Why use 11-point scales?

Annette Scherpenzeel*

Despite the number of well-known problems, category scaling is one of the most commonly used measurement methods in questionnaires. Largely out of habit and because there are no alternative scales. For many questions of the Swiss Household Panel questionnaire, the 11point scale has been chosen instead of a category scale. The 11point scale is used in many other ongoing surveys, for example the Gsoep and World Value Study, and seems to be well handled by respondents. Respondents are asked to indicate the strength of their attitude or opinion in a number between 0 and 10, with the endpoints 0 and 10 being defined by verbal labels. This type of scale is often called a "number production scale" and exists in various forms: In addition to numbers between 0 and 10, people can also be asked to use numbers between 1 and 7, between 0 and 100, and even between 0 and 1000. Number production scales are sometimes considered to belong to the class of magnitude estimation methods, to which also belong graphical scales and line production scales (Lodge, 1981; Van Doorn, Saris, and Lodge, 1983).
The main arguments in favour of this type of scale are:

1. Minimisation of categorisation effects.
2. Improvement of data analysis
3. Reliability of the data (less measurement error)

In addition, the reasons why this type of scale is especially suitable for CATI are:
4. Time saving
5. No response-order biases

We will now give a short description of each of these arguments.

1. Prevention of categorisation effects

We assume that attitudes fall along a single, latent continuum, ranging from positive to negative. The larger the number of points on a response scale, the better it represents this underlying, latent continuum and the more accurate it reflects the variation. Scales with relatively few response alternatives force respondents to categorise their reaction towards an attitude object instead of directly mapping it onto the response continuum, thus causing information loss. Early research has already shown that respondents differentiate more between objects when offered response scales with greater numbers of categories (Bendig, 1954; Garner, 1960). The larger the number of points, the more powerful the scale is in discriminating, but at a certain point respondents become unable to make fine distinctions and thus round off. Thus, on 100 point number production scales, most answers are at 25, 50 and 75 or end in zero (e.g., 30, 40, 60).

* Correspondence to: Swiss Household Panel, FORS, University of Lausanne, Bâtiment Vidy, CH-1015 Lausanne
Tel. +41 (0)21 692 37 30 Fax. +41 (0)21 692 37 35 E-mail: swisspanel@fors.unil.ch

2. Improvement of data analysis

Basically there are two ways to improve the quality of data analysis: adapting statistical data analysis techniques to a low level of data measurement or improving the measurement procedures. In the last decades many efforts have been made to apply the aforementioned strategy in order to develop appropriate analysis techniques for categorical or lower level data. The latter strategy involves a completely different, non-statistical approach, which focuses on the optimisation of the measurement procedures (Batista and Saris, 1992). In their investigation of the possibilities to optimise measurement procedures in social science, Van Doorn, Saris, and Lodge (1983) did not simply enlarge the number of scale points, but used psychophysical scaling (see also Lodge, 1981). Respondents expressed their answers on continuous scales by drawing lines or assigning numbers to their opinions, thus creating interval level measures. Line production is in this respect the best response modality currently available. However, it requires some training of respondents, it has a bothersome coding process in the case of none computer assisted interviews, and cannot be used in CATI. Graphical scales can give very good results, but their use is also restricted to certain modes of datacollection. In conclusion, the best alternative to category scales within the class of magnitude estimation scales are the number production scales.

Nevertheless, it is essential that a magnitude estimation scale has fixed anchors, or reference points. The 11 point number scale used in the panel questionnaire has, for example, two reference points, 0 and 10. These reference points have been given labels that clearly indicate the end point of the scale, for example: 'completely satisfied' and not, for example: 'very satisfied'. Terms like 'very' or 'good' and 'bad' have no fixed position on a subjective scale. Some people might see 'good' as an extreme of the scale, while others think of it as a middle position at the positive side of the scale. Consequently, the use of these terms as reference points can cause individual variation in response functions (Saris and De Rooij, 1988). The terms 'completely' and 'entirely', on the contrary, clearly indicate the end point of the scale for everybody. There can be no doubt about their position on the scale. Scales with two or more reference points and clear labels that fix the end points have proven to decrease the measurement error that can result from variation in response functions (Saris and De Rooij, 1988).

3. Reliability of the data (less random error)

Another argument is the effect of measurement error, or the reliability of the data. Consider an example described by Alwin and Krosnick, (1991):

> "A question offers respondents three response alternatives indicating their favorability toward a government policy: "favor," "neither favor nor oppose," and "oppose". A respondent whose attitude is extremely favorable or unfavorable should readily select one of the extreme alternatives. And a respondent who has neither favorable nor unfavorable feelings would presumably choose the middle alternative. However, a respondent with a relatively weak favorable or unfavorable attitude is confronted with a difficult decision. She or he must choose either the middle alternative, thereby giving the incorrect impression that she or he has no preference or is uncertain, or she or he must choose one of the extreme alternatives, giving the impression that she or he has stronger feelings than is in fact the case. Choices made by such

> respondents when confronted with too few response categories may very likely be random".

This reasoning supports the claim that scales with more response alternatives will be more reliable than those with fewer. It is often stated that the reliability of scales increases with the number of points used. There is probably a limit to the benefit of adding response categories or scale points. The relation between the number of scale points and reliability may be curvilinear. The existing research literature on this issue indicates that 7- to 10-point scales may be the most reliable (Andrews and Withey, 1976; Andrews 1984; Alwin and Krosnick, 1991; Rodgers et al. 1992; and Költringer, 1993) An international study of satisfaction, across 10 different countries, showed that the 11 point scale was the most valid and reliable scale of all scales included in the study (Scherpenzeel and Saris, 1995).

Alwin and Krosnick (1991) found the 2-category scale to be a major exception to this pattern, having relatively reliable responses. They suspected this was because 2-category questions usually measure the direction of attitudes only, with no pretension of measuring intensity, whereas 4 and more category response scales presumably are intended to measure both direction and intensity. The direction of attitude responses may in fact be more reliably assessed than the intensity.

4. Time-saving in telephone interviews

The number production scales do not consist of lists of alternatives that all have to be read aloud in a telephone interview. Instead, *only* the first and end point are read aloud and respondents are asked to produce a response alternative themselves. This takes considerably less time than reading lists of fully labelled categories, especially when we choose long category scales to prevent the categorisation effects described earlier. In fact, fully labelled category scales were originally designed for self-administered questionnaires, where respondents have enough time to read and process each category (see also point 5).

5. No response-order effects in telephone interviews

A great deal of past research has documented these effects, which are defined as changes in answers to closed-ended (category) survey questions produced by varying the order in which the response categories are presented (Krosnick and Alwin, 1987). Two sorts of response-order effects have been discovered: *primacy* and *recency* effects. Primacy effects occur when placement of an response category at the beginning of a list increases the likelihood that it will be selected. Recency effects are those that occur when placement of a category at the end of a list increases the likelihood that it will be chosen (Schuman and Presser, 1981). The nature of effects depends in part on whether response scales are presented visually or orally to respondents.

When response scales are read aloud to respondents, they are not given the opportunity for extensive processing of the first categories offered. Presentation of the next category terminates processing of the one before relatively quickly. Under these circumstances, respondents are able to devote most processing time to the *final* category(s) read, since interviewers usually pause most after reading them (Krosnick and Alwin, 1987). As a result, a recency effect can be expected. When response options are read aloud to respondents, memory biases may also influence responses. Categories presented early in a list are most likely to enter long-term memory, and

categories presented at the end of a list are most likely to be in short-term memory immediately after the list is heard (Krosnick and Alwin, 1987). So response alternatives presented at the beginning and end of a list may be more likely to be recalled and therefore perhaps selected more often. When no visual aids are presented and when the list is long, memory effects may be important (Schuman and Presser, 1981). Krosnick and Alwin (1987) expect these effects to be more pronounced among individuals whose memories are less effective or who concentrate less on what the interviewer says. We expect these effects to be also more pronounced for verbally labeled response categories than for number production scales. The verbal category scales constitute lists of alternatives which often seem nearly equally suitable. The number production scales do not consist of lists of alternatives. Instead, *only* the first and end point are read aloud and respondents are asked to produce a response alternative themselves. Since CATI is exclusively oral, verbal category scales are likely to suffer from the response-order biases. Therefore, number production scales are more appropriate in CATI.

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.

Andrews, F.M. (1984). Construct validity and error components of survey measures: a structural modelling approach. *Public opinion quarterly*, 48, 409-422.

Andrews, F.M. and Withey, S.B. (1976). Social indicators of well-being: American's perceptions of life quality. New York, Plenum Press.

Batista-Foguet, J.M. and Saris, W.E. (1992). A new measurement procedure for attitudinal research. Analysis of its psychometric and informational properties. Quality and Quantity, 26, 127-146.

Bendig, A.W. (1954). Transmitted information and the length of rating scales. *Journal of Experimental Psychology*, 47, 303-308.

Garner, W.R. (1960). Rating scales. Discriminability and information transmission. *Psychological Review*, 67, 343-352.

Költringer, R. (1993). Messqualität in der sozialwissenschaftlichen Umfrageforschung. Endbericht Project P8690-SOZ des Fonds zur Förderung der wissenschaftlichen Forschung (FWF), Wien.

Krosnick, J.A., Alwin, D.F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51, 201-219.

Lodge, M. (1981). Magnitude scaling: quantitative measurement of opinions. Sage University Paper series on Quantitative Application in the Social Sciences, 07-025, Beverly Hills: Sage.

Rodgers, W.L., Andrews, F.M. and Herzog, A.R. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, 8, 251-275.

Saris, W.E. and De Rooij, K. (1988). What kind of terms should be used for reference points? In: Saris, W.E. (ed.). *Variation in response functions: A source of measurement error in attitude research*. Amsterdam, Sociometric Research Foundation.

Scherpenzeel, A.C. and Saris, W.E. (1995). The quality of indicators of satisfaction across Europe: A meta-analysis of multitrait-multimethod studies. In: Scherpenzeel, A.C. *A Question of Quality. Evaluating survey questions by multitrait-multimethod studies*. Dissertation. Royal PTT Nederland NV, KPN Research.

Schuman, H and S. Presser (1981). *Questions and answers in attitude surveys: experiments on question form, wording and context*. Academic Press, New York.

Van Doorn, L., Saris, W.E. and Lodge, M. (1983).Discrete or continuous measurement: What difference does it make? *Kwantitatieve Methoden*, 10, 104-120.