How Many Response Categories Best Scale Stimuli?

Adam Finn, University of Alberta
Ling Peng, Lignan University

Abstract

Multi-item rating scales are the accepted solution for achieving reliable and valid measures in the social sciences. But how many response categories to use is less clear. Past research has employed classical test theory criteria and generally addressed the scaling of respondents, whereas marketers also use multi-item scales to scale stimuli. We suggest generalizability theory criteria better identify how many response categories to use to best scale marketing stimuli. G study data for websites are collected using five and seven category responses to compare their effects on the observed variance components and G-coefficients for websites. Seven category responses outperform five category responses when scaling marketing stimuli.

How Many Response Categories Best Scale Stimuli?

Introduction

Itemized rating scales are the most widely used measuring instruments in marketing, where multi-item scales are the recommended solution for achieving reliable and valid measures. Such rating scales not only require valid item content (Rossiter 2002), they require a decision on an appropriate response format, including how many categories to use. There is plenty of research on these issues, but it has taken a classical test theory perspective and has yet to produce any consensus. This article approaches the number of response category issue from a generalizability theory (hereafter G-theory) perspective (Cronbach, et al. 1972), as it has been identified as a superior approach to marketing measurement issues (e.g., Peter 1979; Rentz 1987). It can be used to determine how many items are needed for a scale to provide data of adequate quality for different managerial decision-making purposes (see Finn and Kayande 1997). G-theory provides specific criteria for assessing the psychometric quality of research data, but it has not yet been accepted in the response category literature.

This paper briefly reviews research on the number of response categories issue. Then, it introduces the G-theory perspective. Then, it reports the results of a study that uses G theory criteria to compare the use of five and seven category ratings in the evaluation of consumer websites. It concludes with the implications for the number of response categories to use in multi-item rating scales in marketing.

Number of Response Categories

Some of the extensive literature on the effect of number of response categories on rating scale performance suggests it has little effect (e.g., Bendig 1953, Jacoby and Matell 1971; Boote 1981; Peterson 1994). Other work reports a positive relationship (e.g., Garner 1960; Green and Rao 1970; Churchill and Peter 1984). Experiments (e.g., Weathers, Sharma and Neidrich 2005, Weng 2004) and meta-analyses (e.g., Churchill and Peter 1984, Peterson 1994) take a classical test theory perspective and focus on performance when scaling respondents. Most research investigates reliability, using coefficient alpha (Churchill and Peter 1984, Peterson 1994, Weathers, Sharma and Niedrich 2005) or test-retest reliability (Jacoby and Matell 1971, Boote 1981). Viswanathan, Sudman and Johnson (2004) argue for the number of categories that are meaningful for respondents. G-theory may make this idea operational. It allows the investigator to identify particular objects of measurement, it disentangles multiple sources of error, and it identifies which ones are of concern for the specific objects of measurement. It also recognizes scale performance can vary across objects of measurement. What is best for scaling respondents may not be what is best for scaling marketing entities (Finn and Kayande 1997). Moreover, how many categories to use might vary with the response format. Most social science research employs either a Likert (1932) or semantic differential (Osgood 1952) format. The Likert dominates academic research in marketing (Bruner 1993), but practitioners prefer various forms of semantic differential (Haley and Case 1979). Moreover, Likert scales can use all positive items or a mixture of positive and negative items.

Response Effects

Responses to itemized rating scales reflect sources other than the intended construct due to a tendency for individuals to respond systematically regardless of content (Tourangeau, Rips and Rasinski 2000). Sources of systematic error, such as acquiescence bias (Couch and

Keniston 1960, Bachman and O'Malley 1984), extremity bias (Hamilton 1968, Shulman 1973, Greenleaf 1992b) and halo (Cooper 1981, Wirtz 2003), can positively bias classical test theory based reliability and validity criteria that correlate response data over respondents. Paradoxically, the more a rating scale induces bias, the better it can appear to perform (Weijters, Schillewaert and Geuens 2007).

Several authors have recommended procedures to reduce response effect bias (e.g., Martin 1964, Wyer 1971, Greenleaf 1992a). Baumgartner and Steenkamp (2001) recommend using balanced scales, eliminating items susceptible to social desirability, and collecting extra data to compute response style indices to obtain residualized substantive scores. However, they address the scaling of respondents, not the scaling of marketing stimuli. Normalization, centering, and standardization can be viewed as remedies for response bias (see Dillon, Frederick and Tangpanichdee 1985). For example, within-respondent standardization can control for both acquiescence and extremity bias (Schimmack, Oishi and Diener 2005), but it eliminates any valid variation due to respondents. G theory allows one to see the extent of variance due to each source.

Generalizability Theory

If respondents r rate marketing stimuli s on items i measuring a construct using a common response format, the observed responses X_{rsi} can be expressed as a linear model of deviations from the main and interaction effect means for the three sources of variation. Following Brennan (2001a) this can be represented compactly as:

(1) $$X_{rsi} = \mu + \nu_r + \nu_s + \nu_i + \nu_{rs} + \nu_{ri} + \nu_{si} + \nu_{rsi,e}$$

where μ is the grand mean for the sampled universe and each υ is an uncorrelated effect, representing a deviation score for its subscripted source of variation (e.g., $\nu_s = \mu_s - \mu$**).** The total observed variance can be decomposed into seven independent variance components:

(2) $$\sigma^2(X_{rsi}) = \sigma^2_r + \sigma^2_s + \sigma^2_i + \sigma^2_{rs} + \sigma^2_{ri} + \sigma^2_{si} + \sigma^2_{rsi,e}$$

Each variance component quantifies the extent to which scores differ, after averaging over other sources. For example, (σ^2_s) quantifies how much responses for stimuli differ after averaging over respondents and items. The G-coefficient for stimuli provides an indicator of the overall adequacy of a relative scaling of stimuli:

(3) $$(\sigma^2_s) / [(\sigma^2_s) + (\sigma^2_{rel})]$$

where for a crossed design with n_r respondents responding to n_i items for each stimulus

(4) $$(\sigma^2_{rel}) = (\sigma^2_{rs}) / n_r + (\sigma^2_{is}) / n_i + (\sigma^2_{rsi,e}) / n_r n_i$$

If categorical G-study data are analyzed as if equal interval, the size of the estimable variance components from Equation 2 will vary with the number of response categories. However, the G-coefficient is determined by the relative sizes of variance components, and so is independent of response format as long as the data satisfy the model assumption of no bias. Bias cannot be completely eliminated when collecting real data. Nor can its contribution to observed variance components be estimated from a single data set. However, when variance components for the same universe of generalization are estimated using a different number of response categories, deviations from invariance can be attributed to the different number of categories. Hence, differences in G-coefficients will indicate the relative quality of the data provided by the number of response categories.

Empirical Study

This study compared five category responses with seven category responses over three item formats, namely semantic differential, all positive Likert and balanced Likert. As shown in Table 1, each was used for a different website scale. The first, referred to as WebSQ+, used 27 website quality items and semantic ratings. The items originated in an adaptation of SERVQUAL to measure website service quality, along lines proposed by Zeithaml, Parasuraman and Malhotra (2000). The second, referred to as SiteQ+, used a balanced Likert format, with two positive and two negative items for each of 11 quality aspects identified by Finn and Kayande (2003). The third, consisted of items required to model brand equity as a signaling phenomenon (Erdem and Swait 1998), used all positive Likert items.

Table 1 Design used for Study

Sample	Scale (Items)	Response categories and labels	Wording
One	WebSQ+ (27)	1-5 Terrible to Excellent Semantic differential	Purely positive
One	SiteQ* (44)	1-5 Strongly agree - strongly disagree Likert	Balanced
One	Br. equity (30)	1-5 Strongly agree - strongly disagree Likert	Purely positive
Two	WebSQ+ (27)	1-7 Terrible to Excellent Semantic differential	Purely positive
Two	SiteQ* (44)	1-7 Strongly agree - strongly disagree Likert	Balanced
Two	Br. equity (30)	1-7 Strongly agree - strongly disagree Likert	Purely positive

Data Collection

In fall 2003, a stratified sample of 20 online retailers was selected. A subsample of 12 small retailers was randomly sampled from a list of 1588 Canadian retail websites. The other eight were randomly sampled from 20 well-known Canadian retail websites. Website raters were respondents to a job posting for part-time research assistants at a major Canadian University. Forty raters, who were paid for providing data for the project, were randomly assigned to the two versions of the data collection instrument. Data were collected at a rate of one website a day for 20 weekdays.

Generalizability Analysis

For each set of data, all 20 raters (r) evaluated all 20 websites (w) on all aspects (a), so raters were crossed with websites and aspects. However items were nested within aspects (i[a]). The negatively worded Likert items were reverse-coded for analysis. Variance components were estimated using urGENOVA (Brennan 2001b). Estimated total variance and the percentage due to each source are shown in Table 2. The percentages can be interpreted as the share of variance to be expected when a single rater evaluates one aspect of one website on a single item. Results for all six sets of data defined in Table 1 are shown in Table 2. The percent of variance due to websites was consistently larger for the seven-category than for the five-category responses.

Table 2 Variance Component Estimates by Scale and Response Categories

Source of variation	WebSQ+ Semantic rating		Brand Equity Likert positive		SiteQ* Likert balanced	
	5-Point	7-Point	5-Point	7-Point	5-Point	7-Point
Total variance	1.055	2.603	.576	1.702	.770	2.494
Percent of variance						
Websites (w)	20.0	21.3	17.7	23.6	10.8	16.9
Raters (r)	6.8	6.9	11.2	9.7	8.3	6.4
Aspects (a)	0	0	0	0	2.5	1.9
Items[aspects] (i[a])	10.4	6.8	2.9	1.4	1.8	1.6
wr	14.7	15.0	25.8	28.8	21.2	22.2
wa	3.9	4.3	1.2	1.6	2.3	2.2
wi[a]	3.2	4.3	1.9	2.2	1.1	1.2
ra	1.8	1.7	0.6	0.7	2.9	2.0
ri[a]	4.9	4.8	3.7	3.1	7.1	6.0
wra	7.9	7.5	3.0	3.4	10.0	8.6
wri[a],error (residual)	27.7	28.0	32.3	25.7	32.1	30.8
Total	100	100	100	100	100	100

Response Category Effects on Website Data Quality

From a managerial perspective, interest in data quality focuses on the expected G-coefficients for websites. Table 3 uses the variance components to estimate the G-coefficients expected for three specific D-study designs that could be used for comparing stimuli. The first is for one rater, one aspect and one item. While not of substantive interest, it provides the simplest direct comparison between the numbers of response categories. The second design is for one rater, using a twenty-item scale consisting of four items for each of five aspects. This is the sort of design used when researchers report traditional reliability measures, such as internal consistency. The third design is for ten raters, using the same twenty-item scale. This is the sort of design that might be necessary for managerial decision-making, yet organizations might still be prepared to fund, given that a fully crossed design requires each rater to visit each website. The better performing number in each comparison is shown in bold.

For the scaling of websites, the G-coefficients for the seven-category responses are always better than the G-coefficients for the five-category responses. As a aside, both all-positive formats, namely the semantic rating for WebSQ+ and the positive Likert for brand equity, always out performed the balanced Likert for SiteQ*. Table 3 also shows quite different results when scaling raters, as the five category responses out performed the seven category responses for both Likert scales, but there is little difference for the semantic differential.

Table 3 Scale Performance Criteria by Scale and Response Categories

Scale performance criteria	WebSQ+ Semantic rating		Brand Equity Likert positive		SiteQ* Likert balanced	
D-study G-coefficients for:	5-Point	7-Point	5-Point	7-Point	5-Point	7-Point
Websites						
1 rater, 1 aspect, 1 item	.258	**.264**	.217	**.276**	.139	**.206**
1 rater, 5 aspects, 4 items	.517	**.529**	.386	**.570**	.299	**.394**
10 raters, 5 aspects, 4 items	.879	**.883**	.843	**.870**	.783	**.849**
Raters						
1 website, 1 aspect, 1 item	.107	.107	**.148**	.136	**.102**	.084
1 website, 5 aspects, 4 items	.271	.271	**.286**	.238	**.244**	.197
10 websites, 5 aspects, 4 items	.735	**.742**	**.783**	.743	**.711**	.664

Discussion

In marketing multi-item responses are not only used to scale respondents, they are also used to scale marketing stimuli, making G theory performance criteria more appropriate than those based on classical test theory. G-coefficients show seven-category ratings perform better than five-category ratings when scaling marketing stimuli. Moreover, they are better for semantic ratings, balanced Likert and all positive Likert items. The same is not true when scaling respondents, as five-categories outperform seven-categories for both Likert formats.

The likely reason for these conflicting conclusions for different objects of measurement is response effects that impact the size of particular variance components observed in marketing studies. For example, response calibration occurs when a respondent's initial response serves as an anchor or reference standard for subsequent responses (Higgins and Lurie 1983, Trabasso 1982). In stimulus evaluation studies, calibration will be more evident with more response categories. It will manifest itself as a disproportionate increase in the variance component for respondents. Thus, G-theory variance components are likely to be more informative than traditional classical test theory based criteria such as coefficient alpha.

This research was limited by the modest samples of 20 websites and 20 raters used for each condition. Larger samples would provide more precise variance component estimates, but for a crossed G-study design, cost is a function of the product of the sample sizes. Second, our universe of generalization consists of research assistants and Canadian retailer websites. While we see no reason to expect the findings to be context specific, more research is needed to see if these results generalize to the broader universe of conditions (e.g., using interviewers, less educated respondents) within which marketers use multi-item scales to scale stimuli.

Conclusion

This research using G theory provides fresh insight into the question of the optimal number of response categories. The number that is best when scaling respondents is not necessarily best for other purposes. In marketing applications, where the purpose of measurement is to differentiate between marketing entities, such as websites, what is of most concern is the G-coefficient for the marketing entities. This research shows seven-category responses are better than five-category responses for both Likert and semantic differential item formats.

References

Alwin, Duane F. (1997), "Feeling Thermometers vs 7-point Scales," *Sociological Methods and Research*, 25 (3), 318-351.

Bachman, Jerald G. and Patrick M. O' Malley (1984), "Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles", *Public Opinion Quarterly*, 48 (Summer), 491-509.

Baumgartner, H. and J-B. E. M. Steenkamp (2001), "Response Styles in Marketing Research: A Cross-National Investigation", *Journal of Marketing Research*, 38 (May), 143-156.

Bendig, A. W. (1953), "The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on a Scale," *Journal of Applied Psychology*, 37 (February), 38-41.

Boote, A. S. (1981), "Reliability Testing of Psychographic Scales: Five-point or Seven-Point? Anchored or Labeled?" *Journal of Advertising Research*, 21, 53-60.

Brennan, Robert L. (2001a), *Generalizability Theory,* New York: Springer-Verlag.

Brennan, Robert L. (2001b), *Manual for urGENOVA Version 2.1*, Iowa Testing Programs Occasional Papers, Number 49.

Bruner, Gordon C. (1993), *A Census of Multi-Item Scales Used in Marketing Research*. Office of Scale Research Technical Report 9305.

Churchill, G. A., Jr., and J. P. Peter (1984), "Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis," *Journal of Marketing Research*, 21 (November), 360-375.

Cooper. W. H. (1981), "Ubiquitous Halo" *Psychological Bulletin*, 90 (2), 218-244.

Couch, A. S. and K. Keniston (1960), "Yeasayers and Naysayers: Agreeing Response Set as a Personality Variable", *Journal of Abnormal and Social Psychology*, 60 (March), 151-174.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam (1972), *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley & Sons.

Dillon, William R., D.G. Frederick and V. Tangpanichdee (1985), "Decision Issues in Building Perceptual Product Spaces with Multi-attribute Rating Data," *Journal of Consumer Research*, 12 (June), 47-63.

Erdem, Tulin and Joffrey Swait (1998), "Brand Equity as a Signaling Phenomenon," *Journal of Consumer Psychology*, 7 (2), 131-157.

Finn, Adam and Ujwal Kayande (2003), *A Parsimonious Website Interaction Scale.* CIRAS Paper. Available at www.ciras.com/research/wrkpaperseries.htm

Finn, Adam and Ujwal Kayandé (1997), "Reliability Assessment and Optimization of Marketing Measurement," *Journal of Marketing Research*, 34 (May), 262-275.

Garner, W. R. (1960), "Rating Scales, Discriminability and Information Transmission." *Psychological Review*, 67, 343-352.

Greenleaf, E. A. (1992a), "Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles," *Journal of Marketing Research*, 56, (May), 176-188.

Greenleaf, Eric A. (1992b), "Measuring Extreme Response Style," *Public Opinion Quarterly*, 56, (Autumn), 328-351.

Green, Paul E. and Vithala R. Rao (1970), "Rating Scales and Information Recovery – How Many Scales and Response Categories to Use?" *Journal of Marketing*, 84 (July), 88-89.

Haley, Russell I. and Peter B. Case (1979), "Testing Thirteen Attitude Scales for Agreement and Brand Discrimination," *Journal of Marketing*, 43 (Fall), 20-32.

Hamilton, David L. (1968), "Personality Attributes Associated with Extreme Response Style", *Psychological Bulletin*, 69 (March), 192-203.

Higgins, E. Tory and Liora Lurie (1983), "Context, Categorization, and Recall: The 'Change-of-Standard' Effect," *Cognitive Psychology* 15 (October), 525-547.

Jacoby, Jacob and Michael S. Matell (1971), "Three-Point Likert Scales Are Good Enough", *Journal of Marketing Research*, 8 (November), 495-500.

Likert, R (1932), "A Technique for the Measurement of Attitudes," *Archives of Psychology*, No. 140.

Martin, John (1964), "Acquiescence – Measurement and Theory", *British Journal of Social and Clinical Psychology*, 3 (October), 216-225.

Osgood, C. E. (1952), "The Nature of Meaning and Measurement of Meaning," *Psychological Bulletin*, 49 (August), 197-213.

Peter, J. Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices", *Journal of Marketing Research*, 16 (February), 6-17.

Peterson, Robert A. (1994), "A Meta-analysis of Cronbach's Coefficient Alpha", *Journal of Consumer Research*, 21 (September), 381-391.

Rentz, Joseph O. (1987), "Generalizability Theory: A Comprehensive Method for Assessing and Improving the Dependability of Marketing Measures," *Journal of Marketing Research*, 24 (February), 19-28.

Rossiter, John R. (2002). The C-OAR-SE Procedure for Scale Development in Marketing. *International Journal of Research in Marketing*, 19 (December), 305-335.

Schimmack, Ulrich, Shigehiro Oishi and Ed Diener (2005), "Individualism: A Valid and Important Dimension of Cultural Differences Between Nations," *Personality and Social Psychology Review*, 9 (1), 17-31.

Shulman, Art (1973), "A Comparison of Two Scales on Extremity Response Bias", *Public Opinion Quarterly*, 37 (Fall), 407-412.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski (2000), *The Psychology of Survey Response*. Cambridge University Press.

Trabasso, Tom (1982), "The Importance of Context in Understanding Discourse", *Question Framing and Response Consistency*, Ed. R. M. Hogarth, San Francisco: Jossey-Bass, 77-89.

Viswanathan, M., S. Sudman and M. Johnson (2004), "Maximum Versus Meaningful Discrimination in Scale Response: Implications for Validity of Representation of Consumer Perceptions about Products," *Journal of Business Research*, 57 (February), 108-124.

Weathers, D., S. Sharma and R. W. Niedrich (2005), "The Impact of the Number of Scale Points, Dispositional Factors, and the Status Quo Decision Heuristic on Scale Reliability and Response Accuracy," *Journal of Business Research*, 58 (November), 1516-1524.

Weijters, B., N. Schillewaert and M. Geuens (2008), "Assessing Response Styles Across Modes of Data Collection," *Journal of the Academy of Marketing Science*, 36 (Fall), 409-422.

Weng, Li-Jen (2004), "Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability," *Educational and Psychological Measurement*, 64 (December), 956-972.

Wirtz, Jochen (2003), "Halo in Customer Satisfaction Measures: The Role of Purpose of Rating, Number of Attributes and Involvement," *International Journal of Service Industry Management*, 14 (1) 96-119.

Wyer, Robert S., Jr. (1971), "The Effects of General Response Style on Measurement of Own Attitude and the Interpretation of Attitude-Relevant Messages", *British Journal of Social and Clinical Psychology*, 8(2). 105-115.

Zeithaml, Valarie A., A Parasuraman and Arvind Malhotra (2000), "A Conceptual Framework for Understanding e-Service Quality: Implications for Future Research and Managerial Practice". MSI Working Paper Report No. 00-115.