UNIVERSIDADE DE SÃO PAULO
ESCOLA DE COMUNICAÇÕES E ARTES
DEPARTAMENTO DE RELAÇÕES PÚBLICAS, PROPAGANDA E TURISMO
ESPECIALIZAÇÃO EM PESQUISA DE MERCADO EM COMUNICAÇÕES

NÚMERO DE CATEGORIAS DE RESPOSTA EM ESCALAS LIKERT:
produção científica do período 2000 a 2014

Tainá Fernandes de Brito
Orientador: Prof. Dr. Leandro Leonardo Batista

SÃO PAULO
2015

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

Profa. Dra. Margarida Maria Krohling Kunsch
Diretor da Escola de Comunicações e Artes

Prof. Dr. Victor Aquino Gomes Correa
Chefe do Departamento de Relações Públicas, Propaganda e Turismo

Profa. Dra. Maria Clotilde Perez Rodrigues Bairon Sant'Anna
Coordenadora do Programa de Especialização em Pesquisa de Mercado em Comunicações

UNIVERSIDADE DE SÃO PAULO
ESCOLA DE COMUNICAÇÕES E ARTES
DEPARTAMENTO DE RELAÇÕES PÚBLICAS, PROPAGANDA E TURISMO
ESPECIALIZAÇÃO EM PESQUISA DE MERCADO EM COMUNICAÇÕES

TAINÁ FERNANDES DE BRITO

NÚMERO DE CATEGORIAS DE RESPOSTA EM ESCALAS LIKERT:
produção científica do período 2000 a 2014

Monografia apresentada ao Departamento de Relações Públicas, Propaganda e Turismo da Universidade de São Paulo como requisito para a obtenção do título de Especialista em Pesquisa de Mercado em Comunicações.

Orientador: Prof. Dr. Leandro Leonardo Batista

SÃO PAULO
2015

BRITO, T. F. **Número de categorias de resposta em escalas Likert**: produção científica do período 2000 a 2014. Monografia apresentada ao Departamento de Relações Públicas, Propaganda e Turismo da Universidade de São Paulo para a obtenção do título de Especialista em Pesquisa de Mercado em Comunicações.

Aprovado em: _____

Banca Examinadora

Prof. Dr. _____ Instituição: _____
Julgamento: _____ Assinatura: _____

_____ Instituição: _____
Julgamento: _____ Assinatura: _____

_____ Instituição: _____
Julgamento: _____ Assinatura: _____

Dedico este trabalho aos meus heróis
(meus pais).

AGRADECIMENTOS

Aos professores e colegas, pelas ideias, histórias e risadas compartilhadas.

À Idalina, pelo apoio durante todo o curso.

RESUMO

BRITO, T. F. **Número de categorias de resposta em escalas Likert**: produção científica do período 2000 a 2014. 2015. 78f. Monografia apresentada ao Departamento de Relações Públicas, Propaganda e Turismo da Universidade de São Paulo para a obtenção do título de Especialista em Pesquisa de Mercado em Comunicações.

O propósito principal desta pesquisa exploratória é analisar quantitativamente a produção científica sobre número de categorias de resposta em escalas Likert, no período 2000 a 2014, e realizar uma síntese dos artigos mais relevantes. Por meio de consultas às bases de dados multidisciplinares Scopus e Web of Science, foram identificados 16 artigos, publicados por 14 periódicos, produzidos por 40 autores, de 20 instituições de ensino em 11 países. Este estudo, então, torna-se importante ferramenta para a construção de instrumentos de coleta de dados quantitativos de pesquisas acadêmicas e de mercado.

Palavras-chave: Likert; escalas; número de categorias de resposta.

ABSTRACT

BRITO, T. F. **Number of response categories in Likert scales**: scienific production for the period 2000-2014. 2015. 78f. Monografia apresentada ao Departamento de Relações Públicas, Propaganda e Turismo da Universidade de São Paulo para a obtenção do título de Especialista em Pesquisa de Mercado em Comunicações.

The main purpose of this exploratory study is to quantitatively analyze the academic production on number of response categories for Likert scales, between 2000 and 2014, and to make a summary of the most relevant articles. Searching the multidisciplinary databases Scopus and Web of Science, 16 articles have been identified. This research, then, becomes an important tool for the development of quantitative data collection instruments for both academic and market researches.

Keywords: Likert; scales; number of response categories.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A escala Likert, desenvolvida por Rensis Likert na década de 1930, é uma ferramenta frequentemente adotada na construção de instrumentos de coleta de dados para pesquisas quantitativas (*surveys*) que têm como objetivo a mensuração de fenômenos sociais – atitudes, opiniões, expectativas, percepções, preferências, disposição física e mental, entre outros (GÖB; McCOLLIN; RAMALHOTO, 2007; LEUNG, 2011; ROSZKOWSKI; SOVEN, 2010).

Caracteriza-se pela exposição de um conjunto de enunciados pertinentes à determinada situação, objeto ou representação simbólica, e pela solicitação ao respondente que externe a sua reação perante cada enunciado proposto selecionando uma das alternativas de resposta, as quais são apresentadas em uma série logicamente ordenada (MARTINS; THEÓPHILO, 2009). Uma analogia proporcionada por ALLIK (2004) ilustra o grande avanço proporcionado pelas escalas Likert às pesquisas em ciências sociais: segundo o autor, as escalas binárias correspondem a fotografias em preto e branco, enquanto as escalas Likert permitem capturar os diferentes tons de cinza intermediários.

Em geral, as escalas Likert são bipolares, medem o grau de concordância do respondente em relação ao enunciado e apresentam cinco ou sete categorias de resposta (PRESTON; COLMAN, 2000), que variam de ‘discordo muito’ a ‘concordo muito’ (ALLEN; SEAMAN, 2007; JAMIESON, 2004). Outras escalas com estrutura similar e diferentes rótulos das categorias de resposta (compostos por variados padrões de verbos e advérbios), são denominadas ‘escalas tipo-Likert’ – exemplos incluem escalas dedicadas à mensuração de frequência, de aprovação, de satisfação, de confiança ou de motivação, entre outros (EDMONSON, 2005; HODGE; GILLESPIE, 2003).

A popularidade das escalas Likert (bem como das escalas tipo-Likert) está relacionada, principalmente, à facilidade de construção e aplicação, bem como à sua capacidade de adaptação às necessidades e especificidades de diferentes projetos de pesquisa (EDMONSON, 2005; HODGE; GILLESPIE, 2003). Tal flexibilidade é atribuída às múltiplas variações de fatores como: o número de categorias de resposta oferecidas, incluindo a decisão de prover um número par ou número ímpar de respostas; a presença ou omissão de um ponto com valor neutro; a inclusão de uma categoria do tipo ‘não sei’; o tipo e a distribuição dos rótulos das categorias de resposta; a direção das categorias de resposta; o balanceamento da escala (proporção de categorias positivas e negativas), entre outros (ROSZKOWSKI; SOVEN, 2010; WEIJTERS; CABOOTER; SCHILLEWAER, 2010). As decisões referentes a

estes fatores podem impactar o valor, a qualidade e a relevância dos dados coletados (PEARSE, 2011; WEIJTERS; CABOOTER; SCHILLEWAERT, 2010), dado o efeito que o formato de uma escala tem sobre as respostas produzidas pelos respondentes e sobre as propriedades psicométricas associadas (WENG, 2004).

1.1 Objetivos Gerais

O principal objetivo deste trabalho é mapear a produção acadêmica dos últimos 15 anos sobre o número de categorias de resposta de escalas Likert. A partir de um levantamento sistemático e abrangente dos artigos publicados em periódicos científicos no período 2000 a 2014 sobre o tema mencionado, serão desenvolvidas uma análise quantitativa das características bibliométricas e breve revisão bibliográfica das publicações mais relevantes.

O escopo desta análise (artigos acadêmicos) foi definido com base no rigor metodológico demandado para publicação em periódicos científicos e a sua delimitação temporal (2000-2014) é adequada aos objetivos de pesquisa, considerando o recente fenômeno de democratização da escala Likert, impulsionado pela difusão das pesquisas online[1] e das pesquisas de satisfação realizadas pelo setor de serviços[2]. Cabe destacar que não foram localizados estudos semelhantes em consultas preliminares às bases de dados Scopus e Web of Science.

1.2 Objetivos Específicos

Os objetivos específicos deste trabalho são: (1) identificar os artigos mais relevantes; (2) identificar os autores mais citados; (3) identificar os periódicos com publicações sobre o tema, no período; (4) identificar as instituições de ensino e países envolvidos; (5) desenvolver uma síntese dos três artigos mais relevantes.

[1] Na década de 1980 foram realizadas as primeiras pesquisas por e-mail. As primeiras pesquisas conduzidas pela Internet foram realizadas na década de 1990 (EVANS; MATHUR, 2005).
[2] Inicialmente, as pesquisas com foco em serviços foram desenvolvidas para avaliar a qualidade de hotéis, de atividades de lazer e de serviços públicos (GÖB; McCOLLIN; RAMALHOTO, 2007).

1.3 Justificativa

Apesar da sua aparente simplicidade, a escala Likert incorpora duas dimensões em seu conjunto de categorias de resposta e, assim, constitui um estímulo complexo, que exige do respondente processamento cognitivo (envolvido no exame do conteúdo de um enunciado e na decisão por uma direção de resposta – concordar ou não concordar) e provoca reação afetiva (a qual permeia a avaliação da intensidade das categorias de resposta disponíveis) (HODGE; GILLESPIE, 2003). Desta forma, é necessário balancear a demanda do pesquisador por informação e a capacidade de discriminação dos respondentes (CHAFOULEAS; CHRIST; RILEY-TILLMAN, 2009) na determinação do número de categorias de resposta que compõe uma escala Likert.

Análises do avanço científico sobre formatos de escalas Likert podem contribuir para melhoria dos resultados de pesquisas futuras, promovendo a difusão do conhecimento disponível, e o decorrente progresso na construção de instrumentos de coleta de dados. Portanto, este estudo visa apresentar um conteúdo particularmente útil para pesquisadores acadêmicos, pesquisadores de mercado, consultores, gerentes de projetos e usuários finais de pesquisas.

Além desta introdução, que contextualiza a proposta deste estudo, o texto está organizado em três partes: metodologia, que aborda a definição de variáveis para a consecução dos objetivos propostos e os procedimentos para a composição de uma amostra de artigos; resultados, que revela as características bibliométricas desta amostra e expõe uma síntese dos três artigos mais relevantes; considerações finais, que incluem uma descrição das limitações deste trabalho.

2. METODOLOGIA

O presente estudo foi desenvolvido em 3 etapas consecutivas: composição da amostra final de artigos, análise descritiva desta amostra e síntese dos artigos mais relevantes. Neste capítulo serão discutidos os procedimentos da etapa de composição da amostra final de artigos e serão definidas as variáveis necessárias para o desenvolvimento da análise descritiva desta.

2.1 Composição da amostra final

Em consonância com o objetivo geral deste estudo, foram coletados registros de artigos científicos publicados no período 2000-2014, por meio de consultas a duas bases de dados multidisciplinares: Scopus (Elsevier) e Web of Science (Thomson-Reuters) – uma vez que a triangulação de fontes de dados aumenta a confiabilidade da pesquisa (MARTINS; THEÓPHILO, 2009). As expressões-chave utilizadas para a primeira fase das consultas às bases de dados foram: 'Likert-*scale*(s)', 'Likert-*type scale*(s)' ou 'Likert *scale*(s)', sendo que estes termos poderiam constar no título, no resumo ou nas palavras-chave do artigo. Além disso, os filtros de busca incluíam o período definido pelo objetivo da pesquisa (anos 2000 a 2014), os idiomas dos textos (português ou inglês) e o tipo de publicação (artigos). Com estes critérios, foram encontrados 8.411 registros na base de dados Scopus e 8.919 na Web of Science.

Os resultados foram refinados com a limitação da área de pesquisa – '*social sciences*' – e com a busca por termos mais específicos: '*number'*, '*response'*, '*categories'* ou '*points*'; no título, no resumo ou nas palavras-chave dos artigos. Imediatamente, foi realizada a padronização dos nomes dos autores e dos títulos dos periódicos dos 1.781 registros, e subsequente tratamento para eliminação de duplicidades.

A última etapa consistiu de exaustiva análise qualitativa dos 1.501 registros remanescentes, com o propósito de garantir o alinhamento da amostra final com o objetivo geral proposto nesta pesquisa. A princípio, a seleção baseou-se na leitura dos títulos e resumos (*abstracts*); em seguida, os artigos remanescentes foram lidos integralmente. Para a composição da amostra final, foram selecionados 16 artigos[3].

[3] Apêndice A.

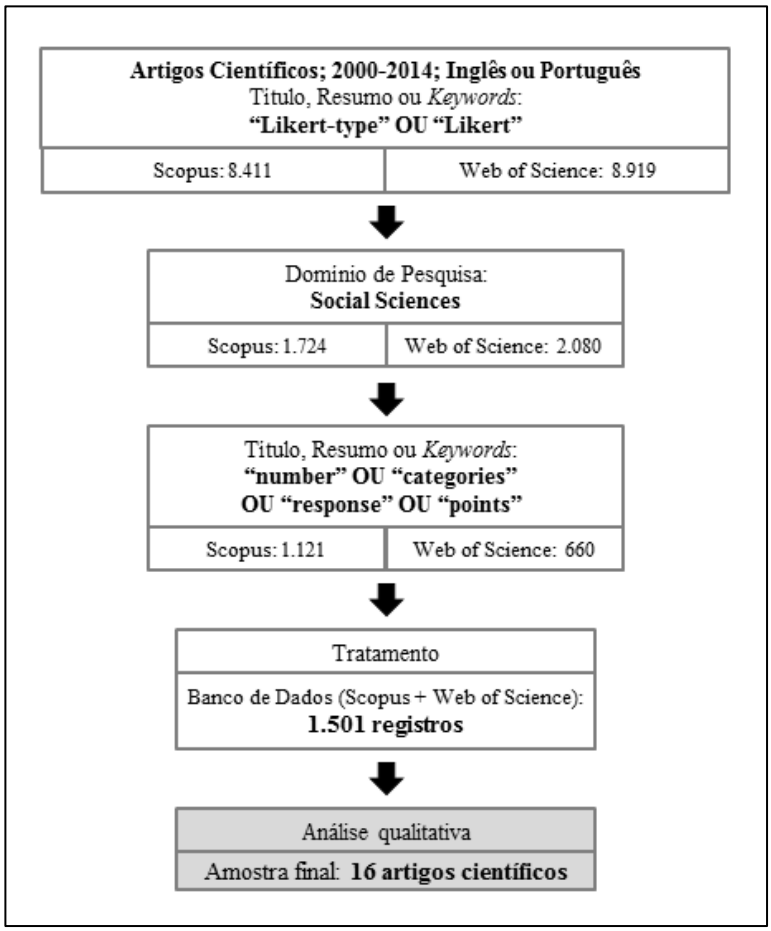

Figura 1 – Coleta, tratamento e análise qualitativa dos registros
Fonte: Elaborada pela autora

2.2 Definição das variáveis

Para cada objetivo específico de pesquisa foram definidas variáveis correspondentes, as quais orientaram a análise descritiva da amostra final de artigos.

A relevância de artigos e autores foi determinada pelo volume de citações em cada base de dados. Assim, para o objetivo específico (1), qualificam-se como mais relevantes os cinco artigos mais citados; os três artigos mais relevantes serão empregados no desenvolvimento da síntese teórica proposta pelo objetivo (5).

Para a identificação dos autores mais citados, como proposto pelo objetivo específico (2), foi empregada a contagem ajustada[4] (fracionamento), sendo cada autor de um artigo

[4] ALVARADO (2002) propõe, segundo o critério de contagem ajustada, a atribuição de **1/(número de autores)** a cada autor de um artigo. Assim, por exemplo, a cada autor de um artigo publicado por dois autores será atribuída 0,5 da contribuição; a cada autor de um artigo publicado por três autores será atribuída 0,33 da contribuição.

creditado com uma fração que multiplica o respectivo volume total de citações[5], em adaptação ao proposto por ALVARADO (2002). O ranking será composto pelos 10 autores com maior fração de citações.

Quanto aos periódicos, objetivo (3), serão apresentados todos aqueles que publicaram os artigos presentes na amostra final.

Quanto à identificação das instituições de ensino e dos países envolvidos – objetivo (4) –, o critério utilizado foi o de contagem completa[6] (ALVARADO, 2002). Para a análise das instituições de ensino, foi considerada a vinculação dos autores às universidades à época da publicação do respectivo artigo. Para a análise dos países, foi considerado o país de origem da universidade com a qual o autor mantinha vínculo à época da publicação.

[5] No método adotado por este estudo, se dará a atribuição de **1/(número de autores)*(volume de citações)** citações a cada autor de um artigo.

[6] Foram atribuídos um ponto para a instituição/país do autor principal e um ponto para a instituição/país do autor secundário, no caso de co-autoria com autores de diferentes instituições de ensino.

3. RESULTADOS

A amostra final é composta por 16 artigos, publicados em 15 periódicos distintos. Estes artigos foram produzidos por 40 autores, os quais mantinham vínculo institucional (à época da publicação de seus respectivos artigos) com 20 universidades distintas, de 11 países.

3.1 Análise Descritiva dos Resultados

3.1.1 Artigos mais relevantes

Os 16 artigos da amostra final têm um total de 715 citações na base de dados Scopus e 514 citações na Web of Science. Segundo o critério estabelecido (volume de citações), foram identificados os 5 artigos mais relevantes, segundo cada base de dados:

Tabela 1 – Artigos mais citados na base de dados Scopus

Artigo	Citações - Scopus
PRESTON; COLMAN (2000)	224
DAWES (2008)	201
WENG (2004)	88
LOZANO; GARCÍA-CUETO; MUÑIZ (2008)	70
WEIJTERS; CABOOTER; SCHILLEWAERT (2010)	38

Tabela 2 – Artigos mais citados na base de dados Web of Science

Artigo	Citações - Web of Science
PRESTON; COLMAN (2000)	189
DAWES (2008)	134
WENG (2004)	81
WEIJTERS; CABOOTER; SCHILLEWAERT (2010)	33
MUÑIZ; GARCÍA-CUETO; LOZANO (2005)	28

Apesar dos rankings serem bastante semelhantes, há dois artigos diferentes em sua composição: Lozano, García-Cueto e Muñiz (2008) foi publicado pelo periódico ‘Methodology: European Journal of Research Methods for the Behavioral and Social Sciences’, o qual não está indexado na base de dados Web of Science e, por esta razão, não

aparece na Tabela 2; assim, o artigo Muñiz, García-Cueto e Lozano (2005) – que, coincidentemente foi publicado pelos mesmos autores daquele – configura a quinta posição na Tabela 5 (sendo que o volume de citações deste na base de dados Scopus é 32).

3.1.2 Autores

Dentre a amostra final de 16 artigos, foram identificados 40 autores. Os dez autores com maior volume de citações fracionadas, por base de dados são:

Tabela 3 – Autores com maior fração de citações na base de dados Scopus

Autor	**Fração de Citações - Scopus**
DAWES, J.	201
COLMAN, A. M.	112
PRESTON, C. C.	112
WENG, L. J.	88
GARCÍA-CUETO, E.	33,66
LOZANO, L. M.	33,66
MUÑIZ, J.	33,66
MOORS, G.	15
CABOOTER, E.	12,54
SCHILLEWAERT, N.	12,54
WEIJTERS, B.	12,54

Tabela 4 – Autores com maior fração de citações na base de dados Web of Science

Autor	**Fração de Citações – Web of Science**
DAWES, J.	134
COLMAN, A. M.	94,5
PRESTON, C. C.	94,5
WENG, L. J.	81
MOORS, G.	13
CABOOTER, E.	10,89
SCHILLEWAERT, N.	10,89
WEIJTERS, B.	10,89
GARCÍA-CUETO, E.	9,24
LOZANO, L. M.	9,24
MUÑIZ, J.	9,24

Apesar da variação na ordem dos autores segundo a fração de citações, as Tabelas 3 e 4 são compostas pelos mesmos autores – os mesmos apresentados na relação dos artigos mais relevantes (Tabelas 1 e 2).

3.1.3 Periódicos

Os 16 artigos da amostra final foram publicados em 15 periódicos:

Tabela 5 – Periódicos

Periódico	**Nº de Artigos**
Educational and Psychological Measurement	3
Acta Psychologica	1
Electronic Journal of Business Research Methods	1
International Journal of Market Research	1
International Journal of Research in Marketing	1
Journal of Business and Psychology	1
Journal of Clinical Epidemiology	1
Journal of Intellectual Disability Research	1
Journal of Psychoeducational Assessment	1
Journal of School Psychology	1
Journal of Social Service Research	1
Methodology: European Journal of Research Methods for the Behavioral and Social Sciences	1
Personality and Individual Differences	1
Quality & Quantity	1

O periódico 'Educational and Psychological Measurement' foi o único que publicou múltiplos artigos da amostra final – Adelson e McCoach (2010), Wakita, Ueshima e Noguchi (2012) e Pearse (2011) –, revelando uma tendência recente da revista, de envolvimento com o tema número de categorias de escala Likert.

3.1.4 Instituições de ensino

Foram identificadas 20 instituições de ensino, a partir da análise dos vínculos institucionais dos autores à época da publicação de seus artigos, pelo método da contagem completa.

Tabela 6 – Instituições de ensino

(continua)

Instituição de Ensino	**Nº de Artigos**
Universidade Federal do Rio Grande do Sul	2
University of Connecticut	2
University of Oviedo	2
Florida State University	1

(conclusão)

Ghent University	1
Kansai University	1
National Taiwan University	1
Rhodes Business School	1
SunYat-sen University	1
Texas A&M University	1
University of Edinburgh	1
University of Georgia	1
University of Jaén	1
University of Leicester	1
University of Louisville	1
University of Macau	1
University of New South Wales	1
University of South Australia	1
University of Tilburg	1
Vlerick Leuven Gent Management School	1

Apenas três instituições de ensino estiveram envolvidas na publicação de múltiplos artigos da amostra final – a Universidade Federal do Rio Grande do Sul/Brasil, representada pelo autor Marcelo Pio de Almeida Fleck; a University of Connecticut/EUA, representada pelos autores D. Betsy McCoach e Dev K. Dalal; e a University of Oviedo/Espanha, representada pelos autores Eduardo García-Cueto e José Muñiz.

3.1.5 Países

Em relação à procedência geográfica, inferida pela identificação do país de origem das instituições de ensino, foram identificados 11 países pelo método de contagem completa.

Tabela 7 – Países

País	**Nº de Artigos**
EUA	4
Austrália	2
Brasil	2
China	2
Espanha	2
Reino Unido	2
África do Sul	1
Bélgica	1
Holanda	1
Japão	1
Taiwan	1

3.2 Síntese dos três artigos mais relevantes

O primeiro estudo sobre construção de escalas foi publicado na década de 1920[7]. Por sua vez, a primeira pesquisa que abordou, especificamente, a questão do número de categorias de resposta foi publicada apenas em 1939[8] (CHAFOULEAS; CHRIST; RILEY-TILLMAN, 2009). Nos últimos 75 anos, diversos estudos foram publicados, com foco no número de categorias de resposta segundo critérios de validade e confiabilidade, bem como sob a ótica de percepção e preferência dos respondentes (LEUNG, 2011; PRESTON; COLMAN, 2000). Diversos autores chegaram à conclusão de que, de fato, não há um número ótimo de categorias de resposta apropriado a qualquer objeto de pesquisa. Alguns acreditam que o número de categorias de resposta deve ser determinado com base no propósito ou na natureza da escala, na discriminação do estímulo ou no número de itens que compõem o instrumento de pesquisa. Outros, que as propriedades do formato de uma escala não são tão importantes quanto o conhecimento, a habilidade e a motivação dos respondentes (CHAFOULEAS; CHRIST; RILEY-TILLMAN, 2009). Em consonância com este argumento, Miller[9] (1956 apud PRESTON; COLMAN, 2000) sugere que a mente humana consegue distinguir, na avaliação de um estímulo, aproximadamente sete categorias diferentes. Assim, para alcançar níveis expressivos de confiabilidade e chegar a conclusões de pesquisa adequadas, o pesquisador deve considerar a capacidade cognitiva de discriminação da população-alvo de sua pesquisa durante o processo de construção das escalas (WENG, 2004).

Como proposto pelo objetivo específico (5) deste estudo, será apresentada uma síntese dos métodos e resultados dos três artigos mais relevantes (com o maior volume de citações nas bases de dados Scopus e Web of Science) da amostra final.

3.2.1 Preston e Colman (2000)

O artigo examinou a validade, a confiabilidade e o poder de discriminação de escalas Likert com diferentes números de categorias de resposta (2, 3, 4, 5, 6, 7, 8, 9, 10 ou 11), bem

[7] SYMONDS, P. M. On the loss of reliability in ratings due to coarseness of the scale. **Journal of Experimental Psychology**, v. 7, n. 6, p. 456-461, dez. 1924.
[8] CHAMPNEY, H.; MARSHALL, H. Optimal refinement of the rating scale. **Journal of Applied Psychology**, v. 23, n. 3, p. 323-331, jun. 1939.
[9] MILLER, G. A. The magical number seven, plus or minus two: some limits on our capacity for processing information. **Psychological Review**, v. 63, n. 2, p. 81-97, mar. 1956.

como a opinião dos respondentes sobre cada uma destas, em termos de facilidade de uso, velocidade de resposta e possibilidade de expressão (por meio de notas para cada atributo, de 0 a 100). O questionário, que tinha como objetivo a avaliação da qualidade do serviço de uma loja ou restaurante (baseada em experiências pessoais reais), foi aplicado a uma amostra de 149 respondentes (sendo 134 estudantes da Universidade de Leicester). Uma semana depois, o questionário foi reaplicado a 129 respondentes, para avaliação da confiabilidade pela técnica de teste-reteste.

Segundo critérios de confiabilidade, validade e poder de discriminação, as escalas com menor número de categorias tiveram desempenho inferior. As escalas com maior confiabilidade foram aquelas com sete e com dez categorias de resposta; e as escalas com maior validade e poder de discriminação foram as escalas com seis ou mais categorias de resposta.

Quanto às opiniões dos respondentes sobre as escalas com diferentes números de categorias de resposta: as escalas com cinco, sete e dez categorias foram indicadas como as mais fáceis de usar; as escalas com menor número de categorias de resposta (duas, três e quatro) foram indicadas como as mais rápidas para responder; e, em termos de possibilidade de expressão, as escalas com nove, dez e onze categorias foram apontadas como as melhores.

De forma geral, as escalas com sete, nove e dez categorias de resposta tiveram o melhor desempenho. No entanto, segundo o autor, a seleção de um número de categorias para a construção de uma escala depende das circunstâncias da coleta de dados, que impõe ao pesquisador *trade-offs* entre validade, confiabilidade, poder de discriminação e preferências dos respondentes.

3.2.2 Dawes (2008)

O artigo examinou a influência do número de categorias de resposta nas estatísticas descritivas dos dados coletados (média, variabilidade, assimetria e curtose). Foram realizadas coletadas de dados sobre consciência de preços via telefone, usando instrumentos compostos por escalas com diferentes números de categorias de resposta (5, 7 ou 10). Para a análise, os dados gerados pelas escalas com cinco e com sete categorias de resposta foram transformados, facilitando comparações com a escala de dez categorias de resposta.

Observou-se que as escalas transformadas (com cinco ou sete categorias de resposta em sua forma original) produzem maiores médias, se comparadas com os dados coletados

com o formato de dez categorias. Além disso, foi detectada a tendência de uso de mais categorias de resposta, quando um número maior de categorias é apresentado. Não foram significativas as diferenças entre variabilidade, assimetria ou curtose (as estatísticas descritivas geradas pelos três formatos de escala são bastante semelhantes).

3.2.3 Weng (2004)

O artigo examinou escalas com diferentes números de categorias de resposta (3, 4, 5, 6, 7, 8 ou 9), segundo dois critérios de confiabilidade: o alfa de Cronbach, que fornece informações sobre a consistência interna, e teste-reteste, que reflete a estabilidade de dados coletados ao longo do tempo. Foram analisadas dados de um questionário aplicado para a seleção de alunos para um programa de qualificação (The Teacher Attitude Test, para o Teacher Educational Program da National Taiwan University). A amostra foi composta por 1.247 estudantes de 13 universidades, que responderam duas vezes o mesmo formato de questionário, com um intervalo de 4 semanas.

A confiabilidade de uma das escalas analisadas revelou-se independente das variações no número de categorias de resposta, sendo que o autor atribuiu este resultado ao poder de mediação da carga fatorial na relação entre formato de escala e confiabilidade.

Ainda assim, o autor recomenda o uso das escalas com seis ou sete categorias de resposta para pesquisas com universitários (ou indivíduos com capacidade cognitiva semelhante), pois estas têm maior chance de atingir superior confiabilidade e gerar resultados de pesquisa consistentes.

4. CONSIDERAÇÕES FINAIS

Os objetivos propostos pelo artigo – a identificação dos artigos mais relevantes, a identificação dos autores mais citados, a identificação dos periódicos, instituições de ensino e países envolvidos na produção acadêmica sobre número de categorias Likert nos últimos 15 anos, além da síntese dos três artigos mais relevantes – foram alcançados, sendo apresentados na seção Resultados.

As realizações mencionadas contribuem para a difusão do conhecimento científico disponível e para a construção de melhores instrumentos de coleta de dados para futuras pesquisas quantitativas, acadêmicas ou mercadológicas.

Dentre as possíveis limitações desta investigação estão as bases de dados consultadas, que não abrangem a produção de diversos países e a publicação de diversos periódicos, e as palavras-chave utilizadas para as consultas a estas bases de dados, podendo ter omitido publicações importantes e condizentes com o objetivo central deste trabalho.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALLEN, I. E.; SEAMAN, C. A. Likert scales and data analyses. **Quality Progress**, jul. 2007. Disponível em: <http://asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html>. Acesso em: 02 dez. 2014.

ALLIK, J. A mixed-binomial model for Likert-type personality measures. **Frontiers in Psychology**, v. 5, maio 2014. Disponível em: <http://journal.frontiersin.org/Journal/10.3389/fpsyg.2014.00371/full>. Acesso em: 02 dez. 2014.

ALVARADO, R. U. A Lei de Lotka na bibliometria brasileira. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 14-20, maio-ago. 2002. Disponível em: <http://revista.ibict.br/cienciadainformacao/index.php/ciinf/article/view/141/121>. Acesso em: 24 nov. 2014.

CHAFOULEAS, S. M.; CHRIST, T. J.; RILEY-TILLMAN, T. C. Generalizability of scaling gradients on direct behavior ratings. **Educational and Psychological Measurement**, vol. 69, n. 1, p. 157-173, fev. 2009. Disponível em: <http://epm.sagepub.com/content/69/1/157>. Acesso em: 11 dez. 2014.

EDMONSON, D. R. Likert scales: a history. In: Conference on Historical Analysis and Research in Marketing, 12., 2005, Long Beach. **Proceedings**… Long Beach: CHARM, 2005. p. 127-133. Disponível em: <http://faculty.quinnipiac.edu/charm/cumulative_proceedings.htm>. Acesso em: 05 dez. 2014.

EVANS, J. R.; MATHUR, A. The value of online surveys. **Internet Research**, v. 15, n. 2, p. 195-219, 2005. Disponível em: <http://www.emeraldinsight.com/doi/abs/10.1108/10662240510590360>. Acesso em: 28 dez. 2014.

GÖB, R.; McCOLLIN, C.; RAMALHOTO, M. F. Ordinal methodology in the analysis of Likert scales. **Quality & Quantity**, v. 41, n. 5, p. 601-626, out. 2007. Disponível em: <http://link.springer.com/article/10.1007%2Fs11135-007-9089-z>. Acesso em: 10 dez. 2014.

HODGE, D. R.; GILLESPIE, D. Phrase completions: an alternative to Likert scales. **Social Work Research**, v. 27, n. 1, p. 45-55, 2003. Disponível em: <http://swr.oxfordjournals.org/content/27/1/45.short>. Acesso em: 18 dez. 2014.

JAMIESON, S. Likert scales: how to (ab)use them. **Medical Education**, v. 38, n. 12, p. 1217-1218, dez. 2004. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2929.2004.02012.x/abstract>. Acesso em: 14 dez. 2014.

LEUNG, S. O. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. **Journal of Social Service Research**, vol. 37, n. 4, p. 412-421, jul.-set. 2011. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true&db=sih&AN=63295558&lang=pt-br&site=ehost-live>. Acesso em: 12 dez. 2014.

MARTINS, G. A.; THEÓPHILO, C. R. **Metodologia da investigação científica para ciências sociais aplicadas**. 2. ed. São Paulo: Atlas, 2009. Acesso em: 10 nov. 2014.

MORAN, M. R. et al. Alianças estratégicas: uma análise bibliométrica da produção científica entre 1989 e 2008. **Revista de Ciências da Administração**, Florianópolis, v. 12, n. 27, p. 63-85, maio-ago. 2010. Disponível em: <https://periodicos.ufsc.br/index.php/adm/article/view/2175-8077.2010v12n27p63/17414>. Acesso em: 27 nov. 2014.

PEARSE, N. Deciding on the scale granularity of response categories of likert type scales: the case of a 21-point scale. **Electronic Journal of Business Research Methods**, vol. 9, n. 2, p. 159-171, set. 2011. Disponível em: <http://www.ejbrm.com/volume9/issue2>. Acesso em: 10 dez. 2014.

PRESTON, C. C.; COLMAN, A. M. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. **Acta Psychologica**, vol. 104, n. 1, p. 1-15, mar. 2000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0001691899000505>. Acesso em: 13 dez. 2014.

ROSZKOWSKI, M. J.; SOVEN, M. Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. **Assessment & Evaluation in Higher Education**, v. 35, n. 1, p. 117-134, jan. 2010. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true&db=afh&AN=49142000&lang=pt-br&site=ehost-live>. Acesso em: 13 dez. 2014.

WEIJTERS, B.; CABOOTER, E.; SCHILLEWAERT, N. The effect of rating scale format on response styles: the number of response categories and response category labels. **International Journal of Research in Marketing**, vol. 27, n. 3, p. 236-247, set. 2010. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167811610000303>. Acesso em: 10 dez. 2014.

WENG, L. J. Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. **Educational and Psychological Measurement**, vol. 64, n. 6, p. 956-972, dez. 2004. Disponível em: <http://epm.sagepub.com/content/64/6/956.short>. Acesso em: 11 dez. 2014.

APÊNDICE A – Amostra final (16 artigos)

ADELSON, J.L.; McCOACH, D.B. Measuring the mathematical attitudes of elementary students: the effects of a 4-point or 5-point likert-type scale. **Educational and Psychological Measurement**, vol. 70, n. 5, p. 796-807, out. 2010. Disponível em: <http://epm.sagepub.com/content/70/5/796>.

CHACHAMOVICH, E.; FLECK, M. P.; POWER, M. Literacy affected ability to adequately discriminate among categories in multipoint Likert Scales. **Journal of Clinical Epidemiology**, vol. 62, n. 1, p. 37-46, jan. 2009. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0895435608000644>.

DALAL, D. K.; CARTER, N. T.; LAKE, C. J. Middle response scale options are inappropriate for ideal point scales. **Journal of Business and Psychology**, vol. 29, n. 3, p. 463-478, set. 2014. Disponível em: <http://link.springer.com/article/10.1007%2Fs10869-013-9326-5>.

DAWES, J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. **International Journal of Market Research**, vol. 50, n. 1, p. 61-77, 2008. Disponível em: <http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2013613>.

FANG, J. et al. The response scale for the intellectual disability module of the WHOQOL: 5-point or 3-point? **Journal of Intellectual Disability Research**, vol. 55, n. 6, p. 537-549, jun. 2011. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2788.2011.01401.x/pdf>.

LEE, J.; PAEK, I. In search of the optimal number of response categories in a rating scale. **Journal of Psychoeducational Assessment**, vol. 32, n. 7, p. 663-673, out. 2014. Disponível em: <http://jpa.sagepub.com/content/32/7/663>.

LEUNG, S. O. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. **Journal of Social Service Research**, vol. 37, n. 4, p. 412-421, jul.-set. 2011. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true&db=sih&AN=63295558&lang=pt-br&site=ehost-live>.

LOZANO, L. M.; GARCÍA-CUETO, E.; MUÑIZ, J. Effect of the number of response categories on the reliability and validity of rating scales. **Methodology: European Journal of Research Methods for the Behavioral and Social Sciences**, vol. 4, n. 2, p. 73-79, 2008. Disponível em: <http://psycontent.metapress.com/content/n2j8027383rg1268/?genre=article&id=doi%3a10.1027%2f1614-2241.4.2.73>.

MOORS, G. Exploring the effect of a middle response category on response style in attitude measurement. **Quality & Quantity**, vol. 42, n. 6, p. 779-794, dez. 2008. Disponível em: <http://link.springer.com/article/10.1007%2Fs10869-008-9064-2>.

MUÑIZ, J.; GARCÍA-CUETO, E.; LOZANO, L. M. Item format and the psychometric properties of the Eysenck Personality Questionnaire. **Personality and Individual Differences**, vol. 38, n. 1, p. 61-69, jan. 2005. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0191886904000820>.

PARKER, R. I.; VANNEST, K. J.; DAVIS, J. L. Reliability of multi-category rating scales. **Journal of School Psychology**, vol. 51, n. 2, p. 217-229, abr. 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022440512001100>.

PEARSE, N. Deciding on the scale granularity of response categories of likert type scales: the case of a 21-point scale. **Electronic Journal of Business Research Methods**, vol. 9, n. 2, p. 159-171, set. 2011. Disponível em: <http://www.ejbrm.com/volume9/issue2>.

PRESTON, C. C.; COLMAN, A. M. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. **Acta Psychologica**, vol. 104, n. 1, p. 1-15, mar. 2000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0001691899000505>.

WAKITA, T.; UESHIMA, N.; NOGUCHI, H. Psychological distance between categories in the Likert scale: comparing different numbers of options. **Educational and Psychological Measurement**, vol. 72, n. 4, p. 533-546, ago. 2012. Disponível em: <http://epm.sagepub.com/content/72/4/533>.

WEIJTERS, B.; CABOOTER, E.; SCHILLEWAERT, N. The effect of rating scale format on response styles: the number of response categories and response category labels. **International Journal of Research in Marketing**, vol. 27, n. 3, p. 236-247, set. 2010. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167811610000303>.

WENG, L. J. Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. **Educational and Psychological Measurement**, vol. 64, n. 6, p. 956-972, dez. 2004. Disponível em: <http://epm.sagepub.com/content/64/6/956.short>.

ELSEVIER

Acta Psychologica 104 (2000) 1–15

acta psychologica

www.elsevier.com/locate/actpsy

Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences

Carolyn C. Preston [a], Andrew M. Colman [b,*]

[a] *Department of General Practice and Primary Health Care, University of Leicester, University Road, Leicester LE1 7RH, UK*

[b] *Department of Psychology, University of Leicester, University Road, Leicester LE1 7RH, UK*

Received 13 April 1999; received in revised form 13 September 1999; accepted 14 September 1999

Abstract

Using a self-administered questionnaire, 149 respondents rated service elements associated with a recently visited store or restaurant on scales that differed only in the number of response categories (ranging from 2 to 11) and on a 101-point scale presented in a different format. On several indices of reliability, validity, and discriminating power, the two-point, three-point, and four-point scales performed relatively poorly, and indices were significantly higher for scales with more response categories, up to about 7. Internal consistency did not differ significantly between scales, but test–retest reliability tended to decrease for scales with more than 10 response categories. Respondent preferences were highest for the 10-point scale, closely followed by the seven-point and nine-point scales. Implications for research and practice are discussed.

PsycINFO classification: 2220

Keywords: Item analysis; Questionnaires; Rating scales; Test reliability; Test validity

* Corresponding author. Tel.: +44-1-16-2522167; fax: +44-1-16-2522067.
E-mail address: amc@le.ac.uk (A.M. Colman).

0001-6918/00/$ - see front matter
PII: S0001-6918(99)00050-5

1. Introduction

Rating scales are among the most widely used measuring instruments in psychology, and it is therefore not surprising that a great deal of research has been devoted to the effects of variations in rating scale format, including differences in the number of response categories. In current practice, most rating scales, including Likert-type scales and other attitude and opinion measures, contain either five or seven response categories (Bearden, Netmeyer, & Mobley, 1993; Peter, 1979; Shaw & Wright, 1967).

In spite of decades of research, the issue of the optimal number of response categories in rating scales is still unresolved. Some investigators have studied response patterns and information retrieval. Schutz and Rucker (1975) found in their study of response patterns that "the number of available response categories does not materially affect the cognitive structure derived from the results" (p. 323), which seems to suggest that the number of response categories has little effect on the results obtained. This conclusion is not in line with the findings of other studies, which have provided support for the use of scales with more than two or three response categories. For example, Garner (1960) suggested that maximum information is obtained by using more than 20 response categories. Green and Rao (1970), on the other hand, found that information retrieval is maximized by using six or seven response categories, with little extra information being gained by increasing the number of categories beyond seven.

Symonds (1924) was the first to suggest that reliability (in this case inter-rater reliability) of scores is optimized by the use of seven categories. This suggestion was contested by Champney and Marshall (1939), who advocated the use of finer scales, but the case for seven-point scales was strengthened by Miller (1956), who suggested in an influential article that the human mind has a span of apprehension capable of distinguishing about seven different items (plus or minus two), which implies a limit of about seven on the number of categories that people are able to use in making judgments about the magnitudes of unidimensional stimuli. This has implications for rating scales: the limit on the human span of apprehension suggests that little if any additional information can be obtained by increasing the number of response categories beyond about seven. The reliability of scores derived from scales with different numbers of response categories was later investigated by Bendig (1953, 1954), who found relatively constant test–retest reliabilities over scales with two, three, five, seven, and nine response categories, relatively constant inter-rater reliability over scales with three, five, seven, and nine response categories, and a decrease in reliability for 11-point scales. A few subsequent researchers confirmed Bendig's finding that reliability is largely independent of the number of response categories (e.g., Boote, 1981; Brown, Wilding, & Coulter, 1991; Komorita, 1963; Matell & Jacoby, 1971; Peabody, 1962; Remington, Tyrer, Newson-Smith, & Cicchetti, 1979).

Some researchers in this area have arrived at different conclusions regarding reliability. In a study based on Monte-Carlo simulation methods, Cicchetti, Showalter and Tyrer (1985) found evidence for an increase in inter-rater reliability from two-point to seven-point scales; beyond this – even up to 100 response categories – no

substantial increase in reliability was found. These researchers concluded that "the differences in scale reliability between a 7-, 8-, 9-, or 10-category ordinal scale on one hand, and a 100-point or continuous scale on the other is trivial . . . 7 ordinal categories of response appear at least functionally interchangeable with as many as 100 such ordered categories" (p. 35). Similar conclusions were drawn by Oaster (1989) with regard to test–retest reliability and inter-item consistency, and a number of other researchers have reported that reliability is maximized with seven-point scales (Finn, 1972; Nunnally, 1967; Ramsay, 1973). These studies provide support for the early findings of Symonds (1924), mentioned above, and more generally for the continued popularity of seven-point scales in practice (see Cox, 1980, for a review). A few researchers have, however, reported higher reliabilities for five-point scales (Jenkins & Taber, 1977; Lissitz & Green, 1975; McKelvie, 1978; Remmers & Ewart, 1941), and a recent study using the multitrait-multimethod approach found evidence for higher monotrait-monomethod (MTMM) reliability in four-point than six-point scales (Chang, 1994).

In a comparatively small number of studies, validity has been used as a criterion for judging the performance of scales with different numbers of response categories. Matell and Jacoby (1971) carried out a thorough empirical study comparing scales with varying numbers of response categories (from 2 to 19) and concluded that as few as two response categories may be adequate in practice. They suggested that both reliability and validity are independent of the number of response categories, and their results implied that collapsing data from longer scales into two-point or three-point scales would not diminish the reliability or validity of the resulting scores. Loken, Pirie, Virnig, Hinkle and Salmon (1987) examined the criterion validity of various scales through their ability to differentiate between different population groups and found 11-point scales to be superior to three-point or four-point scales. Hancock and Klockars (1991) found that nine-point scale scores correlated better than five-point scale scores with objective measures of the original stimuli. In a comparison using a MTMM covariance matrix, Chang (1994) found approximately similar criterion validity coefficients for four-point and six-point scales but higher convergent validity coefficients for the six-point scales. Taken together, these studies tend to suggest that validity increases with increasing numbers of response categories or scale points.

Respondent preferences have not been investigated in depth in previous studies of rating scales. However, Jones (1968) examined respondents' preferences for scales with two or seven response categories and a graphic rating scale and reported that the dichotomous scale was judged to be less "accurate", less "reliable", less "interesting", and more "ambiguous" than both the seven-point and the graphic rating scales, but the two-point and graphic rating scales were judged to be easier to use. Respondents expressed a clear preference for multiple-category over dichotomous scales.

The aim of the investigation reported below is to provide a thorough assessment, using multiple independent criteria, of the reliability, validity, and discriminating power of scores from rating scales varying widely in number of response categories. A secondary aim is to investigate the other important issue in rating scale design,

namely respondent preferences. Our research is more detailed and thorough than most previous studies in the area, and its design allows us to examine not only several objective indices of reliability, validity, and discriminating power, but also subjective measures of respondents' opinions about the scales. For example, if a scale is too difficult to use, or too simple to allow respondents to express themselves, then respondents may become frustrated and demotivated, and the quality of their responses may decrease. In addition to considerations of reliability, validity, and discriminating power, a test designer may wish to take respondent preferences into account when constructing a rating scale.

1.1. Materials and methods

A questionnaire was administered to 149 respondents (45 males and 104 females, aged from 18 to over 60 with a mode of 20), the majority of whom (134) were undergraduate students at the University of Leicester. The respondents were recruited by a form of snowball sampling in which students who volunteered to participate as respondents also recruited additional respondents in return for course credits. The sample thus consisted of undergraduate students and their friends (some of whom were also undergraduate students) and relatives.

Each respondent filled in a questionnaire consisting of a number of rating scales with varying numbers of response categories relating to the quality of service provided by either a store or a restaurant of which he or she had personal experience. These two service categories were chosen on the assumption that all respondents would have visited a store or restaurant in the recent past, and this turned out to be the case. Respondents were first required to provide a global rating of their chosen store or restaurant for overall quality from 0 (*extremely bad*) to 100 (*extremely good*); this global rating served as a criterion measure for later assessments of validity. The rest of the questionnaire consisted of 11 sets of five questions each, in which respondents rated the quality of five service elements for their chosen store or restaurant. To avoid order-of-presentation effects that might have arisen from using a single random order of presentation, the 11 sets were presented in a different randomized order to each respondent. For the restaurant, the five service elements were "competence of staff", "promptness of service", "range of choice", "cleanliness of premises", and "individual attention"; for the store, the service elements were "helpfulness of staff", "promptness of service", "range of products", "tidiness of premises", and "individual attention".

The 11 sets of questions to which the participants responded were identical apart from the number of response categories in the rating scales associated with each of the five service elements in each set. The number of response categories of the rating scales were 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11. Each scale was end-anchored with *very poor* on the left and *very good* on the right and was presented as a series of numbers spaced out across the page representing the response categories. Respondents were required to circle the appropriate number in each case depending on their opinions or judgments. A 101-point scale was also included, with respondents being asked to rate each service element from 0 (*very poor*) to 100 (*very good*), but in this case no

numbers were displayed: respondents were required simply to write down appropriate numbers between 0 and 100 reflecting their ratings.

In order to measure respondents' preferences for the various scales, they were also asked to rate each scale on three different aspects of scale performance. Using the 101-point scale described above, from 0 (*very poor*) to 100 (*very good*), they rated each scale on the following three criteria: "ease of use", "quick to use", and "allowed you to express your feelings adequately".

In order to provide data for an assessment of test–retest reliability, each respondent completed an identical retest questionnaire between one and three weeks after the original testing – the second questionnaire was issued to the respondents one week after the first and had to be returned within two weeks from that date. Of the 149 respondents who completed the first questionnaire, 129 completed the second within the two-week time limit, and the retest results are based on their responses. This response rate of 86% is high for research of this type.

2. Results

2.1. Preliminary checks

The psychometric criteria that are used to compare scales with different numbers of response categories make sense only if such scales measure the same construct. We therefore carried out three preliminary checks.

First, we used a procedure described by Green (1992) to determine whether the 12 correlation matrices representing the associations between the five items in each scale, one correlation matrix per scale length, differed significantly from one another. The weighted least squares chi-square value was 144.09, $df = 155, p = 0.72$, suggesting no evidence of significant differences between the correlation matrices.

Second, the unnormed and normed fit indices (Bentler & Bonnet, 1980) that were achieved were 1.00 and 0.96, respectively; according to Bentler and Bonnet (1980), this indicates that little additional increase in goodness of fit would be likely to be achieved by allowing some of the correlation matrices to differ from others.

Third, we performed a maximum-likelihood factor analysis on each of the 11 correlation matrices and then tested whether the ordinal structure of the factor loadings was the same across matrices. The rank-order of the factor loadings within each scale length (matrix) turned out to be identical across all 11 matrices, yielding a Kendall coefficient of concordance of 1.00.

All three of these preliminary checks confirm that the scales with different numbers of response categories all measure the same underlying construct.

2.2. Reliability

Ratings derived from each scale were evaluated for test–retest reliability (stability) and also, using Cronbach's alpha, for consistency over the five questions within each scale type (internal consistency reliability). Table 1 shows the reliability coefficients

Table 1
Reliability of rating scales with different numbers of response categories

Test–retest	Response categories										
	2	3	4	5	6	7	8	9	10	11	101
Reliability	0.88	0.86	0.89	0.91	0.92	0.93	0.94	0.94	0.93	0.92	0.90
Cronbach's α	0.81	0.79	0.82	0.82	0.83	0.85	0.85	0.85	0.85	0.86	0.85

for the test–retest reliability analysis and the alpha coefficients for the internal consistency reliability analysis.

The reliability coefficients are all relatively high (0.79 or above) and statistically significant beyond $p < 0.05$, and the effect sizes are all large according to Cohen's (1988, 1992) criterion. Test–retest reliability coefficients were lowest for two-point, three-point, and four-point scales and highest for scales with about 7 to 10 response categories; there was a slight decline in test–retest reliability coefficients for scales with 11 and 101 response categories. Statistical tests for the significance of differences between correlations were carried out using Fisher's r to z transformation (Howell, 1992, p. 251). Statistically significant at $p < 0.05$ were the differences between the two-point scale and the scales with 6, 7, 8, 9, and 10 response categories; between the three-point scale and the scales with 6, 7, 8, 9, 10, and 11 response categories; between the four-point scale and the eight-point and nine-point scales; and between the 101-point scale and the eight-point and nine-point scales. All other differences between the test–retest reliability coefficients were statistically nonsignificant.

Cronbach alpha coefficients were lowest for two-point and three-point scales, and like the test–retest reliability coefficients they increased with increasing numbers of response categories up to seven. Little further increase in reliability was found above seven response categories: alpha coefficients for scales with 7, 8, 9, 10, 11, and 101 response categories were all very similar. Using Feldt's test for the significance of differences between more than two related alpha coefficients (Woodruff & Feldt, 1986), it was determined that none of the differences between the alpha coefficients were statistically significant: using either of the recommended statistics HAN2 or UX1, $\chi^2(10) < 18.31$, $p > 0.05$.

2.3. Validity and discriminating power

Validity and discriminating power were assessed in several different ways. First, an index of the criterion validity of scores derived from each scale was created by computing an aggregate score for each respondent over the questions relating to the five service elements and then correlating these aggregate scores with the respondents' scores on the criterion global rating of overall quality of service. The global rating of overall service quality was included in the same questionnaire as the ratings of particular service elements, but it was separated from these other more specific measures. There are many precedents for the use as validity criteria of scales that were included in the same questionnaire as the scales in question (Althauser,

Heberlein, & Scott, 1971; Campbell & Fiske, 1959; Messick, 1993). The global rating that was used in this study was chosen because there was no genuinely *external* criterion available, and because it is reasonable to assume that ratings of particular service elements, if they measure what they purport to measure, ought to correlate with global ratings of overall service quality.

Next, an index of the intertertile discriminating power of scores from each scale was obtained by calculating a t statistic relating to the mean difference in aggregate scores for the scale's five questions between the tertile of respondents who provided the highest global ratings of overall service quality (over 80 on the scale from 0 to 100) and the tertile who provided the lowest global quality rating (0–25). To provide a second index of the discriminating power of each scale, item-whole correlations were calculated between the ratings of each of the five service elements and the aggregate score from all scales relating to the corresponding service element. For each scale, the index of discriminating power was provided by the mean item-whole correlation over the five questions in the scale. For these calculations, ratings from all scales apart from the 101-point scale were rescaled to make them comparable using the following formula:

$$(\text{rating} - 1)/(\text{number of response categories} - 1) \times 100.$$

Finally, convergent validity (Campbell & Fiske, 1959) was evaluated by examining the correlations of scores on each scale with scores on each of the others. Scores from a scale were assumed to show convergent validity to the extent to which they correlated with scores from other scales measuring the same underlying construct (see also Althauser et al., 1971; American Psychological Association, 1985, pp. 9–10; Chang, 1994; Messick, 1993).

Data concerning criterion validity, intertertile discriminating power, and item-whole correlations are presented in Table 2, and data concerning convergent validity are presented in Table 3.

The results in Table 2 show that the pattern for validity and discriminating power is similar to the pattern that was found for reliability. Correlations between scale scores and the criterion variable, t values for intertertile discriminating power,

Table 2
Criterion validity, intertertile discriminating power, and item-whole correlations[a]

	Response categories										
	2	3	4	5	6	7	8	9	10	11	101
Criterion											
Validity (r)	0.83	0.82	0.85	0.87	0.88	0.87	0.87	0.89	0.87	0.88	0.89
Intertertile											
Discrim. (t)	19.2	16.8	18.6	20.8	21.5	20.8	22.2	23.7	22.3	23.0	23.4
Item-whole (r)	0.87	0.90	0.92	0.95	0.96	0.97	0.97	0.96	0.96	0.96	0.96

[a] All t statistics are significant at $p < 0.0001$.

Table 3
Convergent validity: intercorrelations between scales with different numbers of response categories

Categories	2	3	4	5	6	7	8	9	10	11	101
2											
3	0.836										
4	0.859	0.878									
5	0.857	0.906	0.921								
6	0.884	0.899	0.928	0.956							
7	0.884	0.907	0.922	0.956	0.969						
8	0.893	0.908	0.923	0.949	0.973	0.964					
9	0.882	0.907	0.913	0.959	0.965	0.969	0.972				
10	0.880	0.903	0.905	0.955	0.959	0.964	0.975	0.977			
11	0.871	0.902	0.905	0.950	0.958	0.963	0.966	0.970	0.974		
101	0.883	0.910	0.902	0.933	0.954	0.956	0.962	0.964	0.952	0.959	

and item-whole correlations were statistically significant for all scales, and the effect sizes are all large according to Cohen's (1988, 1992) criteria. As regards criterion validity, scores from the scales with two, three, and four response categories produced the lowest correlations (below $r = 0.86$) with the criterion. Correlations for scores from 5-point to 101-point scales were higher and very similar to each other (around $r = 0.88$). The scales that yielded scores with the highest criterion validity coefficients ($r = 0.89$) were those with 9 and 101 response categories. However, none of the criterion validity coefficients shown in Table 2 differs significantly at $p < 0.05$ from any of the others according to Williams's t statistic (Howell, 1992, p. 254).

Turning to intertertile discriminating power, scales with two, three, and four response categories again performed least well, with scores from the three-point scale showing the lowest intertertile discriminating power of all $g < 20$). Scores from the other scales showed similar discriminating power to one another, with nine-point, 11-point, and 101-point scales performing best. The significance of the differences between the intertertile discriminating power t values was tested by transforming the t values to z scores (Howell, 1992, p. 251). The only statistically significant differences at $p < 0.05$ were between the t values for the three-point scale on the one hand and the scales with nine and 101 response categories on the other.

Item-whole correlations were also lowest for the scales with two, three, or four response categories and increased with increasing numbers of response categories up to about six. There was no substantial change in item-whole correlation coefficients for the scales with more than six response categories. Fisher's r to z transformation was used to evaluate the significance of differences between the item-whole correlations. The coefficient for the two-point scale differed at $p < 0.05$ from all of the others. Coefficients for the three-point and four-point scales also differed at $p < 0.05$ from all the others but did not differ statistically significantly from each other. All other differences were nonsignificant.

Table 3 shows the intercorrelations between scores from scales with different numbers of response categories. Every scale correlated highly and statistically significantly with each of the others, and all of the correlation coefficients represent large effect sizes according to Cohen's (1988, 1992) criterion. These results, which provide evidence of convergent validity, also indicate that the scales with relatively more response categories (six or more) correlated best with one another, and that the two-point and three-point scales correlated less highly with the longer scales.

2.4. Respondent preferences

Using a 101-point scale described above, respondents rated each scale for its "ease of use", whether it was "quick to use", and whether it "allowed you to express your feelings adequately". Analysis of variance was carried out on the ratings given in response to each of these questions. The mean score for each scale on each of the three questions designed to measure respondent preferences, and the associated standard deviations, are shown in Table 4, along with the corresponding F values. The effect sizes were all large according to Cohen's (1988, 1992) criterion: "ease of

Table 4
Respondents' preferences for scales: mean ratings (0 = very poor, 100 = very good), and standard deviations

	Response categories											
	2	3	4	5	6	7	8	9	10	11	101	*F*
Ease of use	78.6	81.4	82.0	83.7	81.3	82.3	81.5	81.0	83.2	76.7	74.1	7.57*
(*SD*)	27.3	18.9	17.9	15.6	16.3	15.7	16.1	17.4	16.2	19.4	21.6	
Quick to use	86.6	86.8	85.5	85.1	84.5	83.5	83.1	82.1	82.9	77.8	70.6	25.23*
(*SD*)	18.9	15.8	15.2	14.2	15.0	15.5	15.9	16.6	17.0	19.6	20.5	
Express feelings	17.8	40.0	52.0	63.7	63.4	69.0	68.8	72.9	76.0	73.1	79.3	219.36*
(*SD*)	20.0	22.6	23.1	20.7	21.4	21.3	20.1	20.4	21.1	21.8	22.5	

* $p < 0.0001$.

use'', $\eta^2 = 34$; ''quick to use'', $\eta^2 = 0.63$; ''allowed you to express your feelings adequately'', $\eta^2 = 0.94$.

Table 5 shows the scales ordered from the one that received the lowest mean score (that is, the one that was rated least favorably) to the one that received the highest mean score (that is, the one that was rated most favorably) on each of the three questions related to respondent preferences. Scales in the same row that share the same subscript are *not* significantly different; all other differences between scales in the same row are statistically significant at $p < 0.05$ according to the Tukey-HSD multiple comparisons.

These respondent preference scores show several statistically significant differences between the scales. For ''ease of use'', the scales with five, seven, and 10 response categories were the most preferred, and the scales with 11 and 101 response categories were rated as least easy to use. The scales that were rated as most ''quick to use'' were the ones with the fewest response categories: the two-point, three-point, and four-point scales were rated most favorably on this criterion, and once

Table 5
Scales ranked in order of increasing respondent preference: statistically significant differences[a]

	Scales (No. of response categories)									
	Lowest									Highest
Ease	101	11	2	9_a	6_{ab}	3_b	8_b	4	7	10	5
Quick	101	11	9	10	8_a	7_a	6	5_b	4_b	2_c	3_c
Express feelings	2	3	4	6_a	5_a	8	7	9_b	11_b	10	101

[a] *Note.* Scales in the same row that share the same subscript do not differ significantly; all other differences between scales in the same row are statistically significant at $p < 0.05$ according to the Tukey-HSD comparison.

again the scales with 11 and 101 response categories were least preferred. Ratings of the degree to which the scales "allowed you to express your feelings adequately" showed the greatest differentiation between the scales. The two-point and three-point scales received extremely low ratings on this criterion (well below 50 on the 0–100 scale). In general, longer scales tended to receive more favorable ratings on this dimension: the scales with 9, 10, 11 and 101 response categories were rated highest. Taking into account all three questions related to respondent preferences, the shortest scales (two, three and four response categories) generally received the lowest ratings, but the longest scales (11-point and 101-point scales) also received relatively low ratings. The scale that scores best overall according to respondent preferences was the 10-point scale, closely followed by the seven-point and nine-point scales.

3. Discussion

The rating scales that yielded the least reliable scores turned out to be those with the fewest response categories. Test–retest reliability (stability) was lowest for two-point, three-point, and four-point scales and was significantly higher for scales with more response categories; the most reliable scores were derived from scales with 7, 8, 9, or 10 response categories. Internal consistency was lowest for scales with two or three response categories and highest for those with seven or more, although on this criterion of reliability the differences between the reliability coefficients were not statistically significant. Our results also provide evidence of a decrease in test–retest reliability for scales with more than 10 response categories, although only the decrease from the eight-point and nine-point scales to the 101-point scale attained statistical significance. Bendig (1954) found a similar decrease in inter-rater reliability, but Cicchetti et al. (1985), using a Monte-Carlo simulation rather than a fully empirical research methodology, found no decrease in reliability for long scales.

According to the indices of validity and discriminating power that we examined in this study, the scales with relatively few response categories performed worst. The criterion validity coefficients were lowest for the scales with two, three, or four response categories and were generally higher – though these differences were not statistically significant – for scales with five or more response categories. Discriminating power was lowest for the scales with two, three, or four response categories and statistically significantly higher for scales with 9 or 101 response categories. Item-whole correlations told a similar story: scales with two, three, or four response categories performed worst, and scales with six or more response categories performed generally better – the coefficient for the two-point scale was significantly lower than the coefficients for all other scales in the investigation. Finally, the table of intercorrelations between scores on all of the scales suggested that the scales with two or three response categories yielded scores with lower overall convergent validity than the others. However, it should be borne in mind that a scale with relatively few response categories tends to generate scores with comparatively little

variance, limiting the magnitude of correlations with other scales (e.g., Chang, 1994; Martin, 1973, 1978; Nunnally, 1970). This restriction-of-range effect tends to depress the convergent validity of scores from scales with few response categories, but it is worth remembering that this arises ultimately from the inherent bluntness of such scales, which also limits their usefulness for many practical psychometric purposes.

These findings provide no corroboration for Matell and Jacoby's (1971) suggestion that reliability and validity of scores are independent of the number of response categories and that nothing is gained by using scales with more than two or three response categories; but neither do they corroborate Garner's (1960) suggestion that scales with 20 or more response categories are necessarily best. As regards reliability, the results reported above tend to confirm the findings of Symonds (1924), Nunnally (1967), Green and Rao (1970), Finn (1972), Ramsay (1973), Cicchetti et al. (1985), and Oaster (1989), which suggested that reliability of scores tends to increase from two-point to six-point or seven-point scales. As regards validity and discriminating power, our results provide a remarkably consistent picture across four independent indices, tending to confirm evidence from a small number of studies using much more restricted criteria of validity (e.g., Chang, 1994; Hancock & Klockars, 1991; Loken et al., 1987) that, statistically, scales with small numbers of response categories yield scores that are generally less valid and less discriminating than those with six or more response categories.

Respondent preference ratings differed substantially between the scales, and the differences were statistically significant. Scales with 5, 7, and 10 response categories were rated as relatively easy to use. Shorter scales with two, three, or four response categories were rated as relatively quick to use, but they were rated extremely unfavorably on the extent to which they allowed the respondents to express their feelings adequately; according to this criterion, scales with 10, 11 and 101 response categories were much preferred. On the whole, taking all three respondent preference ratings into account, scales with two, three, or four response categories were least preferred, and scales with 10, 9, and 7 were most preferred.

From the multiple indices of reliability, validity, discriminating power, and respondent preferences used in this study, a remarkably consistent set of conclusions emerges. Our results provide no support for the suggestion of Schutz and Rucker (1975) that the number of response categories is largely immaterial. On the contrary, scales with two, three, or four response categories yielded scores that were clearly and unambiguously the least reliable, valid, and discriminating. The most reliable scores were those from scales with between 7 and 10 response categories, the most valid and discriminating were from those with six or more response categories or-in the case of intertertile discriminating power – those with nine or more. The results regarding respondent preferences showed that scales with two, three, or four response categories once again generally performed worst and those with 10, 9, or 7 performed best. The superiority of scales with around seven response categories is in line with Miller's (1956) theoretical analysis of human information-processing capacity and short-term memory, subsequently refined by Simon (1974) in his characterization of information "chunks". For several decades, the vast majority of

rating scales and related psychometric instruments have used five or seven response categories (Bearden et al., 1993; Peter, 1979; Shaw & Wright, 1967). In the light of our findings, there is some support for seven-point scales, but the popularity of five-point scales seems to be less justified.

Taken together, the results reported above suggest that rating scales with 7, 9, or 10 response categories are generally to be preferred. In this study, participants responded to the scales consecutively, and this may have led to factors other than scale format affecting their ratings. Respondents may, for example, have been inconsistent through carelessness or boredom arising from having to respond to similar questions again and again. However, the satisfactory levels of scale reliability, ranging from 0.79 to 0.94 (see Table 1), and the relatively high correlations between scores from different scales, ranging from 0.84 to 0.98 (see Table 3), suggest that the respondents rated carefully and consistently across scales. Also, the fact that the scales were presented in a different random order to each respondent ensures that any inconsistencies in responding could not contribute to significant differences found between scales. Caution should, however, be exercised in generalizing the conclusions beyond the types of respondents and scales used in this study.

The research reported here examined responses to real-life experiences in recent visits to stores or restaurants. Some previous studies have focused on ratings of real-life experiences (e.g., Neuman & Neuman, 1981), whereas others have involved ratings of hypothetical experiences or abstract concepts (e.g., Matell & Jacoby, 1971; Oaster, 1989) or have used computer simulations (e.g., Cicchetti, Showalter, & Tyrer, 1985). These methodological differences probably explain, in part at least, the discrepant findings reported in these different studies. However, our aim was to investigate the effects of scale length on participant ratings using real-life experiences in order to produce findings that would have practical relevance to ratings in everyday situations such as those studied in market research.

A careful reading of our results suggests that different scales may be best suited to different purposes. Circumstances may, for example, require respondents to use a rating scale under conditions of time pressure, and in such cases it may be necessary, in order to prevent the respondents from becoming frustrated and demotivated, to use five-point or even three-point scales, because our findings show that these scales are likely to be perceived by the respondents as relatively quick and easy to use. On the other hand, where considerations of face validity are regarded as paramount, it may be important for the respondents to perceive the scales as allowing them to express their feelings adequately, and in such cases 10-point scales may be most appropriate. Before deciding on the optimal number of response categories for a rating scale, researchers and practitioners may therefore need to perform a trade-off, in the light of the prevailing circumstances, between reliability, validity, discriminating power, and respondent preferences.

The findings reported in this article relate to ratings of service quality in restaurants and stores; further research in a similar vein, using objective ratings of behavior by others, self-ratings of behavior, ratings of personality traits, ratings of the quality of products, and so on, would be needed to determine the extent to which the conclusions generalize to other domains.

Acknowledgements

Preparation of this article was supported by research awards K38 and M71 from BEM Research. We are grateful to Gareth G. Jones for drawing our attention to the problem addressed in this article and to Barry Denholm, Caroline Gaynor, and Laura Gracie for help with the collection of data. We wish to thank David Stretch and Jeremy Miles for help with the data analysis.

References

Althauser, R. P., Heberlein, T. A., & Scott, R. A. (1971). A causal assessment of validity: the augmented multitrait-multimethod matrix. In H. M. Blalock (Ed.), *Causal models in the social sciences* (pp. 374–399). Chicago, IL: Aldine.

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

Bearden, W. O., Netmeyer, R. G., & Mobley, M. F. (1993). *Handbook of marketing scales: multi-item measures for marketing and consumer behavior research*. Newbury Park, CA: Sage.

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and the number of categories on the scale. *The Journal of Applied Psychology*, *37*, 38–41.

Bendig, A. W. (1954). Reliability and the number of rating scale categories. *The Journal of Applied Psychology*, *38*, 38–40.

Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Boote, A. S. (1981). Reliability testing of psychographic scales: five-point or seven-point? Anchored or labeled? *Journal of Advertising Research*, *21*, 53–60.

Brown, G., Wilding, R.E., II, & Coulter, R.L. (1991). Customer evaluation of retail salespeople using the SOCO scale: A replication extension and application. *Journal of the Academy of Marketing Science*, *9*, 347–351.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Chang, L. (1994). A psychometric evaluation of four-point and six-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement*, *18*, 205–215.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology*, *23*, 323–331.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: a Monte-Carlo investigation. *Applied Psychological Measurement*, *9*, 31–36.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, (2nd ed.). Hillsdale, NJ: Elrbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, *17*, 407–422.

Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, *34*, 885–892.

Garner, W. R. (1960). Rating scales, discriminability and information transmission. *Psychological Review*, *67*, 343–352.

Green, J. A. (1992). Testing whether correlation-matrices are different from each other. *Developmental Psychology*, *28*, 215–224.

Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery: How many scales and response categories to use? *Journal of Marketing*, *34*, 33–39.

Hancock, G. R., & Klockars, A. J. (1991). The effect of scale manipulations on validity: targeting frequency rating scales for anticipated performance levels. *Applied Ergonomics*, *22*, 147–154.

Howell, D. C. (1992). *Statistical methods for psychology*. Boston, MA: Duxbury Press.
Jenkins, Jr., G. D., & Taber, T. D. (1977). A Monte-Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*, *62*, 392–398.
Jones, R. R. (1968). Differences in response consistency and subjects' preferences for three personality inventory response formats. In *Proceedings of the 76th Annual Convention of the American Psychological Association* (pp. 247–248).
Komorita, S. S. (1963). Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology*, *61*, 327–334.
Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: a Monte-Carlo approach. *Journal of Applied Psychology*, *60*, 10–13.
Loken, B., Pirie, P., Virnig, K. A., Hinkle, R. L., & Salmon, C. T. (1987). The use of 0–10 scales in telephone surveys. *Journal of the Market Research Society*, *29* (3), 353–362.
Martin, W. S. (1973). The effects of scaling on the correlation coefficient: a test of validity. *Journal of Marketing Research*, *10*, 316–318.
Martin, W. S. (1978). Effects of scaling on the correlation coefficient: additional considerations. *Journal of Marketing Research*, *15*, 314–318.
Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study 1: reliability and validity. *Educational and Psychological Measurement*, *31*, 657–674.
McKelvie, S. J. (1978). Graphic rating scales: How many categories? *British Journal of Psychology*, *69*, 185–202.
Messick, S. (1993). Validity. In R. L. Lin, *Educational measurement* (3rd ed.) (pp. 13–103). Phoenix, AZ: Oryx Press.
Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63* (2), 81–97.
Neuman, L., & Neuman, Y. (1981). Comparison of six lengths of rating scales: students' attitude toward instruction. *Psychological Reports*, *48*, 399–404.
Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.
Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, *68*, 549–550.
Peabody, D. (1962). Two components in bipolar scales: direction and extremeness. *Psychological Review*, *69*, 65–73.
Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, *16*, February, 6–17.
Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, *38*, 513–533.
Remington, M., Tyrer, P. J., Newson-Smith, J., & Cicchetti, D. V. (1979). Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. *Psychological Medicine*, *9*, 765–770.
Remmers, H. H., & Ewart, E. (1941). Reliability of multiple-choice measuring instruments as a function of the Spearman–Brown prophecy formula. *Journal of Educational Psychology*, *32*, 61–66.
Schutz, H. G., & Rucker, M. H. (1975). A comparison of variable configurations across scale lengths: an empirical study. *Educational and Psychological Measurement*, *35*, 319–324.
Shaw, M. E., & Wright, J. M. (1967). *Scales for the measurement of attitudes*. New York: McGraw-Hill.
Simon, H. A. (1974). How big is a chunk? *Science*, *183*, 482–488.
Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, *7*, 456–461.
Woodruff, D. J., & Feldt, L. S. (1986). Tests for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika*, *51*, 393–413.

Do data characteristics change according to the number of scale points used ? An experiment using 5 point, 7 point and 10 point scales.

Abstract

To what extent does the number of response categories in a Likert-type scale influence the resultant data ? Surprisingly little attention has been paid to the issue of whether the response category format has any influence on data characteristics such as the mean, coefficient of variation, skewness and kurtosis. This issue is important for several reasons. The first is that decisions are made based on outcomes such as the mean score. For example, marketing organizations and research providers use Likert type scales to measure constructs such as customer satisfaction. In this situation a higher score is better. Could the score have been comparatively better if a different scale format had been used ? There is an absence of evidence on this issue. The second reason is that scale formats that are used in on-going market research projects such as tracking studies occasionally change. Can the old results be re-scaled or transformed to be comparable to data from a new scale format ? Again, little is known about this. The third reason concerns data characteristics such as variation about the mean, skewness and kurtosis. Analysis tools such as regression are often used on data of this type to explain the variation in certain variables. If there is little variance in the data, this is harder to do. How does scale format affect these characteristics ? The answers would be useful to both market researchers as well as academics. A literature review found that little work has been done on this issue. Therefore, this study set out to investigate the impact of scale format on data characteristics. It examined how using Likert-type scales with varying numbers of response categories affects the resultant data in terms of mean scores, and measures of dispersion and shape. Three groups of respondents were administered a series of eight questions (group n's = 300, 250,185). Respondents were randomly selected members of the general public. A different scale format was administered to each group – either a five-point, seven-point or ten-point scale. The surveys were conducted by a professional market research organisation via telephone interview. Data characteristics of mean score, standard deviation, skewness and kurtosis were analysed according to scale format. The five and seven-point scales were re-scaled to a comparable mean score out of ten. The study found that the five and seven-point scales produced the same mean score as each other, once they were re-scaled. However the ten-point format tended to produce slightly lower relative means than either the 5 or 7-point scales (after the latter were re-scaled). The overall mean score of the eight questions was 0.3 scale points lower for the 10-point format compared to the 5 and 7-point format. This difference was statistically significant at $p=0.04$. In terms of the other data characteristics, there was very little difference among the scale formats in terms of variation about the mean, skewness or kurtosis. Therefore each of the three formats appears comparable for the type of research project in which multiple-item scales are analyzed with multivariate statistical methods. This study is also 'good news' for research departments or agencies who ponder whether changing scale format will destroy the comparability of historical data. Five and seven-point scales can easily be re-scaled with the resultant data being quite comparable. In the case of comparing five or seven-point data to 10-point data, a straightforward re-scaling and arithmetic adjustment easily facilitates the comparison. Finally, it appears that indicators of customer sentiment – such as satisfaction surveys – may be partially dependent on the choice of scale format. A five or seven-point scale is likely to produce slightly higher mean scores relative to the highest possible attainable score, compared to that produced from a ten-point scale.

Introduction

Rating scales are one of the most widely used tools in marketing research and commercial market research. Rating scales are used to capture information on a range of phenomena. In consumer research, respondents may be asked about their attitudes, perceptions, or evaluations of products, brands, or messages - among many other possibilities. In other marketing research streams, respondents such as managers or marketing personnel may be asked to rate their company's performance, type of strategic focus, personnel, degree of marketing excellence, training regimes and so forth using such scales.

Rating scales typically require the respondent to select their answer from a range of verbal statements or numbers. Scales that use verbal statements include semantic differential scales and Likert scales. An example of the semantic differential scale is *very good … very bad,* or *pleasant … unpleasant*. An example of the Likert response scale is as follows: *strongly disagree, disagree, neither disagree or agree, agree, strongly agree*. This particular example is a five-point Likert scale utilising verbal response descriptors. Likert scales may also use numerical descriptors where the respondent selects an appropriate number to denote their level of agreement. For example, a question could be worded like this: "indicate your agreement from 1 to 5 where 1 equals strongly disagree and 5 equals strongly agree".

The range of possible responses for a scale can vary. Textbooks on the subject typically portray five or seven-point formats as the most common (e.g. Malhotra and Peterson 2006 ch. 10). Ten or eleven-point scales are also frequently used (Loken, Pirie et al. 1987). Hereafter in this study the term 'scale format' is used to refer to scales with differing numbers of response categories.

In terms of the interface between the respondent and the interviewer in a telephone survey, there are some advantages and disadvantages of each scale format. With a five-point scale, it is quite simple for the interviewer to read out the complete list of scale descriptors (1 equals strongly disagree, 2 equals disagree ….). This clarification is lengthier for the seven-point format. Such a verbal clarification becomes quite impractical for a 10-point format as the gradations of agreement become too fine to easily express in words. In this case, the interviewer normally reads out the verbal meaning of the end points. The 10-point format therefore places greater reliance on the respondent using a numerical response for which the precise meaning has not been precisely defined. However, this disadvantage is balanced against the fact that many people are familiar with the notion of rating 'out of ten'.

There have been numerous studies on the topic of how scale format affects scale reliability and validity. Far less attention has been paid to how it influences data characteristics such as mean and variance. The issues of reliability and validity are outside the scope of this study. Suffice to say, simulation studies and empirical studies have generally concurred that reliability and validity are improved by using five to seven-point scales rather than coarser ones (those with fewer scale points). But more finely graded scales do not improve reliability and validity further.

The next section presents some theoretical reasons for why the scale format might influence the mean score, variance and skewness. The small number of empirical studies that have examined this issue are then reviewed.

Why *would* scale format influence data characteristics ?

One of the most basic summary data characteristics is the mean. Scores for Likert-type questions are often 'negatively skewed' (e.g. Dawes 2002b; Peterson and Wilson 1992). This term is counterintuitive and refers to the fact that more responses are at the positive end of the scale and the 'tail' is at the negative end. If more respondents tend to give positive responses, then a finer scale, with more response options, could result in a slightly lower mean score. This can be illustrated by considering the range of positive response options for five, seven and ten-point formats. Firstly consider a five-point scale. There are only two options for a positive response: points four and five. If we average those two responses and re-scale to the equivalent score on a ten-point scale (using the method described and used later under 're-scaling') the result is 8.9 /10. If we undertake the same procedure for a seven-point scale the positive responses are 5, 6, and 7 for an average of 6, which re-scales to a score of 8.3/10. The positive responses for a ten-point scale are 6, 7, 8, 9, and 10 which averages to 8/10. Therefore, based on the arithmetic properties of the scales, the three scale formats would produce somewhat different comparative mean scores if the majority of responses were on the positive side of the mid-point. The potential of the different formats to produce comparatively different mean scores seems worthwhile to investigate.

In relation to the distribution of data about the mean, more scale points, by definition, provide more options for the respondent. Therefore, finer scales could result in a greater spread of the data. This would result in a larger standard deviation, and possibly more positive kurtosis as kurtosis is related to, although not the same thing as, variance.

More scale response options may also conceivably result in less skewed data. This is illustrated using the situation whereby a scale is used to measure a construct that most respondents give a particularly positive response for. A coarse scale will provide few options for this positive sentiment and so the responses may be 'bunched up' at the positive end of the scale. A finer scale could reduce this negative skew by allowing for more gradations of positive response. This could also reduce the overall mean score, for the reasons outlined above.

The empirical studies examining scale format and its association with data characteristics are now reviewed.

Studies examining level and shape of data

There are only a small number of studies on this issue. One is by Finn (1972) which reported means and variances for 3, 5, 7 and 9-point scales. They were 1.6, 2.2, 4.1 and 4.9 for means and .32, .60, 1.32 and 4.0 for variances respectively. Applying a re-scaling formula from Preston and Colman (to be discussed in more detail later in the analysis section), I transformed these reported means to a score out of 100. The transformed scores are 30, 30, 52 and 49 respectively. This suggests the 7 and 9-point formats produced comparatively higher scores. This is counter to the theoretical expectation outlined above. In terms of the variance, taking its square root and dividing this by the original mean score gives the coefficient of variation. This is a standardised measure of variance that controls for the differing number of scale points. The coefficient of variation for Finn's four scale formats is calculated to be .35, .35, .28 and .41 for the 3, 5, 7 and 9-point scales respectively. It appears the nine-point format produced higher comparative variance in that study compared to the coarser scales.

Two other studies are pertinent to the issue of how the number of scale points affects data characteristics such as the mean score. One of these was many years ago, in which Ghiselli (1939) conducted an experiment using undergraduate students who were asked to indicate whether they thought the advertising for 41 different brands was *sincere*. One group answered using a two-point (yes / no) scale, the other group answered using a four-point scale (very sincere … very insincere). The four-point scale resulted in higher ratings for the perceived sincerity of the advertising than the two-point scale.

Another study was by Dawes (2002a) who analysed two split-sample experiments in which groups of respondents were administered questions with either 5-point or 11-point scales. He found that once the 5-point scale was re-scaled to 11-point equivalence, that the means from the 11-point scale were slightly higher by an average of 0.25 points, although no inferential test was applied. This result could be partially attributable to the eleven-point scale having an anchor value of zero (i.e., a zero to ten scale). This characteristic may have artificially lowered the mean score for the 5-point data from the rescaling process. For example, a score of one out of five was rescaled to zero out of ten. In that study, the 11-point scale also produced slightly more dispersion in the data as measured by the coefficient of variation, but there was no difference in skewness or kurtosis between the two scales. Only one other study has examined the issue of scale format and skewness, which was by Johnson, Smith and Tucker (1982). They found that a 2-point format produced more skewness than a five-point format.

This review makes it apparent that basic issues to do with the mean and distribution of the data, and how they are affected by scale format have not been closely studied. There also seems to be some variation in the results from previous studies. While Dawes (2002a) found that re-scaled means from five to eleven-point scales were almost the same; an inspection of the data reported in Finn (1972) showed more marked differences. Likewise, one prior study found that coarse scales resulted in more skewness (Johnson, Smith and Tucker 1982), albeit between two and five-point scales, the former of which is rarely used in marketing studies. Another study found no appreciable difference between five and eleven-point scales in this regard (Dawes 2002a).

Research questions and rationale

We know that scales are ubiquitous in both market research and academic marketing research. But there is a less than comprehensive amount of documented knowledge on the topic. Therefore further investigation of the way scale format might influence the data is warranted. There are at least three reasons for this.

First, the sophistication of analytical methods is increasing. Techniques such as confirmatory factor analysis and structural equation modeling are now commonplace in marketing research. These tools are sensitive to the characteristics of the data, such as variance, kurtosis and skewness (e.g. Bentler 1995). Therefore more knowledge about how scale format affects these characteristics would be desirable.

Second, in many cases the data from a survey is not just reported, it is analysed with the objective of 'explaining' or accounting for the variance in a dependent variable. Examples of the dependent variable might be overall customer satisfaction, probability of purchase, or attitudes towards a brand or organization. The analyst wishes to find out what other variables might be strongly related to higher or lower scores on the dependent variable. In this situation, more variance in the dependent variable is desirable. This is illustrated with an

example. If, hypothetically, all respondents gave the same score for customer satisfaction there would be no variance to explain. If there was very little variance, for example all responses were either 6 or 7 on a seven-point scale then normal OLS regression is not an appropriate analysis method. More complex techniques such as logistic regression would be needed.

The third reason is that in industry, many organizations periodically track consumer sentiment, and often, scales of the type discussed here are a major part of the research. For example, many service organizations such as banks, telecommunication companies or insurance companies routinely survey customers about their perceived levels of service quality or customer satisfaction. For a variety of reasons, the choice of scale is sometimes changed, say, from a five-point scale to a seven-point scale. The reasons for this could be personnel changes, the appointment of a different research provider, department mergers and so on. Obviously the information gleaned from the data, such as mean scores, is based on the number of scale responses used. But is the data dependent on the scale to the extent that the mean score relative to the highest possible score is different for one scale compared to another ? There are some theoretical grounds for thinking scale format might affect the data, as outlined earlier. Also, there is little guidance on this apparently practical and important issue and indeed some conflict in prior results. More knowledge in this area would therefore seem desirable.

This study therefore sought to compare the aggregate-level data characteristics derived from attitudinal questions with either 5, 7 and 11-point numerical scales as the response categories. The specific research question is:

> If data on the same construct is gathered using three scale formats (5-point, 7-point and 10-point numerical scales) and the data from the 5 and 7-point formats are re-scaled to a common 10-point format, are there any differences in terms of mean, variance, kurtosis and skewness ?

This question presumes to treat the data as if they were at least interval quality. There is some evidence that the psychological 'distances' between Likert-type scale points are not equal, for example Bendixen and Sandler (1994) and Kennedy, Riquier and Sharp (1996). That said, the relation between the original scale values and the 'real' identified scale values is very close in these studies. For example in Kennedy, Riquier and Sharp (1996) the notional scale values of 1, 2, 3, 4 and 5 equated to 1, 2.2, 3.1, 4.1 and 5 respectively. The leading texts in the field support the treatment of such scales as if they are equal-interval (e.g. Aaker, Kumar and Day 2004 p. 285; Burns and Bush 2000 p. 314 ; Dillon, Madden and Firtle 1993 p. 276; Hair, Bush and Ortinau 2006 p. 365-366). Based on the empirical studies showing a reasonably close approximation to equal-interval, and the apparent precedent shown in the leading texts, I analysed the data as if it were equal-interval.

Data

In accordance with the research objective, data were gathered via a survey of consumers drawn at random from the electronic telephone directory. The survey was conducted over the 2005-2006 period by a professional market research organisation using CATI (Computer Assisted Telephone Interviewing).

The questionnaire items were derived from existing 'price consciousness' scales (Bruner and Hensel 1992). Price consciousness is an example of a subject-centered scale, and it appeared to contain content that respondents could readily understand and easily answer. The scale comprised eight items, which are shown below. Respondents were asked to answer the questions with the instruction 'please answer using the scale from 1 to *X* where 1 equals strongly disagree and *X* equals strongly agree'. *X* was either 5, 7 or 10 depending on the treatment group. The precise meaning of each scale point was not read out to respondents for any of the three scale formats, whereas normally one would do so for the 5 or 7-point formats. This potentially lowered the utility of those two scale formats, but I wished to have a consistent approach to administering all three of the scales.

TABLE 1 HERE

The number of respondents in each experimental group was: 10-point scale: n=300; 7-point scale n=185; 5-point scale n= 250. The reason for the varying sample sizes for each group is that the study had other unrelated objectives, and the questionnaire programming used to direct respondents into the three scale format groups was also used to direct sample numbers to other question sets, and those groups required different sample sizes. The other survey content did not affect the results reported here.

I considered whether the sample sizes were adequate by calculating how large the difference in mean scores across groups would need to be to achieve statistical significance. I set a difference of half a scale point as the magnitude of difference that the experiment should be able to identify as statistically significant. I had three treatment groups, with the smallest group numbering 185 respondents. To ascertain their adequacy, I conducted a conservative inferential test using simulated data of three groups of n=185. I first generated a series of 185 scores with a mean of 6.0 and a standard deviation of 2.0 using Microsoft excel. These data characteristics were taken from the results of a previous study (Dawes 2002a). I then generated two other series, such that I had three data series that differed by 0.5 scale points to each other. This process was repeated with data series that exhibited progressively smaller differences in mean scores. I found that if there was a mean difference of 0.3 scale points (or more) between each of three groups of this size, an analysis of variance would be statistically significant at the p=0.05 level. Since two of my groups had larger number of respondents than this, the sample sizes appeared to be adequate for the purpose.

The survey sample was broadly representative of the general population, excepting that younger respondents were under-represented. The age breakdown of the sample is shown below. The gender breakdown was 42% male and 58% female. Ideally the survey would have obtained an age and gender breakdown closer to the general population, but in order to do so data collection costs would have increased, which was not feasible for this study. Later I discuss whether the results are comparable for age and gender sub-groups to ensure the results are not biased by the sample.

TABLE 2 HERE

Analysis

Re-scaling

To examine the various data characteristics of interest, it is convenient to re-scale the data so that each scale format is comparable, each with the same upper limit such as out of ten or out of 100. Note that the purpose of this re-scaling is to facilitate comparison between the scale

formats, not to find a specific functional transformation that will minimize any re-scaled differences.

There are a number of straightforward methods by which this could be done. One method is based on a formula used by Preston and Colman (2000). They used the formula (rating -1) / (number of response categories -1)*100. This re-scales to a common score out of 100. For the purpose of this paper we could use the same formula but adapted to be (rating -1) / (number of response categories -1)*10 which re-scales all scale formats to a score out of ten. A feature of this method is that any score using the lowest scale point of any scale becomes zero. For example a score of 1 on a five-point scale would become (1-1) / (5-1) *10 = zero.

Another method is one employed by Dawes (2002a). This is a simple arithmetic procedure whereby the scale end points for the 5 and 7-point versions are anchored to the end points of the ten-point scale. The intervening scale values are inserted at equal numerical intervals. For example, to re-scale the 5-point scale to ten points, one remains as one, five is re-scaled to ten, the mid-point of 3 on the 5-point scale is adjusted to be as per the mid-point between 1 and 10 (namely 5.5) and so on. This is shown below in Table 3.

The second approach has the appealing feature for the present research that the ten-point scale remains unchanged, and the other scales are altered to be comparable to it. However, it results in a slight biasing effect for the lowest scale point. This is because it takes a score of 1 out of 5 or 1 out of 7 and re-scales it to be equivalent to 1 out of 10 – the latter being a lower score in proportional terms. Therefore if there are any responses using these lowest scale points for the five or seven-point formats, the re-scaled score expressed as a mean out of ten will be lower than it was originally. However, preliminary analysis showed that the method based on Preston and Colman (2000) and Dawes (2002a) produced virtually identical results. The Dawes (2002a) method was used because it was slightly simpler.

TABLE 3 HERE

Results

Mean scores
The re-scaled mean scores for each item are shown below, for each of the three scale formats. The data are ordered according to the mean score on the 10-point scale for clarity.

TABLE 4 HERE

The re-scaled 5-point and 7-point scales produced more instances of higher scores compared to the ten-point format. For seven out of the eight questions, the 5-point format (once re-scaled) produced slightly higher scores than the 10-point format. For six out of the eight questions, the 7-point format (once re-scaled) produced slightly higher scores than the 10-point format. There appeared to be little difference between the 5-point and 7-point format.

To test if the overall mean scores from the eight items were statistically significantly different according to scale format, I ran a one-way ANOVA. Since there was virtually no difference between the 5-point and 7-point formats I combined them as one factor. The average re-scaled scores of the eight scale items were the dependent variable, and the factors were the scale formats (5-point re-scaled and 7-point re-scaled combined as one factor, and 10-point

scale comprising the other factor). The result was statistically significant (F=4.1; d.f. 1,733; p=0.04).

Based on this result, it seems that a 10-point scale format will produce slightly lower scores compared to the scores generated from 5-point or 7-point formats, once the latter are re-scaled for comparability.

Variance

I next examined the standard deviation for the re-scaled 5 and 7-point data compared to the 10-point data. If the data is not dependent on the choice of scale format, then once the data is re-scaled to a score out of ten, all three scale formats should exhibit the same standard deviation.

TABLE 5 HERE

Looking across the three scale formats in Table 5, the differences in standard deviation for the individual scale items are quite small, in the order of zero to 0.2. The average difference is only -0.1 when comparing either 7 to 10-point or 5 to 7-point data (with 5 and 7-point formats rescaled). It appears that scale format does not have a marked influence on variation about the mean. To clarify this formally, I tested the overall average score for each format using the Levene test for homogeneity of variance. The test was not significant (Levene Statistic=0.21; d.f. 2,732, p=0.80). Scale format therefore did not have an association with variance in this experiment.

An examination of the standard deviation tells us about the dispersion of scores about the mean for a particular questionnaire item, or variable. It does not, however, tell us about how individual respondents have used the scale. For example, if we ask respondents to answer eight questions using a one to five scale, how many different scale points will they use ? Obviously the precise answer depends on what the questions pertain to. However, researchers would generally want respondents to use more response options over a series of questions, rather than less. The reason is that this indicates those questions are generating discrimination in responses. Therefore, as a supplementary analysis I also examined how many different scale points respondents actually used, and whether this differed according to the scale format. I found that over the eight questions, the average number of scale points used for the five-point scale was 2.9, for the seven-point scale it was 3.6 and for the ten-point scale, respondents used 4.0 different scale points on average. An analysis of variance confirmed that there was a statistically significant difference between the scale formats in terms of the number of scale points used (F=54; d.f. 2,732; $p<0.01$). Therefore, there is evidence that respondents do use more scale points when given a scale format with more response options.

Skewness

Data may be normally distributed, or may be positively skewed or negatively skewed. If the data is negatively skewed this means the data tends to cluster at the ‘high’ end of the scale with a long tail to the lower scale values. The figures for skewness are shown below in Table 6.

TABLE 6 HERE

The data from all three scale formats is negatively skewed. There are some differences among the individual scale items according to scale format, but nothing systematic. In terms

of the skewness of the overall mean score, there is less than one standard error difference between each scale format, therefore this is not statistically significant.

Kurtosis

Kurtosis refers to the shape of the data around the mean and the tails of the distribution. A normal distribution has a kurtosis value of zero. Data that exhibit positive kurtosis are more clustered about the mean ('peaked') and the tails of the distribution are longer. A negative kurtosis score occurs when the data are clustered less around the mean and have shorter tails. A distribution may have the same mean and standard deviation but exhibit different levels of kurtosis. Hypothetical examples of distributions with the same mean and standard deviation but with either zero, positive and negative kurtosis are shown below to help elaborate the term.

FIGURE 1 HERE

The analysis of kurtosis is shown in Table 7. All three scale formats tend to produce data with negative kurtosis scores. There are only minor differences between the scale formats for the individual scale items. The overall score from each scale format exhibits negative kurtosis, and the differences between them are not managerially or statistically significant.

TABLE 7 HERE

Sub-group analysis

As mentioned earlier, the sample used for this experiment was biased somewhat towards older respondents and females. To ensure the results are not influenced by this sample bias, I re-ran the analysis for two sets of subgroups: older vs. younger respondents and male vs. female respondents. There were no significant differences in mean score, variance, skewness or kurtosis within these subgroups. Therefore there is no reason to think that the slight gender bias in the composition of the sample has influenced the overall results.

Discussion and Conclusions

This study conducted a split-sample experiment to assess the impact of scale categories on responses to questions. The study compared data obtained from using 5-point, 7-point and 10-point numerical scale formats. The 5-point and 7-point data were re-scaled to scores out of ten. Once rescaled, the five and seven-point formats tended to produce more instances of higher mean (rescaled) scores compared to the ten-point format. Indeed, an analysis of the aggregated score over the eight question items found the ten-point format produced a 0.3 point lower score, which was statistically significantly different to the other two formats at under the $p=0.05$ level. In terms of the other data characteristics, the three different scale formats exhibited no appreciable differences in terms of standard variation, skewness or kurtosis. The study also found that if a scale with more response options was administered, respondents used more response options.

Based on these findings it seems reasonable to conclude that data gathered from a five-point format can be readily transferred to seven-point equivalency using a simple re-scaling method. If the analyst wishes to compare data from five or seven-point formats to data in a ten-point format, a simple arithmetic adjustment and re-scaling using the method described here produces comparable data. This outcome may be welcome news to those market research departments who ponder whether data gathered using one scale format can be

transformed to make it comparable to another. It also answers a potential question regarding whether results might have conceivably been better (e.g. a higher relative score) had a different scale format been used. The answer appears to be a scale with more response options produces slightly lower scores relative to the upper limit of the scale.

In terms of the other data characteristics, no scale format produced data with markedly lower variances about the mean. This suggests that none of the three formats is less desirable from the viewpoint of obtaining data for that will be used for regression analysis. Kurtosis and skewness were likewise all very similar for each format, therefore either 5, 7 or 10-point scales are all comparable for analytical tools such as confirmatory factor analysis or structural equation models in this respect.

Directions for Future Research

This study examined scale formats that differed in the number of response categories but were all numerical scales. They all required respondents to nominate a number within a specified range. Such numerical scales are but one type of response scale, it is also common for market researchers and academics to ask respondents to use scales that employ only verbal anchors. This paper, therefore, has tackled only one aspect of a wider issue pertaining to the use and comparability of rating scales in market research. More insight into the effect of the number of response categories on the resultant data when using scales that are only verbally anchored would also be a useful addition to current knowledge. Likewise, this study only examined the effect of scale format using telephone survey methodology. There is scope to examine whether the results found here would generalize to other data collection methods such as self-completion or face-to-face.

Acknowledgements

I thank the field team at the Ehrenberg-Bass Institute for their dedicated effort in data collection.

References

Aaker, D.,V. Kumar, and George S. Day. 2004. Marketing Research. New York: John Wiley and Sons.

Bendixen, Mike, and Merle Sandler. 1994. Converting Verbal Scales to Interval Scales Using Correspondence Analysis. Johannesburg: University of Witwatersrand.

Bruner, G. C., and P. J. Hensel. 1992. Marketing Scales Handbook. Chicago: American Marketing Association.

Burns, Alvin C., and Ronald F. Bush. 2000. Marketing Research: Prentice Hall.

Dawes, John G. 2002a. Five point vs. eleven point scales: does it make a difference to data characteristics? Australasian Journal of Market Research 10 (1): 39 - 47.

Dawes, John G. 2002b. Survey Responses Using Scale Categories Follow a "Double Jeopardy" Pattern. In *ANZMAC*. Eds. R. N. Shaw, S. Adam and H. McDonald. Melbourne.

Dillon, William R, Thomas J Madden, and Neil H Firtle. 1993. Essentials of Marketing Research. Homewood: Irwin.

Finn, R.H. 1972. Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings. Educational and Psychological Measurement 32: 255-265.

Ghiselli, Edwin E. 1939. All or none versus graded response questionnaires. Journal of Applied Psychology 23 (June): 405-415.

Hair, Joseph F., Jr. , Robert P. Bush, and David J. Ortinau. 2006. Marketing Research. Boston: McGraw Hill.

Johnson, Steven M. , Patricia C. Smith, and Susan M. Tucker. 1982. Response format of the job descriptive index: Assessment of reliability and validity by the multitrait-multimethod matrix. Journal of Applied Psychology 67 (4): 500-505.

Kennedy, Rachel, Christopher Riquier, and Byron Sharp. 1996. Practical Applications of Correspondence Analysis to Categorical Data in Market Research. Journal of Targeting, Measurement and Analysis for Marketing 5 (No. 1): 56-70.

Loken, Barbara, Phyllis Pirie, Karen A. Virnig, Ronald L. Hinkle, and Charles T. Salmon. 1987. The Use of 0-10 Scales in Telephone Surveys. Journal of the Market Research Society 29 (No. 3, July): 353-362.

Malhotra, Naresh, and Mark Peterson. 2006. Basic marketing research: A decision-making approach. Upper Saddle River, NJ: Prentice Hall.

Peterson, Robert A., and William R. Wilson. 1992. Measuring Customer Satisfaction: Fact and Artifact. Journal of the Academy of Marketing Science 20 (No. 1): 61-71.

Preston, Carolyn C., and Andrew M. Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychologica 104: 1-15.

Figure 1. Examples of distributions with same mean but different kurtosis

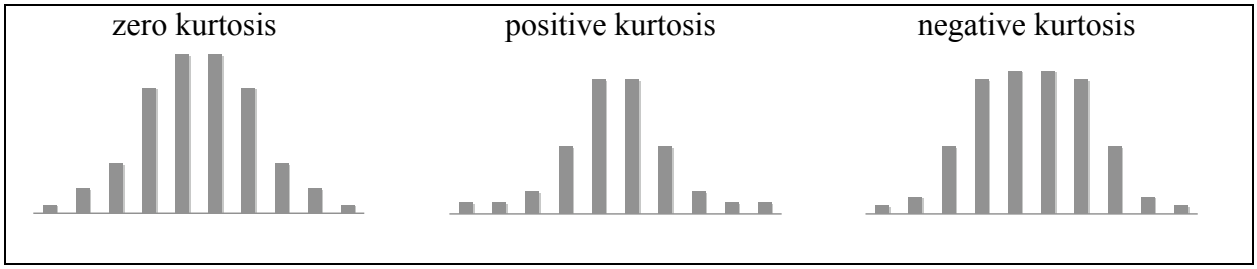

Table 1. Scale items

Item #	Statement
1	When I am in a shop I will always check prices on alternatives before I buy
2	When I buy or shop, I really look for specials
3	I usually watch ads for announcements of sales
4	I believe a person can save a lot of money by shopping around for bargains
5	In a store, I check the prices, even when I am buying inexpensive items
6	I pay attention to sales and specials
7	Clothing, furniture, or appliances ... whatever I buy, I shop around to get the best prices
8	I often wait to purchase items, so I can get them on sale

Table 2. Survey Breakdown

Age category	Sample N	% of sample
Under 21 years	33	5
21 to 30 years	85	12
31 to 40 years	136	19
41 to 50 years	184	25
51 to 60 years	148	20
Over 60 years	149	20
Total	**734**	

Table 3. Re-scaling method for this study

Five-point scale		Seven-point scale		Ten-point scale	
Original value	Re-scaled value	Original value	Re-scaled value	Original value	Scale value
1	1.0	1	1.0	1	unaltered
2	3.25	2	2.5	2	”
3	5.5	3	4.0	3	”
4	7.75	4	5.5	4	”
5	10	5	7.0	5	”
		6	8.5	6	”
		7	10	7	”
				8	”
				9	”
				10	”

Table 4. Mean Scores according to Scale Format

Scale Item	Mean score: 5-point data re-scaled to /10	Mean score: 7-point data re-scaled to /10	10-point data	Mean score 5 point rescaled minus 10 point	Mean score 7 point rescaled minus 10 point	Mean score 5 point rescaled minus 7 point
1	7.8	8.1	7.4	0.4	0.7	-0.3
2	7.4	7.3	6.9	0.5	0.4	0.1
3	5.1	4.6	4.8	0.3	-0.2	0.5
4	7.9	8.1	7.4	0.5	0.7	-0.2
5	6.8	6.9	6.6	0.2	0.3	-0.1
6	7.0	6.9	6.6	0.4	0.3	0.1
7	7.1	7.3	7.6	-0.5	-0.3	-0.2
8	5.9	6.0	5.3	0.6	0.7	-0.1
Overall score (average of all eight items)	**6.9**	**6.9**	**6.6**	**0.3**	**0.3**	**0.0**

** statistically significant difference to the other two formats at p=0.04.*

Table 5. Standard Deviation according to Scale Format

Scale Item	Standard Deviation: 5 point re-scaled /10	Standard Deviation: 7 point re-scaled /10	Standard Deviation: original 10-point data	Std. Dev. 5 point rescaled minus 10 point	Std. Dev. 7 point rescaled minus 10 point	Std. Dev. 5 point rescaled minus 7 point rescaled
1	2.7	2.4	2.5	0.2	-0.1	0.3
2	2.7	2.9	2.7	0.0	0.2	-0.2
3	3.2	3.1	3.1	0.1	0.0	0.1
4	2.4	2.3	2.6	-0.2	-0.3	0.1
5	3.0	2.9	2.8	0.2	0.1	0.1
6	2.7	2.7	2.7	0.0	0.0	0.0
7	2.7	2.6	2.4	0.3	0.2	0.1
8	2.9	2.9	2.8	0.1	0.1	0.0
Overall score (average of all eight items)	**2.0**	**1.9**	**2.0**	**0.0**	**-0.1**	**0.1**

Table 6. Skewness according to Scale Format

Scale Item	Skewness 5 point re-scaled /10 (std error = .14 all items)	Skewness 7 point re-scaled /10 (std error = .16 all items)	Skewness original 10-point data (std error = .16 all items)	Skewness 5 point rescaled minus 10 point	Skewness 7 point rescaled minus 10 point	Skewness 5 point rescaled minus 7 point rescaled
1	-1.2	-1.4	-0.8	-0.4	-0.6	0.2
2	-0.8	-0.8	-0.6	-0.2	-0.2	0.0
3	0.2	0.4	0.3	-0.1	0.1	-0.2
4	-1.1	-1.4	-0.9	-0.2	-0.5	0.3
5	-0.7	-0.7	-0.5	-0.2	-0.2	0.0
6	-0.7	-0.6	-0.5	-0.2	-0.1	-0.1
7	-0.6	-0.8	-1	0.4	0.2	0.2
8	-0.1	-0.2	0.0	-0.1	-0.2	0.1
Overall score (average of all 8 items)	**-0.5**	**-0.4**	**-0.4**	**-0.1**	**0.0**	**-0.1**

Table 7. Kurtosis according to Scale Format

Scale Item	Kurtosis 5 point re-scaled /10 (std error = 0.47 all items)	Kurtosis 7 point re-scaled /10 (std error = 0.47 all items)	Kurtosis original 10- point data (std error = 0.28 all items)	Kurtosis: 5 point rescaled minus 10 point	Kurtosis: 7 point rescaled minus 10 point	Kurtosis: 5 point rescaled minus 7 point rescaled
1	0.6	1.3	-0.1	0.7	1.4	-0.7
2	-0.4	-0.5	-0.7	0.3	0.2	0.1
3	-1.3	-1.1	-1.3	0.0	0.2	-0.2
4	0.3	1.3	-0.2	0.5	1.5	-1.0
5	-0.7	-0.6	-0.8	0.1	0.2	-0.1
6	-0.4	-0.5	-0.8	0.4	0.3	0.1
7	-0.8	-0.2	0.2	-1.0	-0.4	-0.6
8	-1.1	-1.0	-1.1	0.0	0.1	-0.1
Overall score (average of all 8 items)	**-0.4**	**-0.5**	**-0.4**	**0.0**	**-0.1**	**0.1**

IMPACT OF THE NUMBER OF RESPONSE CATEGORIES AND ANCHOR LABELS ON COEFFICIENT ALPHA AND TEST-RETEST RELIABILITY

LI-JEN WENG
National Taiwan University

A total of 1,247 college students participated in this study on the effect of scale format on the reliability of Likert-type rating scales. The number of response categories ranged from 3 to 9. Anchor labels on the scales were provided for each response option or for the end points only. The results indicated that the scales with few response categories tended to result in lower reliability, especially lower test-retest reliability. The scales with all the response options clearly labeled were likely to yield higher test-retest reliability than those with only the end points labeled. Scale design that leads to consistent participant responses as indicated by test-retest reliability should be preferred.

Keywords: *Likert-type rating scales; number of response categories; anchor labels; test-retest reliability; internal consistency; coefficient alpha*

Since Likert's (1932) introduction of the summative method, Likert-type rating scales have enjoyed great popularity among social science researchers (Likert, Roslow, & Murphy, 1934; Wang & Weng, 2002), and they have also drawn much research attention to the effects of scale format on participants' responses and associated psychometric properties. One intensively examined topic is the effect of the number of response categories on scale reliability, especially coefficient α, an estimate of internal consistency reliability.

I would like to thank all the individuals who participated in this study. The assistance of L.-F. Li in data collection and analysis is gratefully acknowledged. I would also like to express my appreciation to the anonymous reviewer for the suggestion to present the graphs of polynomial trend lines. This research was supported by Grant NSC 88-2413-H-002-010 from the National Science Council of Taiwan. Correspondence concerning this article should be addressed to Li-Jen Weng, Department of Psychology, National Taiwan University, Taipei 106, Taiwan, Republic of China; e-mail: ljweng@ntu.edu.tw.

Educational and Psychological Measurement, Vol. 64 No. 6, December 2004 956-972
DOI: 10.1177/0013164404268674

The stability of measurement scores over time is also critical, and the impact of the number of response options on test-retest reliability should be assessed as well. The objective of the present research was to investigate the effects of the scale properties of Likert-type scales on test-retest reliability in addition to internal consistency reliability. The scale properties studied included the number of response categories and the anchor labels attached to the scale.

Reliability evaluates the influence of measurement errors on participants' responses. Types of reliability differ in the sources of measurement errors considered. Internal consistency reliability considers the degree of interrelatedness among individual items, whereas test-retest reliability is concerned with the stability of scale scores across occasions. The evaluation of internal consistency reliability alone can often be inadequate, because no information on the stability of participants' responses is provided (Cortina, 1993; Crocker & Algina, 1986). Inconsistent measures of participants' responses may result in misleading scientific conclusions (Krosnick & Berent, 1993). It is therefore essential to investigate the effect of scale design on test-retest reliability in addition to internal consistency reliability. Coefficient α has received more attention than test-retest reliability in the past (e.g., Aiken, 1983; Bandalos & Enders, 1996; Halpin, Halpin, & Arbet, 1994; Johnson, Smith, & Tucker, 1982; Jenkins & Taber, 1977; Ko, 1994; Komorita & Graham, 1965; Lissitz & Green, 1975; Masters, 1974; Matell & Jacoby, 1971; Oaster, 1989; Preston & Colman, 2000; Wong, Chuen, & Fung, 1993). Hogan, Benjamin, and Brezinski (2000) found that coefficient α was used in over two thirds of the tests they reviewed, and fewer than 20% of the tests reported test-retest reliability. The scarcity of previous studies on test-retest reliability is probably due to the necessity of repeatedly administering the same scale to identical participants. In contrast, the estimation of coefficient α requires only one administration of the measure.

Previous findings on the relationship between the number of response categories and coefficient α have been inconsistent. Some researchers concluded that the number of response categories has no effect on coefficient α (e.g., Aiken, 1983; Wong et al., 1993); others found coefficient α to be affected by the number of options provided but offered different recommendations. For example, Matell and Jacoby (1971) suggested 2 or 3 response categories, Johnson et al. (1982) recommended a 3-point format, Ko (1994) and Oaster (1989) recommended 6- and 7-point scale designs, and Preston and Colman (2000) recommended 7 to 10 points. Bendig (1953) recommended 9 response categories, and Champney and Marshall (1939) recommended as many as 18 response categories. Churchill and Peter (1984) focused on internal consistency reliability and found increasing reliability with more response categories in their meta-analysis of 108 marketing research studies. Jenkins and Taber (1977) and Lissitz and Green (1975) conducted simulation studies and suggested that 5 response categories are sufficient, because no substantial gains in reliability were observed after 5 catego-

ries. However, the applicability of their results to empirical data might be limited because of their assumption of uniformly distributed item scores in the simulations (Micceri, 1989).

The relationship between the number of scale points and test-retest reliability is inconclusive as well. Among the few related studies, Jenkins and Taber (1977) and Lissitz and Green (1975) used simulated data, and Oaster (1989) used alternate forms measured in two sessions, which was in essence an alternate-form reliability estimate. Matell and Jacoby (1971) used scales with numbers of response categories ranging from 2 to 19 and suggested 2 or 3 options to be adequate. However, to incorporate the wide range of numbers of response categories in one single study, only 20 participants completed each form of the scale. Johnson et al. (1982) used the Job Descriptive Index with 3 and 5 response options. Fifty students participated in each condition. Although the 5-point scale discriminated among individuals better than the 3-point format, no substantial differences in test-retest reliability between the two formats were found. Preston and Colman (2000) instead asked a group of 149 participants to respond to the same scale, with the number of responses options ranging from 2 to 11 and 101, and they recommended at least 7 response categories to ensure stable participant responses.

Why did the influence of the number of response categories on reliability differ among studies? Item homogeneity may be a plausible explanation. Komorita and Graham (1965) studied the relationship between the number of scale points and the internal consistency reliability of scales under varying degrees of item homogeneity. The degree of item homogeneity was defined by the sizes of factor loadings on one single factor. Coefficient α was found to be independent of the number of scale points with homogeneous items. The sizes of factor loadings played a role in mediating the relationship between scale design and reliability. Scales of homogeneous items were likely to be less affected by the format used. When heterogeneous items are asked, a respondent would be likely to have differential degrees of propensity for each question. Increasing response options enables the respondent to map his or her response to the appropriate category and thus reduces inconsistent random errors and raises reliability. This hypothesis was to be tested in the present study by including scales of different sizes of factor loadings. The scale with higher loadings was expected to be less affected by the number of response categories used.

In sum, most empirical research on the effects of response categories on the reliability of Likert-type rating scales has focused on internal consistency reliability. Simulation studies might be limited in providing appropriate suggestions to scale design because of the assumption of uniformly distributed scores. The effects of the number of response categories on the stability of scale scores as assessed by test-retest reliability, though important, have been less empirically researched and call for further investigation. Furthermore,

among the few empirical studies of test-retest reliability, different numbers of response categories have been recommended. The present study was therefore designed to examine empirically how many response categories were needed to raise test-retest reliability and internal consistency reliability to desired levels simultaneously.

Why would the number of response categories be expected to affect test reliability? In considering the coarseness of categorization, Symonds (1924) described a parallel between psychological measurements and physical measurements. A physical measurement that used a scale finer than the limits of one's eyesight is useless. Likewise, a psychological measurement is very likely to be of limited value if it uses a scale finer than a judge's ability to discriminate. A scale that requires finer discrimination than the target respondents usually can accomplish may easily add measurement errors to test scores. Increasing the number of scale points does not necessarily lead to better discrimination of participants' judgment on personal attitudes or traits. For instance, an individual may have difficulty discriminating the difference between 8 and 9 on an 11-point scale. The respondent may check 8 on one occasion but 9 on another for an identical item. The inconsistency between the two sets of scores is then due to scale design rather than the trait being measured. On the other hand, a scale with few scale points may lose information on individual differences and lower the reliability estimates.

More recently, Tourangeau, Rips, and Rasinski (2000) proposed a model of survey response processes, and they share a similar view to the proposition of Symonds (1924). Four major components were identified in the model: comprehension of the item, the retrieval of relevant information, the use of the information to make judgments, and the selection and reporting of an answer. The specific process of mapping judgment onto response categories in the last component was related to the influences of scale format on participant responses. With too few categories, the rating scales may fail to discriminate between respondents with different underlying judgments; with too many, respondents may fail to distinguish reliably between adjacent categories. In either case, inconsistent random errors are likely to be introduced and lead to lower reliabilities.

In constructing a rating scale, in addition to how many response options to use, a researcher also determines what verbal labels are to be presented with the scale. Each response option of the scale can be labeled with a verbal description. For example, a 5-point rating scale can include labels such as *strongly agree*, *agree*, *undecided*, *disagree*, and *strongly disagree*. A rating scale can also be anchored with verbal labels of *strongly agree* and *strongly disagree* at only the end points of the scale. Participants' responses on the basis of these two formats of anchor labels have been compared (e.g., Dickinson & Zellinger, 1980; Dixon, Bobo, & Stevick, 1984; Finn, 1972; French-Lazovik & Gibson, 1984; Frisbie & Brandenburg, 1979; Klockars &

Yamagishi, 1988; Lam & Klockars, 1982; Landrum, 1999; Newstead & Arnold, 1989; Wildt & Mazis, 1978). Most previous research has focused on the comparison of score distributions obtained from two forms, and some studies have investigated respondents' preferences of label format. Landrum's (1999) participants showed no difference in their confidence in answering a scale with every response option clearly specified and a scale with only end points defined. Dickinson and Zellinger's (1980) respondents reported being more satisfied when more rating scale points were verbally labeled. Although Finn (1972) compared the reliability obtained from different forms, the effect of scale format on the stability of participants' responses across occasions was not examined.

Churchill and Peter's (1984) meta-analysis suggested that rating scales with each response category clearly defined or with only end points labeled yielded similar reliabilities. However, whether their conclusions could be generalized to test-retest reliability was unclear because of their focus on internal consistency reliability. Moreover, Krosnick's (1999) review of survey research suggested that reliability could be significantly improved if all points on a scale are labeled with words, because they clarify the meanings of the scale points. Krosnick's conclusion was drawn largely from his previous study with Berent (1993), in which data were collected through telephone interviews, face-to-face interviews, and self-administrated questionnaires. This study was insufficient in evaluating the effects of verbal labeling on rating scales for two reasons. First, only one of the eight experiments conducted used self-administrated questionnaires, a frequently adopted method of data collection in educational and psychological research. Second, the experiment that used self-administrated questionnaires compared score consistency between a rating scale with two ends labeled and a branching scale with all response options defined. The branching format asked respondents to respond in two steps, first the direction and then the intensity. A comparison of partially and fully labeled rating scales would be more informative for educational and psychological research. With the limitations and the conflicting conclusions from previous studies, these two frequently used formats of Likert-type rating scales were constructed in the present study to examine the effects of labeling on reliability. If each response category were clearly specified, participants would be less likely to change their interpretations of each response option from one occasion to another. Therefore, a rating scale with all the response categories verbally specified was expected to yield more stable participant responses and higher test-retest reliability than a scale with only end points labeled.

The purpose of the present research was twofold. First, the relationship between the number of response categories on rating scales and reliability was empirically studied. The types of reliability included test-retest reliability and internal consistency reliability, with the former being emphasized

because of the sparse past research on the topic. It was hypothesized that test-retest reliability would be lower with few response categories because of the loss of information. Test-retest reliability might also decrease with a large number of response categories if the degree of discrimination demanded were beyond the abilities of the participants. The mediating effect of item homogeneity on the relationship between the number of response categories and reliability, as suggested by Komorita and Graham (1965), was also tested. The reliability of scales composed of more homogeneous items, as indicated by higher factor loadings, was expected to be less affected by the number of response options used.

Second, the effect of two common forms of anchor labels on scale reliability was investigated. Churchill and Peter's (1984) meta-analysis did not support the hypothesis that scales on which all points were labeled would have higher reliability than scales on which only end points were labeled. Because their analysis gave primary emphasis to internal consistency reliability, and their results contradicted the conclusions of Krosnick (1999), these two forms of scales were constructed in this study to test the effects of anchor labels on test-retest reliability in addition to coefficient α. It was hypothesized that a scale with each anchor label clearly stated should lead to more stable participant responses over time, as represented by higher test-retest reliability, than a scale with only end points labeled.

Method

Instrumentation

Two subscales of the Teacher Attitude Test were used in the present study. The Teacher Attitude Test is used to select eligible students for the Teacher Education Program of National Taiwan University. Over the years, fewer than 20% of the applicants have been admitted to the program. Only those who successfully complete the program are qualified to apply for teaching positions in high schools and junior high schools. Therefore, finding the optimal scale format for the test was of practical importance in selecting the best candidates for potential schoolteachers.

Two of the essential dimensions of the Teacher Attitude Test were chosen for the present study: the Concern for Others scale (CO) and the Determination scale (DE). The 12-item CO scale consisted of two related subcomponents: being sensitive to others' emotions (e.g., "I can easily detect the emotional changes of my family members") and being willing to share experiences with others (e.g., "I enjoy sharing my experiences with others"). The DE scale included 13 items and measured applicants' determination to carry on in times of frustration and difficulty (e.g., "I take on obstacles as chal-

lenges to be overcome"). The original design used a 5-point scale with all response categories clearly labeled. The test was initially administered to more than 1,000 university students applying to the program. Preliminary factor analysis found the items of the two scales to load mainly on one factor, respectively. The factor loadings for the CO scale ranged from 0.43 to 0.74 ($M = 0.58$, $SD = 0.09$). The DE scale had factor loadings ranging from 0.56 to 0.81 ($M = 0.67$, $SD = 0.08$). The loadings for the DE scale were higher than the CO scale, suggesting that the items on the DE scale were more homogeneous than those on the CO scale (Komorita & Graham, 1965).

Every participant was asked to rate how well each statement described him or her. Two forms of each scale were constructed, one with each response category clearly labeled (the ALL form) and the other with only end points labeled (the END form). The labels were selected according to the scale values estimated by Weng (1998). Weng computed the scale values for anchor labels commonly used in Chinese rating scales. The scale values based on the successive interval method (Edwards, 1957) were used to select appropriate anchor labels for the present study. Anchor labels that were approximately equally spaced and had small dispersion were selected for the ALL forms, with the number of scale points ranging from 3 to 8. For example, the 5-point scale adopted the labels *does not describe me at all*, *does not describe me in general*, *can't say*, *describes me in general*, and *describes me completely.* The scale values based on the successive interval method for the 5 anchor labels were 0.066, 2.598, 5.163, 7.632, and 9.931, respectively. They were about equally spaced, with the differences between any two successive labels being 2.466, 2.532, 2.565, 2.469, and 2.299. The labels attached to two ends of the ALL forms were used to define the end points of the END forms. The number of response categories for the END forms ranged from 4 to 9. There was therefore a total of 12 combinations of number of response categories and anchor label format.

Participants

A total of 1,247 college students had complete data from two testing sessions in this study. This sample consisted of 78.6% of the original 1,587 students who filled out the questionnaires at the initial contact. Students received partial credit for introductory psychology courses or gifts for their participation in the experiment. The sample was homogeneous with regard to ethnicity and included 459 men and 788 women from 13 colleges in Taiwan. The number of participants who responded to each form of the scales is presented in Table 1. About 100 participants took each scale form. The 8-point END form had 80 participants, being the smallest among the 12 samples. The sample for the 4-point END form was the largest, consisting of 121 individuals.

Table 1
Mean, Standard Deviation, Coefficient α, and Test-Retest Reliability of the Concern for Others Scale

		Session 1			Session 2			
Number of Categories	*n*	*M*	*SD*	α	*M*	*SD*	α	*r*
ALL form								
3	98	28.74	3.07	.64	28.68	3.31	.70	.77
4	93	36.73	3.79	.70	36.58	3.43	.64	.76
5	101	45.59	4.91	.75	45.98	4.87	.78	.76
6	105	52.06	6.14	.73	51.86	6.32	.79	.82
7	91	60.11	8.49	.85	59.20	7.47	.81	.90
8	89	68.57	8.09	.81	67.91	8.67	.85	.86
χ^2				21.92*			20.23*	16.95*
END form								
4	121	37.36	4.60	.72	37.58	4.48	.79	.72
5	116	45.81	4.77	.58	45.16	5.59	.74	.62
6	119	53.90	6.69	.77	52.99	6.86	.80	.83
7	117	60.95	8.56[a]	.78	58.96	7.34	.75	.77
8	80	71.35	9.49	.76	71.03	8.37	.72	.70
9	117	79.52	11.58	.79	78.61	11.08	.82	.80
χ^2				15.61*			7.04	15.74*

Note. The *t* statistic for the test of the equality of reliabilities represents the *k*-sample significance test for independent α coefficients and the test of equality of multiple independent correlations for *r*.
a. Means between two sessions were significantly different statistically ($p < .01$).
*$p < .01$.

Design and Procedures

All participants responded to both the CO and DE scales, with items of the two scales mixed. The questionnaires were administered to the participants in groups. Each form of the scales was administered to the same group of participants twice to evaluate score stability over time (test-retest reliability). The two testing sessions were scheduled to be at least 4 weeks apart, ranging from 29 to 43 days, to avoid memory effects. The traits measured, determination and concern for others, should not change dramatically within 4 to 6 weeks for college students.

Analyses

The equivalence of test scores across two administrations and between the ALL forms and the END forms was first evaluated prior to the assessment of scale reliability. Dependent *t* tests were performed to test if the means of the same form from two testing sessions were significantly different. Independent *t* tests were performed to test whether the means of the ALL forms and the END forms were different. Internal consistency reliability was estimated by coefficient α. Test-retest reliability was assessed by Pearson's correlation

Table 2
Mean, Standard Deviation, Coefficient α, and Test-Retest Reliability of the Determination Scale

		Session 1			Session 2			
Number of Categories	*n*	*M*	*SD*	α	*M*	*SD*	α	*r*
ALL form								
3	98	30.42	4.54	.86	30.76	4.69	.89	.83
4	93	37.12	5.67	.89	37.20	5.63	.90	.85
5	101	44.24	7.00	.88	44.76	7.10	.91	.81
6	105	52.61	7.73	.86	52.97	8.09	.90	.78
7	91	57.81	9.98	.89	57.77	9.72	.91	.81
8	89	69.29	11.03	.91	68.82	11.87	.93	.89
χ^2				6.19			5.30	7.89
END form								
4	121	37.60	6.39	.89	38.03	6.11	.88	.78
5	116	45.66	7.16	.84	44.92	7.71	.88	.73
6	119	53.31	9.11	.88	53.49	8.79	.90	.83
7	117	60.10	11.02	.89	59.53	9.73	.88	.84
8	80	67.13	13.13	.88	67.79	13.28	.90	.71
9	117	76.94	17.24	.91	78.58	17.10	.93	.86
χ^2				9.02			12.64	13.97

Note. The *t* statistic for the test of the equality of reliabilities represents the *k*-sample significance test for independent α coefficients and the test of equality of multiple independent correlations for *r*.

between scores of the same form from two testing sessions. The *k*-sample significance test for independent α coefficients proposed by Hakstian and Whalen (1976) was adopted to test the effects of the number of response categories and anchor labels on coefficient α. The effects of these two manipulated factors on test-retest reliability were tested by the test of the equality of multiple independent correlations (Hays, 1994, p. 651). Because of the great number of statistical tests conducted in this research, the probability of making a Type I error on each test was constrained at .01 to guard against the possible inflation of the experiment-wise error rate of the study.

Results

The means and standard deviations of each form of the CO and DE scales at two testing sessions are summarized in Tables 1 and 2. The means and standard deviations of both scales increased as more response categories were used, regardless of the format of the anchor labels used. The standard deviations for the DE scale were larger than for the CO scale, indicating greater individual variation on the DE scale among college students. The standard deviations for the END forms were slightly larger than for the ALL forms in some cases.

Statistical tests and associated measures of effect size were used to compare scale scores across various conditions. Dependent *t* tests were used to test the equivalence of scale means across two testing sessions. The results indicated that the means from two administrations of the same scales were not significantly different statistically, except for the CO scale on the 7-point END form, $t(116) = 3.90$, $p < .01$, $d = .36$, suggesting stable participant responses over time. Independent *t* tests were used to compare the first-session scale means between the ALL form and the END form that consisted of the same number of response options. The scale means from these two forms were also nonsignificantly different statistically, although some of the *d* measures reached small effect sizes (Cohen, 1988). The format of the anchor labels appeared to have no effect on average participant responses on both scales.

CO Scale

Coefficient α at two testing sessions and test-retest reliability on both forms of the CO scale are presented in Table 1. The test statistics for testing equal reliability against differing numbers of response categories are also presented. The *k*-sample significance test for independent α coefficients proposed by Hakstian and Whalen (1976) was adopted to test the effects of the number of response categories on coefficient α statistically. This method for testing α coefficients under six independent conditions was distributed as a χ^2 distribution with 5 degrees of freedom under the null hypothesis of equal reliability. The null hypothesis was rejected statistically by three out of the four tests, suggesting that except for the END form at the second testing session, coefficient α values varied with the number of response categories used. The scales with more response categories tended to yield higher coefficient α values than those of fewer response options, as the fitted polynomial trend lines in Figure 1 suggest. The discrepancy of coefficient α values between two testing sessions decreased as the number of response options increased, suggesting less sampling variability of estimates of coefficient α with more response categories.

The *k*-sample significance test for independent α coefficients (Hakstian & Whalen, 1976) was again used to test the equality of coefficient α values between the ALL form and the END form of the same number of response options. The coefficients α values of these two forms were not significantly different statistically, except for the 4-point scale and the 8-point scale at the second testing session, indicating a similar size of coefficient α for both forms. Labeling each response option or labeling only the end points had no effect on coefficient α of the CO scale.

The test for the equality of multiple independent correlations (Hays, 1994, p. 651) was used to test the effect of the number of scale points and anchor

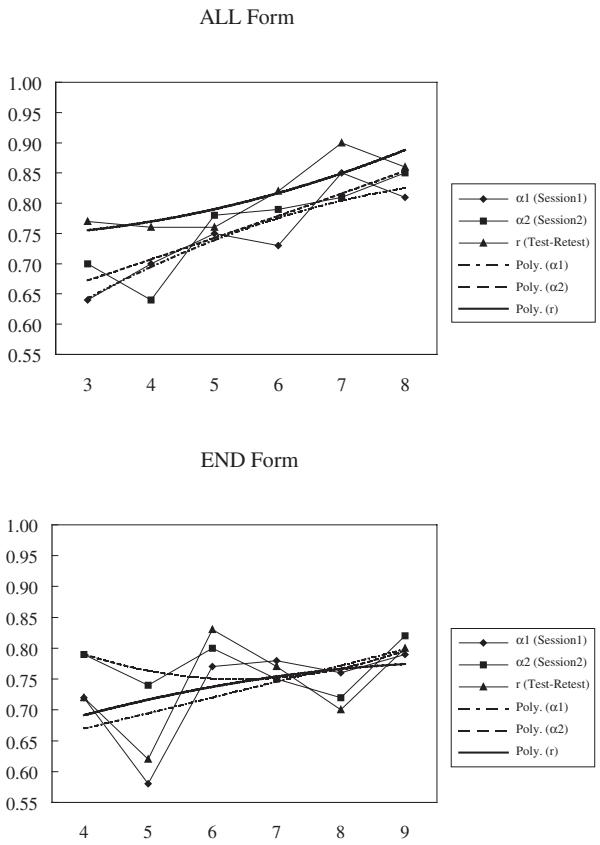

Figure 1. Reliability estimates and fitted polynomial trend lines as a function of the number of response categories for the Concern for Others scale.

labels on test-retest reliability, as assessed by Pearson's product–moment correlation coefficient. This test under six independent conditions of the number of response categories was distributed as a χ^2 distribution with 5 degrees of freedom under the null hypothesis of equal test-retest reliability. The null hypothesis was rejected for both forms of the CO scale, indicating a significant effect of the number of response categories on test-retest reliability. The polynomial trend lines fitted to the data suggested a rising trend of test-retest reliability with more response categories, especially for the ALL form, on which each response option was explicitly expressed. Scales with fewer than 6 response categories tended to yield low test-retest reliability.

The same statistical test (Hays, 1994, p. 651) was also adopted to test the equality of test-retest reliability between two scale formats, yielding a χ^2 distribution of 1 degree of freedom under the null hypothesis of equal reliability

between the ALL form and the END form. The null hypothesis was rejected with 7 and 8 response options and retained for scales of 4 to 6 points. The 7- and 8-point ALL forms yielded significantly higher test-retest reliability than the END forms, suggesting that with more response categories, participants' responses on scales with each anchor label clearly specified were more stable than those on scales with only end points labeled.

In sum, for the CO scale, the number of response categories had an effect on coefficient α and test-retest reliability, whereas the format of the anchor labels showed no effect either on coefficient α, regardless of the number of scale points, or on test-retest reliability with 4 to 6 scale points. For scales with 7 and 8 response options, clearly specifying each response option yielded higher test-retest reliability than specifying the end points only.

DE Scale

The reliability estimates for the DE scale with associated test statistics are summarized in Table 2. All the reliability estimates appeared high and of similar sizes, regardless of the number of response categories used and the design of the anchor labels. The identical statistical tests conducted for the CO scale were applied to the DE scale. None of the null hypotheses of equal coefficient α values across different numbers of response categories was rejected by the four tests, as indicated in Table 2, suggesting that the number of scale points had no effect on coefficient α. The 10 tests of differences between α coefficients obtained from the ALL from and the END form composed of the same number of response options all yielded statistically nonsignificant results, indicating equal coefficient α values across two forms of anchor labels for the DE scale. The results suggested that coefficient α of the DE scale, which had higher factor loadings than the CO scale, was independent of the number of response options offered and the format of anchor labels used.

The test of the equality of multiple independent correlations (Hays, 1994, p. 651) failed to reject the null hypothesis of equal test-retest reliability across different numbers of response categories for both forms. The test was again applied to test the equality of test-retest reliability between the ALL form and the END form of the same number of response options. The null hypothesis was rejected only when 8 response options were used. Test-retest reliability appeared independent of the number of options offered and the anchor labels provided, except for the 8-point scale. With the 8-point scale, the ALL form yielded higher test-retest reliability than the END form, indicating more stable participant responses when all the scale points were clearly defined.

In sum, the internal consistency α coefficient and test-retest reliability for the CO scale tended to increase with more scale points. With 7 and 8 response options, test-retest reliability for the ALL form was higher than for the END form. The reliability of the DE scale, which had higher factor loadings than

the CO scale, was less affected by the number of scale points and the anchor label format adopted. Increasing the number of response categories raised both types of reliability of the CO scale but had no effect on the reliability of the DE scale. The mediating effect of item homogeneity, as indicated by differential sizes of factor loadings (Komorita & Graham, 1965), on the relationship between number of scale points and reliability was supported in this study. The format of anchor labels had no effect on coefficient α for both the CO and DE scales. However, if more response options were used, the ALL form might yield higher test-retest reliability than the END form.

Discussion

Likert-type rating scales have been widely used for the assessment of participants' responses and attributes. Careful scale design is essential for achieving satisfactory scale reliability and appropriate research conclusions (Krosnick & Berent, 1993). Compared with the vast research on the impact of scale design on reliability in the past, the present study is one of the few that have researched the relationship between scale design and test-retest reliability empirically. Because high internal consistency reliability, as commonly measured by coefficient α, does not necessarily guarantee high test-retest reliability (Cortina, 1993; Crocker & Algina, 1986), the impact of scale design on test-retest reliability warrants direct investigation. The present study empirically investigated the impact of scale properties on test-retest reliability in addition to internal consistency reliability on two subscales of the Teacher Attitude Test, used for selection purposes.

The present study suggests a differential impact of scale design on reliability. The reliability of the CO scale was affected by scale design, whereas the reliability of the DE scale was independent of the format of the scale used in general. The original hypothesis of the effect of the number of response categories being mediated by factor loadings was supported. The present study replicated the findings of Komorita and Graham (1965) that the reliability of scales with homogeneous items is less affected by the design of the scale format. In the present study, the DE scales had higher factor loadings than the CO scales and were less affected by not only the number of response categories but also the form of the anchor labels attached. The size of factor loading seemed to play a role in mediating the relationship between scale design and reliability. Scales of more homogeneous items were less affected by the format used. However, an examination of the relative sizes of the standard deviations of both scales might offer another explanation for the differences in reliability. Masters (1974) found that the reliability of the questionnaire with larger individual variation was less affected by the number of response categories used. And in the present study, the DE scale of higher factor loadings happened to have larger standard deviations than the CO scale. The differences between reliability might result from the differential

degrees of individual differences demonstrated on the two traits measured. Further research is needed to clarify the influences of these two factors on the relationship between scale design and reliability.

The results of the present research indicated that reliability did not necessarily level off after 5 categories, as the simulations of Lissitz and Green (1975) and Jenkins and Taber (1977) demonstrated. The recommendations based on the simulated uniformly distributed scores should not be routinely applied in designing Likert-type scales for empirical investigations.

Halpin et al. (1994) suggested that the best choice of the number of scale points depends largely on the content measured in the scale. However, a researcher can avoid adopting a scale format that is likely to yield low reliability. The suggestion proposed by Matell and Jacoby (1971) that 2 or 3 categories should be enough for sufficient reliability needs qualification. The results of this study favor the conclusions of Preston and Colman (2000) that more than 3 response options are needed to achieve stable participant responses. The small sample size in Matell and Jacoby's study might have led to unstable conclusions. The present investigation with larger samples indicates that although fewer scale points is not necessarily paired with lower reliability, scales with more categories have a better chance of attaining higher reliability.

A larger number of scale points was shown to pair with smaller standard errors of reliability estimates in Lissitz and Green's (1975) simulation. Reliability estimates of more categories have less dispersion and result in better stability across samples. Decreasing dispersion with an increasing number of response categories was observed in the present study with the CO scale. Reliability estimates with few categories tend to fluctuate from sample to sample. A rating scale with fewer than 5 scale points should therefore be discouraged if possible. As for the optimal number of scale points, researchers have to take into account the cognitive discriminating ability of the target population (Andrich & Masters, 1988; Komorita & Graham, 1965; Krosnick, 1999). Some suggestions are in line on the basis of the results of the present study. If the cognitive ability of the participants is close to that of college students, an odd-numbered, 7-point scale and an even-numbered, 6-point scale should be able to provide consistent and reliable participant responses. Whether these suggestions apply to other populations of less cognitive sophistication requires further investigation.

The hypothesis that the full specification of response options improves reliability was only partially supported in the present study. The effect of anchor labels on scale reliability depended on the type of reliability estimated. The α coefficients of both forms were similar, replicating Churchill and Peter's (1984) findings of no difference between the reliability estimates with the ALL forms and END forms when internal consistency reliability was considered. Considering the stability of scores as represented by test-retest reliability, the ALL form scale outperformed the END form scale when

more response options were given, as Krosnick and Berent (1993) demonstrated in their study of the consistency of political attitudes over time. Although coefficient α appeared independent of the format of verbal labels attached, a scale with each anchor label clearly specified should be preferred to achieve consistent and stable participant responses. Moreover, with the frequent use of 7-point scales in psychological and management studies (Wang & Weng, 2002), a full specification of response options increases the likelihood of inducing stable participant reaction on the measures.

Clearly stating each anchor label has another advantage of enhancing interpretation of measurement results. Take a group mean of 2.1 on a 7-point scale as an example. The ALL form gives a better sense of what the average represents than the END form. Interpreting the meaning of an average of 2.1 on the END form is a tricky task for readers. Readers may have different interpretations. Take a mean of 2.5 as another example. A mean of 2.5 indicates that on average, participant responses locate in between the labels attached to 2 and 3 on the ALL form. The interpretation of a mean of 2.5 will be difficult and ambiguous on an END form scale. Therefore, clearly specifying each response category verbally should be encouraged, as suggested by Krosnick (1999), to improve test reliability and to facilitate the interpretation of study results.

The evaluation of only internal consistency reliability in scale development is apparently inadequate for understanding scale reliability. Coefficient α, the frequently used internal consistency reliability estimate, provides little information on the stability of participant responses (Cortina, 1993; Crocker & Algina, 1986; Jenkins & Taber, 1977). Test-retest reliability, which assesses the stability of participants' responses over time, should be evaluated in addition to internal consistency reliability. Although the assessment of test-retest reliability requires at least two administrations of the scale and demands more time and effort than the assessment of internal consistency reliability, a scale with low test-retest reliability may lead to inappropriate conclusions (Krosnick & Berent, 1993). Scale developers should make every effort to evaluate test-retest reliability over and above internal consistency reliability.

References

Aiken, L. R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement*, *43*, 397-401.

Andrich, D., & Masters, G. N. (1988). Rating scale analysis. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 297-303). Oxford, UK: Pergamon.

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education*, *9*, 151-160.

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories. *Journal of Applied Psychology*, *37*, 38-41.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology*, *23*, 323-331.
Churchill, G. A., Jr., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, *21*, 360-375.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98-104.
Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.
Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology*, *65*, 147-154.
Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement*, *44*, 61-66.
Edwards, A. L. (1957). *Techniques of attitude scale construction.* New York: Irvington.
Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, *32*, 255-265.
French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement*, *8*, 49-57.
Frisbie, D. A., & Brandenburg, D. C. (1979). Equivalence of questionnaire items with varying response formats. *Educational and Psychological Measurement*, *16*, 43-48.
Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, *41*, 219-231.
Halpin, G., Halpin, G., & Arbet, S. (1994). Effects of number and type of response choices on internal consistency reliability. *Perceptual and Motor Skills*, *79*, 928-930.
Hays, W. L. (1994). *Statistics* (5th ed.). Orlando, FL: Harcourt Brace.
Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523-531.
Jenkins, G. D., Jr., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*, *62*, 392-398.
Johnson, S. M., Smith, P. C., & Tucker, S. M. (1982). Response format of the Job Descriptive Index: Assessment of reliability and validity by the multitrait-multimethod matrix. *Journal of Applied Psychology*, *67*, 500-505.
Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement*, *25*, 85-96.
Ko, Y.-H. (1994). A search for a better Likert point-scale for Mental Health Questionnaires. *Psychological Testing*, *41*, 55-72.
Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, *15*, 987-995.
Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.
Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, *37*, 941-964.
Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement*, *19*, 317-322.
Landrum, R. E. (1999). Scaling issues in faculty evaluations. *Psychological Reports*, *84*, 178-180.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 5-55.
Likert, R., Roslow, S., & Murphy, G. (1934). A simplified and reliable method of scoring the Thurstone attitude scales. *Journal of Social Psychology*, *5*, 228-238.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, *60*, 10-13.

Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, *11*, 49-53.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657-674.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Newstead, S. E., & Arnold, J. (1989). The effect of response format on ratings of teaching. *Educational and Psychological Measurement*, *49*, 33-43.

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, *68*, 549-550.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1-15.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, *7*, 456-461.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Wang, C.-N., & Weng, L.-J. (2002). Evaluating the use of exploratory factor analysis in Taiwan: 1993-1999. *Chinese Journal of Psychology*, *44*, 239-251.

Weng, L.-J. (1998). Scale values of anchor labels in Chinese rating scales: Responses on frequency and agreement. *Chinese Journal of Psychology*, *40*, 73-86.

Wildt, A. R., & Mazis, M. B. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research*, *15*, 261-267.

Wong, C.-S., Chuen, K.-C., & Fung, M.-Y. (1993). Differences between odd and even number of response scale: Some empirical evidence. *Chinese Journal of Psychology*, *35*, 75-86.