Journal of Official Statistics, Vol. 11. No. 3, 1995, pp. 295–307

Weighing Anchors: Verbal and Numeric Labels for Response Scales

Colm O'Muircheartaigh, George Gaskell, and Daniel B. Wright[1]

Among the most commonly used measurement instruments in social and psychological research is the response scale – a question or statement with an accompanying set of response alternatives arranged on a numeric or verbal scale. Research has accumulated which demonstrates that the construction of the response scale may influence substantially the way in which respondents answer questions. We examine the effects on the responses of (i) numeric and verbal anchors – labels at the endpoints of the scale – and (ii) numeric labelling of scale positions. Two experiments, both conducted in large-scale surveys, are described. In the first we compare two sets of numeric labels (the ranges −5 to 5 and 0 to 10) for 11-point scales, and test whether explicit mention of the numeric anchors in the question stem modifies the effect of the labels. We find a clear effect for the scales, and a possible effect for mentioning the numeric anchors. In the second we vary both the numeric and the verbal anchors and partition the total effect according to source. We find significant effects for each factor but no interaction between them. For both experiments we observe that the characteristics of the scale (the numeric scale, uni- or bipolar anchors) lead to differences in the use of the midpoint and endpoints of the scales.

Key words: Response alternatives; measurement error; CASM; question wording; rating scales.

1. Introduction

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 instruments in social and psychological research. There is an implicit assumption that the attitude, belief, opinion, or behavioral frequency/intensity can be measured on a single latent or manifest continuum. Dawes and Smith (1985, p. 540) suggest that these response scales (rating scales) may be justified because they are "compatible with the ways in which people using them think"; for instance, since people may use a left-right continuum to describe political attitudes, 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 people's thinking. Studies by deSoto, London, and Handel (1965) on spatial paralogic show that people do think about some social phenomena in spatial terms. They argue, for example, that in some cases a unipolar (top-down) verbal ordering on a vertical axis may be easier to understand than a bipolar ordering from a central position.

[1] London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom.

Nisbett and Kunda's (1985) data provide support for the idea that respondents tend to think of distributions in general as being approximately normal, and that furthermore, respondents expect the actual distribution in the population to conform to the description of the scale that is presented. In their research when highly dispersed scale descriptions were presented, respondents guessed somewhat higher concentrations at the extremes than would be warranted by the empirical distribution. Whether this should be explained in terms of the availability heuristic (in that extreme examples will be relatively more available than their actual frequency would warrant), or in terms of a response alternative effect, cannot be determined using their data.

Alwin and Krosnick (1991) consider the properties of scales in terms of the reliability of measurement; among the characteristics examined are the number of scale points, the inclusion or omission of a midpoint, the extent and nature of the verbal labelling of response alternatives, and the explicit inclusion of a "don't know" response option. Schwarz, Strack, Müller, and Chassein (1988) demonstrated how, in controlled settings, the response alternatives presented can change the response profile substantially. They hypothesize that this may be the consequence of a *meaning shift* due, in the case of questions using vague terms, to the response alternatives' being used as information to disambiguate the question. Gaskell, O'Muircheartaigh, and Wright (1994) examined this finding in the context of national sample surveys, showing that the results apply, though somewhat diluted, to plausibly interchangeable sets of response alternatives. Schaeffer (1991) contrasted the results of offering apparently interchangeable verbal and non-verbal categories as response alternatives, and showed that for some sub-populations, the two gave contradictory results. Schwarz, Knäuper, Hippler, Noelle-Neumann, and Clark (1991) introduced a further element by considering the effect of the numeric labels assigned to the response options; their results are discussed in detail below. In this paper we consider the nature of the influence of both the verbal and numeric labels which appear on the response scale.

While the numeric values are often included for coding and response convenience, Schwarz et al. (1991) have demonstrated that they carry additional, 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 equivalent to a scale whose values ranged from −5 to +5. The verbal anchors they used 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 the characteristic, 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* (i.e., being a failure).

In our first experiment we explored the observation made by Schwarz et al. (1991) that, while the numeric labels are often chosen merely for administrative convenience, for various reasons some respondents may become aware of them and react to them.

In keeping with Grice's cooperative principle of conversation (Grice 1975) and applying the *quantity maxim*, respondents who notice the numeric labels may choose to use the information implied by the numeric labels in generating their answers. The information may be visual (on a showcard, for instance) or aural (if mentioned by the interviewer) or both. The more the labels are drawn to the attention of the respondents the greater is the proportion of the respondents who are likely to process the information and, having processed it, the more likely they are to use it in constructing their responses. Krosnick and Alwin's (1987) suggestion that respondents employ strategies of just sufficient cognitive effort in answering questions implies that respondents might treat scale numbering as merely ancillary information and ignore it. Cannell, Marquis, and Laurent (1977) show, however, that given the same content, a longer question form elicits more complete reporting than a shorter question. Filling in details of the response alternatives might therefore signal to the respondent that the question requires more careful consideration.

We presented respondents with a showcard containing the response scale for all our questions; the response scales carried numeric labels on all the scale points. We test whether explicitly mentioning the numeric anchors in the question stem changes the effect of different numeric scales; this permits quantifying the effect of the oral/aural signal. The specific form of the lower verbal anchor used by Schwarz et al. (1991) – *not at all successful* – might be considered to be relatively ambiguous in that it is susceptible to interpretation either as "lacking success" or as "being a failure" and might thus be vulnerable to change in interpretation with small changes in format. We used a question with explicitly bipolar verbal anchors – *much more* and *much less*.

In a second experiment we compared the effect of the verbal descriptions with the effect of the numeric anchors. The information that suggests to the respondent whether the response scale is unipolar or bipolar is of two types. The usual way to denote a unipolar scale is to label one anchor of the scale with a null verbal label and the other with a clearly directional verbal label: (*no more power, much more power*), (*not having any success, having great success*), (*none, a lot*), etc. An alternative would be to use a null numeric label as one anchor and a positive or negative number as the other: (*0, 10*), (*0, 5*), (*0, 6*), (*−5, 0*), etc. Similarly a bipolar scale can be signalled either by the use of verbal opposites as the anchors – (*much more power, much less power*), (*much success, much failure*), (*strongly agree, strongly disagree*) – or by using positive and negative numeric anchors: (*+5, −5*), (*+3, −3*), (*+2, −2*). Our interest was in the possibility of an interaction between the verbal and numeric cues presented to the respondent.

Much of the literature in this area treats the responses (not unreasonably) as metric, and consequently presents the results in the form of mean scores or as the proportion above or below some arbitrary scale point for the different scale versions. Part of the reason for this is the small sample sizes used in many of the experiments; the published results are in general inadequate for a detailed examination of the response distributions. We feel that it is important to examine the whole distribution of responses as the influence of differences in the labels may be more subtle than a simple shift in the mean score.

2. Fieldwork Methodology

These experiments were embedded in British Market Research Bureau International's (BMRBI) face to face omnibus survey (ACCESS) and were conducted with BMRBI's assistance. Each week BMRBI 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 BMRBI, typically about 15 minutes into the 25–30 minute interview. We receive for each week from BMRBI 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 version of the questionnaire within the week. In these cases our experimental conditions were randomised within the versions. For the experiments we report there were no topics preceding our questions which might be expected to affect our results.

BMRBI's omnibus survey, which draws respondents aged 15 years and older from Great Britain (with the exception of some 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 sub-classes 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 analyses 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 correctly implemented. The nature of the sample design makes it impossible to calculate response rates, but the distribution of the sub-samples 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.

The fieldwork for each experiment was carried out over two weeks with a separate (balanced) sample interviewed in each week. In order to provide an additional check on the stability of the results the data were analyzed separately for each week. The frequency distributions were stable across weeks and the same conclusions would have been reached on the basis of either of the weeks taken alone; we therefore present the analysis combining the two weeks.

In both the studies the response scale was presented on a showcard. Therefore our results apply only in situations where the response scale is presented visually to the respondents, either as a showcard in a face to face interview or possibly as a response scale in a self-completion context. This contrasts with the use of unfolding or branching techniques, in telephone interviews or in face to face interviews without showcards, in which the position of the response on the scale is arrived at by a sequence of successive approximations.

3. Does the Scale Need to Be Mentioned?

Respondents ($n = 2124$) in BMRBI's July and August 1992 omnibus survey (a face to face general population survey of Great Britain) 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 had the same verbal anchors: *much more entertaining than the programmes* 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 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 the analysis the data were recoded so that the range of scores for both scales was from 0 to 10, where 10 corresponds to *much more*. The means and standard deviations of the four conditions and the significance tests comparing the means are shown in the lower part of Table 1. There is a clear main effect for the numeric scale: respondents given the 0 to 10 scale were more likely to favour the adverts ($F(1,2057) = 14.16$; $p < 0.01$). 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 and the significance level is marginal ($F(1,2057) = 3.52$; $p = 0.06$). There was no evidence of an interaction between the use of different anchors and whether they were mentioned ($F(1,2057) = 0.05$; $p = 0.82$).

The frequency distribution of the responses for each of the four conditions is also given in Table 1. Though the four distributions show some interesting variations, one cell in particular deserves attention. For the {0..10} numeric labels where the numeric anchors are **not** explicitly mentioned by the interviewer there were **no** responses in the bottom category (clearly this result held in both weeks). This contrasts with about 10% (about 50 people) for each of the other three conditions. It is possible that this phenomenon 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 caused the respondents to neglect the zero in favour of what mathematicians call the "natural" numbers. A post hoc χ^2 test (since we had not hypothesized this difference in advance) gives a value of $\chi_1^2 = 35.63$; $p < 0.001$.

Though there are some other suggestive differences among the distributions of the responses for the four conditions the distorting effect of the empty cell makes it difficult to reach any firm conclusion. Two points may, however, be worth noting. First, there is a tendency for the midpoint of the scale to be treated differently for the two numeric scales; when the midpoint is labelled 0, fewer people choose it. We may hypothesise that this is because people wish to avoid the appearance of sitting on the fence by choosing a zero point on a negative-positive scale, whereas they are happy to be "moderate," i.e., near the centre of a continuous 0,10 scale. Second, the variance of the responses is lower in the case of the numeric unipolar scale (Cochran's $C = 0.59$; $p < 0.001$ and Bartlett-Box $F = 31.6$; $p < .001$); this results partly from the different treatment of the midpoint and partly from the non-use of the lowest category for the unipolar numeric scale where the anchors are not signalled.

4. Comparing the Effects of Verbal and Numeric Anchors

Our next experimental question was also embedded in BMRBI'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 to 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 mentioned explicitly by the interviewer to all respondents.

Here we compare the relative effects of numeric and verbal anchors on the use of a response scale. Each of these may be signalling to the respondent that the scale is unipolar or bipolar. A range of values from 0 to 10 may suggest that the researcher's

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	ϕ	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
SD ($\hat{\sigma}$)	2.41	2.96	2.47	2.86	2.67

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

interest is in the amount of the attribute present, with a lower bound meaning the absence of the attribute – a unipolar conceptualisation. A range of values from -5 to $+5$, on the other hand, may suggest that the researcher conceives of the presence either of an attribute or its opposite – a bipolar conceptualisation. Verbal anchors such as *much more, much less* or *agree, disagree* similarly suggest bipolarity, whereas anchors such as *much more, no more* or *agree, do not agree* suggest unipolarity. One might expect that certain combinations of verbal and numeric anchors would be more natural or less awkward than others. Explicitly bipolar verbal anchors might be expected to fit more easily with bipolar numeric anchors; similarly, unipolar verbal and unipolar numeric anchors might be expected to combine well. Conversely, a mixture of bipolar and unipolar cues might be expected to cause some difficulty to the respondent. Therefore, we might expect to find an interaction between the polarities of the two sets of cues.

Table 2 gives the frequency distributions, means, standard deviations, and an analysis of variance of the results. As with experiment 1, responses for the comparative analysis were recoded so that each was based on a comparable 0 to 10 scale.

When measured in terms of the mean scores obtained on the different scale versions, each of the factors (numeric and verbal) had a significant effect on the responses. The magnitudes of the effects are similar and, more interestingly, the effects appear to be additive. This suggests that the verbal and numeric anchors may be tapping different but complementary aspects of the scale. It is somewhat surprising that there is no evidence of an interaction effect.

The mean score (for a 0..10 scale) differs by about 0.5 for the two sets of verbal anchors; in percentage terms, we can see that the percentage giving answers below the midpoint is 30% for the unipolar scale, but 20% for the bipolar scale. For the numeric anchors, the corresponding mean difference is about 0.6; the percentage giving answers below the midpoint is 28% for the unipolar anchors, and 21% for the bipolar anchors. This representation of the results, however, conceals the particular way in which the differences are brought about by the two factors.

Looking at the variances of the responses across the four conditions both Cochran's C and the Bartlett-Box F tests reject the null hypothesis of homogeneity of variances (Cochran's $C = 0.29$; $p < 0.005$; Bartlett-Box $F = 9.2$; $p < 0.001$). In this experiment, however, there is no difference between the variances obtained for the two conditions with the $\{0..10\}$ numeric labels and the two conditions with the $\{-5..+5\}$ numeric labels (Cochran's $C = 0.50$; $p = 0.72$ and Bartlett-Box $F = 0.128$; $p = 0.72$). This contrasts with the result from experiment 1, but is not unexpected given that the numeric anchors were mentioned in the question stem. The difference in variances is entirely accounted for by the contrast between the two different verbal anchors (Cochran's $C = 0.57$; $p < 0.001$ and Bartlett-Box $F = 25.7$; $p < 0.001$). To explain this, and to understand the nature of the influence of the numeric and verbal anchors, a more detailed examination of the full frequency distributions is needed. A Kolmogorov-Smirnov test shows significant differences between the cumulative distributions, both for the contrast of the numeric scales ($\chi^2_2 = 16.9$; $p < 0.001$) and the contrast of the verbal anchors ($\chi^2_2 = 10$; $p < 0.01$).

The comparison of numeric scales is fairly straightforward. The respondents who were presented with the {0..10} scale were more likely to choose the lower scale points and less likely to choose the higher scale points than the respondents who were presented with the {−5..+5} scale. This difference in preferences is fairly evenly distributed across the scale; as there is scope for movement throughout the scale (there is no heaping at the boundaries) the variance is unaffected. The overall pattern of effects is compatible with the results of experiment 1, in line with the results of other research, and is evidenced in the mean scores and the ANOVA results.

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 that the percentage frequencies in the comparable scale categories are *almost* identical. There are two notable exceptions. For the midpoint (5 or 0) the frequency for the bipolar verbal scale is about 30%; the frequency for the unipolar verbal scale is about 20%. Conversely, for the lower endpoint (0 or −5) the percentage frequency is about 6% for the bipolar verbal scale and about 15% for the unipolar scale.

The implications of the distributional findings are disquieting. The effect 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 effect of the unipolar scale, we would

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

	Conditions				
	"not any more"		"much less"		
	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
SD ($\hat{\sigma}$)	3.27	3.24	2.80	2.73	3.05

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

say that the lower endpoint is favoured at the expense of the midpoint. As the bipolar anchors increase the concentration at the midpoint of the distribution the variance of the responses is less in these conditions. These observations are consistent with the findings of Wildt and Mazis (1978) who show that in addition to response labels, category position may influence scale response.

There are two important qualifications to note. First, these results are obtained from samples and are of course subject to sampling error. Second, the analyses are between-subject comparisons and consequently we cannot say what the effect of the change would be on a particular individual. 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.

5. Conclusions

Underlying any scale which is presented visually to the respondent – either as a show-card in a face to face interview or as a response scale in a self-completion context – is an implicit assumption that the spatial positioning of the points on the scale conveys some information to the respondent. The conventional presentation of the points on the scale as *equidistant* from each other accentuates this impression. There is a further implicit assumption in the numeric labels attached to the scale points, realized in the way in which such data are analysed; an underlying metric is assumed, and the data are often treated as interval measures.

This is sometimes justified on the basis that "any statistic computed on a set of numbers correctly is, in fact, correct as a description of these numbers" (Dawes and Smith 1985, p. 533), or that "the validity of the statistical test cannot depend on the type of measurement scale used" (Anderson 1961, p. 309). The value of the numbers as predictors or representations is separate from the statistical assumptions underlying the analyses in which they are used. Even if we relax the assumptions, however, and treat the data as merely ordinal, the results may well be interpreted as conveying some quantitative meaning. In Schwarz et al. (1991, p. 570), for instance, different scale labels are contrasted by observing that "... 34% of the respondents endorsed values between 0 and 5 ... only 13% endorsed *formally equivalent* values [for the other scale labels] between −5 and 0" (our italics). Our results suggest that it is important to go beyond this and to consider the *whole distribution* of responses on a scale rather than to confine analysis to summary measures of the distribution.

Both the numeric and the verbal labels convey information to the respondent about the meaning of the scale points. In our experiments we have used a full set of numeric labels (11) but only two verbal labels (the upper and lower anchor points); this is a common convention in market and social research interviews.

The first experiment confirmed that a change in the numeric labels could induce a shift in the frequency distribution of the responses. Furthermore, we found some evidence of an additional effect when the interviewer explicitly mentioned the numeric anchors while asking the question. On inspection we discovered that, for the particular combination of words and numbers we presented, none of the respondents used the zero value on the scale except when the numeric anchors were explicitly mentioned.

We conjecture that a {0..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. In a separate experiment using {0..6} and {1..7} scales we found a comparable effect, though a weaker one that than found for the {0..10} vs {1..11} comparison (O'Muircheartaigh, Wright, and Gaskell 1993).

Grice's (1975) cooperative principle of conversation implies that if the numeric labels appear on the showcard and are noticed by respondents, they may choose to use the information in generating their answers. Cannell et al. (1977) have demonstrated, for a given content, longer questions signal greater importance and may encourage a greater degree of attention by the respondent. Together with our results this suggests that if the labels appear on a showcard and *may* be seen by respondents, it is important to ensure that they are seen by all respondents, so that all the respondents are subjected, insofar as possible, to the same influences. We would argue that the numeric anchors should therefore be mentioned explicitly by the interviewer.

In the second experiment we addressed the issue of the appropriate combination of numeric and verbal labels. Statistically, each of the factors – verbal anchors and numeric labels – was found to have a main effect. We considered that the numeric and verbal labelling systems could each be thought of as providing a unipolar or bipolar framework to the respondent. We felt that there would be a natural concordance or congruence between certain verbal and numeric anchors; in particular, a bipolar verbal scale would match best with a bipolar numeric scale, and the unipolar scales would similarly reinforce each other. A contrast between the form of the two scales would lead to the respondent's receiving conflicting messages and might lead to confusion, a form of *cognitive dissonance*. We would then expect to find a statistical interaction. The data showed no evidence of any such interaction. This suggests that the words and numbers are being processed independently by the respondents, and either that the words and numbers are not being checked for consistency, or that they do not appear inconsistent to the respondents.

In the case of the verbal labels the distinction between the two unipolar and bipolar scales is clear. We argue that positions between the second pole and the neutral point are incorporated in the lower anchor for the unipolar scale and thus the other unipolar labels are spread along the other half of the possible response space. The situation for the numeric labels is less clear a priori; these have typically been seen as merely convenient codes or recodes of the data to facilitate processing and analysis. The Schwarz et al. (1991, p. 570) conjecture is that the "respondents use the numeric labels to disambiguate the meaning of scale labels, resulting in different interpretations and, accordingly, different subjective scale anchors."

When we examined the distributions of responses we discovered a rather interesting picture. The contrast between the two sets of numeric labels suggested that there was a consistent difference across categories. The proportion in each of the four highest categories was consistently greater and the proportion in all but one of the other six categories was consistently less for the (−5, +5) scale than for the (0, 10) scale. This suggests a shift in location for the whole scale. This is in keeping with other results that show that respondents are more willing to grant positive values than negative values to labels (Bartram and Yelding 1973); the different implications of positive and negative verbal labels was confirmed by Worcester and Burns (1975). The findings of Schwarz et al. (1991, p. 572) on the effect of negative numeric labels in relation to success in life and childhood happiness are similar, but may reflect also the additional reluctance of respondents to apply negative scores to themselves or to other people. They also suggest that "... the numeric labels ... influenced respondents' interpretation of the endpoint labels." If this were the case we would expect a change in the verbal endpoint labels (the verbal anchors) to have the same type of effect as a change in the numeric labels.

The change in verbal anchors, however, appeared to affect the frequencies in only two scale positions – the midpoint (which had no verbal label) and the lower endpoint (which was the anchor that was changed). There was what appeared to be a transfer of a proportion of the respondents from the lower endpoint to the midpoint as a result of the change in the lower anchor (subject to the qualifications expressed in Section 4). This suggests that the cues provided by the verbal labels are very different from those provided by the numeric labels. This selective effect of the verbal anchors may in part be due to the failure to provide a full set of eleven verbal labels. If so, the implications for the labelling of response scales in general are serious. Krosnick and Berent (1993) demonstrate for a variety of political questions that fully labelled branching (unfolding) questions provide consistently more reliable (test-retest) measures than partially labelled non-branching questions (a term that describes the type of question we used in this study).

Though full verbal labelling *may* be preferable in terms of variance explanation (Peters and McCormick (1966) and Zaller (1988) support this, though Andrews (1984) found otherwise), it has generally been assumed that verbal anchors should be sufficient in terms of identifying the range of the response space. Furthermore, there is a serious practical problem in establishing agreed verbal labels for the intermediate points in all but the simplest (three-point) bipolar agreement/disagreement scales.

A possible explanation for the asymmetric nature of the findings is that the full set of numeric labels prevented the unipolarity expressed by the verbal anchors from being taken on board by the respondents. It would appear that the unipolar numeric labels do not truncate the scale (or alternatively that the bipolar numeric labels do not extend the scale).

Finally, the scales used in our questions, and indeed the scales used in most survey questions, are not in any sense absolute or unambiguous. It may be that the terms used – such as much less power, no more power, unsuccessful, not at all successful – are interpreted by the respondents to represent reasonable limits for the responses. They assume that the interviewer is presenting them with the appropriate range of choices, and therefore look for cues to guide them towards a sensible response. The numeric labels, and the positions on the scale, may then assist in the fine tuning of responses rather than lead to a transformation of the underlying scale. We will explore this issue in future work.

6. References

Alwin, D.F. and Krosnick, J.A. (1991). The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes. Sociological Methods and Research, 20, 139–181.

Anderson, N.H. (1961). Scales and Statistics: Parametric and Nonparametric. Psychological Bulletin, 58, 305–316.

Andrews, F.M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. Public Opinion Quarterly, 48, 409–442.

Bartram, P. and Yelding, D. (1973). The Development of an Empirical Method of Selecting Phrases Used in Verbal Rating Scales: A Report on a Recent Experiment. Journal of the Market Research Society, 15, 151–156.

Cannell, C.F., Marquis, K.H., and Laurent, A. (1977). A Summary of Studies of Interviewing Methodology. Vital and Health Statistics (DHEW Publication No. HRA 77-1343, Series 2, No. 69). Washington, DC: U.S. Government Printing Office.

Dawes, R.M. and Smith, T.L. (1985). Attitude Opinion and Measurement. In G. Lindzey and E. Aronson (eds). Handbook of Social Psychology: Volume I, Theory and Method (509–566). New York: Random House.

deSoto, C.D., London, M., and Handel, S. (1965). Social Reasoning and Spatial Paralogic. Journal of Personality and Social Psychology, 2, 513–521.

Gaskell, G.D., O'Muircheartaigh, C.A., and Wright, D.B. (1994). Survey Questions about the Frequency of Vaguely Defined Events: The Effects of Response Alternatives. Public Opinion Quarterly, 58, 241–254.

Grice, H.P. (1975). Logic and Conversation. In P. Cole and J.L. Morgan (eds.). Syntax and Semantics: Speech Acts (Vol. 3, 41–58). New York: Academic Press.

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.

Krosnick, J.A. and Alwin, D.F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. Public Opinion Quarterly, 51, 201–219.

Nisbett, R.E. and Kunda, Z. (1985). Perception of Social Distributions. Journal of Personality and Social Psychology, 48, 297–311.

O'Muircheartaigh, C.A., Wright, D.B., and Gaskell, G.D. (1993). The Logic and Paralogic of Numeric, Verbal and Spatial Response Scale Construction. LSE Methodology Institute Technical Report No. 10.

Peters, D.L. and McCormick, E.J. (1966). Comparative Reliability of Numerically Anchored Versus Job-Task Anchored Rating Scales. Journal of Applied Psychology, 50, 92–96.

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., and 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., and 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.

Wildt, A.R. and Mazis, M.B. (1978). Determinants of Scale Response: Label Versus Position. Journal of Marketing Research, 15, 261–267.

Worcester, R.M. and Burns, T.R. (1975). A Statistical Examination of the Relative Precision of Verbal Scales. Journal of the Market Research Society, 17, 181–197.

Zaller, J. (1988). Vague Minds vs Vague Questions: An Experimental Attempt to Reduce Measurement Error. Paper presented at the annual meetings of the American Political Science Association, Washington, DC.

Received September 1993
Revised July 1995