bloco** inclui variáveis que visam descrever as principais características das escalas desenvolvidas, como o nome dos autores, título da escala, o nome do evento e ano de publicação da escala, o tipo de escala (se escala proposta ou revalidada no Brasil), e outros dados específicos sobre a escala, como seu número de dimensões, número total de itens por dimensão, tipo de escala de avaliação (Likert, Diferencial Semântico, etc) utilizada e número de pontos da escala.

O **segundo bloco** consiste de variáveis que procuram identificar os principais procedimentos qualitativos utilizados para a caracterização do construto, geração de itens e validade de conteúdo. As variáveis são as seguintes: se foi realizada revisão na literatura; se foi utilizado o método de entrevistas em profundidade, o número de entrevistas realizadas e o perfil dos entrevistados; se foi realizado grupo de discussão, o tamanho deste grupo e o perfil dos participantes; se foi realizada avaliação dos itens da escala por juízes, o número e o perfil dos juízes utilizados; e por fim, também foi observado se foi realizado pré-teste do questionário, isto é, se uma pequena amostra de respondentes foi selecionada para avaliar a compreensão dos enunciados dos itens e do questionário como um todo. Observou-se ainda o tamanho dessa amostra e o perfil do participante, se fazia parte do público-alvo de futuros potenciais respondentes, como recomenda a literatura.

O **terceiro bloco** de variáveis inclui as informações sobre os procedimentos de análise preliminares em dados multivariados. Foram levantados dados sobre a verificação de dados faltantes (*missing values*), dados atípicos (*outliers*) e se foram feitos testes que verificassem as principais suposições de análises multivariadas (normalidade, homocedasticidade e linearidade).

Também neste bloco foram incluídas informações sobre a realização de um estudo piloto quantitativo. O termo estudo piloto se refere à primeira amostragem realizada que seja suficiente para iniciar os primeiros procedimentos de purificação da escala. Conforme Vieira (2011), antes de se realizar o teste de uma escala em um estudo oficial, deve-se realizar uma

pesquisa pré-teste com uma amostragem piloto do tipo não probabilística com o objetivo de purificar os indicadores gerados em etapas anteriores, que podem ter um tamanho entre 130 e 200 respondentes, dependendo da opção pessoal de cada pesquisador. As informações coletadas aqui incluíram dados sobre a realização ou não de testes preliminares na amostra, o tamanho da amostra piloto, a realização ou não de uma análise fatorial exploratória na amostra do estudo piloto e da realização ou não da análise de confiabilidade (alpha de Cronbach).

O **quarto bloco** inclui variáveis que procuram avaliar as características da amostra do estudo oficial. São levantadas informações sobre o método de levantamento de dados segundo o modo de aplicação, a técnica de amostragem, o tamanho da amostra e a razão entre tamanho da amostra e número de itens avaliados.

O **quinto bloco** inclui as variáveis que caracterizam os procedimentos metodológicos de análise fatorial exploratória utilizados pelos pesquisadores, considerando a amostra de um estudo oficial. Para os casos em que apenas uma amostra foi utilizada, esta amostra foi considerada como a amostra do estudo oficial. Foram levantadas informações sobre a realização ou não da análise fatorial exploratória em estudo oficial; o método de extração de variância utilizado (ACP – Análise de Componentes Principais; AFC – Análise de Fator Comum); a realização ou não dos testes preliminares da análise fatorial (KMO e teste de Bartlett); o total de variância extraída pelos fatores; método de rotação utilizado; se foi realizada análise das comunalidades e das cargas fatoriais; se foi realizada validação cruzada.

O **sexto bloco** inclui as variáveis que avaliam as técnicas de confiabilidade empregadas. Verificou-se quais dos tipos de teste de confiabilidade citados na literatura foram realizados pelos autores das escalas analisadas: teste reteste; equivalência; consistência interna (alpha de Cronbach); análise das correlações inter-itens; análise das correlações item-total; valor de alpha se o item for retirado e confiabilidade composta.

O **sétimo bloco** de variáveis inclui as informações sobre os procedimentos de validação das escalas. São levantados dados sobre a utilização ou não das técnicas de análise fatorial confirmatória (AFC), a verificação da validade de critério (concorrente e preditiva), da validade de construto convergente (e a técnica de validação convergente utilizada), da validade de construto discriminante (e a técnica de validação discriminante utilizada), da validade nomológica e da validade entre grupos conhecidos.

7.3.5 Etapa 5 - Análise e interpretação dos resultados

A etapa 5 aborda a discussão sobre os textos analisados na revisão integrativa. O pesquisador deve aqui fazer a interpretação dos dados, e visando validar seu estudo, apontar quais são as lacunas de conhecimento existentes, sugerindo pautas para futuras pesquisas. A análise e interpretação dos resultados foram discutidas no capítulo 8 desse estudo. Os apontamentos de pesquisas futuras foram tratados na conclusão dessa tese.

7.3.6 Etapa 6 - Apresentação da revisão/ síntese do conhecimento

A última etapa (etapa 6) consiste na elaboração do documento contendo a descrição de todas as fases percorridas pelo pesquisador, de forma criteriosa, e os principais resultados obtidos.

Ao longo desse capítulo foram descritas todas as fases percorridas pelo autor desta tese no processo de revisão sistemática integrativa. Os principais resultados obtidos e as propostas para estudos futuros foram realizadas nos capítulos 8 e 9 desta tese.

7.4 Resumo da Metodologia de Pesquisa Empregada

O Quadro 12 traz uma síntese da metodologia de pesquisa empregada, abordada ao longo desse capítulo.

Quadro 12: Resumo da Metodologia de Pesquisa Empregada

Unidade de análise	<ul style="list-style-type: none"> • Eventos promovidos pela Anpad, período 2012 a 2014. • Número de artigos analisados: 24
Abordagem da pesquisa (Tipo de Pesquisa)	<ul style="list-style-type: none"> • Revisão Bibliográfica Sistemática
Método de procedimento de pesquisa	<ul style="list-style-type: none"> • Revisão Integrativa
Período de realização do estudo	<ul style="list-style-type: none"> • Início: 1º trimestre de 2014 • Término: 4º trimestre de 2014

Fonte: elaborado pelo autor

8 ANÁLISE DOS RESULTADOS

Neste capítulo, serão discutidos os principais resultados relativos à análise das escalas de marketing desenvolvidas e validadas em anais de eventos da Anpad (período 2012-2014) encontradas segundo a metodologia de pesquisa utilizada e discutida anteriormente. A análise dos resultados será realizada em 7 blocos:

- **Bloco 1 - identificação e caracterização das escalas encontradas:** é feita a identificação e caracterização das escalas selecionadas e analisadas nesta tese, no que se refere aos nome dos autores, nome da escala, o evento da Anpad em que ela foi publicada, o ano e o tipo de escala (se proposta ou revalidada), nº de dimensões por escala, nº de itens por dimensão, tipo da escala de avaliação e o número de pontos.
- **Bloco 2 – procedimentos qualitativos para desenvolvimento das escalas:** discute-se os resultados dos procedimentos qualitativos utilizados pelos autores para o desenvolvimento das escalas, incluindo revisão da literatura, entrevista em profundidade, grupo focal, avaliação de especialistas e análise semântica do questionário.
- **Bloco 3 – análises exploratórias em dados multivariados:** são apresentados as frequências em que os autores dos artigos realizaram as análises preliminares usualmente recomendadas na análise multivariada, como a análise de *outliers*, *missing values* (dados faltantes), normalidade, linearidade, homocedasticidade, multicolinearidade e singularidade; apresenta-se ainda a síntese dos procedimentos utilizados em estudo piloto, incluindo a frequência em que foram realizados tais testes, o tamanho da amostra, entre outros.
- **Bloco 4 – informações sobre a amostra em estudo oficial:** discutem-se alguns pontos sobre o tamanho de amostra utilizado nos estudos oficiais, bem como o método de amostragem.
- **Bloco 5 – procedimentos de análise fatorial exploratória em estudo oficial:** apresenta-se aqui uma discussão dos procedimentos de análise fatorial exploratória empregados pelos autores dos artigos.
- **Bloco 6 – procedimentos de verificação de confiabilidade:** é apresentada a síntese dos principais métodos utilizados pelos autores das escalas para se atestar a confiabilidade, bem como das técnicas estatísticas empregadas.

- **Bloco 7 – procedimentos de verificação da validade:** é apresentada a síntese dos principais métodos utilizados pelos autores das escalas para se atestar a validade de construto e critério, bem como os procedimentos utilizados de análise fatorial confirmatória.

Destaca-se ainda que em alguns momentos da discussão dos resultados, foram extraídos partes dos textos dos autores das escalas para apresentar exemplos dos assuntos abordados, mas por uma questão de ética, os nomes dos autores foram preservados, substituindo-os pelo termo “ARTIGO” e um código indicador (este código foi criado após o primeiro processo de refinamento da escala, que varia de 1 a 210).

8.1 Bloco 1 - Identificação e caracterização das escalas encontradas

O processo de busca sistemática e seleção de artigos resultaram na coleta de 24 publicações que atenderam aos critérios definidos no protocolo de pesquisa e que foram selecionados para análise. Destes artigos, emergiram 28 escalas, as quais são mostradas no quadro 13 abaixo. O EnAnpad foi o evento em que foram encontradas mais escalas, 18 no total (64,3%), enquanto que o EMA (Encontro da Divisão de Marketing da Anpad) contribuiu com 10 escalas (35,7%). Quanto aos autores das publicações, foi possível identificar 67 pesquisadores diferentes, destacando-se entre eles o autor Mauri Leodir Löbner, com 4 escalas publicadas.

Quadro 13: Identificação das escalas encontradas

Autores	Nome da escala	Evento	Ano	Tipo
BOLZAN, L.M.; VIEIRA, K.M.; LÖBLER, M.L.	Escala para medir a inclusão digital individual	EnAnpad	2012	Escala proposta
ABRAHÃO, R. de S.; LOPES, J.E.F.	Escala de atributos de compra de TV por assinatura	EnAnpad	2014	Escala proposta
ARAUJO, C.F.; GRILLO, T.L.H.; ESPARTEL, L.B.	Escala OCP para mensurar a percepção dos outros clientes na entrega de serviços	EnAnpad	2013	Escala revalidada
BRESSAN, V.P.; BASSO, K.; BECKER, L.C.B.	Escala para avaliar a qualidade percebida em serviços educacionais	EMA	2014	Escala proposta
CARVALHO, C.E.; ROSSETO, C.R.	Escala para mensurar o ambiente percebido pelos gestores	EnAnpad	2012	Escala proposta
COELHO, T.R.; FREGA, J.R.	Escala de Agência no contexto de ICT4D	EnAnpad	2014	Escala proposta
COELHO, T.R.; FREGA, J.R.	Escala de Estrutura no contexto de ICT4D	EnAnpad	2014	Escala proposta
COELHO, T.R.; FREGA, J.R.	Escala de Emponderamento no contexto de ICT4D	EnAnpad	2014	Escala proposta

COSTA, A.P.; REZENDE, D.C.de; MESQUITA, D.L.	Escala para avaliação de academias de ginástica	EMA	2012	Escala proposta
DEMO, G.; ALVARENGA, B.; GUARNIERI, P.	Escala de avaliação do e-commerce	EnAnpad	2014	Escala proposta
GARCIA, E.; LOPES, E.L.	Escala para medir a Responsabilidade Social Corporativa (RSC) na visão do consumidor	EnAnpad	2013	Escala proposta
GARRIDO, I.L.; CALIXTO, C.V.; PEREIRA, A.; MONTICELLI, J.M.; VASCONCELLOS, S.L.de	Escala para medir desempenho internacional	EnAnpad	2012	Escala proposta
INACIO, R. de O.; BATTISTELLA, L. F.; XAVIER, T.R.; WITTMANN, M. L.	Escala para medir o desenvolvimento regional sustentável	EnAnpad	2013	Escala revalidada
ISONI FILHO, M.M.; ISONI, M.M.; PINHEIRO, P.R. de L.; ABREU, N.R. de; PEREIRA, R. de C. de F.	Escala multidimensional de valor de marca - Multidimensional brand equity (MBE)	EMA	2012	Escala revalidada
KUMMER, A.A.; SANTOS, G.D.; OLIVEIRA, G.A.; VASCO, A.P.D.	Escala para a avaliação da orientação para a sustentabilidade nas ações de inovação	EnAnpad	2013	Escala proposta
LÖBLER, M.L.; REIS, E. dos; VISENTINI, M. S.; KLIMECK, K. A.	Escala para mensurar os fatores influenciadores na aquisição de Sistemas de Informação em Micro e Pequenas Empresas	EnAnpad	2012	Escala proposta
LÖBLER, M.L.; REIS, E. dos; VISENTINI, M. S.; KLIMECK, K. A.	Escala para mensurar os fatores influenciadores na implantação de Sistemas de Informação em Micro e Pequenas Empresas	EnAnpad	2012	Escala proposta
LOPES, C.; DEMO, G.	Escala de Relacionamento com Clientes Skol	EMA	2012	Escala proposta
LOPES, C.; DEMO, G.	Escala de Relacionamento com Clientes Guaraná Antarctica	EMA	2012	Escala proposta
MAFFEZZOLLI, E.C.F.; PRADO, P.H.M.	Escala para medir identificação com a marca	EnAnpad	2012	Escala proposta
PARENTE, E.S.; LEOCÁDIO, A.L.	Escala de valor percebido pelo correntista no varejo bancário (VPCVB)	EMA	2014	Escala proposta
PINTO, M. de R.; LEITE, R.S.; VIEIRA, C.A.	Experiência de serviço	EnAnpad	2014	Escala revalidada
RAMOS, R.R.; MAZZA, I.; COSTA, F.J. da	Escala para medir a experiência do cliente no ponto de venda varejista.	EMA	2012	Escala proposta
RAMPELOTTO, A.; LÖBLER, M.L.; VISENTINI, M.S.	Escala para avaliar a efetividade do e-Gov na Prestação de Informações e Serviços aos Cidadãos	EnAPG	2012	Escala proposta
SANTOS, M.J. dos; PERIN, M.G.; SAMPAIO, C.H.	Escala para medir a capacidade de inovação de produtos	EMA	2014	Escala proposta
SILVA, R.G.S.; FRIO, R.S.	Escala e-SELFQUAL - Qualidade do autosserviço	EMA	2014	Escala revalidada
SOUZA, E.M. de.; BATISTA, P.C. de S.	Escala para medir modelo de negócios	EnAnpad	2014	Escala proposta
STREHLAU, S.; LOPES, E. L.; FREIRE, O.B. de L.	Escala BLI - Brand Luxury Index	EMA	2012	Escala revalidada

Fonte: autor (2014)

Em relação ao ano de publicação das escalas, 2012 e 2014 produziram respectivamente um total de 13 e 11 escalas, enquanto que o ano de 2013 apenas 4. Isto pode ser explicado em parte pelo fato de que o EMA é realizado a cada dois anos, e os últimos eventos aconteceram exatamente em 2012 e 2014.

O tipo de escala mais usual foi a escala proposta (n=22), em detrimento da revalidada (n=6). Por escala revalidada, considerou-se nesta pesquisa como o processo de avaliação de um instrumento criado em outra língua e/ou culturas diferentes da brasileira, que passou por um processo de validação no Brasil sem a alteração da sua estrutura de itens original. Escalas que sofreram modificações realizadas pelos autores do artigo, como por exemplo incluir outros itens e dimensões na escala, foram consideradas como escalas propostas.

As escalas com até 4 dimensões predominaram entre as escalas (55,5% de frequência acumulada), sendo as escalas com exatamente 4 dimensões a mais usual (32,7%). O quadro 14 apresenta os dados sobre o número de dimensões das escalas. De fato, em marketing são raros os construtos latentes que podem ser considerados puramente unidimensionais e esse resultado pode indicar a tentativa dos pesquisadores de incorporar as múltiplas facetas desses construtos.

Quadro 14: Número de dimensões das escalas

Nº Dimensões	Freq.	Freq. %	Freq. Acum.%
1	4	3,6	3,6
2	3	5,5	9,1
3	5	13,6	22,7
4	9	32,7	55,5
5	1	4,5	60,0
6	2	10,9	70,9
7	2	12,7	83,6
8	1	7,3	90,9
10	1	9,1	100,0
Total	28 escalas	100,0%	

Fonte: Autor (2014)

Porém, é importante notar ainda um grande número de escalas com mais de 4 dimensões ($\cong 45\%$ do total), resultado este que pode talvez ser explicado pelos seguintes motivos:

- a. A medição de construtos mais complexos, que possuem múltiplas dimensões. Foi o caso da escala construída para mensuração de modelos de negócios, que está dividida em 4 pilares diferentes, cada um com dimensões específicas;

- b. A falta de embasamento teórico sólido na literatura e validade de conteúdo dos itens propostos na escala por parte dos autores, que os guiassem nos processos de geração e purificação de itens.

Em relação a este último motivo (b), Hair *et al.* (2009) dizem que os pesquisadores preferem muitos indicadores como uma tentativa de representar completamente um construto e maximizar confiabilidade, mas entretanto, nem sempre mais itens são necessariamente melhores. De acordo com os autores, mais itens produzem estimativas de maior confiabilidade e generalidade, mas também demandam amostras maiores e podem tornar difícil a tarefa de produzir fatores verdadeiramente unidimensionais, sendo maior a possibilidade de que fatores artificiais sejam produzidos. Este problema se torna particularmente presente quando o conteúdo dos itens não foi cuidadosamente analisado com antecedência.

Desta maneira, isto mostra a relevância de se realizar uma boa revisão da literatura em busca das possíveis dimensões do construto pesquisado e gerar os itens em função dessas dimensões, como também do pesquisador assegurar a validade de conteúdo dos itens (validade de face com *experts* e validade semântica dos itens).

Quadro 15: Frequência do nº de itens por dimensão

Nº Itens/Dim	Freq.	Freq. %
1	2	1,8
2	11	10,0
3	25	22,7
4	36	32,7
5	20	18,2
6	6	5,5
7	3	2,7
8	1	0,9
9	1	0,9
10	2	1,8
11	1	0,9
16	1	0,9
21	1	0,9
Total	110	100,0

Fonte: Autor (2014)

Considerando um total de 110 dimensões existentes nas 28 escalas analisadas, avaliou-se ainda o número de itens por cada dimensão das escalas (ver quadro 15 e gráfico 3). O número mais frequente foi o de 4 itens na dimensão ($n=36$), seguido pela de 3 itens ($n=25$) e 5 itens ($n=20$). Ainda sim, 13 dimensões foram propostas considerando um número de 1 ($n=2$) ou 2 itens ($n=11$), o que de acordo com a literatura remetem à vários problemas. Os benefícios das escalas múltiplas em relação às escalas de um item individual já foram discutidos anteriormente (tópico 7.3.2), mas vale reforçar que, de acordo com Churchill (1979), as escalas múltiplas apresentam a possibilidade de reduzir erro com a combinação dos itens e assim, melhorar a confiabilidade da medida. Também permitem a melhor distinção de indivíduos, a possibilidade de atenuar a singularidade e especificidades dos itens com a combinação entre eles e por fim, apresentam ainda melhor validade preditiva.

Gráfico 3: Número de itens em uma mesma dimensão

Fonte: Autor (2014)

A recomendação de se utilizar pelo menos 3 itens por construto tem ainda respaldo dentro das práticas de modelos de equações estruturais, que são bastante utilizados para se atestar a validade convergente e discriminante de construtos. Segundo Hair *et al.* (2009), escalas com apenas dois itens geram mais parâmetros a serem estimados do que variâncias e covariâncias de itens, o que gera um número negativo de graus de liberdade, caracterizando assim um problema de identificação do modelo de mensuração. Em modelos sub-identificados, uma

solução única não pode ser determinada, uma vez que existem mais parâmetros para serem estimados do que valores únicos na matriz de covariância.

Por isso e outros motivos, Hair *et al.* (2009) recomenda que um construto seja refletido por um número mínimo de três itens, preferencialmente quatro, mas não existe um limite teórico máximo do número de itens por fator. Entretanto, os autores indicam que escalas que contenham mais do que 12 itens tornam a análise mais complexa e frequentemente conduzem a problemas de fornecimento de evidência de unidimensionalidade. É o que aconteceu, por exemplo, com as escalas que possuem 11, 16 e 21 itens. Avaliando a variância total extraída dessas escalas (respectivamente 40,2%, 35,5% e 26,4%), como também outros indicadores (comunalidades e cargas fatoriais baixas) percebe-se claramente que a solução unidimensional são seria a mais indicada nesses casos. Apesar de todos os indicativos, os autores dessas escalas mantiveram suas estruturas unidimensionais.

Em relação ao tipo de escala de avaliação (ou escala de verificação) utilizada para mensurar os escores dos respondentes, há supremacia pelo uso da escala Likert (n=16; 57,1% dos casos) e pelo uso de escalas contendo 5 (n=14) e 7 (n=8) pontos (ver quadros 16 e 17). Destaca-se que foi considerado como escala Likert apenas as situações em que os itens da escala foram escritos em forma de afirmações sobre a qual os respondentes deviam expressar seu grau de concordância, que varia entre discordância total e concordância total. Em alguns casos, os autores indicaram ter utilizado escalas do tipo Likert quando de fato, outro tipo de escala foi utilizado, como nestes dois exemplos: “as questões deveriam ser respondidas utilizando uma escala do tipo Likert, com 4 pontos, as alternativas eram “nunca”, “raramente”, “frequentemente” ou “sempre” (ARTIGO 10); “adotou-se uma escala tipo Likert para mensurar cada uma das questões, com variação em cinco pontos, sendo 1 “nada importante” e 5 “extremamente importante” (ARTIGO 11).

De fato, a grande utilização da escala Likert nesta pesquisa vem de encontro com o que pesquisadores também observaram em outras áreas das Ciências Sociais. De acordo com Costa (2011), seguramente a escala Likert é a escala mais amplamente utilizada devido a suas inúmeras vantagens e características, que justificam sua fama e sua constante presença nas pesquisas e nas atividades de formação de novos pesquisadores.

Percebe-se ainda o uso de escalas de importância em 3 dos instrumentos analisados, utilizadas para indicar o grau de importância do respondente para cada um de uma bateria de itens. Este tipo de procedimento causa um viés de mensuração conhecido na literatura como efeito *ceiling* (OLIVER, 1996; GHISI; MERLO; NAGANO, 2004) ou efeito teto, que ocorre quando respondentes são instados a atribuir notas a atributos segundo sua importância em

escalas não comparativas. O resultado desse viés é a ocorrência de escores com distribuição leptocúrtica e assimétrica no extremo superior da escala, e baixa variância para a maioria dos itens. Em pelo menos 1 dos textos, ficou evidente a ocorrência desse viés, o que ainda permitiu talvez uma conclusão equivocada do pesquisador: “Em relação ao total de respostas, a maioria dos respondentes (60,6%) apresentou valores 4 e 5, que indicam a relevância dos itens para um profícuo comércio eletrônico, contra 17,8% de respostas 1 e 2” (ARTIGO 48). As médias dos itens nesse estudo variaram entre 4,70 (máximo) e 2,59 (mínimo).

Quadro 16: Tipo de escala de avaliação utilizada nas escalas

Tipo de escala	Freq.	Freq. %
Likert	16	57,1
Não informado	4	14,3
Escala de importância	3	10,7
Mista: Likert e gradação	2	7,1
Escala de frequência	1	3,6
Diferencial Semântico	1	3,6
Escala de avaliação	1	3,6
Total Geral	28	100,0

Fonte: Autor (2014)

Quadro 17: N° de pontos da escala de avaliação

N° de pontos	Freq.	Freq. %
5	14	50,0
7	8	28,6
6	2	7,1
4	1	3,6
Não informado	3	10,7
Total Geral	28	100,0

Fonte: Autor (2014)

Quanto ao número de pontos das escalas, a literatura sugere que o número de categorias da escala está correlacionada positivamente à magnitude do coeficiente de confiabilidade do alpha de Cronbach (CHURCHILL; PETER, 1984; PETERSON, 1994), sendo que as escalas com 2 e 3 pontos são as que apresentaram os menores níveis de confiabilidade. Avaliando os resultados do quadro 17, é possível perceber que não houve nenhum pesquisador que utilizou escalas com menos do que 4 pontos, portanto, entende-se que os pesquisadores seguiram as recomendações usuais da literatura.

8.2 Bloco 2 - procedimentos qualitativos para desenvolvimento das escalas

Este bloco procura analisar os principais métodos utilizados pelos autores das escalas quanto aos procedimentos qualitativos que visam verificar a validade de conteúdo das escalas. Churchill (1979) sugere uma série de opções que poderiam ser utilizadas para isso, como a pesquisa na literatura, entrevistas com pessoas experientes no assunto, técnica do incidente crítico, grupo de foco, entre outras.

Nesta tese, optou-se por investigar a utilização dos seguintes procedimentos de validade de conteúdo e suas características: (1) realização de pesquisa na literatura; (2) realização de entrevistas, número de entrevistas realizadas e tipo de pessoa entrevistada; (3) grupo de foco, tamanho do grupo de foco e tipo de pessoas participantes; (4) apreciação ou não do instrumento para especialistas na área, número de especialistas pesquisados e tipo de especialistas considerado. Também foi verificado a realização de pré-teste do questionário.

Quadro 18. Síntese dos procedimentos qualitativos realizados pelos autores das escalas

Ocorrência	Rev. Literatura	Entrevistas	Grupo focal	Especialistas	Pré-teste
Não	0	15	28	10	7
Sim	28	13	0	18	21
Total	28	28	28	28	28

Fonte: Autor (2014)

O quadro 18 apresenta os principais resultados deste bloco. É uma recomendação entre os autores pesquisados na literatura de desenvolvimento de escala (CHURCHILL, 1979; COSTA, 2011; HAIR *et al.*, 2009) que o instrumento seja criado com forte embasamento teórico e conceitual, tendo clara definição do domínio do construto mensurado e das possíveis dimensões que ele pode apresentar. O primeiro indicador do quadro 18 revela que todos os autores (100%) realizaram alguma forma de revisão teórica sobre o tema abordado, mas muitas vezes essa revisão foi superficial. O que pode ser observado ainda é que nem todos os pesquisadores desenvolveram uma definição clara do construto medido conforme preconizado pela literatura especializada, quanto menos indicaram suas possíveis dimensões.

A prática de se realizar entrevistas em profundidade foi procedimento realizado no desenvolvimento de 13 escalas. O número de entrevistados variou entre 2 e 20, sendo mais usual a realização de até 7 entrevistas nas pesquisas. O perfil de entrevistado preferido pelos pesquisadores foi o público alvo de respondentes da escala (9 usaram esta abordagem, das 13 escalas que realizaram entrevistas), seguido de especialistas da área (6/13) e acadêmicos

(3/13). Em alguns artigos, foi citado o uso da técnica de análise de conteúdo para investigação dos dados das entrevistas, mas este não foi o procedimento usual de tratamento desses dados. Quanto à realização de grupo focal (forma de entrevistas em grupo), não foi verificado em nenhum dos artigos pesquisados o uso desta técnica.

Já a submissão dos itens para apreciação de especialistas (*experts*, também chamados de juízes) foi o procedimento qualitativo usualmente utilizado pelos autores, sendo empregado em 18 das 28 escalas analisadas (64,3%). O número de juízes variou entre 2 e 7, sendo formados por acadêmicos (em 11 escalas das 18 que passaram por validação de juízes) e/ou profissionais especialistas da área (8/18). Destaca-se aqui que em 5 situações não foi relatado o perfil dos juízes, e na grande maioria dos casos, muito pouco foi comentado por parte dos autores sobre os procedimentos utilizados para a operacionalização da análise dos juízes. Estima-se até, pela forma como os autores tratam este procedimento no texto, que na maior parte dos casos não tenha sido utilizada nenhum procedimento mais sistemático, como sugere a verdadeira análise de especialistas.

Por fim, ainda foi verificado se os autores realizaram o chamado pré-teste do questionário, que visa verificar a boa compreensão e entendimento dos itens por pessoas que fazem parte do grupo de potenciais pesquisados. De acordo com Malhotra (2001), depois de redigido, o questionário precisa ser testado por meio de um pré-teste adequado, devendo ser aplicado em um ambiente e contexto semelhante ao da pesquisa definitiva. Como mostra o quadro 18, a grande maioria das escalas passou pelo procedimento de pré-teste (21 de 28 escalas) para avaliação de semântica, sendo que 15 delas foram realizados com futuros potenciais pesquisados (público alvo da escala), como sugere a literatura. Entretanto, 1 dos pré-testes foi realizado com especialistas e em outras 5 ocasiões não há informação suficiente para qualificar o perfil do grupo pré-teste. O tamanho da amostra variou entre 4 e 133 respondentes.

8.3 Bloco 3 – Análises preliminares em dados multivariados

Como é muito comum no processo de desenvolvimento de escalas a utilização de técnicas estatísticas, como a análise fatorial exploratória (AFE) e confirmatória (AFC), entre outras, é recomendado por muitos autores na literatura um exame mais minucioso dos dados antes que as análises multivariadas sejam realizadas.

De acordo com Hair *et al.* (2009, p. 49), “o exame de dados é um passo inicial que consome tempo, mas necessário, que às vezes é ignorado por pesquisadores. Segundo os

autores, os pesquisadores devem avaliar o impacto dos dados perdidos (*missing values*), identificar observações atípicas (*outliers*) e testar as suposições inerentes à maioria das técnicas multivariadas.

Nesse sentido, procurou-se verificar nesta pesquisa a frequência em que os autores das escalas realizaram algum tipo de análise dos dados faltantes (dados perdidos ou *missing values*) e dados atípicos (*outliers*), como também avaliar as suposições subjacentes das análises multivariadas (normalidade, homocedasticidade e linearidade). Os resultados são mostrados no quadro 19.

Quadro 19: Frequência de realização dos testes preliminares das análises multivariadas

Análise	Sim	Sim %	Não	Não %	Total
Dados faltantes	11	39,3%	17	60,7%	28
Dados atípicos	9	32,1%	19	67,9%	28
Normalidade	6	21,4%	22	78,6%	28
Homocedasticidade	4	14,3%	24	85,7%	28
Linearidade	2	7,1%	26	92,9%	28

Fonte: Autor (2014)

Como é possível perceber pelo quadro, a prática de realização de testes preliminares não é algo comum nos artigos pesquisados. Não é possível afirmar, entretanto, se esses testes não foram realmente realizados ou se apenas não foram citados no texto, muito embora é de se esperar que se tivessem sido realmente realizados, seriam citados. Em algumas situações ainda, os pesquisadores apenas afirmaram terem realizados alguns desses testes (computados nas estatísticas “sim”), sem indicar efetivamente os resultados deles, o que pode indicar um quadro ainda menos animador.

Entre as poucas análises realizadas pelos autores, destacaram-se a análise dos casos faltantes (n=11), dados atípicos (n=9) e normalidade (n=6). A prática de tratamento de dados faltantes foi a de eliminar o caso. Quanto à análise dos dados atípicos, foram verificados casos de utilização da distância de Mahalanobis como procedimento de verificação de casos atípicos multivariados².

²A informação sobre o tipo de teste utilizado pelos autores para verificação de dados faltantes e casos atípicos não foi inicialmente incluída no protocolo de pesquisa, e por isso estas informações não foram tabuladas. Assim, não há estatísticas sobre elas. O mesmo acontece quanto aos testes estatísticos dos pressupostos da análise multivariada.

Também foi analisado a frequência com que os autores das escalas realizaram ou não estudos pilotos de suas escalas, com o objetivo de purificar os indicadores gerados em uma etapa anterior. Esta atividade se refere basicamente aos passos 3 e 4 sugeridos por Churchill (1979) e ao passo 4 do modelo de Vieira (2011) previamente discutidos no capítulo 6 desta tese. O estudo piloto é realizado em uma amostra anterior e independente do estudo oficial.

A análise dos dados do quadro 20 mostra que em apenas 5 das escalas (17,86% do total) foram realizadas um estudo piloto preliminar seguindo as recomendações da literatura de desenvolvimento de escalas. Destas, a análise fatorial exploratória foi realizada em quatro oportunidades e o indicador do alpha de Cronbach foi calculado em apenas 3. Em uma das escalas, foi utilizada diretamente a análise fatorial confirmatória (AFC). Estes dados mostram que a grande maioria dos pesquisadores ignorou esta etapa no desenvolvimento da escala, e deixou a etapa de purificação dos itens apenas para o estudo oficial.

Quadro 20. Estatísticas do estudo piloto

Ocorrência	Estudo Piloto	AFE - Estudo Piloto	Alpha - Estudo Piloto
Não	23	24	25
Sim	5	4	3
Total	28	28	28

Fonte: Autor (2014)

Para a realização de análise fatorial e outras técnicas multivariadas, a literatura recomenda uma amostra mínima de 5 respondentes para cada item da escala, sendo 10 o ideal (Hair *et al.*, 2009). Foram levantados dados sobre o número de itens das escalas e o tamanho da amostra que foram utilizadas no estudo piloto. Das 5 escalas em que o estudo piloto foi realizado, a relação entre o tamanho da amostra e o número de itens foi menor do que o recomendado (5) em duas oportunidades (2,7 e 3,2). Em nenhuma delas a razão foi superior a 10. Estes dados indicam sinais de que não é prática entre os pesquisadores utilizar amostras grandes para etapas de estudo piloto.

8.4 Bloco 4 – informações sobre a amostra em estudo oficial

Neste bloco, são feitas considerações sobre a amostra utilizada pelos pesquisadores no chamado estudo oficial, que corresponde à etapa 5 dos modelos de desenvolvimento de escalas de Churchill (1979) e Vieira (2011). Como muitos dos pesquisadores realizaram apenas uma coleta de dados a qual basearam todas as suas análises de confiabilidade e

validade, para fins de estatística nesta tese, esta única coleta foi considerada como pertencente ao estudo oficial. Neste sentido, todos os artigos analisados apresentaram pelo menos uma coleta de dados.

Quadro 21: Método de levantamento segundo o modo de aplicação (estudo oficial)

Método de levantamento	Freq.
Eletrônico via web	10
Pessoal de auto-administrado	6
Misto	3
Pessoal por entrevista	1
Telefônico	1
Não informado	7
Total	28

Fonte: Autor (2014)

Como apresentado no quadro 21 acima, os principais métodos de levantamento de dados utilizados pelos autores das escalas, quanto ao modo de aplicação, foram os métodos de questionário eletrônico via web ($n=10$) e o pessoal auto-administrado ($n=6$). Este método, segundo Malhotra (2001), possui como vantagens o baixo custo, rapidez na coleta e análise de dados, nenhum problema com equipe de campo e entrevistador, porém, sofre de tendenciosidade na seleção do entrevistado, uma vez que somente as pessoas que possuem acesso ao correio eletrônico ou à internet podem ser incluídas na estrutura da amostragem. Destaca-se ainda que no desenvolvimento de 7 das escalas avaliadas, os autores não informaram o modo de aplicação utilizado.

Quadro 22. Técnica de amostragem mais utilizada

Técnica de amostragem	Freq.	Freq. %
Por conveniência	22	78,6
Por julgamento	1	3,6
Por quotas	1	3,6
Aleatória simples	1	3,6
Não informada	3	10,7
Total	28	100,0

Fonte: Autor (2014)

Quanto ao tipo de amostragem realizada, como era de se esperar, a amostragem por conveniência foi a técnica de amostragem mais utilizada pelos pesquisadores (aplicada em 22 das 28 escalas avaliadas), enquanto que a amostragem aleatória simples foi utilizada em apenas 1 oportunidade. O quadro 22 apresenta estas estatísticas.

Em relação ao tamanho da amostra válida do estudo oficial, os valores variaram de um mínimo de 38 à um máximo de 950 respondentes, sendo a mediana de 278,5 e média de 353,6 participantes por amostra. Para uma melhor avaliação do tamanho de amostra utilizado pelos pesquisadores, procedeu-se novamente com a divisão do número de respondentes pelo número de itens total das escalas. Em três das escalas analisadas não havia informações suficientes para poder analisar o número de itens submetido à avaliação dos respondentes, sendo que então os resultados apresentados a seguir se referem a uma base de 25 escalas. Essa relação é apresentada a seguir no quadro 23 e gráfico 4 abaixo.

Quadro 23: Relação nº de respondentes por item (estudo oficial)

Razão Nº de Respondentes / nº de itens da escala	Freq.
0 - 5	1
5,1 - 10	9
10,1 - 15	5
15,1 - 20	4
20,1 - 25	1
25,1 - 30	2
30,1 - 35	1
35,1 - 40	0
40,1 - 45	1
45,1 - 50	0
50,1 - 55	1
Total	25

Fonte: Autor (2014)

Gráfico 4: Histograma da razão do nº de respondentes por item da escala

Fonte: Autor (2014)

A questão do tamanho da amostra no processo de desenvolvimento de escalas é algo que deveria ser considerado crucial para o pesquisador em seu processo de planejamento. Hair *et al.* (2009) sugerem que tamanhos de amostras com um número elevado de respondentes tem o efeito de aumentar poder estatístico por redução de erro de amostragem. Também indicam um efeito semelhante quanto à violação da suposição de normalidade, no sentido de que amostras maiores reduzem os efeitos nocivos da não-normalidade.

Costello e Osborne (2005) também sugerem que amostras maiores reduzem a probabilidade de ocorrência de soluções incorretas na análise fatorial exploratória. Nas pesquisas realizadas pelos autores, o tamanho da amostra se mostrou como elemento crítico para a definição correta do número de fatores quando a relação respondentes/por item da escala foi analisada. Numa proporção 2:1, 90% das amostras produziram uma solução com número de fatores incorretos; numa proporção 5:1, esse número declinou para 60%; na proporção 20:1, as taxas de erros foram bem menores, mas ainda sim ficaram bem acima dos valores padrão aceitos de erros ($\alpha = 0,05$). O fato é que com o aumento da proporção de respondentes por item, a porcentagem de erro na especificação do número correto de fatores caiu dramaticamente. Além disso, o número de itens agrupados em dimensões erradas também caiu com o aumento da amostra, sendo que o erro médio nos *eigenvalues* e nas cargas fatoriais também decaiu.

De fato, a importância do tamanho da amostra é tal para a análise fatorial que Tabachnick e Fidell (2001) indicam que uma amostra com menos de 300 sujeitos torna inadequada a análise fatorial.

Em relação às informações apresentadas no quadro 23, nota-se que a grande maioria dos pesquisadores utilizou uma relação de pelo menos 5 respondentes por item, conforme recomendado pela literatura. Cerca de 15 escalas (60% do total de 25) foram analisadas com amostras superiores à razão 10:1, e 6 (24%) acima da razão 20:1, o que mostra um quadro mais animador em relação ao visto na amostra piloto. Desta forma, o tamanho de amostra utilizado pelos pesquisadores nas escalas analisadas pode ser considerado como razoável.

8.5 Bloco 5 – procedimentos de análise fatorial exploratória em estudo oficial

Neste bloco, são analisados os procedimentos metodológicos utilizados pelos pesquisadores na análise fatorial exploratória, considerando a amostra do estudo oficial. Cerca de 24 das 28 escalas analisadas passaram por algum procedimento de verificação do número de dimensões e/ou purificação de itens por meio da análise fatorial. Como pode ser observado nos quadros 24, 25 e 26, os procedimentos de análise fatorial normalmente empregados pelos pesquisadores foram o método de extração por Componentes Principais (n=13), utilizando como critério de extração de fatores o autovalor maior do que 1 (n=19) e rotação ortogonal Varimax (n=13). Observa-se ainda que algumas dessas informações não foram relatadas nos artigos pelos pesquisadores (n=6).

Quadro 24. Método de extração de variância utilizado na AFE

Método de extração	Freq.	Freq. %
ACP - Análise de Componentes Principais	13	54,2
AFC - Análise dos Fatores Comuns	1	4,2
Ambos ACP e AFC	4	16,7
Não informado	6	25,0
Total	24	100,0

Fonte: Autor (2014)

Quadro 25: Principais critérios utilizados para definição do número de fatores na AFE

Ocorrência	Autovalor>1	VTE>60%	Gráfico Screeplot	Nº de fatores pré-definido	Outro
Não	5	20	17	21	20
Sim	19	4	7	3	4
Total	24	24	24	24	24

Fonte: Autor (2014)

Quadro 26: Método de rotação utilizado na AFE

Método de rotação	Freq.	Freq. %
Ortogonal Varimax	13	54,2
Oblíquo	4	16,7
Dúbio (Varimax ou oblíquo?)	1	4,2
Não especificado	6	25,0
Total	24	100,0

Fonte: Autor (2014)

Segundo Costello e Osborne (2005), enquanto a Análise de Componentes Principais (ACP) com rotação Varimax e o critério de Kaiser (*eigenvalues* ou autovalor maior que 1) são a norma, eles não são ideais, especialmente quando os dados não atendem a algumas suposições estatísticas, como é frequentemente o caso nas Ciências Sociais.

A literatura sobre análise fatorial normalmente recomenda, para os casos em que o objetivo é prioritariamente identificar fatores ou dimensões latentes que reflitam o que as variáveis tenham em comum, que o método de extração de variância utilizado seja a Análise da Fator Comum (HAIR *et al.*, 2009; FAVERO *et al.*, 2009; COSTELLO; OSBORNE, 2005). Segundo Pasquali (1997, p. 173):

A variância comum representa o que as variáveis do teste têm em comum (expressa pelas intercorrelações entre elas) e é expressa pelas cargas fatoriais no fator comum F, e é esta que constitui a questão da validade do teste, ou seja, quanto do traço latente (fator F) é representado empiricamente pelas variáveis (itens). O restante da variância dos itens é recolhido na chamada unicidade (U) de cada item que representa tanto o que é específico de cada um deles quanto os erros de medida. Estes dois últimos aspectos da variância (especificidade e erro) são agrupados num conceito só – a unicidade –, porque eles não contribuem para a validade do teste, pois é a parte do item que não constitui representação do traço latente.

Quanto à avaliação dos indicadores da análise fatorial (ver quadro 27), a grande maioria dos pesquisadores realizou os testes para verificar a adequação da amostra de itens para a análise fatorial: o teste de KMO (Kaiser-Meyer-Olkin) foi realizado em 22 das escalas

e o teste de esfericidade de Bartlett em 18. Também foi elevada a frequência com que foram analisadas as comunalidades ($n=18$; valor de referência normalmente entre 0,4 e 0,5) e as cargas fatoriais ($n=21$; valor de referência utilizado acima de 0,5). Já o procedimento de validação cruzada, sugerido por Vieira (2011), foi utilizado apenas uma única vez.

Quadro 27: Procedimentos realizados na análise fatorial exploratória – estudo oficial

Ocorrência	Análise Fatorial Exploratória	Teste KMO	Teste de Bartlett	Análise das comunalidades	Análise das cargas fatoriais	Validação cruzada
Não	4	2	6	6	3	23
Sim	24	22	18	18	21	1
Total	28	24	24	24	24	24

Fonte: Autor (2014)

Finalmente, o gráfico 5 abaixo apresenta os dados da variância total extraída nas análises fatoriais realizadas nas escalas avaliadas. Para esta estatística, considerou-se apenas os dados da última análise fatorial do pesquisador, realizada após todos os processos de purificação da medida. Escalas que tiveram suas dimensões analisadas individualmente em AFE separadas, não entraram nessas estatísticas.

Gráfico 5: Variância total extraída pelos fatores da AFE

Fonte: Autor (2014)

Com base nos dados do gráfico, observa-se que a maior parte dos pesquisadores procurou purificar as medidas e atingir o piso de 60% de variância total extraída, que é um valor base normalmente recomendado na literatura (HAIR *et al.*, 2009). Entretanto, observou-se pelo menos 3 escalas onde os pesquisadores extraíram um número de fatores com baixíssima variância total e mesmo assim, após avaliação do coeficiente alpha, afirmaram ter atestado a validade de construto do instrumento.

8.6 Bloco 6 – procedimentos de verificação de confiabilidade

Neste bloco, são analisados os dados de confiabilidade e os procedimentos utilizados pelos pesquisadores para sua mensuração. Foram levantadas informações sobre os três tipos de confiabilidade: teste-reteste, equivalência e consistência interna. Os resultados são apresentados no quadro 28 e discutidos a seguir.

Quadro 28: Indicadores de confiabilidade utilizados

Ocorrência	Teste-reteste	Equivalência	Alpha de Cronbach	Confiabilidade Composta	Análise das correlações inter-itens	Análise das correlações item-item total	Análise do indicador "alpha if item deleted"
Não	28	28	4	19	27	27	27
Sim	0	0	24	9	1	1	1
Total	28	28	28	28	28	28	28

Fonte: Autor (2014)

Conforme o esperado, em nenhuma das 28 escalas (n=0) se utilizou de algum procedimento de verificação da confiabilidade teste-reteste ou equivalência. A verificação da confiabilidade recaiu exclusivamente no cálculo do alpha de Cronbach (n=24) e confiabilidade composta (n=9), que é um coeficiente alternativo de confiabilidade. De acordo com Zambaldi, Costa e Ponchio (2014), este coeficiente indica a “proporção de variância do escore verdadeiro de um construto em relação à variância total do escore calculado”, e tem como vantagem de uso em relação ao alpha de Cronbach o fato de não apresentar o inconveniente de se inflar com a inclusão de itens na escala. Considerando as duas abordagens conjuntamente, em apenas uma oportunidade (ARTIGO 208) não houve algum teste de confiabilidade.

Notou-se ainda que as abordagens complementares de análise de confiabilidade (Análise das correlações inter-itens; análise das correlações item-item total; e análise do

indicador "alpha if item deleted"), que deveriam ser de uso mais amplo, foram citadas apenas uma única vez nos artigos (n=1). Essas abordagens auxiliam bastante o pesquisador na fase de purificação do instrumento, quando muitos itens por dimensão são elaborados e o pesquisador precisa identificar quais dos itens possuem maior consistência com os demais.

Especificamente sobre o alpha de Cronbach, notou-se ainda em duas oportunidades o procedimento de se calcular um coeficiente alpha “global” para escalas multidimensionais, como na seguinte passagem:

A análise fatorial exploratória realizada com as 26 variáveis de influência de aquisição de um SI em MPEs indicou a existência de sete fatores. Todas as variáveis elencadas foram contempladas nos fatores [...] A confiabilidade do instrumento como um todo foi igual a 0,889, indicando uma boa consistência interna. (ARTIGO 11)

Este procedimento não é considerado correto, haja vista que nas escalas multidimensionais, a análise de consistência interna, tendo o alpha de Cronbach como o método mais utilizado para sua verificação, deve ser realizada em blocos, considerando apenas os itens que fazem parte da dimensão, e não com todos os itens da escala. E especificamente no caso particular apresentado aqui (ARTIGO 11), mesmo sendo calculado sobre uma estrutura multidimensional, o valor do coeficiente alpha ainda foi elevado, e permitiu ao pesquisador tirar conclusões equivocadas.

Gardner (1995) apresentou e exemplificou uma série de confusões que muitos pesquisadores fazem entre unidimensionalidade e consistência interna. Segundo o autor, se uma escala é unidimensional, então espera-se que ela seja internamente consistente. O contrário, porém, não é necessariamente verdadeiro. Uma escala pode ser internamente consistente e apresentar um alpha alto, mas isso não significa necessariamente que ela seja unidimensional. Assim, a consistência interna não fornece evidência suficiente de unidimensionalidade. Um alpha alto não constitui evidência que todos os itens estão medindo a mesma coisa, mesmo porque, segundo Pasquali (2011), o próprio alpha é influenciado pelo comprimento do teste. Por conta disso, calcular o coeficiente alpha considerando conjuntamente todos os itens de uma escala multidimensional pode resultar em um alto valor de alpha, visto que este coeficiente é influenciado pelo número de itens, não significando, necessariamente, que todos os itens estão medindo a mesma coisa.

8.7 Bloco 7 – procedimentos de verificação da validade

Finalmente, neste último bloco serão verificados e discutidos os procedimentos de validação das escalas utilizadas pelos pesquisadores. Neste sentido, verificou-se a frequência de utilização da técnica de análise fatorial confirmatória (AFC) e de alguns de seus procedimentos, tais como o teste das cargas fatoriais, a verificação da variância média extraída (AVE, do inglês *Average Variance Extracted*), do ajuste do modelo de mensuração e seus indicadores, do teste de modelo de segunda ordem e das cargas estruturais. Dentro desse contexto, verificou-se em quais oportunidades o pesquisador conseguiu atestar as validades de construto convergente, discriminante e nomológica. Também verificou-se a validade de critério concorrente e preditiva.

Em 12 de 28 oportunidades (42,9%), a análise fatorial confirmatória (AFC) teve seu lugar entre os procedimentos utilizados pelos pesquisadores para o desenvolvimento de suas escalas. Muita pouca informação foi relatada nos textos quanto ao método de AFC utilizado. Em 3 oportunidades, foi referido o método de Máxima Verossimilhança, enquanto que em outras 2 foi citado o uso do método de Mínimos Quadrados Parciais (Partial Least Square - PLS). Nas outras 7 ocasiões, não havia nenhuma informação sobre o método utilizado.

Essa é uma informação importante a se considerar. De acordo com Garcia e Lopes (2013), a decisão sobre qual método de modelagem de equações estruturais (MEE) utilizar depende da análise da normalidade dos itens das variáveis dependentes. Caso seja observada a normalidade dos dados, a matriz de covariância seria a mais indicada, mas em caso contrário, a matriz de correlação com estimação dos Mínimos Quadrados Parciais (*Partial Least Square* – PLS) deveria ser a preferida.

Entretanto, Hair *et al.* (2009) alega que a técnica da Máxima Verossimilhança (MV) tem se mostrado bastante robusta diante de violações da suposição de normalidade, apresentando ainda ser mais estável e precisa em termos de ajuste empírico e teórico em relação a outros estimadores.

Quanto aos processos de validação, de acordo com Hair *et al.* (2009), os itens que são indicadores de um construto específico devem convergir ou compartilhar uma elevada proporção de variância em comum para se evidenciar a validade convergente. Pasquali (1997) afirma que a validade de construto é determinada pela grandeza das cargas fatoriais (que são correlações que variam de -1 a +1) das variáveis no fator, sendo aquelas a representação comportamental deste fator, que por sua vez, é o traço latente para o qual elas foram inicialmente elaboradas como representação empírica.

Conforme Hair *et al.* (2009), no mínimo é esperado que todas as cargas fatoriais sejam estatisticamente significantes. Entretanto, como uma carga significativa ainda poderia ser relativamente fraca, assume uma regra prática e mais conservadora de que as estimativas das cargas padronizadas devam ser de 0,5, preferencialmente 0,7. Este procedimento também é sugerido por Bagozzi e Philips (1991), que ainda citam que a variância extraída deve ser superior a 0,5 para atestar a validade convergente. Variância média extraída (AVE), de acordo com Hair *et al.* (2009), é uma medida resumida de convergência em um conjunto de itens que representa um construto latente, sendo o percentual médio de variação explicada entre os itens. Estudos que realizaram análise fatorial exploratória para verificar a validade convergente do construto sem realizar os testes de significância das cargas não foram considerados como procedimentos válidos de validade convergente.

Costa (2011) ainda afirma que se a variância extraída de um número de fatores considerados for menor do que 50%, provavelmente a escala terá problemas nas etapas posteriores. Neste caso, deve-se considerar a possibilidade de exclusão de itens, e se mesmo assim não houver melhorias, recomenda-se fortemente retornar para as fases anteriores de construção da escala, em particular sobre a definição da natureza formativa ou reflexiva do construto e sobre a validação de face e conteúdo.

Baseado nesses dois indicadores (teste *t* das cargas e medida AVE), observou-se em 11 oportunidades a utilização da variância média extraída, e em 8 situações o uso de testes *t* das cargas fatoriais, sendo 7 os casos em que as duas abordagens foram usadas conjuntamente.

Para a verificação da validade discriminante, tomou-se como ponto de partida os procedimentos citados por Vieira (2011) comentados anteriormente no capítulo 6. O método de Fornell e Larcker (1981), em que se compara a variância média extraída (AVE) com o quadrado da correlação dos construtos (r^2), verificando se o r^2 é menor do que AVE, foi o método preferido pelos pesquisadores em 10 oportunidades. O método de Bagozzi e Philips (1982) *apud* Vieira (2011) que compara a diferença de qui-quadrado entre dois modelos, o primeiro em que a correlação entre os dois modelos estruturais (via estruturas de covariância) é fixada em 1 e outro no qual ela é deixada livre, foi utilizado em 1 oportunidade.

Nota-se ainda que em uma oportunidade (ARTIGO 33), os autores afirmaram terem verificado a validade discriminante pela análise da matriz de correlação entre os fatores. Utilizaram como parâmetro de correlação máxima permitida o valor de 0,9, mas foi verificado uma correlação acima desse teto (0,91): “Como pode-se observar na tabela 7, a correlação entre fator social e econômico atingiu esse valor máximo aceitável e as demais correlações

foram significativas e inferiores a 0,90, atestando-se a validade discriminante do modelo proposto” (ARTIGO 33, p.12). Neste caso em particular, não foi creditado a validade discriminante para esta escala, uma vez que até os próprios autores reconhecem que as correlações apresentadas foram elevadas: “Destaca-se que as correlações que variam entre 0,7 e 0,9 são consideradas fortes pela literatura” (ARTIGO 33, p. 13).

O ajuste do modelo de mensuração proposto foi testado em 9 das 12 aplicações de análise fatorial confirmatória. De acordo com Hair *et al.* (2009), as medidas de ajuste absoluto de qualidade de ajuste são a estatística do qui-quadrado da razão da verossimilhança (χ^2), o índice de qualidade de ajuste (GFI) e a raiz do erro quadrático médio (RMSR). Medidas de Ajuste Incremental avaliam a medida de qualidade de ajuste comparado com um modelo nulo. Um modelo teorizado com apenas um fator e sem erro de mensuração. O cálculo é feito a partir do Índice de Tucker-Lewis e do Índice de Tucker-Lewis Normado (NFI). As últimas medidas compreendem a Medida de Ajuste Parcimonioso que avaliam a parcimônia do modelo proposto pela análise do ajuste versus o número de coeficientes estimados necessário para atingir aquele nível de ajuste. As medidas para avaliação direta são o índice ajustado de qualidade (AGFI) e o qui-quadrado normado ($\chi^2 / G.L.$).

Os índices de ajuste χ^2/gl (n=8), RMSEA (n=8) e CFI (n=7) foram os índices mais calculados pelos pesquisadores, seguidos por GFI (n=6), NFI (n=5), AGFI (n=4) e TLI (n=3). Alguns outros índices ainda foram citados. Destaca-se ainda, que foram verificados problemas em alguns dos modelos, especificamente nos indicadores NFI (n=3), CFI (n=3), RMSEA (n=3) e χ^2/gl (n=2), que nem sempre foram citados pelos autores. Para esta análise dos índices de ajustamento, utilizou-se como parâmetro as recomendações e diretrizes de Hair *et al.* (2009), considerando ainda os valores do tamanho da amostra (N) e número de variáveis (m) do modelo. Outra prática verificada é a de afirmar que o modelo conseguiu obter bons ajustes mesmo quando indicadores apresentaram-se ligeiramente acima das recomendações da literatura. Para a tabulação desses casos, considerou-se uma posição mais conservadora em incluir o caso como problema de ajuste.

Nota-se ainda que em duas oportunidades os autores propuseram modificações no modelo de mensuração a fim de tentar obter ajustes aceitáveis de seus modelos, como neste exemplo:

[...] O resultado inicial demonstrou que pelo critério de exclusão das cargas padronizadas inferiores a 0,50 nenhuma variável foi excluída. No entanto, para melhoria do modelo, foram implementadas correlações indicadas pelos outputs do software Amos com suas respectivas justificativas teóricas (Figura 3). (ARTIGO 33)

As recomendações da literatura sugerem que este tipo de procedimento não deva ser realizado. De acordo com Hair *et al.* (2009), mesmo que a adição desses novos caminhos leve a um ajuste significativamente melhor, o pesquisador não deve liberar (teorizar) cargas cruzadas, pois a existência de cargas cruzadas significantes é evidência de uma falta de validade de construto. Segundo os autores, quando uma carga cruzada significativa é descoberta, a validade menor é evidenciada em melhor ajuste e quando elas são estimadas, qualquer melhora potencial de ajuste é artificial, no sentido de que é obtida com a admissão de uma correspondente falta de validade de construto.

Outra forma de liberação de caminhos é a covariância entre termos de erro de duas variáveis medidas, que podem ser entre dois itens de um mesmo construto (covariância interna de erro de construto) e entre itens de diferentes construtos (covariância de erro entre construtos). Segundo Hair *et al.* (2009, p. 595), “liberar esses caminhos pode apenas reduzir o χ^2 ou deixá-lo igual. Quanto maiores ficam essas estimativas de relação, mais os resultados sugerem problemas de validade de construto, particularmente no caso de correlações entre construtos”.

A validade nomológica, esta definida como a relação entre construtos teóricos, de forma a confirmar as correlações significativas entre eles, originadas por uma teoria (MALHOTRA, 2005), foi verificada em 4 oportunidades nas escalas avaliadas. Satisfação, lealdade, qualidade, intenção de boca a boca e recompra foram as medidas usualmente utilizadas nesses modelos para atestar as relações entre os construtos pesquisados. A verificação nesse caso foi realizada pela verificação das cargas estruturais do modelo completo, não apenas do modelo de mensuração. Os resultados indicaram em todos os casos bom ajustamento do modelo estrutural e correlações entre as medidas criadas e os construtos conforme hipotetizadas na teoria.

Finalmente sobre a validade de critério, observa-se que não houve nenhum autor que avaliou efetivamente a validade preditiva da escala desenvolvida. De acordo com Costa (2011), a validade preditiva refere-se à capacidade da variável de efetivamente prever os resultados no futuro. Envolve a tomada de evidência em diferentes momentos, e a comparação de resultados com a medida tomada na escala.

Em duas oportunidades em que autores afirmaram ter conseguido atestar a validade preditiva, com base na definição de Costa (2011) expressa acima, a percepção é a de que esses autores não avaliaram efetivamente este tipo de validade:

- “As validades nomológica e preditiva também foram satisfatórias, tendo em vista o impacto da qualidade na intenção de compra” (ARTIGO 43);
- “Nota-se que o cumprimento de serviço é algo ligado à companhia, onde a participação do consumidor é menor, e assim a mesma apresentou validade preditiva em todos os constructos testados” (ARTIGO 39).

Esta constatação está de acordo com o que afirma Costa (2011), que dada a dificuldade de verificação da validade preditiva, que requer mais de uma tomada de informações, ela costuma ser negligenciada no desenvolvimento de escalas em administração.

Já a validade de critério concorrente foi verificada em 5 oportunidades. De acordo com Costa (2011), a validade concorrente refere-se à capacidade da escala de efetivamente se comportar como esperado em uma coleta simultânea das medidas da escala e de um dado critério, a partir das quais se verifica uma relação esperada. A confirmação da relação como esperado dá evidências de validade simultânea. Segundo Malhotra (2001), as variáveis de critério podem incluir características demográficas e psicográficas, medidas de atitude e comportamento, ou escores obtidos de outras escalas.

Desta forma, considerando que os modelos nomológicos apresentaram critérios de teste simultâneo da escala e as correlações entre os construtos indicadas pelas cargas estruturais se apresentaram significantes, é possível dizer que os pesquisadores também atestaram a validade concorrente nessas análises. Observou-se ainda que em uma das pesquisas os autores atestaram a validade de critério comparando a escala desenvolvida com uma variável critério, baseado em teste de correlação padrão. Mesmo assim, ficou evidenciado o pouco uso deste tipo de validade pelos pesquisadores da área de administração.

9 CONSIDERAÇÕES FINAIS

Este trabalho buscou avaliar a qualidade das escalas de mensuração que foram desenvolvidas e publicadas na área de marketing nos principais congressos da Anpad nos últimos três anos.

Como etapa inicial, foi estudada na literatura a teoria sobre mensuração, para entender os problemas e erros relacionados ao processo de medir e identificar os modelos que poderiam ser utilizados para a construção de instrumentos de medida confiáveis e válidos. Em seguida, foram compiladas as publicações em anais de congressos nacionais que desenvolveram e/ou validaram escalas de mensuração na área de marketing. O resultado da busca sistemática retornou um conjunto de 602 artigos, os quais foram refinados baseados em critérios pré-definidos pelo pesquisador. Após o refinamento, emergiram um conjunto de 24 artigos contendo 28 escalas que foram o foco de análise desta pesquisa.

Sequencialmente, no capítulo 8 desta tese, foi verificado quais dos procedimentos de desenvolvimento de escalas foram os mais utilizados nos artigos levantados, apontando ainda alguns erros encontrados nas análises desses pesquisadores.

Posteriormente, no capítulo 9 desta tese, é elaborado um relatório geral sintetizando a qualidade dos instrumentos desenvolvidos nesses artigos levantados quanto a sua confiabilidade e validade, assim como são feitas recomendações e apontado as limitações para o desenvolvimento e uso de escalas de marketing.

O quadro 25 abaixo apresenta os dados em síntese do processo de validação e teste de confiabilidade das escalas analisadas.

Quadro 29. Frequência verificada dos tipos de confiabilidade e validade encontradas nas escalas analisadas

Conteúdo		Confiabilidade			Critério		Construto		
Face (especialistas)	Análise Semântica	Teste-reteste	Equivalência	Consistência Interna	Concorrente	Preditiva	Convergente	Discriminante	Nomológica
18	21	0	0	27	5	0	12	11	4
64%	75%	0%	0%	96%	18%	0%	43%	39%	14%

Fonte: elaborado pelo autor (2014)

Como mostra o quadro 1, a qualidade das escalas propostas e/ou validadas analisadas nesta tese pode ser considerada ainda em um baixo patamar. O processo de validade de conteúdo, aqui englobando a validade de face e a análise semântica do instrumento, foi realizado pela maioria dos pesquisadores, entretanto, ainda existe uma parcela considerável de pesquisadores que não trataram sobre este assunto em suas escalas.

Esta é uma constatação preocupante, visto que muitos autores comentam não apenas da grande importância da validade de conteúdo para o desenvolvimento de escalas, mas também sugerem que dentre todas é esta a validade mais importante em todo o processo de validação. Por exemplo, autores como Hair *et al.* (2009) recomendam que a validade de expressão (conteúdo) deva ser estabelecida antes de qualquer teste teórico quando se usa a análise fatorial confirmatória, pois sem uma compreensão sobre o conteúdo ou significado de cada item, fica impossível expressar e especificar corretamente uma teoria de mensuração: “assim, sob um ponto de vista realista, validade de expressão é o mais importante teste de validade” (HAIR *et al.*, 2009, p. 593).

Por esta razão, recomenda-se fortemente que os pesquisadores dediquem um tempo considerável no processo de revisão da literatura, elaborando uma definição clara do que é e o que representa o construto. Devem ainda, discutir o que o construto não representa (diferenças conceituais entre o construto pesquisado e outros próximos dele relacionados) e procurar identificar dentro da própria literatura pesquisada, variáveis que poderiam ser utilizadas como critério para uso em análise de validade posterior. Recomenda-se ainda que a validade de face, a qual é realizada com especialistas, seja seguida de forma sistemática, como sugere a literatura, haja vista que houve indícios nos artigos analisados de que ela tenha sido tratada de uma maneira mais informal.

Outro ponto importante que merece destaque é o fato de que a maioria das medidas propostas carece de validade de construto, do ponto de vista mais restrito de análise. Neste sentido, o que ficou bastante evidente foi a prática de se usar a análise fatorial exploratória como ferramenta para atestar a validade de construto da medida, quando essa técnica deveria ser tratada apenas para fins exploratórios. Segundo Costello e Osborne (2005), a análise fatorial é exploratória por natureza, não possui inferência estatística, sendo projetada e mais apropriada para uso na exploração de dados. Assim, não foi projetada para testar hipóteses ou teorias, como é o caso da modelagem de equações estruturais e da análise fatorial confirmatória.

Além do grande uso da análise fatorial exploratória nos artigos analisados, foi verificada também a possibilidade de ocorrência de vieses nas pesquisas, relacionados aos

principais procedimentos adotados pelos pesquisadores no uso dessa ferramenta. O dogma de se utilizar a análise de componentes principais, extraindo fatores ortogonais (Varimax) por meio de autovalores maiores do que 1 não é apropriada para os casos em que se tem como objetivo primário identificar fatores ou dimensões latentes. Isso ainda toma uma dimensão mais crítica em função da análise fatorial exploratória representar um processo sujeito a muitos erros, mesmo com amostras muito grandes e dados ideais, como mostrou Costello e Osborne (2005) em suas simulações de Monte Carlo.

Lembra-se ainda que muitos dos construtos utilizados em marketing possuem algum grau de correlação entre eles, então assumir ortogonalidade dos fatores pode representar um erro. Como afirma Vieira (2011, p. 15):

Embora a rotação ortogonal seja a mais utilizada, em conjunto com a extração VARIMAX, em ciências sociais, especificamente em marketing e comportamento do consumidor, existe uma forte suposição de que os conceitos nessas áreas são relacionados em algum grau ao invés de não relacionados. Diante disso, uma rotação oblíqua – a qual examina essa suposição – pode ser mais bem empregada.

Recomenda-se assim, que os pesquisadores utilizem a análise fatorial exploratória dentro de sua própria natureza, isto é, de explorar os dados considerando métodos diferentes de extração e rotação de fatores.

De acordo com Pasquali (1997), as técnicas de validação apresentam dificuldades graves, mas nem por isso se justificaria o abandono das mesmas, primeiramente pelo fato de que em ciência empírica nada existe de perfeito e isento de erro e, em segundo lugar, a consciência dessas dificuldades deve servir para melhorar e não para abandonar as técnicas. De acordo com o autor, é recomendável o uso de mais de uma das técnicas para demonstrar a validade de construto, dado que a convergência de resultados das várias técnicas constitui garantia para a validade do instrumento.

De fato, como afirma Costa (2011), não há sentido em admitir como válida uma escala que mantenha uma boa evidência de validade de translação sem que sejam evidenciadas também as validades de critério e de construto, o que poderia levar à mesma cilada das escalas do passado, sepultando toda a motivação para a proposição de métodos rigorosos de trabalho.

Desta forma, recomenda-se que os pesquisadores da área utilizem das técnicas confirmatórias para verificar a significância das relações entre os itens dos construtos latentes (validade interna), mas também procurem relacionar a escala com outras medidas critério dentro de uma rede teórica de pesquisa (validade externa).

9.1 Limitações do estudo

Em relação às limitações desse estudo, o período analisado na revisão sistemática pode ser considerado um aspecto relevante, já que na pesquisa de campo foram considerados apenas os artigos publicados nos últimos três anos (2012 a 2014), e, portanto, não entraram os trabalhos de escalas anteriores ao ano de 2012. Além disso, não foram consideradas como amostra as publicações de periódicos em administração, que poderiam talvez apresentar artigos de qualidade superior devido aos procedimentos mais rigorosos de submissão.

Uma outra limitação relacionada à primeira se refere ao tamanho da base de artigos e escalas analisadas, que pode ser considerado baixo. O fato de terem sido avaliados poucos artigos nesta tese talvez não tenha evidenciado com efetividade a real situação da qualidade das métricas propostas.

Um outro aspecto importante quanto à limitação advém da própria metodologia adotada – a revisão integrativa –, que seleciona os artigos para comporem a base de dados de acordo com um procedimento padrão. No caso desse estudo, foi verificado a partir do título e do resumo do artigo se ele se encaixava ou não dentro do contexto de escalas de marketing. Isso significa que podem ter sido excluídos da amostra artigos relevantes para o trabalho, e que poderiam ter sido considerados, no caso de uma leitura dos mesmos por completo. No entanto, verificar artigos na íntegra, “um a um” seria um processo muito demorado, e é por essa razão que existe a metodologia de revisão integrativa. De todo modo, é importante reforçar que essa limitação foi em parte resolvida, pois quando havia dúvida da sua abordagem, o autor desta tese verificava de forma mais detalhada o artigo como um todo, para uma análise mais clara sobre incluí-lo ou não na base de dados.

Outra limitação verificada nesta pesquisa é que não houve a intenção de se fazer uma análise aprofundada sobre a etapa “definição de construto e revisão da literatura” nos artigos analisados. Sugere-se assim que futuras pesquisas nessa área procurem identificar essa característica de forma mais acentuada e qualificada.

Por fim, pode-se citar a limitação desse estudo quanto à análise dos dados. Embora existam procedimentos para avaliar se a escala foi bem construída (testes de confiabilidade e validade), existe, obviamente, um grau de subjetividade nesta análise; isto é, depende em parte da forma como este pesquisador avaliou o que o autor do artigo fez e relatou no seu trabalho, o que também não está livre de erros!

Salienta-se também que em muitos dos artigos pesquisados, nem sempre a informação necessária estava disponível, isto é, em muitos casos os autores não relataram no texto as

informações necessárias ou ainda, havia falta de clareza ou concisão entre textos e tabelas, o que dificultou o processo de entendimento e tabulação dos dados. Por conta disso, não foi possível dizer com toda certeza se o pesquisador não realizou determinado procedimento ou se apenas não o descreveu no corpo do artigo. Por uma questão de espaço no texto, limitação natural de páginas de um artigo, é possível que isso tenha acontecido algumas vezes. Desta forma, as estatísticas indicadas nesta tese devem ser analisadas sob um prisma mais conservador.

9.2 Sugestões de pesquisas futuras

Considerando as sugestões de pesquisas futuras, tem-se a possibilidade de ampliar a pesquisa sobre escalas de marketing para um intervalo mais abrangente, isto é, incluir um período mais antigos ao que foi analisado. Por exemplo, considerar os estudos realizados entre 1997 a 2014.

Também, um novo trabalho sobre escalas de marketing poderia considerar uma maior abrangência de eventos, incluindo ainda periódicos, como as revistas com classificação Capes (2014) A1 e A2. É importante lembrar que alguns autores muitas vezes possuem preferência em publicar seus artigos em revistas menos impactantes, devido ao tempo demandado para um artigo ser aprovado e publicado em revistas mais conceituadas. Por essa razão, sugere-se que novos estudos explorem também as revistas B1 e B2. Ainda assim, seria interessante comparar a qualidade dos instrumentos propostos e validados entre diferentes estratos da Capes. É possível ainda estender a coleta para outras bases de dados, tais como monografias, dissertações de mestrado e teses de doutorado, o que não foi considerado nesse trabalho.

Por fim, este estudo, levando-se em conta a metodologia empregada, pode servir de parâmetro para que sejam desenvolvidos estudos similares para outras áreas, como as áreas de recursos humanos, gestão empresarial, empreendedorismo, etc. A área de recursos humanos é uma área em que existe, de forma intensa, o uso de testes e escalas de mensuração para avaliar atitudes. Logo, desenvolver um inventário de escalas na área de recursos humanos parece ser uma proposta bastante interessante de pesquisa futura.

O fato de ter sido encontrada dificuldades para identificar o conjunto de informações relevantes dos artigos avaliados nesta tese, seja pela própria falta de informação nos artigos, seja pela falta de clareza no texto elaborado, é recomendado que em futuras pesquisas deste tipo se utilize um processo de avaliação por pares, em que dois especialistas avaliem os artigos e entrem em concordância para a tabulação das informações.

REFERÊNCIAS BIBLIOGRÁFICAS

- AAKER, D. A.; KUMAR, V.; DAY, G.S. **Pesquisa de Marketing**. São Paulo: Atlas, 2001.
- AIKEN, L.R. Number of response categories and statistics on a teacher rating scale. **Educational and Psychological Measurement**, v. 43, p. 397-401, 1983.
- ALDERSON, J. C.; CLAPHAM, C; WALL, D. **Language test construction and evaluation**. Cambridge: Cambridge University Press, 1995.
- ALEXANDRE, N. M. C.; COLUCI, M. Z. O. Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. *Ciência & Saúde Coletiva*, v. 16, n.7, p. 3061-3068, 2011.
- ALLEN, M.J.; YEN, W.M. **Introduction to measurement theory**. Monterey: Brooks/Cole Publishing Company, 1979.
- ALVES, J.N.; FLAVIANO, V.; LÖBLER, M.L.; PEREIRA, B.A.D. Os caminhos da produção brasileira sobre o processo de validação de instrumentos de pesquisa em administração. In: Encontro de Ensino e Pesquisa em Administração e Contabilidade da ANPAD, Anais... 4, 2013, Brasília, DF.
- AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. American Psychological Association. National Council on Measurement in Education. **Standards for Educational and Psychological Testing**. Washington, D.C.: AERA, 1999.
- ARMSTRONG, B.K.; WHITE, E.; SARACCI, R. **Principles of exposure measurement in epidemiology**. Oxford University Press, Oxford, New York, Tokio, 1994.
- ATALLAH, A.N.; CASTRO, A.A. **Medicina baseada em evidências**. São Paulo: Lemos, 1998.
- BAGOZZI, R. P. An examination of the validity of two models of attitude. **Multivariate Behavioral Research**, v.16, p. 323-359, 1981.
- _____; PHILIPS, L. W. Representing and Testing Organizational Theories: A Holistic Construal. **Administrative Science Quartely**, v. 27, n. 3, p. 459-489, 1982.
- BAKER, B. O.; HARDYCK, C. D.; PETRINOVICH, L. F. Weak measurements vs strong statistics: an empirical critique of S. S. Stevens's proscriptions on statistics. **Educational and Psychological Measurement**, v. 26, p. 291-309, 1966.
- BALLARD, B.; CRINO, M. D.; RUBENFELD, S. Social desirability response bias and the Marlowe-Crowne Social Desirability Scale. **Psychological Reports**, v. 63, n. 1, p. 227-237, 1988.
- BANDALOS, D. L.; ENDERS, C. K. The effects of nonnormality and number of response categories on reliability. **Applied Measurement in Education**, v. 9, p. 151-160, 1996.

BEARDEN, W.; NETEMEYER, R.G.; HAWS, K.L. **Handbook of marketing scales: multi-item measures for marketing and consumer behavior research**. 3 ed. Thousand Oaks: Sage Publications Inc., 2011.

BERGKVIST, L.; ROSSITER, J. R. The predictive validity of multiple-item versus single-item measures of the same constructs. **Journal of Marketing Research**, v. 44, n. 2, p. 175-184, 2007.

BERGKVIST, L.; ROSSITER, J. R. Tailor-made single-item measures of doubly concrete constructs. **International Journal of Advertising**, v. 28, n. 4, p. 607-621, 2009.

BERWANGER, O.; SUZUMURA, E.A.; BUEHLER, A.M.; OLIVEIRA, J.B. Como avaliar criticamente revisões sistemáticas e metanálises? **Revista Brasileira de Terapia Intensiva**, São Paulo, v.19, n.4, p.475-480, out/dez. 2007.

BESSA, N. M. Validade - o conceito, a pesquisa, os problemas de provas geradas pelo computador. **Estudos em Avaliação Educacional**, v. 18, n. 37, maio/ago., 2007.

BEYEA, S. C.; NICOLL, L.H. Writing an integrative review. **AORN Journal**, Philadelphia, v. 67, n.4, p. 877-880, Apr. 1998.

BLOCK, G. A. Review of validation of dietary assessment methods. **American Journal of Epidemiology**, v. 115, p.492-505, 1982.

BOLLEN, K. A. **Structural equations with latent variables**. New York: John Wiley & Sons, 1989.

BOLZAN, L. M.; VIEIRA, K. M.; LÖBLER, M. L. Validação de um instrumento capaz de identificar o nível de inclusão digital individual. In: XXXVI Encontro da Anpad, 2012, Rio de Janeiro. **Anais...** Rio de Janeiro: ANPAD, 2012.

BORGATTA, E.F.; BOHRNSTEDT, G.W. Level of measurement: once over again. **Sociological Methods and Research**, v. 9, p. 147-160, 1980.

BOTELHO, L. L. R; CUNHA, C. C. de A.; MACEDO, M. O método da revisão integrativa nos estudos organizacionais. **Gestão e Sociedade**, Belo Horizonte, v.5, n.11, p. 121-136, maio/ago., 2011.

BOUDREAU, M. C.; GEFEN, D.; STRAUB, D. W. Validation in IS Research: A State-of-the-Art Assessment. **MIS Quarterly**, p. 1-24, march, 2001.

BOYLE, G. J. Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? **Personality and individual differences**, v. 12, p. 291-294, Mar., 1991.

BRANDÃO, M. F. R. Um modelo de avaliação de programa de inclusão digital e social. 2009. 138f. Tese (Doutorado em Psicologia Social do Trabalho e das Organizações) – Programa de Pós-Graduação em Psicologia Social do Trabalho e das Organizações. Universidade de Brasília, Brasília, 2009.

BREWER, J.; HUNTER, A. **Foundations of multimethod research**. Thousand Oak: Sage, 2006.

BROOME, M. E. Integrative literature reviews for the development of concepts. In: RODGERS, B. L.; CASTRO, A. A. **Revisão sistemática e meta-análise**. 2006. Disponível em: <www.metodologia.org/meta1.PDF>. Acesso em: 20 ago. 2012.

BRUNER, G. C.; HENSEL, P. J.; JAMES, K. E. **Marketing scales handbook: a compilation of multi-item measures (Vol.4)**. Cincinnati: South-Western College Publishing, 2005.

CAMPBELL, D. T. Recommendations for APA test standards regarding construct, trait, or discriminant validity. **American Psychologist**, v. 15, p.546-553, 1960.

CHAPELLE, C. A. Validity in language assessment. **Annual Review of Applied Linguistics**, 19, p. 254-272, 1999.

CHAMPION, D.J. Some observations on measurement and statistics: comment. **Social Forces**, v. 46, n.4, p. 541-541, 1968.

CHERRYHOLMES, C. Power and criticism: Poststructural investigations in education. New York: Teachers College Press, 1988.

CHURCHILL, G.A. Jr. A paradigm for developing better measures of marketing constructs. **Journal of Marketing Research**, v. 16, p. 64-73, Feb., 1979.

_____; PETER, J. P. Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis. **Journal of Marketing Research**, v. 21, p. 360-375, Nov., 1984.

_____. **Marketing research: methodological foundations**. 7 ed. Orlando: Harcourt College Publishers, 1999.

CLEMMENS, D. The Relationship Between Social Support and Adolescent Mothers' Interactions With Their Infants: A Meta-Analysis. **Journal of Obstetric, Gynecologic, & Neonatal Nursing**, v. 30, n. 4, p. 410-420, 2001.

CONFORTO, E.C; AMARAL, D.C. SILVA, S.L da. Roteiro para revisão bibliográfica sistemática: aplicação no desenvolvimento de produtos e gerenciamento de projetos. **Anais**. 8º Congresso Brasileiro de Gestão de Desenvolvimento de Produto, 12 à 14 de setembro de 2011, Porto Alegre.

COOK, D. J.; MULROW, C. D.; HAYNES, R. B. Systematic reviews: synthesis of best evidence for clinical decisions. **Annals of Internal Medicine**, v.126, n.5, p.376-380, 1997.

COOK, T. D.; CAMPBELL, D. T. Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand McNally, 1979.

COOPER, D. R.; SCHINDLER, P.S. **Métodos de pesquisa em administração**. trad. Luciana de Oliveira da Rocha. 7 ed. Porto Alegre: Bookman, 2003.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. Análise multivariada para os cursos de administração, ciências contábeis e economia. 1º ed. São Paulo: Atlas, 2009.

CORTINA, J. M. What is Coefficient Alpha? An examination of theory and applications. **Journal of Applied Psychology**, v. 78, p. 98-104, feb., 1993.

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

COSTELLO, A. B.; OSBORNE, J. W. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, v. 10, n. 7, p. 1-9, 2005.

CROCKER, L.; ALGINA, J. **Introduction to classical and modern test theory**. New York: Holt, Rinehart e Winston, 1986.

CRONBACH, L. J.; MEEHL, P. E. Construct validity in psychological tests. **Psychological Bulletin**, v.52, p.281-302, 1955.

_____. Coefficient Alpha and the internal structure of testes. **Psychometrika**, v. 16, p. 297-334, Sep., 1951.

CROSSETTI, M. da G.O. Revisão integrativa de pesquisa na enfermagem: o rigor científico que lhe é exigido. **Revista Gaúcha de Enfermagem**. Porto Alegre, v.33. n.2, p.8-9, jun. 2012.

DANE, F. **Research methods**. Brooks/Cole Publishing Company: California, 1990.

DANTAS, R. A. S. Adaptação cultural e validação do Questionário de Senso de Coerência de Antonovsky em uma amostra de pacientes cardíacos brasileiros. 2007. Tese (Doutorado) - Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo, São Paulo, 2007.

DAVIS, F.B. **Educational measurements and their interpretation**. Belmont/CA: Wadsworth, 1964.

DE-LA-TORRE-UGARTE-GUANILO, M., TAKAHASHI, R.F. BERTOLOZZI, M.R. Revisão sistemática: noções gerais. **Revista Escola de Enfermagem, USP**, v.45, p. 1260-1266.

DEVELLIS, R.F. **Scale development: theory and applications**. 3 ed. Thousand Oaks: Sage, 2012.

DIAMANTOPOULOS, A.; WINKLHOFER, H. M. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, v. 38, p. 269-277, May 2001.

_____; SARSTEDT, M.; FUCHS, C.; WILCZYNSKI, P.; KAISER, S. Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. **Journal of the Academy of Marketing Science**, v.40, p. 434-449, 2012.

DICIONÁRIO PRIBERAM DA LÍNGUA PORTUGUESA. Disponível em: <http://www.priberam.pt/dlpo/>. Acesso em: 15/07/2012.

DING, D. X.; HU, P. J.; SHENG, O. R. L (2011). e-SELFQUAL: A scale for measuring online self-service quality. **Journal of Business Research**, 64, 508-515.

DOLLAGHAN, C.A. **The handbook for evidence-based practice in communication disorders**. Baltimore: Brookes Publishing Co.; 2007.

DUNNETTE, M. D. **Handbook of industrial and organizational psychology**. New York: Wiley, 1983.

EDWARDS, A. L. **The social desirability variable in personality assessment and research**. New York: Dryden, 1957.

ERTHAL, T.C.S. **Manual de psicometria**. 8 ed. Rio de Janeiro: Jorge Zahar Ed., 2009.

EVANS, D.; PEARSON, A. Systematic reviews: gatekeepers of nursing knowledge. **Journal of Clinical Nursing**, Oxford, v. 10, n. 5, p. 593–599, Sep. 2001.

FAGARD, R.H.; STAESSEN, J.A.; THIJS, L. Advantages and disadvantages of the meta-analysis approach. **Journal of Hypertension**, v.14, n.(S2), p.S9-S13, 1996.

FARIAS, E. **Dicionário escolar latim-português**. 6 ed. Rio de Janeiro: FAE, 1985.

FÁVERO, L.P.; BELFIORE, P.; SILVA, F.L. da; CHAN, B.L. **Análise de dados: modelagem multivariada para tomada de decisões**. Rio de Janeiro: Elsevier, 2009.

FERRIS, T.L.J. A new definition of measurement. **Measurement**, v. 36, p. 101-109, 2004.

FIGUEIREDO, O.H.dos SANTOS; ROCHA, A.M.C. da; SILVA, J.F. da. Mensuração do Construto da Distância Psíquica: desenvolvimento, comparação e validação de instrumentos multi-items. XXXIII Encontro da ANPAD. **Enampad**, São Paulo, 19 a 23 de setembro de 2009.

FIRAT, A. F.; GUMMESSON, E.; LEVY, S. J.; TADAJEWSKI, M.; BRODIE, R. J.; HACKLEY, C.; LEE, N.; BAKER, M. J.; PELS, J.; ALPERT, F.; VAREY, R. J. Commentaries on the state of journals in marketing. **Marketing Theory**, v.10, n.4, 437-455, 2010.

FONSECA, R.M.P. Revisão integrativa da pesquisa em enfermagem em centro cirúrgico no Brasil: trinta anos após o SAEP. 135f. **Tese**. Dissertação em Enfermagem, Universidade de São Paulo. São Paulo, 2008.

FORNELL, C., LARCKER, D.F. Evaluating structural equations models with unobservable variables and measurement error. **Journal of Marketing**, v.18, n.1, p.39-50, 1981.

FREGE, J. G. Os fundamentos da aritmética – uma investigação lógico-matemática sobre o conceito de número. In: Victor Civita (ed.), **Os pensadores: Peirce e Frege**. São Paulo: Victor Civita, p. 195-276, 1884.

FROEMMING; L. M. S.; LUCE, F. B.; PERIN, M. G.; SAMPAIO; C. H.; BEBER, S. J. N.; TREZ, G. Análise da qualidade dos artigos científicos da área de marketing no Brasil: as pesquisas survey na década de 90. **Revista de Administração Contemporânea**, v.4, n.3, p. 201-219, Set./Dez. Curitiba, 2000.

FUMAGALLI, F. Validação de questionário quantitativo de frequência alimentar para crianças de 5 a 10 anos. 2007. **Dissertação** (Mestrado em Ciências Nutricionais) – Faculdade de Ciências Farmacêuticas da Universidade Estadual Paulista, Araraquara, 2007.

GAITO, J. Measurement scales and statistics: resurgence of an old misconception. **Psychological Bulletin**, v. 87, n. 3, p. 564-567, 1980.

GALVÃO, C.M.; SAWADA, N.O.; TREVIZAN, M.A. Revisão sistemática: recurso que proporciona a incorporação das evidências na prática da enfermagem. **Revista Latino-Americana de Enfermagem**, v.12, n.3, Ribeirão Preto, p. 549-556, maio/jun. 2004.

GANONG, L. H. Integrative reviews of nursing research. **Research in Nursing & Health**, Hoboken, v. 10, n. 1, p. 1-11, Mar. 1987.

GARDNER, P.L. Measuring attitudes to science: unidimensionality and internal consistency revisited. **Research in Science Education**, v.25, n.3, p. 283-289, 1995.

GARVER, M. S.; MENTZER, J. T. Employing structural equation modeling to test for construct validity. **Journal of Business Logistics**, v. 20, n. 1, p. 33-58, 1999.

GERBING, D.W.; ANDERSON, J.C. An updated paradigm for scale development incorporating unidimensionality and its assessment. **Journal of Marketing Research**, v. XXV, p. 186-192, May, 1988.

GHISELLI, E.E.; CAMPBELL, J.P.; ZEDECK, S. **Measurement theory for the behavioral sciences**. San Francisco: W. H. Freeman and Company, 1981.

GHISI, M.A.; MERLO, E.M.; NAGANO, M.S. A mensuração da importância de atributos em serviços: uma comparação de escalas. **XXVIII EnAnpad**, 2004.

GRESSLER, L.A. **Pesquisa educacional**. São Paulo: Loyola, 1989.

GUILFORD, J.P. **Psychometric methods**. 2 ed. New York: McGraw-Hill, 1954.

_____. **Personality**. New York: McGraw-Hill, 1959.

GUTTMAN, L. What is not what in statistics. **The Statistician**, v. 26, p. 81-107, 1977.

HAIR, J. F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **Multivariate Data Analysis**. 5 ed. Upper Saddle River, New Jersey: Prentice Hall, 1998.

_____; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L.. **Análise multivariada de dados**. 6 ed. São Paulo: Bookman, 2009.

_____.; WOLFINBARGER, M.; ORTINAU, D. J.; BUSH, R.P. **Fundamentos de Pesquisa de Marketing**. 3 ed. Mc Graw Hill Education. São Paulo: Bookman, 2013.

HAYS, W. J. **Quantificação em psicologia**. São Paulo: Herder, 1970.

HENNING, G. **A guide to language testing**. Cambridge: Cambridge University Press, 1987.

HENSEL, P.J.; JAMES, K.E.; BRUNER, G.C. **Marketing scales handbook: a compilation of multi-item measures for organizational behavior and advertising (Vol.4)**. Cincinnati: South-Western Educational Publishing, 2008.

HOFFMAN, D. Automatic testing with intelligent sensor system: measuring or classification? In: Technological and Methodological Advances in Measurement Acta – **IMEKO**, 3, Amsterdam, 1982. Proceedings... Amsterdam: 1983.

HUGHES, A. **Testing for language teachers**. Cambridge: Cambridge University Press, 1989/1994.

JACOBY, J. Consumer research: a state of the art review. **Journal of Marketing**, v.42, n.2, p.87-96, Apr., 1978.

JARVIS, C. B.; MACKENZIE, S. B.; PODSAKOFF, P.M. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, v. 30, n. 2, p. 199-218, Sep. 2003.

JENKINS, G.D.; TABER, T.D. A Monte Carlo study of factors affecting three indices of composite scale reliability. **Journal of Applied Psychology**, v. 62, p. 392-398, 1977.

KAACKS, R.; RIBOLI, E.; ESTEVE, J.; VAN KAPPEL, A.L.; STAVEREN, W.A. Estimating the Accuracy of Dietary Questionnaire Assessment: Validation in Terms of Structural Equation Models. **Statistics in Medicine**, v.13, p. 127-42, 1994.

KAPLAN, R.W.; SACCUZZO, D.P. **Psychological testing: principles, applications, and issues**. Monterey/CA: Brooks/Cole, 1982.

KERLINGER, F. M. *Foundations of Behavioral Research*. 2 ed. Nova Iorque: Holt, Rinehart and Winston, 1973.

_____. **Metodologia da pesquisa em ciências sociais: um tratamento conceitual**. São Paulo: Editora Pedagógica e Universitária, 2003.

KLINE, R. B. *Principles and practice of structural equation modeling*. New York: Guildford Press, 1998.

KNAPP, T.R. Treating ordinal scales as interval scales: an attempt to resolve the controversy. **Nursing Research**, v. 39, n. 2, p. 121-123, 1990.

LABOVITZ, S. Some observations on measurement and statistics. **Social Forces**, v. 46, n. 2, p. 151-160, Dec., 1967.

_____. Some observations on measurement and statistics: reply to Champion and Morris. **Social Forces**, v.46, n.4, p. 543-544, 1968.

LADEIRA, W. JR. Três décadas do modelo de Churchill: utilização da análise fatorial e do alpha de Cronbach na validação de instrumentos de coleta de dados no marketing. *Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia*, nº 5, Set. 2010.

LARUS, J.; PUENTE-PALACIOS, K. E. Validação cruzada de uma escala de clima organizacional. **Estudos de Psicologia**, v. 9, n. 1, p. 113-119, 2004.

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

LISSITZ, R.W.; GREEN, S.B. Effect of the number of scale points on reliability: a Monte Carlo approach. **Journal of Applied Psychology**, v. 60, p. 10-13, 1975.

LÖBLER, M. L.; BOLZAN, L. M.; KUBOTA, F. I.; FLORES, S. A. M.; KICH, T. G. F. Inclusão digital: mapeamento de publicações sobre o tema, na área de Administração. *Revista Novas Tecnologias na Educação*, Porto Alegre/RS, v. 8, n. 3, 2011.

LORD, F.M. On the statistical treatment of football numbers. **The American Psychologist**, v.8, p. 750-751, 1953.

LOVATTO, P.A.; LEHNEN, C.R.; ANDRETTA, I.; CARVALHO, A.D.; HAUSCHILD, L. Meta-análise em pesquisas científicas – enfoque em metodologias. **Revista Brasileira de Zootecnia**, v.36, suplemento especial, p.285-294, 2007.

MACINNIS, D. J. A framework for conceptual contributions in marketing. **Journal of Marketing**, v.75, n.4, p.136-154, 2011.

MACKENZIE, S. B.; PODSAKOFF, P. M.; JARVIS, C. B. The problem of measurement model mis specification in behavioral and organizational research and some recommended solutions. **Journal of Applied Psychology**, v. 90, n. 4, p. 710-730, Jul 2005.

MALHOTRA, N.K. et al. **Introdução à Pesquisa de Marketing**. São Paulo: Prentice Hall, 2005.

_____. **Pesquisa de Marketing: uma orientação aplicada**. 3. ed. Porto Alegre: Bookman, 2001.

MARI, L. Epistemology of measurement. **Measurement**, v.34, p. 17-30, 2003.

_____. Foundations of measurement. **Measurement**, v. 38, p. 259-266, 2005.

MARTINS, G. A. Sobre confiabilidade e validade. **Revista Brasileira de Gestão de Negócios**, São Paulo, v. 8, n. 20, p. 1-12, jan./abr., 2006.

MATARAZZO, D.C. **Análise Financeira de Balanços**. 4 ed. São Paulo: Atlas, 1997.

MATELL, M.; JACOBY, J. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. **Educational and Psychological Measurement**, v. 31, p. 657-674, 1971.

MATTAR, F.N. **Pesquisa de Marketing**. Vol. 1, 2º ed. São Paulo: Atlas, 1994.

MCDANIEL, C.; GATES, R. **Pesquisa de Marketing**. trad. James F. Suderland Cook. São Paulo: Pioneira Thomson Learning, 2003.

MENDES, A. L. L.; FRACOLLI, L. A. Revisão sistemática de literatura e metassíntese qualitativa: considerações sobre sua aplicação na pesquisa em enfermagem. **Texto Contexto Enfermagem**, Florianópolis, v. 17, n. 4, p. 771-779, out./dez. 2008.

MENEZES, P.R. Validade e confiabilidade das escalas de avaliação em psiquiatria. Revista de Psiquiatria Clínica. Disponível na URL< <http://www.hcnet.usp.br/ipq/revista/vol26/vol25/n5/conc255b.htm>>. Acesso em 10 de setembro de 2014.

MESSICK, S. Validity. In R. L. Linn (Ed.), **Educational measurement**. 3 ed. New York: American Council on Education & Macmillan, p. 13-103, 1989.

MONDADORI, M. G.; LADEIRA, W. JR. Validação de um instrumento quantitativo em pesquisa de empreendedorismo e inovação: um estudo no contexto dos recursos tangíveis e intangíveis. In: Encontro da Associação Nacional dos Programas de Pós-Graduação em Administração, 32, 2007. Rio de Janeiro, RJ. **Anais...** Rio de Janeiro: Anpad, 2007.

MORALES VALLEJO, P. M.; SANZ, B. U.; BLANCO, A. **Cuadernos de Estadística: construcción de escalas de actitudes tipo Likert**. Madrid: Editorial La Muralla S.A, 2003.

MORRIS, R.N. Some observations on measurement and statistics: further comment. **Social Forces**, v. 46, n. 4, p.541-542, 1968.

MOSS, P. A. Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, v. 62, n. 3, p. 229-258, 1992.

MOSTELLER, F.; TUKEY, J.W. **Data analysis and regression: a second course in statistics**. Boston: Addison-Wesley, 1977.

MURPHY, K. R.; DAVIDSHOFER, C.O. **Psychological testing: principles and applications**. Englewood Cliffs/NJ: Prentice-Hall, 1988.

_____.; JAKO, R.; ANHALT, R.L. Nature and consequences of halo error: a critical analysis. **Journal of Applied Psychology**, v. 78, p. 218-225, Apr., 1993.

NETEMEYER, R. G.; BEARDEN, W.O.; SHARMA, S. **Scaling procedures: issues and applications**. Thousand Oaks: Sage, 2003.

NORMAN, G. Likert scales, levels of measurement and the “laws” of statistics. **Advances in Health Sciences Education**, p. 625-632, Dec., 2010.

NUNES, C. H. S. S.; PRIMI, R. Aspectos técnicos e conceituais da ficha de avaliação dos testes psicológicos. (Org.). Em: Conselho Federal de Psicologia (CFP). **Avaliação Psicológica: Diretrizes na regulamentação da profissão**, 2010, n.1, p. 101-128.

NUNNALLY, J.C. **Psychometric theory**. 1 ed. New York: McGraw-Hill, 1967.

_____. **Psychometric theory**. 2 ed. New York: McGraw-Hill, 1978.

_____.; BERNSTEIN, I. H. **Psychometric theory**. 3 ed. New York: McGraw-Hill, 1994.

OLIVEIRA, D. de; WALTER, S. A.; BACH, T. M. Critérios de validade em pesquisas em estratégia: uma análise em artigos publicados no EnANPAD de 1997 a 2010. In: **III Encontro de Ensino e Pesquisa em Administração e Contabilidade (EnEPO)**. João Pessoa, 20 a 22 de novembro de 2011.

OLIVER, R.L. **Satisfaction: a behavioral perspective of the consumer**. Boston: Irwin/McGraw-Hill, 1996.

PARASURAMAN, A., ZEITHAML, V., MALHOTRA, A. E-S-QUAL a multiple-item scale for assessing electronic service quality. **Journal of Service Research**, v.7, n.3, p.213-233, 2005.

PASQUALI, L. Psicometria. **Revista da Escola de Enfermagem da USP**, v. 43 (Esp.), p. 992-999, Dec., 2009.

_____. **Psicometria: teoria dos testes na psicologia e na educação**. Petrópolis: Vozes, 2003.

_____. **Psicometria: teoria e aplicações**. Brasília: UNB, 1997.

_____. **Psicometria: teoria dos testes na psicologia e na educação**. 4 ed. Petrópolis: Vozes, 2011.

PEDHAZUR, E. J.; SCHMELKIN, L.P. **Measurement, design and analysis: an integrated approach**. New York: Psychology Press, 1991.

PEREIRA, J. C. R. **Análise de Dados Qualitativos: estratégias metodológicas para as ciências da saúde, humanas e sociais**. São Paulo: Edusp, 2001.

PETERSON, R. A. A meta-analysis of Cronbach's Coefficient Alpha. **Journal of Consumer Research**, v. 21, p. 381-391, Sep., 1994.

RAMPELOTTO, A.; LÖBLER, M. L.; VISENTINI, M. S. Construção e Validação de um Instrumento para Avaliar a Efetividade do e-Gov na Prestação de Informações e Serviços aos Cidadãos. In: V Encontro da Divisão de Administração Pública/APB da ANPAD – EnAPG, 2012, Salvador. **Anais...** Salvador: ANPAD, 2012.

REDEKER, N. S. Sleep in acute care settings: An integrative review. **Journal of Nursing Scholarship**, v. 32, n. 1, p. 31-38, 2000.

REIBSTEIN, D. J.; DAY, G.; WIND, J. Guest editorial: is marketing academia losing its way? **Journal of Marketing**, 73, n.4, p. 1-3, 2009.

RIBANI, M.; BOTTOLI, C. B. G.; COLLINS, C. H.; JARDIM, I. C. S. F.; MELO, L. F. C. Validação em métodos cromatográficos e eletroforéticos. **Revista Química Nova**, vol.27, n.5, São Paulo, sept./oct., 2004.

ROBINSON, J. P.; SHAVER, P. R.; WRIGHTSMAN, L. S. **Measures of Personality and Social Psychological Attitudes (Vol. 1)**. San Diego/CA: Academic Press, 1991.

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

ROTHER, E. T. Revisão sistemática x revisão narrativa. **Acta Paulista de Enfermagem**, São Paulo, v. 20, n. 2, p. v-vi, jun., 2007.

RUBIN, R. B.; PALMGREEN, P.; SYPHER, H. E. **Communication research measures: a sourcebook**. New York: The Guilford Press, 1994.

RUBIO, D. M.; BERG-WEGER, M.; TEBB, S. S.; LEE, S.; RAUCH, S. Objectifying content validity: conducting a content validity study in social work research. **Social Work**, v.27, n.2, p.94-105, 2003.

SAMPAIO, C. H.; PERIN, M. G.; LUCE, F. B.; SANTOS, M. J.; SANTINI, F. O.; OLIVEIRA, M. O. R.; LENZ, G. S. **Pesquisa científica da área de Marketing no Brasil: Uma revisão da primeira década do século 21**. Revista de Administração Contemporânea, v.16, n.3, p. 459-478, 2012.

SAMPAIO, R. F.; MANCINI, M.C. Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. **Revista Brasileira de Fisioterapia**, v.11, n.1, p.83-89, Jan/Fev., 2007.

SANTOS, D. B. Avaliação de habilidades de inclusão digital: uma proposta de instrumento de medida. 2007. 66p. Monografia. Universidade de Brasília, Brasília. 2007.

SCARAMUCCI, M.V.R. Validade e consequências sociais das avaliações em contextos de ensino de línguas. **Lingvarvmarena (REVISTA DO PROGRAMA DOUTORAL EM DIDÁTICA DE LÍNGUAS DA UNIVERSIDADE DO PORTO)**, v. 2, ano 2011, pg. 103 – 120.

SELLTIZ, C.; WRIGHTSMAN, L. S.; COOK, S. **Métodos de pesquisas nas relações sociais**. 4. ed. São Paulo: Editora Pedagógica e Universitária, 1987.

SHARON, A. T.; BARTLETT, C. J. Effect of instructional conditions in producing leniency on two types of rating scales. **Personnel Psychology**, v. 22, p. 251-263, 1969.

SILVA, R. G. S.; FRIO, R. S. Qualidade do Autosserviço Online: Validação da Escala e-SELFQUAL no Contexto Brasileiro. In: VI Encontro de Marketing da Anpad, 2014, Gramado. **Anais...** Gramado: ANPAD, 2014.

SILVA, R. P. A.; MACEDO, L. C. B. de.; SILVA, I. L. R. da. Avaliação das características psicométricas dos questionários utilizados nos periódicos da área contábil: um estudo longitudinal compreendido no período 2003-2012. XX Congresso Brasileiro de Custos. **Anais...** Uberlândia, MG, Brasil, 18 a 20 de novembro de 2013.

SLACK, N.; CHAMBERS, S.; HARLAND, C.; HARRISON, A.; JOHNSTON, R. **Administração da produção**. São Paulo: Atlas, 1997.

SOUSA, F. F.; SILVA, J. A. A métrica da dor (dormetria): problemas teóricos e metodológicos. **Revista Dor**, v.6, n.1, p.469-513, 2005.

SPECTOR, P.E. **Summated rating scale construction**. Newbury: Sage Publications, 1992.

STEVENS, S.S. On the theory of scales of measurement. **Science**, v. 103, n. 2684, p. 677-680, jun. 1946.

STRAUB, D. W.; CARLSON, C. L. Validating instruments in MIS research. *MIS Quarterly*, p. 147-169, June, 1989.

TABACHNICK, B.; FIDELL, F. S. Using multivariate statistics. 3 ed. New York: Harper Collins, 1996.

TEIXEIRA, A. C. Formação docente e inclusão digital: a análise do processo de emersão tecnológica de professores, no Estado do Rio grande do Sul. 2005. 126p. Tese (Doutorado em Informática na Educação) – Programa de Pós-Graduação em Informática na Educação. Universidade Federal do Rio Grande do Sul. Porto Alegre. 2005.

_____. Internet e democratização do conhecimento: repensando o processo de exclusão social, no Estado do Rio grande do Sul. 2001. 132f. Dissertação (Mestrado em Educação) – Programa de Pós-Graduação em Educação. Universidade de Passo Fundo. Passo Fundo. 2001.

TOWNSEND, J.T.; ASHBY, F.G. Measurement scales and statistics: the misconception misconceived. **Psychological Bulletin**, v. 96, p. 394-401, 1984.

TRANFIELD, D.; DENYER, D. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. **British Journal of Management**, v.14, n.3, p.207-22, 2003.

TUKEY, J. W. Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. 1961. In: **The collected works of John W. Tukey**. Belmont, CA: Wadsworth, v. III, p. 391-484, 1986.

_____. The future of data analysis. 1962. In: **The collected works of John W. Tukey**. Belmont/CA: Wadsworth, v. III, p. 187-389, 1986.

URBINA, S. **Fundamentos da testagem psicológica**. Porto Alegre: Artmed, 2007.

VELLEMAN, P.F.; WILKINSON, L. Nominal, ordinal, interval and ratio typologies are misleading. **The American Statistician**, v. 47, n. 1, p. 65-72, 1993.

VIANNA, H. M. Validade de Construto em Teses Educacionais. **Educação e Seleção**, São Paulo, n.8, p. 35-44, 1983.

VIEIRA, V. A.; SLONGO, L. A. Na inventory of the characteristics of the marketing scales created and tested in Brazil. *Revista de Administração Mackenzie – RAM*, v. 8, n. 4, p. 11-34, 2007.

_____; TORRES, C. V. Verificação do Instrumento E-S-Qual para Mensuração da Qualidade Eletrônica no Varejo. In: III Encontro de Marketing da Anpad, 2008, Curitiba. **Anais...** Curitiba: ANPAD, 2008.

_____. **Escalas em marketing: métricas de resposta do consumidor e de desempenho empresarial**. São Paulo: Atlas, 2011.

WEB OF KNOWLEDGE. Disponível em: www.webofknowledge.com. Acesso em: 30/11/2014.

WEBSTER, J.; WATSON, J.T. Analyzing the past to prepare for the future: writing a literature review. **MIS Quarterly & The Society for Information Management**, v.26, n.2, p.13-23, 2002.

WHITTEMORE, R.; KNAFL, K. The integrative review: updated methodology. *Journal of Advanced Nursing*, Oxford, v. 52, n. 5, p. 546-553, Dec. 2005.

WIGGINS, G. P. **Assessing student performance: Exploring the purpose and limits of testing**. San Francisco: Jossey-Bass Publishers, 1993.

WIRTZ, J. Improving the measurement of customer satisfaction: a test of three methods to reduce halo. **Managing Service Quality**, v. 11, n. 2, p. 99-111, 2001.

WONG, C.S.; CHUEN, K.C.; FUNG, M.Y. Differences between odd and even number of response scale: some empirical evidence. **Chinese Journal of Psychology**, v. 35, p. 75-86, 1993.

YADAV, M. S. The decline of conceptual articles and implications for knowledge development. **Journal of Marketing**, 74, n.1, p.1-19, 2010.

ZAMBALDI, F.; COSTA, F. J. da; PONCHIO, M.C. Mensuração em marketing: estado atual, recomendações e desafios. *Revista Brasileira de Marketing – ReMark*, v. 13, n. 2, mai, 2014.

ZANDONAI, A. P. A busca de evidências para os fatores de risco alimentares do câncer colorretal: revisão integrativa da literatura. 183f. **Tese**. Dissertação em Enfermagem, Universidade de São Paulo. Ribeirão Preto, 2010.

ZIMMER, M. V.; FERREIRA, L.; HOPPEN, N. Validação e confiabilidade em pesquisas na área de sistemas de informação: uma análise dos artigos publicados no Enanpad entre 1998 e 2006. **XXXI Encontro da ANPAD**. Rio de Janeiro, 22 a 26 de setembro de 2007.

ZWICK, W. R.; VELICER, W. F. Comparison of five rules of determining the number of components to retain. *Psychological Bulletin*, v. 99, n. 3, p. 432-442, 1986.