

**Five Point vs. Eleven Point Scales:
Does It Make A Difference To Data Characteristics ?**

(published in Australasian Journal of Market Research Vol 10 No. 1 2002)

John Dawes

Marketing Science Centre

University of South Australia

Email John.Dawes@marketingsciencecentre.com

Keywords: Scales, split-sample experiment.

Five Point vs. Eleven Point Scales: Does It Make A Difference To Data Characteristics ?

Abstract

This study examines whether the number of scale points used in a market research survey affects the resultant data. It uses two 'split sample' surveys, one using face to face interviews and one gathered using telephone interviews. In each case, a subset of the sample was administered questions using a five point scale and another subset was administered an eleven point 'zero to ten' scale. The results show that the eleven point scale produces data that is essentially the same as that produced by the five point scale in terms of mean, after allowing for the five point scale to be re-scaled for comparability. However, the eleven point scale produced data with more variance (coefficient of variation) than the five point scale. There were some differences between the scale types in terms of kurtosis and skewness, but these were not systematic.

Background & Previous Research

One of the most ubiquitous tools of the marketing or academic researcher is the use of multiple category numerical scales. The question of finding the 'optimal' number of scale points has been discussed from a variety of perspectives. Several studies have used simulations to examine issues of information recovery and the precision of data (Givon and Shapira 1984; Green and Rao 1970; Lehmann and Hulbert 1972). Others have used survey data. Hulbert (1975) examined the degree to which respondents used different response categories in a very lengthy questionnaire, where the respondents were not provided with pre-set response categories, but simply allocated numbers to each question. He found that the

mean number of different responses was between six and ten. This suggests that ten may be an upper limit in terms of the degree to which respondents can discriminate between questionnaire stimuli, confirming earlier research from the 1950's.

Cox (1980) extensively reviewed previous studies to conclude that scales with two or three response categories were inadequate - (rebutting Jacoby & Mattell (1971)), but there was little marginal gain from using more than nine categories. Clarke (2000) found that increasing the number of scale categories from three to five reduced extreme responses but beyond five categories there was little effect. Holmes (1974) conducted an experiment comparing the responses for unstructured versus structured scales and other derivations of scale construction, but did not test for differences according to the number of response categories. Schertzer & Kernan (1985) investigated the semantic intensity, ordinality and interval level qualities of response labels according to the number of labels used (such as three, four, five categories of response label etc.), however respondents did not rate objects *per se*, they matched each response label to a point on a 100 point scale. Alwin (1997) found that eleven point scales performed better than seven point scales in terms of reliability and validity. Grigg (1980) found that an eleven point scale produced more dispersion in responses than a seven point scale. Grigg's result is closer to the issue examined in this paper, however was conducted twenty years ago and used a self completion method. More analysis using data gathered via telephone as well as by self completion would be desirable.

Several studies have also examined whether the inclusion of scale *mid points* affect responses. Worcester & Burns (1975) found that descriptors such as *tend to agree* represented more positive sentiment on a four point scale compared to a five point scale. Spagna (1984) found that the omission of a mid point had most effect on increasing the frequencies of the neighbouring scale responses. Si & Cullen (1998) found that mid points did not affect mean

response levels. Dawes (2001) found a scale mid point resulted in fewer positive responses. Garland (1991) found the opposite - a mid point resulted in fewer *negative* ratings. So there is mixed evidence on this issue.

In general there is support for the view that more scale categories reduce the incidence of extreme responses, and may produce more dispersion in the data. However, as this review demonstrates, not much is known in descriptive terms about the effect of different numbers of scale response categories on basic data characteristics such as the relative mean, variance, shape and so on - and certainly not for data gathered via telephone. Yet these are important issues:

- The mean (average) forms the basis of many evaluations contained in market research reports. Would the mean (relative to the highest possible score obtainable from the scale used) have been different if a different number of response categories had been used ?
- The variance (ie standard deviation) is integral to many tests of statistical inference. For example, if one has a hypothesis that responses may differ according to a certain grouping variable, ideally support for this would come in the form of little within-group variance and considerable across-group variance. Does the number of scale points affect the amount of variance in the data ?
- Some statistical techniques are sensitive to kurtosis. Do scales with different numbers of response categories produce different degrees of kurtosis ?
- The degree of skewness in the data may affect various statistical analyses. Do scales with different numbers of response categories produce different degrees of skewness ?

This study sets out to address these four issues using results from two survey studies.

Data

Two data sets are used in the analysis. The first comprised a convenience sample of 301 undergraduate students who were interviewed at various campuses of an Australian university. The topic was the campus cafeteria at that location. Respondents were interviewed face-to-face at various locations on each campus by IQCA accredited interviewers. A voucher for a free cup of coffee was offered as an incentive to participate. There was no criteria for inclusion except that the respondent was a student at that campus.

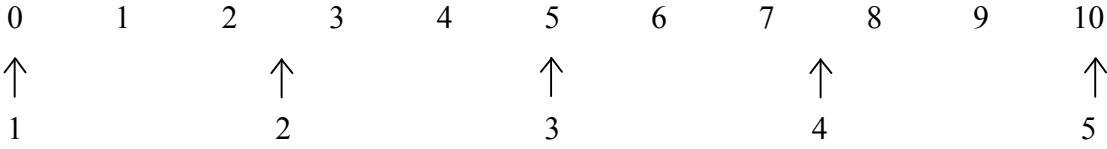
Of the 301 respondents, 146 were administered a series of questions and asked to respond using a five point "1 = strongly disagree ... 5 = strongly agree" scale. The other 155 were administered the same questions but were asked to respond on an eleven point zero to ten scale, where zero meant no agreement and ten denoted 100% agreement. Allocation of respondents to the respective scale type was random. The response rate for this survey was not recorded but qualitative information from the interviewers was that it was approximately 50%.

The second data set was part of a telephone survey of 751 people who had used taxicabs in the past six months. The sample was selected from a random generation based on the electronic telephone directory. Of the 751 respondents, 151 were administered a series of questions and asked to respond using a five point "strongly disagree - strongly agree" scale. The other 592 were administered the same questions but were asked to respond on an eleven point zero to ten scale, where zero meant no agreement and ten denoted 100% agreement. The difference in sample sizes for the two groups was due to commercial considerations. The sponsor of the study was particularly interested in obtaining scores "out of ten" to retain comparability with previous research, but was prepared to allow a sub group of respondents to be administered the other five point scale to determine how it might affect results. As with the first study, allocation of respondents to the respective scale was random. The response

rate for this survey, based on automated call management software was 49%. Therefore we have confidence that the results are likely to be typical of such surveys since they did not suffer from unduly low response rates.

Analysis

The first issue concerns the possible effect on the mean value of the variables. In order to determine whether the two different scale types had any effect on the level, the mean score for each question was tabulated. To compare the mean scores for the two scale types, the 5 point scale data was re-scaled, by recoding the five point scores to a score out of ten as shown in the diagram below:



1 and 5, the extreme points on the five point scale were transformed to 0 and 10, the corresponding extreme points on the eleven point scale. 3 was transformed to the corresponding mid-point on the eleven point scale, namely 5. 2 and 4 were fixed mid-way between the extreme point and the mid point. This meant that 2 was transformed to 2.5 and 4 was transformed to 7.5. Obviously the analysis based on this re-scaling depends on the assumption of equal intervals for both scales. This presumption is arguable with some writers findings showing a lack of intervality (eg. Schertzer and Kernan 1985) and others much more supportive (Crask and Fox 1987). The assumption of equal interval qualities was assumed for this analysis, and if the different scale types produced similar results this assumption would be justified. Table 1 shows the differences between the eleven point and five point (rescaled) data for both data sets; so all scores are “out of 10”. The Mean Absolute Difference (MAD)

is shown at the bottom of the table, which is the average of differences, not accounting for the sign of the difference. The data are ordered by mean score.

INSERT TABLE 1 HERE

On average, the use of a five point scale compared to an eleven point scale results in quite a small difference in mean scores (when re-scaled). The MAD (Mean Absolute Difference) is 0.25 of a scale point on a zero to ten scale (once the 5 point data has been re-scaled) for both data sets. If we do take the sign of the differences into account, the eleven point scale produces mean scores that are around 0.1 of a scale point higher than the re-scaled five point scale. So the difference is minor.

It is also noteworthy that there are more cases where the use of the eleven point scale results in a higher comparative score than a lower comparative score. In Table 1, for data set 1 the eleven point scale produces higher mean scores than the re-scaled five point scale in five out of seven instances. In data set 2 this occurs in eight out of thirteen instances.

There is also a tendency for this effect to be more marked where the particular question attracted a lower average score across respondents. We see in Table 1, for data set 1 the difference between the 11 point score and the 5 point re-scaled score is larger for question 8 (the question with the lowest mean score). This occurs again in data set 2 (still in Table 1) for the two lower scoring questions, namely Q 11 and Q7.

Variance

To examine the impact of the different scales on the variance, we report on the coefficient of variation for each variable. The coefficient of variation is the standard deviation divided by the mean score, then multiplied by 100. It facilitates the comparison of variances for items with different mean levels.

Table 2 shows the results for both data sets. The results suggest that the use of an eleven point scale results in a greater amount of variance (relative to the mean score) compared to a five point scale - note how the "difference" column is mostly positive. The results from data set 2 confirm this, as shown in the right-hand side "difference" column for data set 2 in Table 2. The rationale for this result is that the eleven point scale simply provides a wider range of responses, therefore produces more dispersion in responses (consistent with Grigg 1980). Note that data set 2 overall has much higher coefficients of variation for most questions than data set 1. This result is not relevant to the major issue examined in the study, (which is *within*-data set differences according to the number of scale points used) but the reason for this result should be clarified. It is simply due to a much higher degree of variation in perceptions of service quality by users of taxicabs compared to student users of cafeteria services.

INSERT TABLE 2 HERE

Kurtosis

Kurtosis is the degree to which the data is peaked, or flat. Kurtosis is also 'scale free' so it is possible to directly compare the results for the five and eleven point scales without any need to transform the data. Table 3 shows that for data set 1, the eleven point scale data has higher levels of kurtosis in four out of six cases. In data set 2 the eleven point scale data produces *lower* levels of kurtosis in eight out of 13 cases. Therefore, there does not appear to be any systematic association between the two scale types and the amount of kurtosis, in these two data sets.

INSERT TABLE 3 HERE

Skewness

Negative skewness denotes the extent to which there is a "tail" below the mean. A positive skew means the opposite - a "tail" above the mean. The data from both surveys is all negatively skewed, with mean scores all above the scale mid-point.

The results for skewness according to scale type are rather mixed. In Table 4, for data set 1 there is no clear pattern in the difference in skewness between the two types of scales. For the second data set, there is a tendency for the 11 point scale to produce a more negative skewness figure (a longer tail below the mean). However, this appears in only one of the data sets. The conclusion is that there is no systematic relationship between the number of scale points and skewness.

INSERT TABLE 4 HERE

Discussion & Conclusions

This study found in two "split sample" experiments that:

- In both cases, the difference in mean scores produced by the eleven point scale was small, compared to the re-scaled scores produced by the five point scale. The eleven point (zero to ten) scale produced more instances of slightly higher mean scores than the five point scale - after allowing for the five point scale to be re-scaled.
- Furthermore, this tendency of the eleven point scale to produce slightly higher comparative scores was more marked for questions that attracted lower overall mean scores to begin with.
- An eleven point (zero to ten) scale produced higher levels of variance (ie coefficient of variation) compared to a five point scale
- There was no systematic relationship between the use of an eleven point vs. a five point scale in terms of kurtosis
- There was no systematic relationship between the use of an eleven point vs. a five point scale in terms of skewness

It would appear that the choice of the number of response categories depends on the purpose of the study. If the major purpose is simply to obtain numerical responses and produce mean scores, the number of response categories does not make a marked difference to the mean score obtained, relative to the upper scale limit. In terms of interpretability, a 'score out of ten' has some appeal in terms of managerial interpretability.

However, if there is an intent to examine dependence relationships between scale variables using tools such as regression, for example, a scale with more response categories could be more useful as it appears to result in more variance in the data.

The results from this study are also 'good news' for those who commission or undertake market research for clients with historical data they wish to preserve comparability. For example, if an organisation had years of data that was gathered using a five point scale, it is likely that it could be converted or transformed to make it comparable with other data that was gathered using more scale categories. Comparability would be more of an issue with items that received lower scores. In this study, the rescaling process had the effect of 'deflating' the questionnaire variables that received lower scores using the five point scale. The author plans to conduct more replications of this result to test the robustness of the findings.

References

- Alwin, Duane F (1997), "Feeling thermometers vs 7-point scales," *Sociological Methods and Research*, 25 (3), 318-51.
- Clarke III, Irvine (2000), "Extreme Response Style in Cross-Cultural Research: An Empirical Investigation," *Journal of Social Behavior & Personality*, 15 (1), 137-52.
- Cox III, Eli P. (1980), "The Optimal Number of Response Alternatives for a Scale: A Review," *Journal of Marketing Research*, 17 (November), 407-22.
- Crask, Melvin R. and Richard J. Fox (1987), "An Exploration of the Interval Properties of Three Commonly Used Marketing Research Scales: A Magnitude Estimation Approach," *Journal of the Market Research Society*, 29 (No. 3), 317-39.
- Dawes, John G. (2001), "The Impact of Mentioning a Scale Mid-Point in Administering a Customer Satisfaction Questionnaire Via Telephone," *Australasian Journal of Market Research*, 9 (No. 1, January), 11-18.
- Garland, Ron (1991), "The Mid-Point on a Rating Scale: Is It Desirable?," *Marketing Bulletin*, 2, 66-70.
- Givon, Moshe M. and Zur Shapira (1984), "Response to Rating Scales: A Theoretical Model and Its Application to the Number of Categories Problem," *Journal of Marketing Research*, 21 (November), 410-19.
- Green, Paul E. and Vithala R. Rao (1970), "Rating Scales and Information Recovery - How Many Scales and Response Categories to Use?," *Journal of Marketing*, 34 (July), 33-39.
- Grigg, A.O. (1980), "Some Problems Concerning the Use of Rating Scales for Visual Assessment," *Journal of the Market Research Society*, 22 (No. 1), 29-43.
- Holmes, Cliff (1974), "A Statistical Evaluation of Rating Scales," *Journal of the Market Research Society*, 16 (No. 2, April), 87-107.
- Hulbert, James (1975), "Information Processing Capacity and Attitude Measurement," *Journal of Marketing Research*, 12 (February), 104-06.
- Jacoby, Jacob and Michael S. Matell (1971), "Three Point Scales Are Good Enough," *Journal of Marketing Research*, 8, 495-500.
- Lehmann, Donald R and James Hulbert (1972), "Are Three-Point Scales Always Good Enough?," *Journal of Marketing Research*, 9 (November), 444-46.
- Schertzer, Clinton B. and Jerome B. Kernan (1985), "More on the robustness of Response Scales," *Journal of the Market Research Society*, 27 (No. 4, October), 261-82.

Si, Steven and John Cullen (1998), "Response Categories and Potential Cultural Bias: Effects of an explicit middle point in cross-cultural surveys," *International Journal of Organizational Analysis*, 6 (3), 218-31.

Spagna, Gregory J. (1984), "Questionnaires: Which Approach Do You Use?," *Journal of Advertising Research*, 24 (No. 1, February/March), 67-70.

Worcester, Robert M. and Timothy R. Burns (1975), "A Statistical Examination of the Relative Precision of Verbal Scales," *Journal of the Market Research Society*, 17 (No. 3), 181-97.

Table 1: Differences in Means for Data Set 1 & 2

Data set 1				Data set 2			
Q no.	11 pt. scale	5 pt. (rescaled)	Diff.	Q. no	11 pt. scale	5 pt. (rescaled)	Diff.
Q12	7.6	7.8	-0.2	Q8	7.7	7.7	0
Q11	7.8	7.6	+0.2	Q9	7.5	7.6	-0.1
Q6	7.3	7.6	-0.3	Q3	7.3	7.2	+0.1
Q10	7.6	7.5	+0.1	Q10	6.9	6.9	0
Q7	7.7	7.5	+0.2	Q4	6.7	6.2	+0.5
Q9	7.3	7.2	+0.1	Q6	6.4	6.3	+0.1
Q8	6.7	5.9	+0.8	Q5	6.3	6.2	+0.1
				Q2	5.9	6.0	-0.1
				Q12	5.9	5.6	+0.3
				Q13	5.8	5.4	+0.4
				Q1	5.7	5.8	-0.1
				Q11	5.6	4.9	+0.7
				Q7	5.5	4.7	+0.8
		MAD	0.25			MAD	0.25

MAD 0.8

MAD 0.3
