2006

Answer Format Suitability - The Interdependence of Answer Format and Construct Measured

Sara Dolnicar
University of Wollongong, sarad@uow.edu.au

Bettina Grun
Vienna University of Technology, Austria

Publication Details

This paper was originally published as: Dolnicar, S & Grun, B, Answer Format Suitability - The Interdependence of Answer Format and Construct Measured, ANZMAC 2006 CD Proceedings (Australia and New Zealand Marketing Academy Conference, Brisbane, Queensland, 4-6 December 2006.

Answer Format Suitability - The Interdependence of Answer Format and Construct Measured

Abstract

The vast majority of surveys use ordinal answer formats independently of the construct under study. We hypothesise that the ordinal scale is not optimal under all circumstances, but that the suitability of answer formats depends on the construct measured. A repeat measurement study is conducted using binary and ordinal answer formats measuring two different constructs: attitudes and behavioural intentions. A clear interaction effect between answer formats and constructs is revealed, supporting the notion that there is not a single optimal answer format, but that some constructs are naturally more suitable for certain answer formats than others. These findings call for increased use of pre-studies to determine the optimal answer format before fieldwork rather then relying on standard answer formats.

Disciplines

Business | Social and Behavioral Sciences

Publication Details

This paper was originally published as: Dolnicar, S & Grun, B, Answer Format Suitability - The Interdependence of Answer Format and Construct Measured, ANZMAC 2006 CD Proceedings (Australia and New Zealand Marketing Academy Conference, Brisbane, Queensland, 4-6 December 2006.

Answer Format Suitability -
The Interdependence of Answer Format and Construct Measured

Sara Dolnicar, University of Wollongong
Bettina Grün, Vienna University of Technology
*Authors listed in alphabetical order.

Abstract

The vast majority of surveys use ordinal answer formats independently of the construct under study. We hypothesise that the ordinal scale is not optimal under all circumstances, but that the suitability of answer formats depends on the construct measured. A repeat measurement study is conducted using binary and ordinal answer formats measuring two different constructs: attitudes and behavioural intentions. A clear interaction effect between answer formats and constructs is revealed, supporting the notion that there is not a single optimal answer format, but that some constructs are naturally more suitable for certain answer formats than others. These findings call for increased use of pre-studies to determine the optimal answer format before fieldwork rather then relying on standard answer formats.

Introduction

There would be little resistance among marketing researchers against the statement that different kinds of questions require different answer formats. Yet the ordinal answer format dominates marketing research (Van der Eijk, 2001). The vast majority of studies undertaken both by market research companies and by academic researchers uses 5 or 7-point ordinal scales in questionnaires. In recent issues of the *JMR, JCR* and the *JM* (Journal of Marketing Research, May 2005, Journal of Marketing 69(3) 2005, Journal of Consumer Research 32(1) 2005), 21 articles reported empirical findings based on consumer responses. Of these, 86 percent used ordinal multi-category scales.

Research work comparing answer formats does not support this apparent agreement in the scientific marketing community that ordinal scales are the optimal choice in questionnaire design. A vast amount of literature exists comparing different answer formats. Typical criteria used to undertake such comparisons are reliability and validity, structural equivalence, user friendliness and the susceptibility to response styles. Prior work typically used artificial data for such comparative studies or collapsed empirical data with more answer options to fewer options. Results are controversial. Some studies conclude that, if means and analyses based on means are of interest, binary scales are sufficient and lead to the same results (Lehmann and Hulbert, 1972; Loken, Pirie, Virnig, Hinkle and Salmon, 1987; Preston and Colman, 2000; Dolnicar, Grün and Leisch, 2004). Furthermore, it was shown that binary scales are not significantly different from multi-category ordinal scales with respect to reliability (Bendig, 1954; Peabody, 1962; Komorita, 1963; Komorita and Graham, 1965; Matell and Jacoby, 1971; Jacoby and Matell, 1971; Remington, Tyrer, Newson-Smith and Cicchetti, 1979; Preston and Colman, 2000) and validity values (Matell and Jacoby, 1971; Jacoby and Matell, 1971; Preston and Colman, 2000) as well as structural equivalence of constructs (Martin, Fruchter and Mathis, 1974; Percy, 1976). Also contrary findings are reported with regard to most of these criteria concluding that more options lead to better values (reliability: Symonds, 1924; Nunnally, 1967; Jones, 1968; Oaster, 1989; Finn, 1972; Ramsay, 1973; validity: Loken, Pirie, Virnig, Hinkle and Salmon, 1987; Hancock and Klockars, 1991; structural equivalence: Green and Rao, 1970). User friendliness and economic efficiency have not been studied as extensively. While Jones (1968) concludes that respondents prefer multiple answer options, Dolnicar (2003) finds that binary format is perceived as easier and quicker by respondents.

For some formats, such as the ordinal one, a number of authors have published analyses discussing the dangers of inappropriate data assumptions and the ambiguity of interpretations based on frequently ill-defined ordinal formats (Kampen and Swyngedouw, 2000). A review article on comparisons of answer formats by Cox (1980), on the other hand, draws the conclusion that the seven-point ordinal scale generally represents a good option, while noting that there is no single optimal scale for all circumstances and that one of the two main challenges of future work is to establish methods of pre-testing to determine which answer format might be most suitable under the given circumstances of the research problem. However, Cox (1980, p. 420) also argues that "scales with two or three alternatives are generally inadequate in that they are incapable of transmitting very much information and they tend to frustrate and stifle respondents."

We assume that the ability of respondents to correctly differentiate between the grey shades of the scale categories offered by an answer format depends on the construct measured. While it is reasonable to ask respondents to distinguish between several levels of agreement for a complex construct in order to be able to correctly measure their true values of agreement, such a fine measurement will not increase the amount of information in the data for a simple or rather vague construct, but will aggravate the amount of the noise which might be due to individual response styles. Based on Cox' conclusions, the response style literature which indicates that multi-category ordinal scales are susceptible to scale usage heterogeneity and the assumption that different constructs implicate a different level of differentiation, we hypothesize that: (H1) overall use of scale categories differs for different constructs, (H2) different people use answer formats differently, (H3) individual answer format use depends on the measured construct, and (H4) ordinal scales are perceived as more user-friendly.

These hypotheses are empirically studied based on a comparison of responses using a binary and a seven-point ordinal scale, respectively, on questions on attitudes and behavioural intentions. In addition, the respondents' own evaluations of the different answer formats are used to investigate user friendliness of the answer formats. The results have major implications for market research: if empirical evidence for the assumption that different answer formats are suited differently well for different constructs can be provided, it would be recommendable from the perspective of saving field cost and reducing respondent fatigue, to use the simplest and quickest possible format that is suitable for the construct under study.

Data

Data was collected at the University of Wollongong in two subsequent tutorials. Student identification numbers were used to match the two questionnaires that contained the same questions using different answer formats: binary (yes-no) and ordinal (seven-point scale). The questionnaires included questions about two different constructs: behavioural intentions (to use recycled water for different purposes) and attitudes (about environmental protection). Attitudes were measured using a shortened version of the scale known as the New Ecological Paradigm (Dunlap *et al.*, 2000) consisting of eight questions. This measurement is referred to as NEP. Behavioural intentions were measured by asking respondents if they would use recycled water for purposes from a list of 13 possible uses. In addition beginning and finishing time were noted and respondents evaluated each questionnaire with respect to its user-friendliness on a 5-point bipolar ordinal scale. In total, 80 fully completed sets (including two repeated measurements) were available. The repeat-measure nature of the survey is of central importance as it assures that any differences in answer format usage in dependence of constructs under study is in fact due to the different answer formats and constructs rather than the nature of the sample.

Results

All computations and graphics for the empirical analysis have been done using the R statistical software package (R Development Core Team, 2006).

H1 Overall use of scale categories differs for different constructs.

To test this hypothesis the aggregated absolute and relative frequency of the scale categories are determined for the two answer formats and the two constructs (Table 1). For the binary scale 619 responses are available on the NEP scale (96.7% of the possible answers) and 1030 for the intentions (99.0% of the possible answers). A two-sample test for equality of proportions does not indicate that there is a significant difference in overall usage between the two constructs (χ^2=2.23, df=1, p-value=0.14). For the seven-point scale 638 (99.7%) responses on the NEP scale and 1038 (99.8%) for the intentions are available. A Pearson's chi-square test for independence indicates that there is a significant difference in usage of the scale categories for the two constructs (χ^2=139.13, df=6, p-value < 0.001). The endpoints are more frequently ticked for the behavioural intentions, whereas scale points indicating slight agreement (categories four to six) are more frequently used for the NEP scale.

Table 1: Use of Scale Categories for the Two Answer Formats and Constructs

		Binary		Seven-Point Scale						
		Yes	No	1	2	3	4	5	6	7
Absolute	NEP	344	275	76	64	82	109	115	121	71
	Intentions	532	498	231	117	123	87	88	119	273
Relative	NEP	0.56	0.44	0.12	0.10	0.13	0.17	0.18	0.19	0.11
	Intentions	0.52	0.48	0.22	0.11	0.12	0.08	0.08	0.11	0.26

The aggregate analysis indicates a difference in usage of the scale categories for the seven-point ordinal scale (H1 not rejected), whereas no difference is detected for the binary scale (H1 rejected).

H2 Different people use answer formats differently.

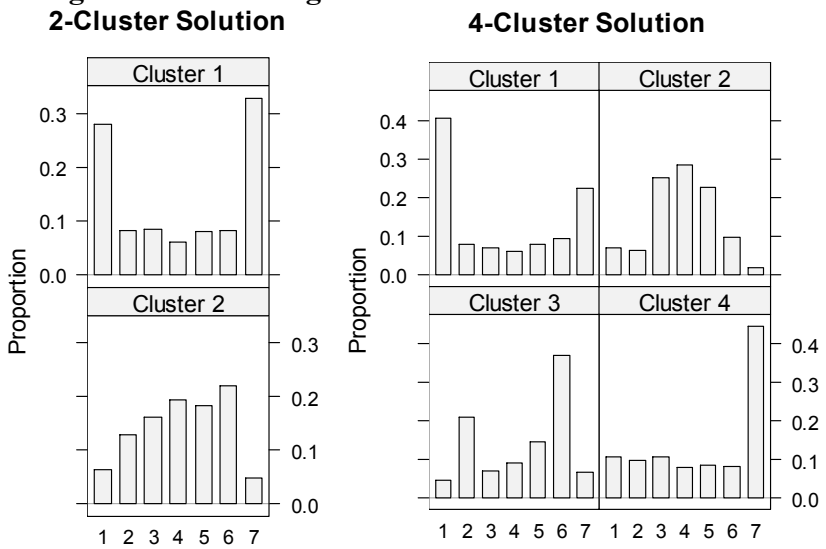
Based on the response style literature one would assume that respondents individually use scales differently. A different use of the scale categories leads to different answer patterns of the respondents, where answer patterns are given for each respondent by his proportional use of the scale categories, i.e. the relative number of times he ticked this category. In order to avoid confounding the effect of individual use with the construct effect the answer patterns of each respondent are determined separately for each construct.

In order to detect segments of respondents who use answer formats in a similar way having the construct fixed respondents' answer patterns are partitioned using the K-means algorithm (Hartigan and Wong, 1979).The K-means algorithm is an iterative grouping procedure that aims at minimising the sum of distances between the answer patterns within each group/ cluster and maximising the sum of distances between groups/ clusters. In order to ensure detection of a global optimum the K-means algorithm is repeated with 20 random initializations. The best solution with respect to the within-sum of distances is reported.

Because natural clusters cannot be expected to exist it is not trivial to choose the optimal number of clusters. A visual inspection of the within-sum of distances for the different number of clusters indicates that a solution with six clusters seems to appropriately represent the structure of the binary responses. For the 7-point format two or four clusters appear to provide the best representation. The prototypes of these solutions are given in Figure 1. As can be seen the two-cluster solution splits respondents into a group which primarily uses the endpoints of the answer format and a second group which prefers the middle categories. The four-cluster solution refines this solution by splitting the endpoint users into the yea-sayers and those using both endpoints rather equally and the middle scale

users into those favouring the scale points next to the endpoints and those nearly only using the three middle points. H2 can consequently not be rejected for both answer formats as heterogeneity in the answer patters can be considerably reduced by segmenting them into groups.

Figure 1: Answering Patterns of the K-Means Solutions



H3 Individual answer format use depends on the measured construct.

The cluster assignments are cross-tabulated with the constructs in order to assess if certain clusters occur more or less frequently for one of the constructs. A Pearson's chi-square test is used to check for significant association. For the binary answer patterns no significant relationship between scale usage and construct is detected (χ^2=3.50, df=5, p-value=0.62). For the 7-point scale the association is significant for both cluster solutions (two-cluster solution: χ^2=32.41, df=1, p-value < 0.001, four-cluster solution: χ^2=37.40, df=3, p-value < 0.001). Table 2 shows which clusters occur more often for which construct: for the two-cluster solution most answer patterns in cluster one occur for the behavioural intentions, i.e. a lot of respondents use the 7-point scale like a binary scale. This conclusion is confirmed by the four-cluster solution where answer patterns from behavioural intentions are in general assigned to cluster one and four.

Table 2: Clusters Assignments Given Constructs for the Seven-Point Scale

		2-Cluster		4-Cluster			
		1	2	1	2	3	4
Absolute	NEP	21	59	14	35	21	10
	Intentions	58	22	33	6	14	27
Relative	NEP	0.26	0.72	0.18	0.44	0.26	0.12
	Intentions	0.74	0.28	0.41	0.08	0.18	0.34

It can be therefore concluded that the individual scale usage differs for the constructs for the ordinal scale (H3 not rejected) while no difference can not be detected for the binary scale (H3 rejected). In addition the results indicate that the ordinal scale is by a lot of respondents used like a binary scale for the behavioural intentions.

H4 Ordinal scales are perceived as more user-friendly by respondents.

While we studied the actual answers and derived our conclusions on this pattern analysis above, we also asked respondents to evaluate themselves how long they perceived the questionnaire to be, how complex, and how well they were able to express their feelings. In addition we noted the time it took them to complete the survey. Unfortunately, these

evaluations are not available separately for the constructs given that the respondents were asked to respond to both construct using one format first and then the second format in the second wave. Nevertheless, this data allows to test the conclusion Cox (1980, p. 420) drew with respect to two- and three-point scales that they "frustrate and stifle respondents".
The differences in duration and evaluation are tested using t-tests. For testing difference in duration only those respondents who needed at least one minute and less than 20 minutes are included (i.e. 95.6% of the observations) as other duration lengths are impossible to occur. The logarithm is taken as otherwise the variable is distributed skewed to the right. This analysis shows that it was significantly faster to complete the binary answer format (t=2.05, df = 150.82, p-value = 0.04) and that the binary answer format is perceived as significantly simpler (t=2.20, df = 157.83, p-value = 0.03). The answer formats are equally perceived with respect to the ability to express the feelings (t=-0.85, df =155.87, p-value=0.39). With respect to pleasantness and quickness the binary scale is more favourably evaluated but no definite conclusions can be drawn as the p-values are insignificant but close to the 5% significance level which might be due to lack of power given the small sample size (pleasant: t=1.82, df = 156.74, p-value =0.08, quick: t=1.86, df=156.41, p-value=0.07). Consequently, H4 has to be rejected.

In sum, while Cox's finding that no single answer format is best under all circumstances is supported by our study, the reasons for his rejection of two-and three-point scales are not. In our study, binary format appeared to be suitable for evaluating behavioural intentions and was not perceived as more frustrating. On the contrary, it took less time to complete and was evaluated as being quicker to complete.

Conclusions and Future Work

The choice of the most suitable answer format for a particular research problem is crucial in market research: it affects both the validity of the research and the fieldwork cost. The present study demonstrates the interaction between response formats and constructs measured and illustrates that selecting the most appropriate answer format is not a commonsense problem that can be decided by a researcher alone. Optimally, answer formats should be pre-tested for suitability.

We investigated the interdependence of the suitability of binary and ordinal answer formats for the evaluation of attitudes and behavioural intentions and found that the same respondents used the same answer formats in a different way when asked to evaluate different constructs. While it appeared that a seven-point ordinal scale was well suited to capture respondents' attitudes, the patterns of responding to the set of behavioural intentions demonstrated a strong binarisation, indicating that the binary scale is suitable to capture those responses and can be used without sacrifice in user-friendliness. On the contrary, the binary format led to substantial efficiency gains through reduced completion times.

This study is limited in sample size as well as by using a student population, although it is not expected that students would demonstrate systematically different answer format effects than general population would. Nevertheless, a replication with a larger sample of general population and including a broader range of answer formats as well as constructs would be desirable.

Acknowledgements: This research was supported by the Australian Research Council (through grants DP0557257 and LX0559628), the University of Wollongong through internal research grant schemes and the Austrian Academy of Sciences (ÖAW) through a DOC-FFORTE scholarship for Bettina Grün.

References

Bendig, A. W., 1954. Reliability and the Number of Rating Scale Categories. Journal of Applied Psychology 38(1), 38-40.

Cox, E. P., 1980. The optimal number of response alternatives for a scale: A review. Journal of Marketing Research 17 (4), 407-422.

Dolnicar, S., 2003. Simplifying three-way questionnaires - Do the advantages of binary answer categories compensate for the loss of information? ANZMAC CD Proceedings.

Dolnicar, S., Grün, B., and Leisch, F., 2004. Time efficient brand image measurement - Is binary format sufficient to gain the market insight required? CD Proceedings of the 33rd EMAC conference.

Dunlap, R. E., Van Liere, K. D., Mertig, A. G., and Jones, R. E., 2000. Measuring Endorsement of the New Ecological Paradigm: A Revised NEP Scale. Journal of Social Issues 56(3), 425-442.

Finn, R. H., 1972. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. Educational and Psychological Measurement 32 (2), 255-265.

Green, P. E., and Rao, V. R., 1970. Rating Scales and Information Recovery---How Many Scales and Response Categories to Use? Journal of Marketing 34, 33-39.

Hancock, G. R., and Klockars, A. J., 1991. The effect of scale manipulations on validity: Targetting frequency rating scales for anticipated performance levels. Applied Ergonomics 22 (3), 147-154.

Hartigan, J. A., and Wong, M. A., 1979. Algorithm AS 136: A K-means clustering algorithm. Applied Statistics 28(1), 100-108.

Jacoby, J., and Matell, M.S., 1971. Three-Point Likert Scales Are Good Enough. Journal of Marketing Research 8, 495-500.

Jones, R.R., 1968. Differences in Response Consistency and Subjects' Preferences for Three Personality Inventory Response Formats. Proceedings of the 76th Annual Convention of the American Psychological Association, 247-248.

Kampen, J., and Swyngedouw, M., 2000. The Ordinal Controversy Revisited. Quality & Quantity 34(1), 87-102.

Komorita, S.S., 1963. Attitude content, intensity, and the neutral point on a Likert scale. Journal of Social Psychology 61, 327-334.

Komorita, S.S., and Graham, W. K., 1965. Number of scale points and the reliability of scales. Educational and Psychological Measurement 25(4), 987-995.

Lehmann, D.R., and Hulbert, J., 1972. Are Three Point Scales Always Good Enough? Journal of Marketing Research 9(4), 444-446.

Loken, B., Pirie, P., Virnig, K. A., Hinkle, R. L., and Salmon, C. T., 1987. The Use of 0-10 Scales in Telephone Surveys. Journal of the Market Research Society 29 (3), 353-362.

Martin, W. S., Fruchter, B., and Mathis, W. J., 1974. An investigation of the effect of the number of scale intervals on principal components factor analysis. Educational and Psychological Measurement 34, 537-545.

Matell, M. S., and Jacoby, J., 1971. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. Educational and Psychological Measurement 31, 657-674.

Nunnally, J.C., 1967. Psychometric Theory. New York: McGraw-Hill, 1st edition
Oaster, T. R. F., 1989. Number of alternatives per choice point and stability of Likert-type scales. Perceptual and Motor Skills 68, 549-550.

Peabody, D., 1962. Two components in bipolar scales: direction and extremeness. Psychological Review 69(2), 65-73.

Percy, L., 1976. An Argument in Support of Ordinary Factor Analysis of Dichotomous Variables. In: Anderson, B. (Ed.), Advances in Consumer Research. Association for Consumer Research, pp. 143-148.

Preston, C.C., and Colman, A.M., 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychologica 104, 1-15.

Ramsay, J. O., 1973. The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values. Psychometrika 37, 513-532.

Remington, M., Tyrer, P. J., Newson-Smith, J., and Cicchetti, D. V., 1979. Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. Psychological Medicine 9, 765-770.

R Development Core Team, 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Symonds, P. M., 1924. On the Loss of Reliability in Ratings Due to Coarseness of the Scale. Journal of Experimental Psychology 7, 456-461.

Van der Eijk, C., 2001. Measuring agreement in ordered rating scales. Quality & Quantity 35, 325-341.

Reply to the reviewers:

Reviewer 1:
Some characters seem not to be translated from your WORD version to the online version. Check these with your resubmission.

> *Thank you for pointing this out. Unfortunately we can only submit the final paper version in Word. We tried to omit all special characters to avoid conversion problems.*

It's not clear how the Proportional use of the scale categories was measured for your testing of H2. Can you expand briefly on how this was done. If the cluster analysis was simply done of the ratings scales alone without some sort of proportional transformation, then I don't see how this is a measure of differences in usage of the scales - it would be simply finding people who gave similar scores across different measures.

> *Response:*
> *The proportional use of the scale categories was determined by checking how often each category was ticked by the respondent for each construct and then the relative number was determined by dividing through the total number of answers.*
> *We consequently did not investigate differences in "scores" as we never aggregated responses across multiple measures. Instead, we investigate systematic patterns of use of groups of respondents.*
>
> *Change:*
> *We have added a sentence in the section "H2 Different people use answer formats differently" in which we explain in more detail how the proportions are computed.*

Reviewer 2:
Good paper based on a small sample to answer interesting research questions on measurement. Research design was good, analyses are appropriate and conclusions are appropriately drawn on the analyses. The paper is well written and conform the prescribed style for the conference.