Articles

Should I label all scale points or just the end points for attitudinal questions?

Aaron Maitland*

Tags: survey practice

DOI: 10.29115/SP-2009-0014

Survey Practice

Vol. 2, Issue 4, 2009

Should I label all scale points or just the end points for attitudinal questions?

The decision about whether to label all response scale points or just the end points for attitudinal questions can be an important and vexing decision for question designers. This short article will first discuss theoretical and practical considerations that should help guide the decision about how to label response scale points. Next, I will discuss some of the important empirical findings on this question from the literature. I will also provide some advice about how to evaluate verbal labels.

The amount of clarity that the labels add to the response scale is the most important consideration in the decision to label scale points. The approach to labeling should be the one that most clearly defines the response scale for respondents. One might argue that labeling of all scale points might offer an advantage in this regard. Several authors concede that it is probably more natural for a person to express his or her opinion using words (Fowler 199 5; Krosnick and Fabrigar 1997). However, there is inherent ambiguity in the verbal labels that are frequently used with response scales. For example, people might have different interpretations of what it means to "somewhat favor" a public policy. Conversely, one might argue that even though numbers might be more abstract for most respondents, they might also be more accurate. For example, it has been noted that numbers in response scales at least convey the idea of equal intervals between points on a response scale (Krosnick and Fabri gar 1997). However, respondents can also vary in how they interpret numbers. Schwarz et al. (1991) present evidence that respondents interpret 10 point scales from –5 to +5 quite differently than 10 point scales ranging from 1 to 10. They found that negative numbers imply the opposite of something, whereas the low end of a scale with all positive numbers merely implies the absence of something.

There are many other aspects of the research design that might influence the decision to label scale points. For example, the length of the scale will determine

* **Institution:** National Center for Health Statistics

the feasibility of labeling all points. It will be much easier to create labels for 5 point scales than it will be for 11 point scales. As Fowler (1995) writes, "it is difficult to think up adjectives for more than 5 or 6 points along most continua." The mode of the interview also influences whether or not all scale points can be labeled. Generally, the use of telephone interviewing encourages shorter scales so that the respondent does not have to listen to a long list of response options before answering a question. Furthermore, when longer scales are used in telephone surveys it is more common to label only the endpoints of the scale (Dillman, Smith, and Christian 2008). Data collection methodologies that utilize visual modes of communication can more easily incorporate a full set of labels for response scales.

There are particular challenges with using verbal labels in cross-cultural research. Adequately translating fully labeled verbal scales into other languages is extremely difficult and can require considerable resources. For this reason, some (e.g., Fowler 1995) have suggested that numeric scales might be easier to translate and thus better suited for cross-cultural surveys since the researcher would only need to translate anchors at the ends of a scale. However, there is a dearth of evidence that numeric scales create more comparable survey data and cultural factors can also influence the interpretation of the numbers in a scale (Smith 2003).

Several empirical studies (Alwin 2007; Krosnick and Fabrigar 1997; Saris and Gallhofer 2007) have examined the effect of fully labeling scales versus partial labeling of scales on the quality of the resulting data. The general consensus from these studies is that fully labeled scales produce better data quality than partially labeled scales. Krosnick and Fabrigar (1997) came to this conclusion based on a review of several studies that examined the reliability and validity of fully versus partially labeled scales. Alwin (2007) compared the longitudinal reliability of 26 seven-point scales with only the endpoints labeled with 11 fully labeled seven-point scales from the National Election Studies. This study found that the fully labeled scales had a reliability of .719, whereas the scales with only the endpoints labeled had a reliability of .506. Saris and Gallhofer (2007) also concluded that labeling all points had a positive effect on reliability based on a meta-analysis of 1023 survey questions from 87 multi-trait multi-method experiments. Labeling did not have a significant effect on validity according to their meta-analysis.

There are a couple of other reasons to believe that labeling might lead to better data quality. First, improvements in both reliability and validity tend to be the greatest amongst respondents with lower levels of education (Krosnick an d Fabrigar 1997). This is a group that frequently encounters comprehension problems in surveys and question designers are often looking for design strategies to improve data quality among this group. Second, fully labeling scales seem to reduce the effects of question features that are unrelated to the response task. For example, a study by Tourangeau et al. (2007) conducted

experiments using Web surveys that varied the color of the shading on response scales and the use of labels. They found that the effect of color on respondents answers – something designers would not intend – disappeared when the response scales were fully labeled. Hence the literature reviewed in this article seems to suggest that fully labeling has a number of advantages over labeling only the endpoints.

The decision about which labels to use is just as important as the decision about whether or not to use the labels at all. Saris and Gallhofer (2007) meta-analysis offers some guidelines for selecting labels. Generally, symmetrical labels tend to yield higher reliability and validity. Verbal labels that match the numbers on the scale lead to higher reliability. For example, it might be better to have bipolar scales with negative labels matching negative numbers and positive labels matching positive numbers. Finally, if one does decide to label only the endpoints, these labels should represent fixed reference points (e.g. completely dissatisfied – completely satisfied, instead of dissatisfied-satisfied) so that the end of the scale is well defined.

Even though there are some guidelines, it is always necessary to evaluate a question to determine the best labeling. There are a number of question evaluation methods that might be useful for designing scale labels. Qualitative techniques such as cognitive interviewing that make use of probing or thinkaloud methods will help to understand how respondents assign meaning to the labels on a response scale and whether that meaning matches the survey designer's intended meaning. Item response theory modeling approaches provide a useful quantitative tool to assess whether the pattern of responses matches the theoretical underpinnings of the response scale. For example, one can assess whether the points on the scale are increasingly difficult to endorse as one expresses more extreme attitudes. It is always good practice to use a combination of qualitative and quantitative methodologies such as these to design good survey questions.

In conclusion, the approach to labeling should be the one that most clearly defines the response scale for respondents. The existing empirical literature generally suggests that fully labeled scales have an advantage over partially labeled scales according to a number of criteria. However, practical considerations such as the mode of the interview, number of scale points, and the population under study can play an important role in deciding the extent of labeling that should be used. Additionally, question designers should rely on appropriate question evaluation methods to determine which labeling approach is best for a specific research design.

NOTE

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

REFERENCES

Alwin, D.F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. New York, NY: John Wiley and Sons, Inc.

Alwin, D.F., and J.A. Krosnick. 1991. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods and Research* 20: 139–81.

Dillman, D.A., J.D. Smith, and L.M. Christian. 2008. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. New York, NY: John Wiley and Sons, Inc.

Fowler, F.J. 1995. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 201–19.

Krosnick, J.A., and L.R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 141–64. New York, NY: John Wiley and Sons, Inc.

Saris, W.E., and I.N. Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. John Wiley and Sons, Inc.

Schwarz, N., B. Knauper, H.J. Hippler, E. Noelle-Neumann, and F. Clark. 1991. "Rating Scales May Change the Meaning of Scale Labels." *Public Opinion Quarterly* 55: 618–30.

Smith, T. 2003. "Developing Comparable Questions in Cross-National Surveys." In *Cross-Cultural Survey Methods*, edited by J. Harkness, F. Van de Vijver, and P. Moher, 69–91. New York, NY: John Wiley and Sons, Inc.