EVALUATING NUMERIC AND VERBAL LABELS FOR RESPONSE SCALES

Colm A. O'Muircheartaigh, George D. Gaskell & Daniel B. Wright
London School of Economics, WC2A 2AE, England

KEY WORDS: Response alternatives, Measurement error, CASM, question wording

A question or statement with an accompanying set of response alternatives arranged on a numeric or verbal scale is among the most commonly used measurement instrument in social and psychological research. There is an implicit assumption that the attitude or behaviour being measured falls along a single, latent or manifest, continuum, ranging from positive to negative. Dawes and Smith (1985) give a general discussion of the properties and justification of such scales. Alwin and Krosnick (1991) consider their properties in terms of the reliability of measurement. Among the characteristics traditionally considered are the number of scale points, the inclusion or omission of a midpoint, the extent and nature of the verbal labelling of response options, and the explicit inclusion of a don't know response option.

Response scales (rating scales) may be justified by arguing that they are "compatible with the ways in which people using them think" (Dawes & Smith, 1985; p 540). For example, people may spontaneously characterise political attitudes in terms of a left-right continuum. A rating scale consisting of a line with the labels *left wing* and *right wing* at the extreme left and right may be compatible with such people's thinking. Clearly not all rating scales are compatible with intuitive thought nor does compatibility imply that rating scales are isomorphic with such thought. It has been argued that in some cases a unipolar verbal ordering on a vertical axis may be easier to understand than a bipolar ordering from a central position.

Schwarz, Strack, Müller and Chassein (1988) raise the issue of how the response alternatives themselves may determine the meaning of the question. Subsequent research has accumulated which demonstrates that the construction of the response scale for a question may substantially influence the way in which respondents answer questions (eg Schaeffer, 1991; O'Muircheartaigh, Gaskell, & Wright, 1992). Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark (1991) introduced a new element by considering the effect of the numeric values assigned to the response options (see also Smith's (1993) results from the General Social Survey). In this paper we consider the nature of the influence of both the verbal and numeric labels which appear with the scale.

While the numeric values are often included only for coding and response convenience, Schwarz et al (1991) have demonstrated that they carry more, sometimes unintended, meanings. For a particular question, "How successful have you been in life, so far?", they showed that a scale with numeric values ranging from 0 to 10 was not the same as a scale whose values ranged from -5 to +5. The verbal anchors were "not at all successful" (0 or -5) and "extremely successful" (10 or +5). They argued that when a 0 to 10 scale is used respondents infer that 0 stands for the absence of any amount; the scale becomes unipolar. In contrast, respondents infer that the scale is bipolar when the numeric values range from -5 to +5. For example, when asking people how successful they had been in their life, if a 0 to 10 scale is offered, they will assume that the low anchor (0) corresponds to *not having any success*. This contrasts with the interpretation of the lowest point on the -5 to +5 scale as being *unsuccessful* (being a failure).

Our research explores two aspects of the issues involved. First, in an experiment in which we replicated, in the setting of a large scale survey, the work reported by Schwarz et al (1991), we also tested whether mentioning the numeric anchors explicitly in the question stem altered the magnitude of the effect of the numeric scales.

In a second experiment we compared the impact of the numeric and verbal anchors. There are two ways in which we may signal to respondents whether we wish them to treat a response scale as unipolar or bipolar. The usual way is by using verbal anchors which are either unipolar (eg *{no more power, much more power}, {not having any success, having great success}*) or bipolar (eg *{much more power, much less power}, {much success, much failure}*). The second way is to use numeric labels which either imply a unidimensional construct (eg *{0 to 10}, {0 to 5}, {0 to 6}, {-5 to 0}*) or bipolar construct (eg *{+5 to -5}, {+3 to -3}, {+2 to -2}*). We compared various combinations of verbal and numeric anchors in order to identify their relative contributions. Our interest was in the possibly complex interplay of the verbal and numeric cues presented to the respondent.

These experiments were embedded in the British Market Research Bureau's (BMRB) face to face omnibus survey (ACCESS) and were conducted with their assistance. Each week BMRB carries out an omnibus survey with questions on a variety of topics which vary from week to week. Our questions were inserted at a point in the questionnaire considered suitable by our colleagues at BMRB, typically about 15 minutes into the 25-30 minute interview. We receive for each week from BMRB a list of the topics included in that week's questionnaire, giving both the order of the topics and the time spent on each. In some cases there was more than one BMRB version of the questionnaire within the week. In these cases our experimental conditions were randomised within their versions. For the experiments we report, there were no topics from which an effect would predicted.

BMRB's omnibus survey, which draws respondents aged 15 years and older from Great Britain (with the exception of offshore islands), uses a sampling technique known as GRID Random Location. This is a probability sample of final stage area units combined with a non-probability quota-controlled selection of individuals. The sample is a cluster sample. However, as we are examining comparisons between subclasses which are distributed fairly uniformly across clusters and as the average cluster take is relatively small (approximately 10), the design effects can be predicted to be close to 1. Thus, the p values obtained from standard statistical programs can reasonably be applied.

To the extent that it is possible, we have checked that the allocation of the sample to the different experimental conditions was properly implemented. The nature of the sample design makes it impossible to calculate response rates, but the distribution of the subsamples across the experimental conditions was compared on a variety of social and demographic characteristics and was found to be within the expected range of variation in all cases.

Does the scale need to be mentioned?

Respondents in the July and August (1992) BMRB omnibus survey (n=2124) were asked

> *"How entertaining do you think the adverts on television are, compared to the programmes?".*

Respondents were randomly allocated to one of four conditions of a 2×2 experimental design. The first factor was whether the numeric values on the scale ranged from 0 to 10 or from -5 to +5. Each scale used bipolar verbal anchors: "much more" and "much less" "entertaining than the programmes". The scale was presented on a showcard as a vertical ladder. The second factor was whether a description of the scale, including the numeric values, was included in the question. The description read: "The scale ranges from 10 (+5), if you think the adverts are much more entertaining than the programmes, to 0 (-5) if you think the adverts are much less entertaining than the programmes". This allows us to test whether explicit signalling (as used by Schwarz et al, 1991) is necessary to produce the response shifts.

For analysis purposes the data were recoded so that the range of scores for both scales was from 0 to 10. The means of the four conditions and the significance tests comparing them are shown at the bottom of table 1. There is a clear main effect for the anchors: respondents given the 0 to 10 scale were more likely to say the adverts were entertaining. This is a confirmation of the Schwarz et al (1991) result. There is a possible effect for explicitly mentioning the scale; the estimated size of the effect is about half that of the numeric labels, but the significance level is marginal at 0.06. There was no evidence of an interaction between the use of different anchors and whether they were mentioned.

Table 1.
The percentages and means for comparing advertisements with programmes (Study I).

	Conditions				
	not mentioned		mentioned		
	0→10	-5→+5	0→10	-5→+5	
scale	Percentages				total %
10 or +5	4	4	2	5	4
9 or +4	2	7	2	6	4
8 or +3	6	13	8	11	9
7 or +2	13	16	9	16	13
6 or +1	12	10	13	9	11
5 or 0	17	14	22	13	17
4 or -1	13	6	9	5	8
3 or -2	13	7	13	10	11
2 or -3	7	9	8	9	8
1 or -4	13	5	7	5	8
0 or -5	∅	9	7	11	7
total n	537	514	527	483	2061
$\bar{x}$ (0-10)	4.73	5.15	4.48	4.96	4.83

Main effects

Mentioning	F(1,2057) = 3.52	p=.06
Scale	F(1,2057) = 14.46	p=.00
Interaction	F(1,2057) = 0.05	p=.82

The frequency distribution of the responses for each of the four conditions is also shown in table 1. Though the four distributions show some interesting variations, one cell in particular deserves attention. When the 0 to 10 numeric labels were used but **not** explicitly mentioned by the interviewer there were **no** responses in the bottom category; respondents did not choose to use the zero value. This contrasts with about 10% (50 or so) for each of the other three conditions. It is possible that this finding has to do with the physical layout of the showcard in that the words in the lower verbal anchor were printed close to the numeral 0 and may have encouraged the respondents to neglect it in favour of what mathematicians call the natural numbers (1,2,3,...).

Though there are some other suggestive differences between the distributions of the responses for the four conditions the distorting effect of the empty cell makes it difficult to reach any firm conclusion.

Comparing the effects of verbal and numeric anchors

The experimental question was embedded also in BMRB's July and August (1992) face to face omnibus surveys, but for different respondents than those used in experiment 1. Respondents (n=2165) were asked

> *"... to what extent do you think the Advertising Standards Authority should be given more power to control advertisements?".*

Respondents were divided among four conditions in a 2×2 design. The first factor was unipolar vs bipolar verbal anchors ({not given any more power, given much more power} vs {given much less power, given much more power}. The second was unipolar vs bipolar numeric anchors -- and the associated intermediate numeric labels for an 11-point scale -- ({0,10} vs {-5,+5}). These were explicitly mentioned to all respondents.

There is no earlier work of which we are aware that compares the relative effects of numeric and verbal anchors on the use of a response scale. Each of these may be signalling to respondents whether the scale is unipolar or bipolar. The 0 to 10 numeric labels and the {not given any more, given much more} verbal labels each imply a unipolar construct. The -5 to +5 numeric labels and the {given much less, given much more} verbal labels imply a bipolar construct. The unipolar numeric labels should fit most easily with the unipolar verbal anchors; similarly, the bipolar numeric labels and bipolar verbal anchors should combine well. Conversely, a mixture of bipolar and unipolar cues might be expected to cause some difficulty to the respondent. Therefore, we might expect some interaction between the two sets of cues.

Table 2 gives the frequency distributions, means and an analysis of variance of the results. As with experiment 1, for analysis purposes responses were recoded so that each was based on a comparable 0 to 10 scale. Each of the factors had a significant effect on responses. The magnitudes of the effects are similar and, much more interestingly, the effects appear to be additive (ie no interaction was observed). This suggests that the verbal and numeric anchors may be tapping different but complementary aspects of the scale.

The comparison of numeric scales is fairly straightforward. The respondents who were presented with the 0 to 10 scale were more likely to choose the lower scale points than the respondents who were presented with the -5 to +5 scale. This is compatible with the results of experiment 1, in line with the results of other researchers, and is what the mean scores and the ANOVA results convey.

Table 2.
The percentages and means for the Advertising Standards question (Study II).

	Conditions				
	'not any more'		'much more'		
	0→10	-5→+5	0→10	-5→+5	
scale	Percentages				total %
10 or +5	16	16	15	17	16
9 or +4	3	6	3	7	5
8 or +3	8	14	10	13	11
7 or +2	10	14	9	14	12
6 or +1	10	5	10	6	8
5 or 0	21	19	30	27	24
4 or -1	4	3	5	1	4
3 or -2	6	3	5	4	4
2 or -3	3	4	2	4	3
1 or -4	3	1	4	1	2
0 or -5	16	15	7	6	11
total n	538	521	509	483	2051
$\bar{x}$ (0-10)	5.31	5.79	5.70	6.38	5.78

Main effects		
Verbal endpoint	$F(1,2047) = 13.4$	p=.00
Scale	$F(1,2047) = 18.6$	p=.00
Interaction	$F(1,2047) = 0.6$	p=.44

Comparing the distributions for the verbal anchors shows an entirely different picture, however. Comparing the two conditions given the unipolar verbal labels with the two conditions given the bipolar verbal labels we see the percentage frequencies are almost identical, with two exceptions. For the midpoint (5 for the 0 to 10 scale and 0 for the -5 to +5 scale) the percentage for those given the unipolar verbal scale is about 20%; the percentage for those given the bipolar verbal scale is about 29%. Conversely, for the lower endpoint (0 or -5) the percentage frequency is about 15% for the unipolar verbal scale and about 6% for the bipolar scale.

The implications of the distributional findings are quite strange. The impact of the bipolar verbal anchors is to increase (at least in the sample we observed) the percentage at the midpoint of the scale (the neutral point) at the expense of the lower endpoint of the scale. (Expressed in terms of the impact of the unipolar scale, we would say that the lower endpoint is favoured at the expense of the midpoint.)

There are two important qualifications to note. First, these are cross-sectional samples and are subject to sampling error. Second, these are between-subject comparisons and consequently we cannot say what the effect of the change would be on a particular individual or set of individuals. What we have is a comparison of the frequency distributions of the responses for two (nearly) independent samples for the two forms of the question. Clearly the reliability of such findings is important. The fieldwork was carried out over two weeks with a separate (balanced) sample interviewed in each week. This allowed us to check on the stability of the results. The frequency distributions were stable across weeks; the same conclusions would have been reached on the basis of either of the weeks taken alone.

Conclusions

The discussion in the literature of the effect of the labels attached to response scales has concentrated on the efficiency or level of discrimination obtained with different labels (in some cases expressed as the proportion of variance on the continuum explained by the scale). Schwarz et al (1991) extended the argument to include the possible impact of the numeric labels, and demonstrated in particular situations that such an effect could occur.

Our results suggest that the way in which numeric and verbal labels affect response patterns is different and that it is important to consider the whole distribution of responses on a scale rather than to confine analysis to summary measures of the distribution. In the first experiment we discovered that for the particular combination of words and numbers we used the respondents did not appear to use the zero value on the scale except when the numeric anchors were explicitly mentioned by the interviewer. We conjecture that a 0 to 10 scale is particularly vulnerable to this effect as the remaining numbers (1...10) would appear to form a perfectly reasonable (perhaps intuitively more plausible) scale. The results suggest that if the numeric labels are to appear on the scale then they should be signalled to the respondents so that all the respondents are subject to the same influences.

In the second experiment we addressed the issue of the appropriate combination of numeric and verbal labels. We considered that both the numeric and verbal labelling systems could be thought of as providing either a unipolar or bipolar framework to the respondent.

We found that both the numeric labels and the verbal anchors had an effect on the responses and their effects were of a comparable size. Further, there was no evidence of interaction; the effect appear additive. This suggests that verbal and numeric anchors may be altering the response processes in independent ways; indeed, when we examined the distributions of responses we discovered their effects on the response patterns were strikingly dissimilar.

The contrast between the two sets of numeric labels suggested that there was a consistent difference across categories. For the unipolar numeric labels the percentage of respondents in the lower half of the scale was larger than for the bipolar numeric labels. There was a consistently lower proportion in each of the lowest four categories and a consistently higher proportion in all but one of the highest categories for one scale in contrast with the other. This suggests a shift in location for the whole scale.

An examination of the distributions for the two verbal anchors revealed a completely different contrast. The percentages in each of the categories for the two versions was effectively identical with the exception of two scale positions - the midpoint (which had no verbal label) and the lower endpoint (which was the anchor which was changed). There was what appeared to be a direct transfer of a proportion of the respondents from the lower endpoint to the midpoint as a result of the change in the lower anchor. This suggests that the cues provided by the verbal labels are very different from those provided by the numeric labels.

References

Alwin, D.F., & Krosnick, J.A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research*, **20**, 139-181.

Dawes, R.M., & Smith, T.L. (1985). Attitude opinion and measurement. In G. Lindzey & E. Aronson (Eds). *Handbook of Social Psychology: Volume I Theory and Method*. Random House: New York. 509-566.

O'Muircheartaigh, C.A., Gaskell, G.D., & Wright, D.B. (1992). *When response alternatives affect survey results for vague behavioral frequency questions*. American Association of Public Opinion Research (AAPOR); Annual Conference, Florida, USA.

Schaeffer, N.C. (1991). Hardly ever or constantly? Group comparisons using vague quantifiers. *Public Opinion Quarterly*, **55**, 395-423.

Schwarz, N., Knäuper, B., Hippler, H.J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, **55**, 570-582.

Schwarz, N., Strack, F., Müller, G., & Chassein, B. (1988). The range of response alternatives may determine the meaning of the question: Further evidence on informative functions of response alternatives. *Social Cognition*, **6**, 107-117.

Smith, T.W. (1993). *An analysis of response patterns to the ten-point scalometer*. American Association of Public Opinion Research (AAPOR); Annual Conference, Peasant Run, Illinois.