Articles

How Many Scale Points Should I Include for Attitudinal Questions?

Aaron Maitland*

Tags: survey practice

DOI: 10.29115/SP-2009-0023

Survey Practice

Vol. 2, Issue 5, 2009

Response scales are frequently used to measure attitudes in survey research. In this short article, I will discuss some theoretical considerations in the measurement of attitudes that influence the number of scale points. Next, I will discuss empirical results regarding the quality of scales with different numbers of scale points. Last, I will provide some question development strategies.

There are various theoretical considerations for determining the number of scale points. Attitudes are abstract constructs that are not directly observable and exist only in the respondent's mind. Response scales allow respondents to express both the direction and intensity of their attitudes. Some attitudes are viewed as bipolar concepts where two opposing sides of a concept are measured, whereas other attitudes are viewed as unipolar concepts where only the level of an attitude or just one side of a concept is measured. Researchers must clearly define the attitude object towards which a respondent can express an attitude. The respondent can then represent his or her stance for or against an attitude object by selecting the appropriate option on the response scale. Researchers usually conceptualize attitudes as existing along an attitude continuum. Hence, response scales that allow respondents to express different shades of an attitude rather than being simply for or against an attitude object will allow for better measurement of that continuum.

However, the difficulty of the response task must also be considered when designing response scales. Although longer scales might seemingly measure the attitude continuum in more detail, the response task might become too demanding, with too many scale points. This might force respondents to make more finely graded distinctions between scale points than might be possible. Respondents then might decide that the question is too demanding and *satisfice* (Krosnick 1991) by choosing the first plausible option that they encounter rather than carefully considering all options along the scale (Krosnic k and Fabrigar 1997). Additionally respondents might resort to rounding their

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

answers.

It is also important to consider the mode in which a question is going to be administered. The key distinction is between modes that rely solely on oral communication such as telephone versus modes that make use of visual communication such as the Web, mail, or face to face surveys with show cards. Although it depends on a number of factors such as how many scale points are labeled, one might generally conclude that longer response scales are easier to administer in a mode that uses visual communication since the respondent does not have to store all of the options in memory. The advantage of visual communication is probably minimal if only the endpoints of a lengthy response scale are labeled. Furthermore, unfolding or two step procedures can be used in telephone surveys to offer more response options without forcing the respondents to store the full range of options in short-term (working) memory. Research has found very few differences between the answers to questions using this unfolding technique over the telephone and those administered with a show card in face to face surveys (Groves 1988).

Finally, one must consider the interpretability of a middle position and whether it is meaningful for a specific concept. One interpretation is that respondents use this option when the middle category accurately describes their position (i.e., neither for nor against). Others suggests that a middle position is often interpreted as a "no opinion" option or an invitation to take an easy out for respondents who actually do have opinions, but are either unwilling or unable to express them due to the cognitive burden of the survey question (Krosnick 1991).

There are data quality standards that can be used to provide some insight into the optimal number scale points. Reliability and validity are two data quality standards most often employed using a quantitative framework. Reliability refers to how consistent answers are over replications. Reliability is measured over replications of the same question at different points in time or over multiple questions measuring the same attitude on a single occasion. Validity in the context of attitude measurement refers to how closely a survey question measures the construct of interest. Validity is difficult to measure, but is often operationalized quantitatively by assessing the extent to which a question converges with other questions measuring similar constructs and diverges from other questions measuring different constructs (Saris and Gallhofer 2007). Qualitative research methods are also useful for assessing the quality of survey questions. For example, in depth *cognitive* interviews can provide a detailed understanding of how survey respondents use the response categories and allow the researcher to assess whether this matches the question designer's intent.

Several empirical studies have examined the effect of the number of scale points on the reliability of questions with response scales. The literature is mixed, probably indicating that the number of scale points depends on the specific

objectives of a research project. Nonetheless, I will highlight some of the important conclusions that have been drawn. A review by Krosnick and Fabrigar (1997) did not find a monotonic increase in reliability as the number of scale points increased. Instead, a curvilinear pattern emerged in their review such that scales between 5–7 points were more reliable than scales with fewer points or more points. This was true for both bipolar and unipolar scales. Another study analyzed the longitudinal reliability of more than 300 survey questions that were repeated at more than two points in time (Alwin 2007). Once again, there was no monotonic increase in reliability as the number of scale points increased. Overall, two point scales were the most reliable followed by four, five, and nine point scales. Reliability was lower for six and seven point scales. The high reliability for two point scales could be due to the fact that two point scales only measure direction, whereas larger scales measure both direction and intensity (Alwin and Krosnick 1991). Interestingly the results were clearer for unipolar than bipolar scales. Four and five point unipolar scales demonstrated superior reliability compared to unipolar scales of other lengths; however, there were smaller differences in reliability between bipolar scales of different lengths.

There is some evidence that a middle position leads to lower reliability for shorter scales. Alwin (2007) found that three category scales are less reliable than two or four category scales. However, there was no clear evidence that five point scales were less reliable than four or six point scales. This suggests that the damaging effect of a middle position on reliability weakens as the number of categories increase.

The issue of validity has been addressed less frequently in the literature. Krosnick and Fabrigar (1997) report evidence that supports their view that 5–7 scale points are optimal. They found that questions using scales in this range tended to correlate the strongest with questions measuring conceptually related variables. Another interesting finding is that context effects – the effects of previous questions on a target question – tend to weaken as the number of scale points increases up to 7 points, after which there is very little change. They also report that even though the proportion of scale points used stays fairly constant up to 19 points, scales longer than 7 points do not seem to convey any additional information to researchers.

Although much of the evidence does seem to converge around the conclusion that 5–7 points might be optimal, others argue that more scale points is better. Based on results from multi-trait, multi-method experiments, Saris and Gallhofer (2007) conclude that up to 11 categories may be optimal. They claim that other authors are mistakenly interpreting variation from longer scales as measurement error. In short, they argue that different people's attitudes are calibrated differently so that similar opinions might be expressed with different values. For example, some respondents might have a tendency to express themselves with extreme words, whereas others express themselves more

moderately. This issue becomes more pronounced with longer scales. To prevent variation due to these individual response differences the authors argue for the use of fixed reference points (e.g., completely disagree, completely agree) at the end points of a scale to help reduce this type of variation.

Despite the principles that have been discussed in this article, many issues are unresolved and the choice of response scales should be driven by research objectives. Pretesting enables researchers to match question design with these objectives. In-depth, qualitative, or *cognitive* interviews making use of think-alouds or probing techniques can help a researcher understand if a response scale resembles how respondents tend to think about and answer survey questions. In other words, these in-depth interviews should lead to better validity by more closely matching the response scales with the respondents own representations of an attitude. This technique also provides an understanding of the burden that a scale places on respondents. Given enough time and budget, field experiments that include different forms with scales of different lengths are another useful technique. Additionally, it is important to include repeated measurements of the response scales within a survey to assess their reliability.

NOTE

The findings and conclusions in this report are those of the author 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.

Groves, R.M. 1988. *Telephone Survey Methodology*. New York, NY: John Wiley and Sons, Inc.

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.