Original Article

An investigation of the relationship between the number of response categories and scale sensitivity

Received (in revised form): 23rd February 2011

Sunil H. Contractor

is an assistant professor at Towson University, and has earned a PhD in Marketing from University of Georgia. Sunil has over 10 years of R&D experience in industry. He is inventor of 34 patents. Sunil has worked on over 20 new product introduction projects. Sunil teaches Principles of Marketing, Marketing Research and Advanced Business Statistics courses.

Richard J. Fox

is an associate professor at Terry College of Business, University of Georgia, and has earned a PhD in Mathematics from Michigan State University. He has over 14 years of Marketing Research experience in industry. He teaches undergraduate, graduate and PhD level courses in Marketing Research. He has served as a director of Coca-Cola Center for Marketing Studies at UGA and consulted with Fortune 500 companies.

ABSTRACT Category rating scales are frequently used in survey research by researchers and practitioners to measure attitudes or reactions to a stimulus. One important scale design issue is whether the number of response categories included in the category rating scale affects scale sensitivity (that is, the ability to statistically separate mean ratings for different stimuli). The results of two experiments to evaluate and compare the sensitivities of commonly used category rating scales, differing in number of response categories, are reported in this article. The results suggest that increasing the number of response categories within the range of commonly used itemized rating scales does not increase sensitivity, and scales with 5 or 6 categories of response may in fact be more sensitive than scales with 7, 9 or 10 response categories.

Journal of Targeting, Measurement and Analysis for Marketing (2011) 19, 23–35. doi:10.1057/jt.2011.4;
published online 25 April 2011

Keywords: scale sensitivity; scale design; measurement; category rating scale

INTRODUCTION

Measurement is defined as the assignment of numbers to objects according to rules.^{1,2} Measurement plays a vital role in social science research where the object of the measurement can vary from an abstract construct, such as the perceived importance of a service attribute, to something more tangible such as the acceptability

of the taste of new soft drink. Researchers commonly use an itemized rating scale, or category rating scale, to facilitate measurement. These scales include a finite set of distinct response categories, from which a respondent is asked to select the category that best represents his or her opinion, attitude or evaluation. Among the many issues considered when designing an itemized rating scale is the fundamental question regarding the number of response categories to include. This question has been addressed by researchers primarily in the context of scale reliability and validity. However, the impact of

---

Correspondence: Sunil H. Contractor
Towson University, 123 Stephen Hall, 8000 York Road, Towson,
Maryland 21252, USA
E-mail: SContractor@towson.edu

the number of response categories on scale sensitivity, that is the ability to statistically detect differences among mean ratings of objects or constructs being measured, has not been addressed, and this is the focus of the research reported in this article. The results of this research dismiss the notion that 'more is better' when it comes to the relationships between the number of response categories and scale sensitivity. In fact, the results suggest that 5 or 6 response categories produce more sensitive scales than scales with 9 or 10 response categories.

Itemized rating scales play prominent roles in many social science research studies, so it is not surprising that much research has focused on how various scale design issues relate to scale utility.³⁻⁷ Researchers have conducted empirical studies to investigate scale design issues such as the inclusion of a scale midpoint,^{6,8} using labeled versus unlabeled response categories,^{9,10} and so forth. Considerable extant research dating back to the works of Bendig¹¹ and Symonds¹² has addressed the fundamental question of how many response categories to include in a scale. However, these investigations have addressed this question in the contexts of scale reliability and validity.^{6,13} Adoption of a valid rating scale ensures that the measurement of objects or constructs reflects the focal characteristics being measured. Scale reliability refers to consistency of results when independent, but comparable, measures of the same object or construct are obtained.¹⁴ Certainly, these are desirable properties for a rating scale to possess, and an understanding of how validity and reliability are affected by the number of response categories is important when designing a scale. Scale sensitivity is also an important design consideration, and the issue of how sensitivity is affected by the number of response categories deserves investigation.¹⁵

BACKGROUND

Mean scale ratings are often compared to determine whether significant differences exist when analyzing survey or experimental research results; and important decisions can be based on the conclusions. For example, a probability sample of consumers might be asked to rate the

appeal of several new product concepts using an itemized rating scale, and the concepts (if any) whose means are rated significantly higher than the others are designated for further study. The choice of itemized rating scale affects the ability of the researcher to detect and identify differences in such studies, underscoring the importance of scale sensitivity in the scale choice decision.

In statistical terms, sensitivity refers to statistical power, which, in the context of comparing sample means, is the probability that the difference between two sample means is statistically significant (at a specified risk level), given that the corresponding population 'true' means differ by a specified amount. For a given level of significance and sample sizes, the power curve associated with a specific rating scale indicates how the probability of rejecting the null hypothesis of equality of two means increases as the magnitude of the true difference increases. Statistical power is directly proportional to the level of significance (α risk) of the statistical test, the magnitude of the difference between the population means on the measurement scale and the sample sizes.

The variance of measurements across a population also changes with the scale. This variance, often referred to as measurement error, is inversely related to statistical power. Measurement error has two components. One component is variance associated with multiple measurements of the same entity made on the same unit (within subject variance in experimental design terminology). This variance is the focus of test-retest reliability.¹⁶ Poor test-retest reliability means that this component of measurement error is relatively large. The second component represents the variance across individuals (between subjects variance) in the construct or object being measured, for example variance among survey respondents in the perception of the importance of nutritional value in the selection of a snack food.

In the case of comparing two means from independent samples of the same size, if the ratio of the absolute difference between actual scales means to the standard deviation (square root of the variance of the individual measurement) is the same

for two scales, then the power is the same for the two scales. A general method for translating the magnitude of the difference between two stimuli or the measurement standard deviation on one scale to a corresponding magnitude of difference or measurement standard deviation on another scale is unavailable because of the nuances introduced by the type of stimuli and specific scales. For this reason, we use empirical evidence rather than an analytical argument based on power curves to investigate how the number of response categories affects scale sensitivity.

There are two conflicting views regarding the relationship between the number of categories in an itemized rating scale and scale sensitivity.¹⁷ On the one hand, increasing the number of scale categories arguably enhances the precision of a scale increasing the separation of the true scale means of the stimuli and therefore increases the statistical power or sensitivity.¹⁸ This assertion is based on the information theoretic approach which posits that higher number of categories in a scale facilitates greater utilization of response categories translating into higher information transmission by the scale.¹⁹ On the other hand, Miller²⁰ suggests that respondents may not be able to make the fine mental distinctions that are demanded by scales having many response categories, which increases within subject variance (decreases test-retest reliability) and hence may increase at least one of the components of measurement variance. This in turn has a negative impact on statistical power or

sensitivity. Therefore, Komorita and Graham²¹ claim that increasing the number of scale categories beyond a certain point may negate the anticipated increase in precision, and actually decrease sensitivity.

The primary objective of this article is to test the two conflicting views regarding the relationship between the number of response categories and scale sensitivity. The results of two studies, in which itemized rating scales that differ in number of response categories are compared in the terms of their ability to statistically 'separate' a common set of stimuli, are reported. As already noted, examining whether means obtained using itemized rating scales are statistically significantly different is often a crucial component of the analysis of the results of a survey or an experiment. However, the extant research in the area of scale sensitivity is quite limited.

In Study 1, the scales shown in Table 1, ranging in number of categories from 5 to 10, were compared in terms of their ability to statistically 'separate' stimuli in three different product categories – writing instruments, orange juice and facial tissue. Experiments of this nature are quite common in marketing research when seeking to screen the 'best' alternative from a diverse set of product options. Specifically, within the three product categories, average ratings of three stimuli were statistically compared, separately for each scale. Scales with 5-10 categories were used in this study because the extant research suggests that 7 ± 2 categories is optimal from the

Table 1: Overall rating scales (Study 1)

5-pt. category	6-pt. category	9-pt. category	10-pt. category scale
5 – Excellent	6 – Excellent	9 – Like It Extremely	Only endpoints labeled
4 – Very Good	5 – Extremely Good	8 – Like It Very Much	10 – Excellent
3 – Good	4 – Very Good	7 – Like It Moderately	9
2 – Fair	3 – Good	6 – Like It Somewhat	8
1 – Poor	2 – Fair	5 – Neither Like/Dislike	7
	1 – Poor	4 – Dislike It Somewhat	6
		3 – Dislike it Moderately	5
		2 – Dislike It Very Much	4
		1 – Dislike It Extremely	3
			2
			1 – Poor

*The 5, 6 and 9-pt. scales are all labeled, and thus can be compared in the context of sensitivity without confounding number of categories with the issue of labeled versus non-labeled categories. The results for the 10-pt. scale are included in the tables for the sake of completeness, and the conclusions are the same with or without the results of this scale.

reliability and validity point of view,¹⁸ and such scales are most commonly used in research and practice. The frequently used 10-pt. scale used in this study is different in nature from the other three scales, which all have labeled categories of response. Comparisons involving the 10-pt. scale must be viewed keeping this in mind. However, the results for this scale are of interest and their inclusion with this caveat is benign, as attention can be focused solely on comparisons involving only the three labeled scales (5, 6 and 9 categories) without the caveat. Of course, there must be some difference in wording between two labeled rating scales having differing numbers of response categories because differing numbers of labels are required. Further, the 9-pt. scale is bipolar, while the other three scales are not. It can be argued that these differences among scales may contribute to some degree to any observed differences among the sensitivities of the scales. For this reason, a second experiment, Study 2, was conducted. Rating scales (5, 6, 7, 9 and 10 categories) having only end point labeled (same pair of labels for all scales) are included eliminating any concerns about confounding due to varying scale type, polarity or wordings of categories. The results from both studies are consistent, and refute the notion that increasing the number of scale categories, within the range of 5–10 categories, increases scale sensitivity.

In a study comparing scale sensitivities, the magnitudes of the differences between stimuli being rated as well as sample size are important considerations. Obviously, if the stimuli are basically all the same, then no differences among stimuli should be detected with any scale. On the other hand, if the stimuli are very different, then all scales are likely to detect significant differences in all pairs of stimuli even for relatively small sample sizes. Keeping these issues in mind, the studies conducted in this research involved ‘reasonable’ (typical in marketing research applications) sample sizes and stimuli that were found to be noticeably different, but not so obviously different as to make comparison trivial, in pilot testing. The selection of stimuli for each study is discussed in the respective study design descriptions.

Finally, another consideration in rating scale design is whether to include a scale midpoint. The designs of the experiments reported in Studies 1 and 2 also allow the investigation of the impact of inclusion of a scale midpoint on sensitivity on a limited basis. At the end of the discussion of each study, we briefly report the findings of the studies in the context of this scale design issue.

STUDY 1

Design

Consumers in six geographically distributed shopping malls were recruited to participate in the study, which was conducted on site at the respective marketing research facilities. The sample was approximately half male and half female, and equally distributed across the three age groups: 18–30, 31–50 and 51–65. Four category rating scales, containing 5, 6, 9 and 10 response categories (Table 1), were included in the study, and compared in terms of their ability to separate stimuli within product categories. Each participant was interviewed by a trained member of the local marketing research staff and was asked to experience and rate nine stimuli, three from each of three product categories, orange juice, writing instruments and facial tissue, using one of the four scales. A show card containing the category rating scale was presented to the respondent by the interviewer at the beginning of the evaluation process, and the respondent used the scale to rate each stimulus immediately after experiencing it.²² When the respondent had completed rating all nine stimuli, the interviewer removed the products from the room, and returned with a ‘new’ set of products (actually the same set of stimuli), and a different show card containing a different rating scale. Respondents were asked to rate the new stimuli (in the same order) using the second rating scale.

Within each product category, the three stimuli were chosen from commonly available brands to represent a broad spectrum of quality. One premium brand (P), one high quality brand (Q) and one low quality economy brand (E) were selected so that significant differences among overall

average ratings could be expected. Participants tasted the orange juices, signed their names with the writing instruments and simulated wiping their noses to form opinions of the overall quality of the stimuli in each product category.

A pilot test was conducted among about 20 graduate students to determine which product categories to use in the research. Use of both orange juice and potato chips, along with facial tissues, resulted in a tedious respondent task involving too many taste evaluations and palate cleansings. As a result, writing instruments were chosen to replace potato chips. A second pilot test was then conducted using 'real' respondents and the mall intercept setting. Changes in the questionnaire and the research protocol were made on the basis of this pilot test to insure that the length of the interview would not exceed 15 min. Also, the three stimuli levels for each product category were validated as producing differences in perceived quality.

The products were presented unidentified (unbranded), and evaluations were obtained sequentially for all three stimuli for one product category before progressing to the next. Two sequences of presentation of the product categories, orange juice-facial tissue-writing instruments ('SEQ' = 1), and writing instruments, facial tissue, orange juice ('SEQ' = 2) were used in the study. The order of presentation of the three stimuli within each product category was rotated across respondents as is typical in practice.

The research design called for 18 respondents to be randomly assigned to each of the 24 groups, 6 pairs of scales (formed from the four scales studied) $\times$ 2 orders of scale presentation (for example, 5-pt. followed by 6-pt. or vice versa) $\times$ 2 product category sequences, (SEQ = 1 or 2). This design would produce a total sample of 432 respondents, each stimulus being evaluated on each scale by a set of approximately 216 respondents (3 pairs containing the particular scale $\times$ 2 orders of scale presentation $\times$ 2 product category sequences $\times$ 18 respondents per group). The data collection, including a request for slight over-sampling to account for unusable interviews, was evenly divided across the six mall locations. A total of 434 usable interviews were obtained

yielding a sample size for each combination of stimulus and scale ranging from minimum of 207 to a maximum of 220. Within each product category and scale, the ratings were analyzed for differences in the average overall mean ratings of the three corresponding stimuli. For each set of stimuli, scales were compared in terms of sensitivity, that is ability to produce statistically significant differences in the mean scores for the three stimuli.

Analysis and results

In order to simplify the analysis, the first step was to determine whether rating results using a particular scale were affected by the other scale used, the position of usage of the particular scale in question (first versus second) and the product category sequence (SEQ = 1 or SEQ = 2). The lack of design effects on scale rating results would permit aggregating results for each scale across design conditions before comparing scales in terms of their respective abilities to separate stimuli. A summary measure of a respondent's ratings using one scale was calculated by averaging the nine ratings, three stimuli in each of three product categories, provided by the respondent using the scale. This average was used as the dependent variable in this analysis because there is no reason for any design effect, that is other scale, position of usage or product category sequence, to be stimulus specific. An analysis of variance (ANOVA) model was applied separately to the individual respondent average ratings obtained for each scale. In each case, the ANOVA model included the main effects of position (first or second), product category sequence (1 or 2) and other scale used (three levels defined by the other three possible scales), as well as second- and third-order interactions. If the overall *F*-test associated with the ANOVA was not statistically significant ($\alpha = 0.05$), then the rating results for the corresponding scale did not appear to be affected by experimental design conditions, and results could be aggregated across these conditions for each of the three product categories.

The overall ANOVA *F*-test was significant only in the case of the 6-pt. scale where a

significant main effect for 'Other' (other scale used) is indicated. Examination of the tables of means, corresponding to the various experimental conditions, shows that this effect stems from the ratings being slightly low when the 6-pt. scale was administered after the 5-pt. scale, and when orange juice was the first product category presented ($SEQ = 1$). Further investigation revealed that this small effect was generally consistent across the nine stimuli, thus having little or no impact on comparisons of individual stimuli mean ratings within product category, obtained with the 6-pt. scale. Hence, rating results for the stimuli in each product category were aggregated across experimental conditions for the 6-pt. scale as well as the other three scales.

To provide additional support for aggregating results, the analyses of stimuli ratings were repeated for each product category and scale using only results obtained when the scales were administered first. Position of usage and other scale used are no longer factors to be considered. In particular, the data corresponding to the condition for which the 6-pt. scale averages across stimuli were slightly low are no longer included in the analyses. The results are essentially the same as those obtained using the aggregated data, but with less statistically significant differences among stimuli average ratings as one would expect due to the reduced sample size.

For each of the scales studied and each of three product categories, an ANOVA model was applied with the respondent treated as a random effect.²³ In each case, an F -test of the hypothesis of equality of mean stimuli ratings was conducted. Table 2 contains the respective F -test results (P -values), the stimuli means along with the ranges of the three stimuli means, the LSD 95 per cent (least significant difference 95 per cent, for example Siegel²⁴), and the estimated effect size, Eta^2 , for each case. If the P -value of the associated F -test is less than 0.05, then the means are significantly different at the 5 per cent risk level. The LSD 95 per cent 'yardstick' is used for pairwise comparison of the stimuli means. Table 2 also contains the index

formed by the range of stimuli means divided by the LSD 95 per cent as another indicator of the effect size; the larger the ratio, the larger the effect.

For the orange juice category, the economy brand was consistently rated lower on average than the other two brands. However, the F -test associated with testing the hypothesis of equality of stimuli means is not significant ($\alpha = 0.05$) for any of the scales. The estimates of effect size, Eta^2 , are all very small.

In the case of facial tissue, the F -tests are consistently significant ($P < 0.0001$), and the economy brand is rated significantly lower than the other two brands across all scales. As expected, the mean rating for the premium facial tissue is higher than the mean rating for the quality facial tissue for all scales, but the difference exceeds the LSD 95 per cent only for the 5-pt. scale results. The estimated effect size, Eta^2 , is the same for all four scales.

In the case of writing instruments, the F -tests are again consistently significant ($P < 0.0001$) for all scales, and the economy writing instrument is rated significantly lower than the other two writing instruments on all scales. Although the premium product is rated higher than the quality product for all scales, the difference exceeds the LSD 95 per cent only for the 5- and 6-pt. scale results. The estimated effect sizes for the four scales are again comparable, ranging from 0.07 to 0.12.

The ratios of the range in means to the LSD 95 per cent are also shown for each scale and each product category in Table 2. Based on this measure of sensitivity, it is again evident that the scales perform quite comparably in terms of separating stimuli within product categories. In particular, the 9-pt. scale (or the 10-pt. scale) does not exhibit a consistent advantage over the two scales with fewer response categories.

After completing both ratings tasks the participants were asked which of the two scales he or she preferred.²⁵ Table 3 contains the results of this preference question for all six pairs of scales among participants stating a preference. The 6-pt. scale is the more preferred scale in each of its head-to-head comparisons, and its overall winning percentage (63.7 per cent) is considerably higher

Table 2: Rating results for all scales and product categories (Study 1)

Stimulus category	Stimulus sample and mean				ANOVA results and effect size					
	Scale pt.	Stimulus level	Sample size (n)	Mean	Comparison of means ^a	F-test (P-value)	LSD 95%	Range	Range/LSD	Eta ^{2b}
Orange juice	5	Premium	219	3.24	P>E	0.13	0.24	0.25	1.03	0.01
		Quality	219	3.13		—	—	—	—	—
		Economy	219	2.99		—	—	—	—	—
	6	Premium	214	3.77	P>E	0.06	0.27	0.33	1.19	0.01
		Quality	214	3.58		—	—	—	—	—
		Economy	214	3.44		—	—	—	—	—
	9	Premium	216	6.34		0.23	0.42	0.37	0.87	0.00
		Quality	216	6.13		—	—	—	—	—
		Economy	216	5.97		—	—	—	—	—
	10	Premium	207	6.42	Q>E	0.06	0.49	0.59	1.21	0.01
		Quality	207	6.68		—	—	—	—	—
		Economy	207	6.09		—	—	—	—	—
Facial tissue paper	5	Premium	220	3.58	P>Q>E	<0.0001	0.23	0.90	3.91	0.09
		Quality	220	3.25		—	—	—	—	—
		Economy	220	2.68		—	—	—	—	—
	6	Premium	213	4.08	P, Q>E	<0.0001	0.28	1.04	3.75	0.09
		Quality	213	3.88		—	—	—	—	—
		Economy	213	3.05		—	—	—	—	—
	9	Premium	219	6.87	P, Q>E	<0.0001	0.39	1.37	3.49	0.09
		Quality	219	6.77		—	—	—	—	—
		Economy	219	5.50		—	—	—	—	—
	10	Premium	210	7.50	P, Q>E	<0.0001	0.47	1.82	3.84	0.09
		Quality	210	7.04		—	—	—	—	—
		Economy	210	5.69		—	—	—	—	—
Writing instruments	5	Premium	221	3.61	P>Q>E	<0.0001	0.24	0.98	4.06	0.09
		Quality	221	3.22		—	—	—	—	—
		Economy	221	2.63		—	—	—	—	—
	6	Premium	216	4.07	P>Q>E	<0.0001	0.30	1.10	3.66	0.08
		Quality	216	3.73		—	—	—	—	—
		Economy	216	2.97		—	—	—	—	—
	9	Premium	219	6.90	P, Q>E	<0.0001	0.43	1.86	4.29	0.12
		Quality	219	6.60		—	—	—	—	—
		Economy	219	5.04		—	—	—	—	—
	10	Premium	210	7.21	P, Q>E	<0.0001	0.51	1.65	3.25	0.07
		Quality	210	6.84		—	—	—	—	—
		Economy	210	5.56		—	—	—	—	—

^aComparison based on LSD 95 per cent; Parentheses indicate that corresponding overall F-test is not significant, so that the noted difference exceeding the LSD 95 per cent is not statistically significant.

^bEta² is the per cent of total variance in the dependent variable accounted for by the variance between groups formed by the independent variable. We used formula provided by Olejnik and Algina²⁷ for computing eta-square statistics in the context of this type of experiment.

than the overall winning percentages of the other three scales. The 5-pt. scale had the second highest overall winning percentage (51.7 per cent) – about 10 percentage points higher than the 9- and 10-pt.

scales which had the same overall ‘winning’ percentages (41.5 per cent).

In practice respondent time is equivalent to money, so the length of time required to conduct

Table 3: Scale preference results^a (Study 1)

	Overall ^b	5–6	5–9	5–10	6–9	6–10	9–10
		(62)	(71)	(68)	(69)	(62)	(65)
	%	%	%	%	%	%	%
5-pt.	51.7	32.3	49.3	72.1***	—	—	—
6-pt.	63.7	67.7***	—	—	68.1***	54.8	—
9-pt.	41.5	—	50.7	—	31.9	—	41.5
10-pt.	41.5	—	—	27.9	—	45.2	58.5

^aPercentages based on those participants stating preference.

^bCalculated by combining results from three pairs involving each scale.

***Statistically significant ($P < 0.05$).

an interview is an important cost consideration. Also, in practice it is desirable to keep interview time minimal in order to hold a respondent's attention, and to reduce respondent fatigue.²⁶ The length of time to complete the first half of the interview (rate all nine stimuli on one scale) was recorded for about 50 randomly selected respondents for each scale. Similar to what Matell and Jacoby²⁸ observed, the average recorded times using the 5- and 6-pt. scales were both at least 1 min less than the average recorded time for the 9-pt. scale, 8.5 and 8.6 min versus 9.7 min.²⁹ Hence, among the scales with labeled categories, the 5- and 6-pt. scales also exhibited an advantage over the 9-pt. scale in the context of time to administer the interview. Interestingly, the average time to complete the rating task using the 10-pt. scale was slightly larger than the 9-pt. scale average despite the fact that the 9-pt. scale is seemingly more complicated as all categories are labeled while only the endpoints are labeled in the 10-pt. scale.

Finally, the 5-pt. scale has a midpoint, '3' or 'Good', while the 6-pt. scale of course does not have a midpoint. These two scales are as close as possible in terms of number of categories while still differing with respect to whether there is a midpoint, so a comparison of the results for these two scales sheds light on the issue of the effect of inclusion of a midpoint on scale sensitivity. While the sensitivity results for these two scales are similar, the 5-pt. scale does separate one pair of stimuli not separated by the 6-pt. scale – premium versus quality facial tissue. Hence, there is some evidence that inclusion of a midpoint,

keeping the number of categories close, enhances sensitivity of the scale.

STUDY 2

Design

Study 1 involves respondents in a realistic marketing research environment, performing a typical research task, that is experiencing and rating stimuli. All scales used in the study, except the 10-pt. scale, had labeled categories of response. The 9-pt. scale is bipolar while the other three scales are not, and the labels for the 9-pt. scale reflected degree of 'liking' while the labels for the 5- and 6-pt. scales reflected perceived 'goodness'. Hence, even if one excludes the 10-pt. scale from the analyses, some concern still lingers regarding possible confounding of the effects of category labels and/or polarity with the number of scale categories. This is the price of using scales popular among practitioners. To address this concern, a second study was conducted, in which five scales having only labeled endpoints and using the same anchoring phrases – 'About the same' and 'A whole lot less', but including differing numbers (5, 6, 7, 9 or 10) of response categories, were investigated.

A sample of approximately 450 business students from a large northeastern university participated in the second study, which was conducted in class. The sample was approximately equally divided by gender, and the vast majority of subjects were between the ages of 18 and 25. The study consisted of three experiments which

differed only in the context of the tasks – monthly rent, annual rainfall and exam scores. All students participated in each experiment. Participants were organized into five groups of about 90 students each. Each group used a different one of the five scales studied. For each context/experiment, about one-fourth of each group of 90 students was assigned to each of four ‘treatments’ defined by the degree of difference between the corresponding stimulus and a benchmark value. Participants rated their perceptions of the degree of difference between the stimulus (see Tables 4–6) compared to a reference value, using the same one of five rating scales. Such perceptions are antecedents to a diverse set of behaviors such as renting an apartment or satisfaction with a course. Box 1 shows one of the four stimuli to be rated for each context. For simplicity, the sequence of context presentation, rent–rain–exam score, was held constant. Similarly, the rating scale (5, 6, 7, 9 or 10 response categories) a respondent used to express his or her opinion in each context was held constant so that the potential for confusion caused by varying the scale was eliminated. In

summary, within each of three contexts, each participant compared a randomly assigned stimulus from a set of four, to a benchmark level, and rated the perceived difference between the stimulus and the benchmark using the same one of five scales.

A pilot study was conducted to identify the stimuli levels to be used in each context of the study. The goal was to choose values that produced ‘noticeable’ changes in perceived differences versus the benchmark. Eighteen students were asked to rate the perceived similarity of six different stimulus levels to the benchmark level within each context using the following four-item scale: ‘Almost Similar’, ‘Slightly Different’, ‘Moderately Different’ and ‘Whole Lot Different’. Frequency counts of responses in each category were used to judiciously select four treatment or stimulus levels for each context.

Analysis and results

Each context was treated as a separate experiment involving a sample of about 450

Table 4: Rating results of stimulus rent (Study 2)

Scale pt.	Sample and mean			ANOVA results and effect size					
	Stimulus level	Sample size (n)	Mean	Comparison of means	F-test (P-value)	LSD 95%	Range	Range/LSD	Eta ²
5	550	24	3.46	Comparison of all pairs significant	$P < 0.0001$	0.41	2.28	5.56	0.61
	650	21	2.90		—	—	—	—	—
	715	22	2.10		—	—	—	—	—
	740	22	1.18		—	—	—	—	—
6	550	24	4.75	Comparison of all pairs significant	$P < 0.0001$	0.60	3.21	5.35	0.59
	650	23	3.87		—	—	—	—	—
	715	21	2.90		—	—	—	—	—
	740	22	1.54		—	—	—	—	—
7	550	24	5.45	Comparison of all pairs significant	$P < 0.0001$	0.61	3.49	5.72	0.59
	650	21	4.00		—	—	—	—	—
	715	22	3.33		—	—	—	—	—
	740	22	1.96		—	—	—	—	—
9	550	23	6.78	Comparison of all pairs significant	$P < 0.0001$	0.84	4.82	5.74	0.63
	650	25	5.40		—	—	—	—	—
	715	22	3.45		—	—	—	—	—
	740	23	1.96		—	—	—	—	—
10	550	22	6.45	Comparison of all pairs significant	$P < 0.0001$	0.95	4.49	4.73	0.54
	650	24	5.08		—	—	—	—	—
	715	20	3.50		—	—	—	—	—
	740	24	1.96		—	—	—	—	—

Reference value=\$750.

Table 5: Rating results of stimulus rain (Study 2)

Sample and mean				ANOVA results and effect size					
Scale pt.	Stimulus level	Sample size (n)	Mean	Comparison of means	F-test (P-value)	LSD 95%	Range	Range/LSD	Eta ²
5	16	22	3.72	Comparison of all pairs significant	$P < 0.0001$	0.42	2.31	5.50	0.64
	18	24	3.00		—	—	—	—	—
	23	21	1.86		—	—	—	—	—
	24	22	1.41		—	—	—	—	—
6	16	22	4.45	16 > 18 > 23, 24	$P < 0.0001$	0.55	3.02	5.49	0.67
	18	24	3.83		—	—	—	—	—
	23	23	1.78		—	—	—	—	—
	24	21	1.43		—	—	—	—	—
7	16	23	4.57	16, 18 > 23 > 24	$P < 0.0001$	0.68	3.15	4.63	0.56
	18	22	4.41		—	—	—	—	—
	23	24	2.75		—	—	—	—	—
	24	24	1.42		—	—	—	—	—
9	16	23	5.78	16, 18 > 23 > 24	$P < 0.0001$	0.77	4.05	5.26	0.62
	18	23	5.04		—	—	—	—	—
	23	25	2.64		—	—	—	—	—
	24	22	1.73		—	—	—	—	—
10	16	24	6.71	16 > 18 > 23, 24	$P < 0.0001$	0.90	4.61	5.12	0.60
	18	22	4.95		—	—	—	—	—
	23	24	2.88		—	—	—	—	—
	24	20	2.10		—	—	—	—	—

Reference Value=25 inches annual rainfall.

Table 6: Rating results of stimulus exam score (Study 2)

Sample and mean				ANOVA results and effect size					
Scale pt.	Stimulus level	Sample size (n)	Mean	Comparison of means	F-test (P-value)	LSD 95%	Range	Range/LSD	Eta ²
5	67	22	3.95	Comparison of all pairs significant	$P < 0.0001$	0.38	2.76	7.26	0.74
	75	22	2.82		—	—	—	—	—
	78	24	1.71		—	—	—	—	—
	79	21	1.19		—	—	—	—	—
6	67	21	4.76	Comparison of all pairs significant	$P < 0.0001$	0.51	3.59	7.04	0.70
	75	22	2.86		—	—	—	—	—
	78	24	2.29		—	—	—	—	—
	79	23	1.17		—	—	—	—	—
7	67	24	5.71	67 > 75, 78 & 75 > 79	$P < 0.0001$	0.70	3.59	5.13	0.59
	75	23	3.26		—	—	—	—	—
	78	22	2.59		—	—	—	—	—
	79	24	2.12		—	—	—	—	—
9	67	22	6.77	67 > 75 > 78, 79	$P < 0.0001$	0.96	4.37	4.55	0.61
	75	23	4.52		—	—	—	—	—
	78	23	3.30		—	—	—	—	—
	79	25	2.40		—	—	—	—	—
10	67	20	7.50	67 > 75 > 78, 79	$P < 0.0001$	0.92	5.33	5.79	0.68
	75	24	5.00		—	—	—	—	—
	78	24	2.09		—	—	—	—	—
	79	22	2.17		—	—	—	—	—

Reference value=80.

Box 1: (Study 2)

The purpose of this survey is to obtain your opinion for a variety of situations. Please read the description of each situation carefully and answer each survey question.

- (1) Normally, the rent for a single bedroom apartment in a particular complex is \$750 per month. Currently there is a special promotion going on to attract new tenants. The apartment rent has been reduced to \$740. Please rate the amount of the rent reduction by comparing the reduced rent (\$740) to the usual rent (\$750) using the scale below. Choose number between 1 and 6 to reflect your opinion, where 1 represents 'About the same rent' and 6 represents 'A whole lot less rent'.

About the same						A whole lot less
1	2	3	4	5	6	

- (2) Imagine the average annual rainfall is 25 inches where you live. This year the total rainfall was only 16 inches. Please rate this year's rainfall (16 inches) by comparing it to the average annual rainfall (25 inches) using the scale below. Choose number between 1 and 6 to reflect your opinion, where 1 represents 'About the same rainfall' and 6 represents 'A whole lot less rainfall'.

About The same						A whole lot less
1	2	3	4	5	6	

- (3) Suppose the class average score on a midterm exam for a class using a newly developed review aid is 75 out of 100. The average score for all other classes taking this same exam is 80 out of 100. Rate how much lower the average score (75) is for the class using the new review aid compared to the average score for other classes (80) not using the aid. Choose number between 1 and 6 to reflect your opinion, where 1 represents 'About the same score' and 6 represents 'A whole lot less score'.

About the same						A whole lot less
1	2	3	4	5	6	

respondents. A one-way ANOVA, involving four groups defined by the stimulus level (amount below the benchmark), was conducted within each context for each of the five scales. Subjects' rating of the stimulus is the dependent variable and stimulus level is the independent variable in each ANOVA. Overall 'F' tests were conducted, and Fisher's least significant difference (LSD 95 per cent) was employed for testing for significant differences at the 5 per cent risk level ($\alpha = 0.05$) between stimuli means. The results of these analyses are presented in Tables 4–6.

First, with one minor exception (the two largest stimuli in the context of exam scores using the 10-pt. scale) the rank order of the means for each scale in each context is consistent with the ordering of the stimuli based on magnitude of difference from the benchmark. The overall *F*-test for the equality of mean ratings is significant ($P < 0.05$) for each context and each scale, suggesting the existence of at least one difference between mean ratings of stimuli for all 15 cases (5 scales $\times$ 3 contexts). However, some differences in scale sensitivity are observed among the scales.

In the context of rent, the overall *F*-test is significant ($P < 0.0001$) and all pairwise comparisons

of stimuli means are significant (LSD 95 per cent) for all scales. Hence, in this context, all five scales were found to be equally sensitive. For the annual rainfall and exam score contexts, all overall *F*-test are again highly significant ($P < 0.0001$), but the scales are not equally sensitive as shown in Tables 5 and 6. In the rainfall context, all pairwise comparisons were significant only when the 5-pt. scale was used. In the context of exam scores all pairwise comparisons are significant only when the 5- and 6-pt. scales are used. The scales having 7, 9 and 10 categories produce similar results for the rainfall and exam score contexts. Finally, in all three contexts, the values of η^2 (measure of effect size) using the 5- and 6-pt. scales are generally as large as, or larger than, the corresponding values for the other three scales. The results of the second study reinforce the finding of the first study that sensitivity does not appear to increase with the number of response categories. As in Study 1, the evidence suggests that sensitivity may be higher for 5- and 6-pt. scales than for scales having increased numbers of categories.

As in Study 1, the results of Study 2 can also be used to investigate the effect of the inclusion of a scale midpoint on scale sensitivity. Comparing the results of the 5-pt.

scale (midpoint) to those for the 6-pt. scale (no midpoint) and the results for the 9-pt. scale (midpoint) to those for the 10-pt. scale (no midpoint) again sheds light on the impact of inclusion of a midpoint on scale sensitivity. The results of the 9- and 10-pt. scales are very similar. However, as in Study 1, slightly more significant differences were observed when the 5-pt. scale was used than when the 6-pt. scale was used. Again, there is some evidence that inclusion of a midpoint, keeping the number of response categories about the same, enhances scale sensitivity.

SUMMARY AND CONCLUSION

The sensitivities of scales with differing numbers of categories were compared. Pilot tests were conducted to ensure that meaningful, but not excessive differences exist among the stimuli used in the studies. The notion that increasing the number of response categories in a scale, within the range of categories typically used in practice, increases sensitivity is not supported by the results of this study. In fact, the evidence suggests that the opposite is true. Although the range of the differences between sample means generally increases with the number of response categories in the scale, the 'yardstick' required for significance (LSD 95 per cent), which is directly proportional to the variance of the associated measurements, also increases. The 5- and 6-pt. scales appear to be more sensitive than their counterparts having more response categories. Note that Krosnick and Fabrigar¹⁸ concluded that 5-7 response categories are optimal from reliability and validity perspectives. Combining these results with the findings of this study regarding scale sensitivity, suggests that, among scales typically used in practice, 5- or 6-pt. scales should be chosen over scales with higher numbers of response categories.

The argument for 5- and 6-pt. scales is reinforced by the overall respondent preference and shorter interview times observed for these scales in Study 1. The practical advantages of shorter interview time are lower costs, lower incidence of respondent refusal to complete the task, which further lowers survey cost, and

reduced respondent fatigue and frustration, which enhances data integrity.

The importance of scale sensitivity is clear given the prevalence of comparisons of means based on itemized rating scales used in the analysis of survey research results and experiments. Researchers can feel confident that they are not compromising sensitivity by yielding to the preferred practical advantages of 5- and 6-pt. scales when adopting a scale for a study. Further, the preferred 5- and 6-pt. scales fall within the recommended range for the number of categories to include in an itemized rating scale based on research focused on scale criteria other than sensitivity such as the important considerations of reliability and validity.

The results of both studies reported in this article suggest that the inclusion of a scale midpoint may enhance scale sensitivity. An area for future research is to thoroughly investigate this phenomenon. It may be that this effect, if it exists, diminishes as the number of scale categories increases as appears to happen in Study 2 reported in this article. Another potential area for future research is the relationship between scale characteristics, and the length of time required to administer an interview. The results of Study 1 reported in this article suggest that increasing the number of response categories increases the time to complete an interview. This issue should be further investigated along with the possible effect of other scale characteristics on interview length. If interview time can be decreased without detracting from the integrity of the data collected, researchers can derive a direct benefit from reduced cost, as well as the indirect benefits described earlier.

ACKNOWLEDGEMENT

The authors thank Jonathan Reynolds and two anonymous reviewers for their valuable comments on an earlier version of this manuscript. We also thank Directions Research, Inc., in particular, Randy Brooks, President and Alec Schendzelos, Vice-President, for sponsoring this project and providing guidance. We wish to acknowledge the assistance of Margaret Hanson, Jill Gardere and Linsey Matthews, former graduate students in the

University of Georgia's Master of Marketing Research Program, who worked on the project. Finally, we thank the University of Georgia's Terry College of Business Center for Marketing Studies for its support.

REFERENCES AND NOTES

- DeVellis, R. (2003) *Scale Development Theory and Applications*. Thousand Oaks, CA: Sage Publication.
- Trochim, W. (2001) *The Research Methods Knowledge Base*, 2nd edn. Cincinnati, OH: Automicdoc Publishing.
- Bergkvist, L. and Rossiter, J.R. (2007) The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research* 44: 175–184.
- Viswanathan, M., Sudman, S. and Johnson, M. (2004) Maximum versus meaningful discrimination in scale response: Implications for validity of measurement of consumer perceptions about products. *Journal of Business Research* 57: 108–124.
- Rossiter, J.R. (2002) The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing* 19(December): 305–335.
- Jacoby, J. and Matell, M.S. (1971) Three-point Likert scales are good enough. *Journal of Marketing Research* 8(November): 495–500.
- Churchill, G.A. (1979) A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research* 16: 64–73.
- Krosnick, J.A. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5: 213–236.
- Krosnick, J.A. and 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.
- Zeller, J. (1988) Vague questions get vague answers: An experimental attempt to reduce response instability. PhD thesis, University of California at Los Angeles, CA.
- Bendig, A.W. (1954) Reliability and the number of rating scale categories. *Journal of Applied Psychology* 38: 38–40.
- Symonds, P.M. (1924) On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology* 7: 456–461.
- Alwin, D.F. and Krosnick, J.A. (1981) The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods and Research* 20: 139–181.
- Churchill Jr, G.A. and Iacobucci, D. (2005) *Marketing Research Methodological Foundations*. Mason, OH: South-Western.
- Zikmund, W.G. (2003) *Exploring Marketing Research*, 8th edn. Mason, OH: South Western.
- Netemeyer, R., Bearden, W. and Sharma, S. (2003) *Scaling Procedures Issues and Applications*. Thousand Oaks, CA: Sage Publications.
- Parasuraman, A., Grewal, D. and Krishnan, R. (2004) *Marketing Research*. Boston, MA: Houghton Mifflin Company.
- Krosnick, J.A. and Fabrigar, L.R. (1997) Designing rating scales for effective measurement in surveys. In: L.E. Lyberg, P. Biemer, M. Collins and E.D. Leeuw (eds.) *Survey Measurement and Process Quality*. New York: John Wiley.
- Garner, W.R. and Hake, H.W. (1951) The amount of information in absolute judgments. *Psychology Review* 58: 446–459.
- Miller, G.A. (1956) The magical seven plus or minus two: Some limits on our capacity for processing information. *Psychology Review* 63: 81–97.
- Komorita, S.S. and Graham, W.K. (1965) Number of scale points and the reliability of scales. *Education Psychology Measurement* 25(4): 987–995.
- For example, below is the text used by the interviewer in collecting ratings for the first two writing instruments (for the product category sequence in which writing instruments were presented first).

As you were told earlier, today we are conducting a study about a few everyday products and would like to include your opinions. To start, we'd like your opinion of three different writing pens. Here is the first pen for you to use. I'd like you to evaluate it by signing your name on this pad of paper. Please choose the answer from the card (interviewer points to show card containing scale) that best represents your overall opinion of the pen. (Respondent signs name and provides rating.) Now, here's the second pen. Please tear off the top sheet of paper and evaluate the pen by signing your name on the pad. Please choose the answer from the card that best represents your opinion of the pen

Also, note that respondents cleansed their palates by taking a bite of a plain cracker and a sip of water between orange juice samples.
- Winer, B.J., Brown, D.R. and Michels, K.M. (1991) *Statistical Principles in Experimental Design*. Boston, MA: McGraw-Hill.
- Siegel, A.F. (2000) *Practical Business Statistics*. Boston, MA: McGraw-Hill.
- Both rating cards used were placed in front of the participant who was then asked: 'which scale did you like using better?'
- Hogg, A. and Miller, J. (2003) Watch out for dropouts. *Quirk's Marketing Research Review*, <http://www.quirks.com/articles/a2003/20030706.aspx?searchID=159236243&sort=5&pg=1>.
- Olejnik, S. and Algina, J. (2003) Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods* 8(4): 434–447.
- Matell, M.S. and Jacoby, J. (1972) Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology* 56: 506–509.
- The mean times for the 5-pt. and 6-pt. scales are not significantly different, and the mean time for the 5-pt. and 6-pt. scales combined is significantly ($\alpha=0.05$) smaller than the 9-pt. scale mean.