University of Wollongong

Research Online

Faculty of Commerce - Papers (Archive) Faculty of Business and Law

1-1-2005

Answer format effects revisited

Sara Dolnicar
University of Wollongong, s.dolnicar@uq.edu.au

Bettina Grun
University of Wollongong, bettina@uow.edu.au

Follow this and additional works at: https://ro.uow.edu.au/commpapers

Part of the Business Commons, and the Social and Behavioral Sciences Commons

Recommended Citation

Dolnicar, Sara and Grun, Bettina: Answer format effects revisited 2005, 1-54.
https://ro.uow.edu.au/commpapers/2264

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Answer format effects revisited

Abstract

The effect of answer formats presented to respondents in written surveys are investigated for two constructs (attitudes and behavioral intentions) and three response scales (binary, ordinal and metric). Results indicate that (1) formats differ in their susceptibility to response styles but lead to the same results with respect to average values and underlying dimensions; (2) binary format is quicker to complete and perceived as quicker while all formats are perceived as equally simple, pleasant, and useful to express feelings; (3) an interaction between the construct measured and the answer format clearly exists which should be investigated more systematically in future research.

Keywords

Answer, Format, Effects, Revisited

Disciplines

Business | Social and Behavioral Sciences

Publication Details

Dolnicar, S. & Grun, B. (2005). Answer format effects revisited. In H. Timmermans (Eds.), European Institute of Retailing and Services Studies (EIRASS) 2005 Conference (pp. 1-54). The Netherlands: Technische Universiteit Eindhoven.

This conference paper is available at Research Online: https://ro.uow.edu.au/commpapers/2264

Answer format effects revisited

Sara Dolnicar*

School of Management & Marketing, University of Wollongong

Wollongong, NSW 2522, Australia

Telephone: (61 2) 4221 3862, Fax: (61 2) 4221 4154

sara_dolnicar@uow.edu.au

Bettina Grün*

Department of Statistics and Probability Theory ,Vienna University of Technology

Wiedner Hauptstraße 8-10/1071, A-1040 Vienna, Austria

Telephone: (43 1) 58801 10716, Fax: (43 1) 58801 10798

bettina.gruen@ci.tuwien.ac.at

*Authors listed in alphabetical order.

Abstract

The effect of answer formats presented to respondents in written surveys are investigated for two constructs (attitudes and behavioral intentions) and three response scales (binary, ordinal and metric). Results indicate that (1) formats differ in their susceptibility to response styles but lead to the same results with respect to average values and underlying dimensions; (2) binary format is quicker to complete and perceived as quicker while all formats are perceived as equally simple, pleasant, and useful to express feelings ; (3) an interaction between the construct measured and the answer format clearly exists which should be investigated more systematically in future research.

Keywords: Answer format, response scale

INTRODUCTION

Multi-category response format represents the most popular option in marketing surveys (Van der Eijk, 2001). Consequently, the optimal number of response options has been extensively researched throughout the twentieth century.

A number of distinct streams of research have developed using different criteria for the evaluation of the "optimality" of a multi-category rating scale: reliability or validity, the interpretational perspective typically using market structure analysis to derive managerial recommendations from data of different scales, the consumer perspective of answering complexity, and the viewpoint of susceptibility to response styles which has been repeatedly demonstrated to cause significant problems when ordinal response scales are used.

None of the studies have, however, adopted a longitudinal approach to answer format comparison, thus implicitly assuming to know what the respondents' transformations from one scale to another would be. It should be noted that some prior studies did collect data from their respondents twice. The aim of these repeat measurements was, however, the computation of test-retest reliabilities, rather than the comparison of results based on different scales including identical respondents.

The present work makes one step towards filling this gap by investigating longitudinal data, which allows investigation of individual-level transformations between scales for two different constructs to determine differences in answer scale effects resulting from the nature of the construct measures.

The aim of this study is to investigate (1) the existence of heterogeneity in respondents' answering patterns due to different person and construct-related response styles, (2) the existence of statistical differences in the answers of respondents to questions of different formats, (3) the existence of differences in the managerial interpretations derived from positioning analyses based on different answer formats, (4) the existence of differences in answering speed, which has direct consequences for the price of the fieldwork as well as indirectly decreasing the data quality through respondent fatigue, and (5) the existence of differences in answering ease, which can be seen as a subjective measure of perceived complexity of the questionnaire and assumed to be indirectly proportional to the resulting data quality.

The results are highly relevant to market research: If the answer time for binary or metric questions is shorter, the questions are perceived as easier, the answers and the managerial implications do not significantly differ, binary or metric scales should be preferred because they are likely to be cheaper and / or generate higher quality data as well as requiring less assumptions about data characteristics to be made in the analysis step.

PRIOR WORK

The effects of different answer scales have been studied extensively for half a century now. The areas of investigation are characterized by different foci of interest and contributions to the body of knowledge emerge from a wide variety of scientific disciplines.

One such research area centers on the measurement paradigms underlying different answer scales and the possible mistakes that result from ignoring the assumptions one can reasonably make about each scale format.

The most comprehensive contribution of this nature was made by Kampen and Swyngedouw (2000) who review a century of controversies regarding the use of ordinal variables in empirical research. They state that ordinal scales would essentially not be viewed as measurement from a classical measurement theory perspective due to a lack of measurement unit, like meters, liters or centigrades. From a representation measurement theoretical view, ordinal scales are capable of representing an attribute. However, without knowing the psychometric characteristics of the attributes, the selection of a scale to represent it is random, as it cannot be checked if good representation actually occurs. Kampen and Swyngedouw (2000) classify ordinal measures in five types of different nature. Type 1 is a categorized metric variable with known thresholds (as, for instance, age groups). For such ordinal variables an objective standard exists. Type 2 is defined as a categorized metric variable with unknown thresholds (for instance, age groups like "young" and "old'). Such ordinal variables are very difficult to calibrate and any analysis of such data is difficult to interpret due to a lack of clear operationalisation. Type 3 is a categorized latent variable with unknown thresholds (low-middle-highly friendly receptionists) and – if it can be calibrated by experts – suffers from typically low inter-experimenter agreement levels. Type 4 is a semi-standardized discrete variable with ordered categories (the example provided by the authors is that of a classification into dead, handicapped and sound mice in an experiment). The quality of such ordinal variables depends on the quality of calibration of the classification. Finally, type 5 is an unstandardized discrete variable with ordered categories (as the agreement with statements or level of satisfaction). Similarly to type 2, type 5 has very undesirable properties best described by the following statement (p. 99) "in many instances the experimenter can only hope that in general respondents or experimentators attach the same meaning to the categories of an ordinal variable."

Essentially Kampen and Swyngedouw (2000) thus see major problems associated with the use of ordinal scales: the problem of subjective measurement where certain scale points mean different things to different people (for instance, "very satisfied"); the lack of equidistance which makes it difficult to justify the use of analytic techniques developed for metric data, thus limiting the available methods to those specifically designed for ordinal data. And even among such methods, Kampen and Swyngedouw (2000) demonstrate differences in methods that claim to measure the same thing, for instance the association of two ordinal variables. And if, ignoring all data assumptions, metric methods are applied to ordinal data, interpretations of results are impossible without substantial understanding of the ordinal steps and the differences between the ordinal steps. Furthermore, distributional assumptions that are typically made for parametric tests cannot be tested, as even the existence of an underlying metric variable cannot be proven. Finally, there is a lack of invariance under groupings of adjacent categories. "Thus, the choice of using a three, five or seven point scale in measuring the ordinal characteristics becomes a crucial decision." (p. 89).

Cox (1980) published a comprehensive review on answer formats from a marketing perspective discussing the contributions of information theory, the absolute judgment paradigm and metric approaches. He comes to the conclusion that – while a democratic vote for the best number of response alternatives would be seven – additional research is needed to replicate prior findings and extend investigations to new areas related to the problem. Specifically he believes that the issue of response error and response bias has not been investigated sufficiently and that "Surprisingly little is known about the process of psychological judgment." (p. 419).

A different approach with a narrower perspective on analytic issues of different scale formats is taken by Lehmann and Hulbert (1972). They conduct simulation studies and conclude that, if mean values of a sample are of interest, dichotomous or trichotomous scales are sufficient, if, however, individual behavior is of interest, five to seven point scales should be used.

Similar points are made by numerous researchers whose main interest was in response style identification and correction as well as researchers investigating response style effects in a cross-cultural setting. These studies are reviewed below.

A second area that has been studied extensively since the early Fifties is the effect of different response scales on reliability and validity of findings.

Studies include different methodological approaches ranging from simulation work to the analysis of empirical data. Overall, it appears that there is substantial evidence for the fact that the number of response options provided in an answer scale is not related to reliability levels (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). Only few studies conclude the opposite (Symonds, 1924; Nunnally, 1967; Jones, 1968; Oaster, 1989; Finn, 1972; Ramsay, 1973).

Controversy also resulted from the studies investigating the effects of answer scales on validity. A number of authors conclude from their empirical studies that no significant difference in validity can be found between different answer scales (Matell and Jacoby, 1971; Jacoby and Matell, 1971; Preston and Colman, 2000). Others (Loken, Pirie, Virnig, Hinkle and Salmon, 1987; Hancock and Klockars, 1991) find increased validity levels for higher numbers of scale points.

An important contribution to this stream of research was made by Chang (1994) who demonstrated that many of the past studies comparing reliabilities and validities did not decompose systematic method variance and trait variance. Therefore larger numbers of answer options have rendered more reliable findings, which, however, is the consequence of the restriction of range effect (see Nunnally, 1970; Cohen, 1983; Martin, 1973;1978) impacting all measures based on Pearson correlation, such as Cronbach's alpha and test-retest measures. Chang used structural equation modeling to decompose these two components and found that criterion related validity was independent of the number of answer options and reliability values were better using a four point scale as opposed of a six point scale.

While validity and reliability dominated the discussion for a long time, the issue of differences in the interpretation of findings based on different scales has not developed to become an equally popular field of research.

Three different approaches were taken in the past to compare interpretations: the use of ordinal-level empirical data that is collapsed to dichotomous or trichotomous levels, followed by multivariate analyses conducted separately on the original and derived data sets. This approach was chosen by Martin, Fruchter and Mathis (1974) and Percy (1976). They collapsed empirical data and computed factor analyses to compare findings using an objective measure of compliance between the two (or more) resulting factor solutions as well as graphical inspection. Both studies conclude that no significant differences exist between the solutions based on different answer formats.

Green and Rao (1970) chose the approach of constructing artificial data in order to control for true data structure recovery. They come to the conclusion that at least 6 points should be used on an ordinal scale and at least 8 attributes should be included in a scale.

Loken, Pirie, Virnig, Hinkle and Salmon (1987) conducted a fully empirically study where respondents were questioned both on an 11 and a 4 point scale using a phone survey. Results emerging from the two different scales seem to be equally good regarding discrimination power between socio-demographic groups and capturing of relationships between variables.

Similarly, Preston and Colman (2000) empirically compared results derived from 10 different scales, including dichotomous and nearly metric (101 scale points) format. They conclude that there are no differences regarding the correlation matrices of the five items; the relation of items to each other is the same on all scales. Scales rendered the same underlying factor structure and the same Cronbach alphas. One difference detected was in discriminating power for certain scales. The binary scale did not significantly differ in this criterion from the scales with larger numbers of scale points. They recommend the use of seven, nine or ten categories, but do acknowledge that (p. 13) "different scales may be best suited to different purposes."

Dolnicar, Grün and Leisch (2004) compare the mean values of the items derived from repeated questioning of students with both binary and ordinal scales and develop a model to predict the binary responses from ordinal responses concluding that there are little differences in managerial interpretation.

A less extensively researched topic is the user-friendliness of different scale formats. With the main focus having been on methodological issues, the respondents perspective was neglected in the past. Only one very early (Jones, 1968) and two recent ones (Preston and Colman, 2000; Dolnicar, 2003) include this dimension in their comparisons of alternate formats. Jones (1968) reveals that respondents have a clear preference for multiple categories. Preston and Colman (2000) investigate different dimensions of user-friendliness and find that individuals can better express their feelings when more categories are offered. By contrast, the perceived speed of questionnaire completion is associated with lower numbers of answer categories. Dolnicar (2003) finds that ordinal scales are perceived as significantly more difficult to answer than binary scales by respondents.

Differences in economic efficiency have rarely been studied directly but are frequently mentioned by various authors. Payne (1951), Dillman (1978), Bradburn and Sudman (1979), Churchill (1979) and Peterson (1982) all make clear recommendations not to use too many answer categories in the context of telephone surveys, for instance. Dolnicar (2003) asked students to repeatedly respond to the same questionnaire using different scales and found a significant difference in completion times with the ordinal version taking on average six minutes and the binary one four. Komorita and Graham (1965, p. 989) after the comparison of reliability and validity measures state economic arguments for scale choice: " the major implication is that, because of simplicity and convenience in administration and scoring, all inventories and scales ought to use a dichotomous, two-point scoring scheme."

Finally, there is a large body of work on the susceptibility of response scales to response styles. These can broadly be divided into studies investigating response styles in general and an extensive body of work investigating the same issue in the cross-cultural context.

The problem of susceptibility of ordinal scales to response styles in market research is as old as market research itself. For instance, Cronbach (1950, p. 21) stated that "Since response sets are a nuisance, test designers should avoid forms of items which response sets infest". Cronbach is so concerned about the contamination of data with response sets that he recommends a reduction to only dichotomous answer options in order to avoid these effects.

Later, a number of researchers investigated response style effects that manifest themselves on ordinal scales and made recommendations for identification and correction of such effects (Cunningham, Cunningham and Green, 1977; Greenleaf, 1992a; 1992b; Heide and Gronhaug, 1992; Watson, 1992; Van de Vijver and Poortinga, 2002; Welkenhuysen-Gybels, Billiet and Cambre, 2003).

Two quotes from these studies illustrate the gravity of the problem for everyday empirical research work: Watson (1992, p. 83) warns that "it is dangerous to assume that acquiescence does not affect responses to attitude questions" and Greenleaf (1992, p. 183) concludes that "Hence, the most appropriate rating scale scores to use in data analysis are "corrected" attitude scores [...] where the correction removes the bias but retains the attitude information".

Most of the studies conducted in the cross-cultural context have major implications for market research in general as well. A subset of these studies aims at empirically demonstrating the differences in response styles in different countries or among respondents from different cultural backgrounds (Chun, Campell and Yoo, 1974; Bachman and O'Malley, 1984; Marin, Gamba, Marin, 1992; Watkins and Cheung, 1995; Albaum, 1997; Byrne and Campbell, 1999; Clarke III, 2000; 2001; Van Herk, Poortinga and Verhallen, 2004). All these studies use ordinal data formats for their studies and determine significant systematic differences in the way different cultures use these scales.

Warnings from this stream of work include Arce-Ferrer and Ketterer's (2003, p. 499) conclusion that "Adapting scales across countries involves more than adapting items linguistically and using judges to appraise items for cultural adequacy. It involves an assessment of generalizability of a construct theory." Byrne and Campbell (1999) recommend to systematically pre-analyze the data with regard to violating normality assumptions before conducting any comparative analyses between groups that are based on means (*t* tests, analyses of variance) and Cheung and Rensvold (2000) who recommend a number of methods for the identification and correction of response styles in cross-cultural research, criticize the predominant way of studying differences between countries (p. 188): "The naïve approach involves taking a scale that has been validated in Culture A, administrating it in Culture B, and then uncritically comparing the scale scores using a *t* test or some similar procedures" warning that comparisons of means are essentially uninterpretable because, for instance, extreme response style affects numerical scores.

Only three studies have so far combined the aspect of cross-cultural response style differences on ordinal scales and the number of scale points. Clarke III (2000; 2001) finds that increasing from three to five options reduced the amount of extreme answers. While this appears an attractive option at first, the differences in response styles between cultures increase with more scale points. A three point scale, although leading to a higher use of extreme values, might consequently be better for cross-cultural comparisons as the differences between cultural groups are smaller. These findings are supported by the results reported by Roster, Rodgers and Albaum (in press) who find that extreme values are used more when a five point scale is presented than in case of a ten point scale.

DATA

The data set was collected at the University of [*name to be added after the review process*] among students attending lectures or tutorials in Commerce subjects. A student sample was chosen to investigate the research questions for two reasons: First, there is no reason to believe that the responses triggered by certain answer scales in questionnaires should systematically differ between students or the general population. Second, data collection on campus enabled highly customized data collection: each respondent included in the final data set had to complete three consecutive surveys using different answer scales and the order in which the answer scales were presented to students was rotated, so that each subject had a unique combination of the exposure to different scales. For instance, students in the Strategic Marketing subject were first presented a questionnaire with binary response options in week 11 of session, followed by an ordinal scale in week 12 and a metric scale in week 13, whereas students in International Marketing received the metric questionnaire first, followed by a binary and an ordinal version. Binary, ordinal (seven point scale) and metric scales were incorporated.

Students were approached in lectures and tutorials and asked to complete a survey on water recycling. They were informed that the fieldwork would be carried out over three consecutive weeks. In the second and third week they were briefed that they would be recognizing the survey, but that their second and third response was crucial to investigate the stability of their responses across different survey conditions.

Two different constructs were included in the survey: behavioral intentions and attitudes. Attitudes were measured using a shortened version of the scale known as the New Ecological Paradigm (Dunlap, Van Liere, Mertig and Jones, 2000). The following statements were included and will be referred to as the NEP scale throughout the article: The balance of nature is very delicate and easily upset, When humans interfere with nature it often produces disastrous consequences, Humans are severely abusing the environment, The so-called "ecological crisis" facing humankind has been greatly exaggerated, If things continue in their present course, we will soon experience a major ecological catastrophe, Humans have the right to modify the natural environment to suit their needs, Humans were meant to rule over the rest of nature, Plants and animals exist primarily to be used by humans. Items were prompted with the words: "Please indicate your agreement with the following statements by ticking the respective box." In its binary version the options to answer were "I disagree" or "I agree", in the seven-point scale all seven scale points had numbers from 1 to 7 and the endpoints were verbally anchored as "Strongly disagree' and "Strongly agree". The metric answer scale was a horizontal line with no division markers. The endpoints were again anchored in the same way as for the ordinal scale.

Behavioral intentions were measured by giving respondents the following list of possible uses of recycled water: Watering the garden, Washing the car, Washing clothes, Cooking, Showering, Taking a bath, Drinking, Toilet flushing, Washing the house, windows, driveways, Watering of garden vegetables and herbs, Swimming pool, Fish pond, Air conditioning. The binary options to the question "Would you personally use recycled water for this purpose?" were "yes" and "no", ordinal options were "Very unlikely[1]", "Unlikely[2]", "Rather unlikely[3]", "Undecided[4]", "Rather likely[5]", "Likely[6]" and "Very likely[7]" where the question was asked as "How likely is it that you personally would use recycled water for this purpose?". Finally, the metric version used the same question, offered respondents a horizontal line to indicate their likelihood of using recycled water for these purposes and anchored the endpoints with "Very unlikely" and "Very likely".

In addition to the behavioral intentions and attitudes, the following information was collected from students: the actual beginning and end time of completing the questionnaire, perceived simplicity, perceived pleasantness, perceived speed and perceived ability to express their feelings. The responses were recorded in the same way for all questionnaire versions, namely using a five-point bipolar ordinal scale. These questions were related to the entire questionnaire, thus including both attitudes and behavioral intentions.

In total, 60 fully completed sets of data were available including three repeated measurements. Given that students did not show up to all classes, the originally balanced design (same number of questionnaires with certain sequences of presenting answer scales) is not reflected in the final data set: 16 respondents completed the ordinal-metric-binary sequence, 43 the binary-ordinal-metric and 1 the metric-binary-ordinal one.

RESULTS

All computations and graphics for the empirical analysis have been done using the R statistical software package (R Development Core Team, 2004).

For the direct comparison of the answers and the results of market structure analyses the answers on the different answer formats were rescaled to have values in the interval [0,1]. For instance, the ordinal answers at levels one to seven were transformed into equidistant values from zero to one. It is also important to note that - due to the longitudinal design - there is no need for the requirement that results for each answer format be representative for a given population in order to legitimately expect comparable results across answer formats.

Heterogeneity in respondents' answering patterns

Multi-category scales are known to be susceptible to scale usage heterogeneity. Different response styles (as, for instance, extreme or mild) or answer tendencies (as, for instance, Yeah-saying) deteriorate the information in the data. It is still an open question if scale usage heterogeneity is not only influenced by the answer format, but also by the construct measured. Response styles and answer tendencies lead to different answering patterns, i.e. the number of times each of the categories is selected.

For each respondent answering patterns can be defined, which indicate how often she or he has used each category. For this purpose the metric scale is split into seven equidistant levels and the answering patterns are determined for these categories. This means that the metric and ordinal results are directly comparable, whereas on the binary scale answering patterns are only available for two categories. A splitting of the metric scale is not only done to achieve comparability with the ordinal scale, but also because by directly using the continuous values the multi-modality of the answering patterns of respondents tending to use only the endpoints of the scale can not be modeled with a normal distribution.

Under the assumption that there are groups of respondents with the same response style or answer tendencies finite mixtures of multinomial distributions are fitted to the answering patterns of the respondents for each answer format. We assume that the answering patterns differ for the two constructs, but the class membership is the same for the respondents.

For the ordinal scale we choose a 2-segment solution because it already indicates which prototypes of answering patterns can be distinguished. The posterior probabilities indicate that the respondents can be very well assigned to the different segments. 98 percent of the respondents have a posterior probability of at least 0.9 for one segment. The probabilities for each category and construct are given in Figure 1 for each segment.

---------- Figure 1----------

Segment 1 contains 68 percent of the respondents, who obviously used the scale differently for the two constructs. While the seven categories have an equal frequency for the attitudes, the end-points are more likely for the intentional behavior. By contrast, respondents in Segment 2 (32 percent) have a mild response style and tend to use the middle categories for both constructs.

For the binary scale a 3-segment solution is suggested by the BIC information criterion. The different segments are well separated with 78 percent of the respondents being assigned to one cluster with a posterior probability of at least 0.9. The probabilities for each category and construct are given in Figure 2 for each segment.

---------- Figure 2 ----------

Clearly, no difference in use of the scale can be found for the attitudes which have a balanced design. Segment 2 (11.7%) and Segment 3 (15.0%) seem to reflect no different use of the scale, but the minorities of the population which reject its use and which are strongly inclined to use recycled water.

For the metric scale we also choose a 2-segment solution. The different segments are well separated with 97 percent of the respondents being assigned to one cluster with a posterior probability of at least 0.9. The probabilities for each category and construct are given in Figure 3 for each segment.

---------- Figure 3 ----------

Segment 1 contains 57 percent of the respondents. These respondents tend to use only the endpoints of the scale for both constructs. In Segment 2 - with 43 percent of the respondents - the middle categories are primarily used for the attitudes, whereas the use of the scale for the intentions do not differ very much from Segment 1.

The mixture solutions indicate that there are different response styles and tendencies present in the data for the ordinal and metric answer format which lead to different answering patterns. It can also be clearly seen by visual inspection that differences in response styles exist in dependence of the constructs investigated. Especially for the ordinal scales a group of respondents emerges that avoids the endpoints and prefers using the middle points.

These findings have numerous implications for marketing research: first, additional support is provided for the fact that ordinal scales are susceptible to capturing systematic response styles and similar findings apply to the metric scale. The consequence is that the existence of such response styles has to be investigated and possible data has to be normalized to eliminate the distortion effect of response styles if ordinal or metric answer scales are chosen. Second, aggregate analysis of such data can in fact hide extreme positions among respondents, potentially leading to misinterpretations of the data. Finally, the difference between constructs indicates that some constructs may be more suited for alternative answer scales than others. Based on the present results, behavioral intentions appear to be better suited for binary scales than attitudes.

Mappings between the answer formats

The longitudinal design enables the estimation of mapping functions between the different answer formats. It can be assumed that the mapping functions are not the same for all respondents and that segments of respondents exist who share similar mapping functions.

Therefore finite mixtures of logit models were fitted using the binary responses as dependent variables and the metric and the ordinal answers as independent variables, respectively. For the relationship between ordinal and metric answers mixtures of proportional odds-models (McCullagh, 1980) are fitted which represent a parsimonious alternative for multinomial logit models for ordinal data. The proportional odds assumption means that the ratio of corresponding odds is independent of the scale category and depends only on the difference between the covariate values. It is chosen for this estimation because the sample size is too small to sensibly estimate a large number of parameters.

The 2-segment solution for the mapping of metric responses to binary responses is shown in Figure 4. The choice of 2-segments is supported by the BIC information criterion. Segment 1 includes nearly all respondents with a size of 92 percent and fulfils the a-priori assumption that the cut-off point is close to 0.5. Segment 2, including 8 percent of the respondents, obviously contains the respondents who did not complete the questionnaires properly two times and appear to have given rather random answers.

---------- Figure 4 ----------

In Figure 5 the 2-segment solution mapping ordinal responses to the binary answers is shown. In order not to fit too many parameters only a linear term was fitted for the dependent variable. The resulting mapping is very similar to the mapping patterns revealed for the binary and metric formats: 92 percent of respondents are assigned to Segment 1 and 8 percent to Segment 2, which seems to collect all the respondents who tend to use category "no" on the binary scale, because the first 4 categories are mapped to "no" for the NEP scale and so are all except for the seventh category for the intentional behavior.

90 percent of the respondents are assigned to the same segment based both on the binary-metric and the binary-ordinal model.

---------- Figure 5 ----------

The most interesting mapping is between ordinal and metric, because it provides an opportunity to investigate whether the assumption generally made when analyzing ordinal data (that they have metric properties) is valid. The 3-segment solution is given in Figure 6. Segment 1 with 8 percent of the respondents contains the students who appear to give random answers. Segment 2 with 73 percent of respondents contains students who tend to use the endpoints of the ordinal scale, whereas Segment 3 representing 18 percent of the sample avoids the end points and prefers levels two and six on the ordinal scale. A comparison with the segmentation of the ordinal answering patterns shows that all respondents who are in Segment 3 are in Segment 2 in the ordinal model, which consists of the respondents exhibiting a mild response style.

---------- Figure 6 ----------

The estimated mixtures of mapping functions where able to identify the groups of respondents who did not complete the questionnaires properly. While interesting in this experimental setting, the advantage of this finding to market researchers and managers is limited, given that repeated measures are not typically undertaken. Of higher interest to practitioners, however, are the results of the mappings of different scales for respondents who did provide reliable answers: while translation to binary format is highly consistent using both metric and ordinal data as starting points, the mapping of ordinal answers to metric answers reveals the influence of response styles: some respondents refused ticking the endpoints on the ordinal scale while using the entire range when presented a continuous metric response format. The estimated mappings between metric and ordinal answers indicate that the ordinal answers are not implicitly constructed by the respondents from an underlying metric latent variable using equidistant cut-off points. Depending on the tendency to either prefer the endpoints of the ordinal scale or the middle points, the cut-off points are completely different. Therefore, it is doubtful if metric properties can be assumed for ordinal scales.

Differences in answers in dependence of answer formats

The estimated mean values across answer formats are compared to each other. Table 1 includes the mean values sorted in decreasing order with respect to the ordinal scale for the behavioral intention items of the questionnaire.

---------- Table 1 ----------

The mean answers for all three formats are very similar. For behavioral intentions only one single item ("washing clothes") demonstrates differences: respondents express lower likelihood of using recycled water for that purpose when using the binary scale then when using either ordinal or metric format. For the attitudinal questions the inspection of Table 1 indicates that the binary average deviates from the ordinal and metric values more strongly than this is the case for behavioral intentions.

The influence of the answer formats on the mean values of the different question is assessed using a Type-II ANOVA given in Table 2. The interaction effect between question and format for the ordinal and metric scale is not significant and therefore indicates that the mean values do not differ for the two answer formats. In fact no interaction between question and format is significant with p-value < 0.01 for ordinal versus metric. Between binary and metric the interaction is significant with p-value < 0.01 for the question on "balance of nature" and "washing clothes", while this is the case between binary and ordinal for the question on "balance of nature". This signifies that the average binary answers differ from the metric and ordinal answers only for a small number of questions (2 respectively 1 out of 19).

---------- Table 2 ----------

Practically these findings mean that, if average responses given by a sample for each question asked is the only information that is of interest to management, it makes no difference which answer format is chosen. In this case one could argue to either offer respondents the scale that is most pleasant to them, or alternatively, the most cost effective scale in terms of time and field cost: the binary scale.

Differences in managerial interpretations of positioning analyses based on different answer formats

Typically, the mean values will not be the only market data interpretation of interest to management. Frequently some form of market structure analysis is applied in order to derive strategic market information as, for instance, positioning. By doing this, further insight into how a brand is perceived as opposed to competitors can be gained or homogeneous market segments can be derived that represent useful target markets for organizations. Frequently this is done by undertaking factor analyses. The water recycling data is thus analyzed using this approach in order to determine whether or not the results from different scales lead to different managerial interpretations.

Factor analysis is conducted separately for the two different constructs, as in general only latent factors of questions for the same construct are of interest. Principal component analysis is applied to the correlation matrix of the answers on each of the different answer formats. The scree plots (Cattell, 1966) suggest two components for each of the answer formats and both constructs. For the NEP scale the cumulative proportion of explained variance is 66.5 for the ordinal, 51.5 for the binary and 74.6 for the metric scale. The cumulative proportion of explained variance for the behavioral intentions is 0.58 for the ordinal, 0.58 for the binary and 0.65 for the metric scale. The two factors which result after *varimax* rotation with Kaiser normalization of the first two principal components for each answer format are given in Table 3 for the NEP scale and in Table 4 for behavioral intentions. The *varimax* rotation is often applied in factor analysis to clarify the structure of the estimated loadings matrix. It maximizes the sum over factors of the variances of the normalized squared loadings. The questions are sorted in ascending order with respect to the loadings of the first factor for the ordinal answer format.

---------- Table 3 ----------

In Table 3 the factor loadings of the NEP scale are given. For the ordinal and metric scale the factor structure is already determined by the wording of the questions: the positively worded questions and the negatively worded questions form a factor respectively. For the binary answers the role of "exaggerated ecological crisis" and "balance of nature" is interchanged. This result is in fact more intuitive because it shows a negative correlation between the questions "exaggerated ecological crisis" and "lead to ecological catastrophe", whereas it might be suspected that the factor structure for the ordinal and metric scale is a mere artefact that the negative part of the scale is used in a different way as the positive part.

As can be seen in Table 4, the structure of the corresponding factors for the behavioral intentions are highly comparable for the three different answer formats. Factor 1 loads primarily on all questions where no direct personal contact is involved (from "Watering the vegetables" up to "Watering the garden") with recycled water, while the other factor loads on the remaining questions ("Drinking" to "Swimming pool") with direct personal contact. Only the question on "Air conditioning" does not to primarily load only on one factor.

---------- Table 4 ----------

As an objective criterion for the congruence between the factors for each answer format, we use Tucker's coefficients of congruence (Harman, 1964). The Tucker coefficients of congruence are defined by

$$CC_{pq} = \frac{\sum_{j=1}^{n} f_{jp} g_{jq}}{\sqrt{\left(\sum_{j=1}^{n} f_{jp}^{2}\right)\left(\sum_{j=1}^{n} g_{jq}^{2}\right)}}$$

where f_{jp} is the jpth element in the with respect to *varimax* rotated loadings matrix of one answer format, g_{jq} the jqth element of the loadings matrix of another answer format and n the number of attributes. The Tucker coefficients lie in the interval [-1,1] and measure the similarity between two factors on a factor-to-factor basis. The results are given in Table 5.

---------- Table 5 ----------

For the NEP scale the correspondence between metric and ordinal principal components is very high with 0.99 on average, whereas it is 0.81 for the first component where the binary scale is involved, which is relatively low in comparison to the other values. The resulting coefficients of congruence for the behavioral intentions are all at least 0.96 or larger indicating a strong correspondence of the rotated principal components. The average congruence is greatest for the formats metric and ordinal scale with 0.99.

From a managerial perspective this means, that interpretations generally do not significantly differ in dependence of the answer format used, although this is true to a higher extent for behavioral intentions than for attitudes.

Differences in reliability

Repeated measurements on the same scale are often used for test-retest reliability. In this case the test-retest reliability can be determined depending on the two different answer formats which are matched. These coefficients do not only indicate the stability of the answers but also the accordance of the answers on different answer formats. The reliability is determined by the correlation between the answer vectors. The results are given in Table 6.

---------- Table 6 ----------

The test-retest reliabilities are relatively high. They are better between the ordinal and the metric scale than where the binary scale is involved and they are generally better for the behavioral intentions than for the NEP scale thus reflecting the findings from the comparison of factor analytic results. The difference in test-retest reliability where the binary answer format is involved is smaller for the behavioral intentions, as in this case respondents on the ordinal and metric scale used the ends of the scale more frequent than in the NEP case where cautious answers in the middle of the answer categories offered were more likely.

As the binary scale has a purely methodological disadvantage in this comparison by offering only two categories, the agreement of the answers are compared using a second approach: collapsing both the ordinal and metric data to binary format and then computing reliability values. For this purpose the midpoints both on the ordinal and metric scale were excluded. The overall agreement using this approach is found to be quite high amounting to 79 percent across all scale comparisons. The overall agreement is higher for the behavioral intentions with 83 percent than the NEP scale with 73 percent. The comparisons between pairs of scales for the different constructs and both together are given in Table 6. It can be clearly seen that the agreement is similar for all three possible combinations. The assumption that the percentage of agreement is the same for each of the three possible combinations can not be rejected using a test for equal proportions (χ^2=0.80, p-value = 0.67 for both constructs, χ^2=0.56, p-value = 0.76 for behavioral intentions and χ^2=0.34, p-value = 0.84 for the NEP scale). This signifies that the answers on two different scales have the same percentage of agreement if the ordinal and metric answers are collapsed to binary.

From a managerial perspective similar conclusions can be derived for market research work as this was the case when factor analytic solutions were compared: first, differences between constructs exist. Behavioral intentions are stated more similarly on different scales than this is the case for attitudes. Nevertheless – when the mathematical disadvantage of binary scales in correlation measures is eliminated by collapsing the multi-category scales to binary format, no significant differences in agreement between pairs of scales could be determined.

Differences in user-friendliness

The duration of the questionnaire in the different answer formats was measured in minutes by subtracting begin time from end time. After eliminating answers with negative durations or durations of more than 20 minutes 174 observations are left (these are 97 percent of the answers). In the analysis of the relationship between duration and answer format the number of repetitions was included as covariate because a balanced design was not achieved with respect to the sequence of answer formats.

As an indicator for the possible influence of answer format and repetition a linear mixed-effects model with the logarithm of duration in minutes as dependent variable is used. The logarithm is chosen because the distribution of duration is slightly skewed to the right. Random effects are modeled for the respondents, which are assumed to be individually faster or slower in completing the questionnaire. Fixed effects are modeled for the answer formats and repetition and the estimated coefficients and standard deviations are given in Table 7.

---------- Table 7 ----------

As can be seen, the questionnaires were completed faster the second and third time the questionnaire was presented. This is plausible even independent of the answer formats given that the respondents are already familiar with the task and do not require the time to study the instructions as carefully anymore.

No significant difference in the time required to complete the questionnaire can be found for the ordinal scale and the metric scale. Questions in binary format, however, are completed significantly faster than items presented with seven response options. For example, if the mean values for binary (4.0 minutes) and ordinal scale (6.3 minutes) in the case where the questionnaire is answered for the first time are compared, the absolute difference is 2.3 minutes indicating that it took 58 percent longer to complete the questionnaire in the ordinal answer format.

With respect to the subjective evaluations of the answer formats it has to be acknowledged that the negative part of the scale is hardly ever used, with 3.9 percent slightly negative answers and 0.6 percent answers where the negative endpoint is ticked. This signifies that any difference between the three different answer formats has to be captured by the three remaining levels. The proportion of use of the negative part of the scale is the same for each of the four items (perceived simplicity, perceived pleasantness, perceived speed and perceived ability to express their feelings), as indicated by Fisher's exact test for count data (p-value = 0.36).

For the analysis of the subjective perceptions proportional odds-models were used due to the ordinal nature of the dependent variable. We test the proportional odds-assumption with a χ^2-test comparing the proportional odds-model with a full multinomial logit model. The p-values of the χ^2-test together with the estimated coefficients and standard deviations for the answer formats and repetitions are given in Table 7. For none of the four items measuring user-friendliness the proportional odds-assumption has to be rejected. With respect to repetition the third time is perceived as the most simple, the most pleasant and the quickest. While it is intuitive that simplicity and quickness increases with repetition, the reason that the third time is the most pleasant one, might be that it is the last time. The answer format has no significant influence with respect to pleasance and the ability to express the feelings as with respect to the AIC information criterion the model only including repetition as dependent variable is preferred to the model including repetition and format. With respect to speed and simplicity the models including repetition and format are suggested by the AIC information criterion, as the binary answer format is perceived as significantly quicker than the other answer formats and it is also perceived as significantly simpler than the metric answer format while the ordinal answer format does neither differ significantly form the binary nor from the metric answer format with respect to simplicity.

The findings on the user-friendliness of questionnaires have major implications for market research practice: if indeed respondents perceive binary scales to be as pleasant and simple as ordinal scales – which they have virtually been trained to use by generations of market researchers – the time-efficiency as well as perceived speed are major arguments to consider making more use of binary scales, in particular for constructs as behavioral intentions, where only few differences can be found with respect to the interpretations of findings.

CONCLUSIONS

The effect of answer formats was investigated using a longitudinal student sample in the context of both the measurement of attitudes and behavioral intentions with three repeated measurements on different scales: binary, ordinal and metric. The criteria used in this investigation were the susceptibility to response styles, the equivalence of responses, construct equivalence, time required to complete questionnaires in different formats and respondents' perceptions of how pleasant, simple and quick the surveys were. The longitudinal design allows comparisons on individual level where past research has typically compared independent samples. This not only enables stronger statements about the findings but also enables the investigation of how individuals internally transform responses to the same items from one scale to another, not requiring assumptions about which answer categories should be merged to form categories on scales with fewer options.

The analysis heterogeneity of answer patterns shows that groups of respondents can be identified who have different patterns of responding to the ordinal scale. One group of respondents who avoided the endpoints was found which might be due to a mild response style. Another group of respondents has a tendency to use the endpoints for expressing behavioral intentions. These findings supports a whole body of prior work where the susceptibility of ordinal scales to response styles was empirically determined under various survey conditions (Cronbach, 1950; Cunningham, Cunningham and Green, 1977; Greenleaf, 1992a; 1992b; Heide and Gronhaug, 1992; Watson, 1992; Van de Vijver and Poortinga, 2002; Welkenhuysen-Gybels, Billiet and Cambre, 2003). This avoidance of the use of the endpoints on the ordinal scale has an influence on the mapping functions from the metric scale, which means that the answers on the two answer formats are not comparable and cannot be transformed from one to the other without knowing the response style of the respondents.

Managerially, such susceptibility to tendencies of answering to certain scales independent of the actual content of the question endanger the quality of the interpretation of data. Scales that are less susceptible to such systematic patterns are preferable, leading back to a conclusion drawn by Cronbach (1950) that binary format might be the preferable option in order to avoid response styles. However, it could be claimed that such styles also manifest itself in binary format, but are not as easy to determine; an issue that has not received much attention in the past and might require more attention in future work on response scales.

The comparison of results of standard methods of analyses for the different answer formats indicated no substantial differences, both when simple means were computed and compared or when multivariate techniques like factor analysis were applied. Regardless of the answer format the main conclusions drawn are the same. Consequently it appears that market researchers are free to select the optimal answer format with respect to other evaluation criteria for scales, as, for instance, the speed of completing a questionnaire or low complexity for the respondents. These findings support conclusions drawn by researchers who have used a wide variety of approaches, including artificial data, to determine differences in interpretations of findings (Lehmann and

Hulbert, 1972; Martin, Fruchter and Mathis, 1974; Percy, 1976) while contradicting the results derived by Green and Rao (1970) who recommend six point scales as superior scale.

With respect to duration the binary answer format is significantly and substantially faster to complete, thus leading to smaller field costs and probably more reliable answers for long questionnaires where respondent fatigue can compromise data quality. For perceptions of the different answer formats no differences between simplicity, pleasantness, and the ability to express the feelings were found. Interestingly, these simple practical criteria are among the least investigated in the past. The findings of this study contradict the results presented by Jones (1968) and Preston and Colman (2000) who report that respondents prefer multiple categories because it enables them to better express their feelings.

The findings from all analyses reported in this study are summarized in Table 8. In conclusion, it seems that with regard to behavioral intentions market researchers have a choice of which scale they wish to present their respondents. The deviation of results will be minimal and other criteria, as for instance the speed of completing a questionnaire, can be used to make such a decision. Although the results of this study indicate that the same is true for attitudes, some evidence has emerged that respondents react differently when asked about attitudes than behavioral intentions. It would consequently be important to conduct more research into comparative studies of answer format effects across constructs to enable clear recommendations of which answer format offers the optimal trade-off between data quality, field work efficiency and mathematical correctness for each construct.

The main limitation of this study is the small sample size which was a consequence of the research design in which each group of respondents was presented with a different sequence of answer scales and three repeated measurements were taken. Future work should include other constructs that are typically measured in the market research context to determine whether the findings for behavioral intentions and attitudes are generalizable.

ACKNOWLEDGEMENTS

Bettina Grün is supported by the Austrian Academy of Sciences (ÖAW) through a DOC-FFORTE scholarship. We thank The Australian Research Council (ARC) for supporting this collaborative research project (LX0559628) under the ARC Linkage International Grant Scheme.

We would furthermore like to acknowledge the work of our research assistants, Petra Arko and Kristie Smith for entering the data and all the students and tutors at the University of X[to be included after review]X who made it possible for us to collect the longitudinal data required for this kind of research.

REFERENCES

Albaum, G., 1997. The Likert Scale Revisited: An Alternate Version. Journal of the Market Research Society 39(2), 331-348.

Arce-Ferrer, A. J., Ketterer, J. J., 2003. The Effect of Scale Tailoring for Cross Cultural Application on Scale Reliability and Construct Validity. Educational and Psychological Measurement 63(3), 484-501.

Bachman, J.G., O'Malley, P.M., 1984. Yea-Saying , Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles. Public Opinion Quarterly 48(2), 491-509.

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

Bradburn, N., Sudman, S., 1979. Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass

Byrne, B. M., Campbell, T. L., 1999. Cross-Cultural Comparisons and the Presumption of Equivalent Measurement and Theoretical Structure - A Look Beneath the Surface. Journal of Cross-Cultural Psychology 30(5), 555-574.

Cattell, R.B., 1966. The scree test for the number of factors. Multivariate Behavioral Research 1, 245-276.

Cheung, G. W., Rensvold, R. B., 2000. Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research using Structural Equation Modeling. Journal of Cross-Cultural Psychology 31(2), 187-212.

Chun, K.-T., Campbell, J. B., Yoo, J. H., 1974. Extreme Response Style in Cross-Cultural Research. Journal of Cross-Cultural Psychology 5(4), 465-480.

Churchill, G.A., Jnr., 1979. A Paradigm for Developing Better Measures of Marketing Constructs. Journal of Marketing Research 16(1), 64-73.

Clarke III, I., 2000. Extreme Response Style in cross Cultural Research: An Empirical Investigation. Journal of Social Behavior and Personality 15(1), 137-152.

Clarke III, I., 2001. Extreme Response Style in Cross Cultural Research. International Marketing Review 18(3), 301-324.

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

Cronbach, L.J., 1950. Further Evidence on Response Sets and Test Design. Educational and Psychological Measurement 10, 3-31.

Dillman, D.A., 1978. Mail and Telephone Surveys: The Total Design Methods, John Wiley, New York.

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., 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., Jones R. E., 2000. Measuring Endorsement of the New Ecological Paradigm: A Revised NEP Scale. Journal of Social Issues 56(3), 425-442

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

Greenleaf, E. A., 1992. Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles. Journal of Marketing Research 29(May), 176-188.

Greenleaf, E. A., 1992. Measuring Extreme Response Style. Public Opinion Quarterly 56(3), 328-351.

Harman, H. H., 1964. Modern Factor Analysis. Chicago: University of Chicago Press.

Heide, M., Gronhaug, K., 1992. The Impact of Response Styles in Surveys: A Simulation Study. Journal of the Market Research Society 34(3), 215-223.

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

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

Marin, G., Gamba, R.J., Marin, B.V., 1992. Extreme Response Style and Acquiescence among Hispanics - The Role of Acculturation and Education. Journal of Cross-Cultural Psychology 23(4), 498-509.

Martin, W.S., 1973. The Effects of Scaling on the Correlation Coefficent: A Test of Validity. Journal of Marketing Research 10(3), 316-318.

Martin, W.S., 1978. Effects of Scaling on the Correlation Coefficient: Additional Considerations. Journal of Marketing Research 15(2), 304-308.

Martin, W. S., Fruchter, B., 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., 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.

McCullagh, P., 1980. Regression Models for Ordinal Data. Journal of the Royal Statistical Society, Series B (Methodological) 42 (2), 109-142

Nunnally, J.C., 1967. Psychometric Theory. New York: McGraw-Hill, 1st edition

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

Payne, S. L., 1951. The Art of Asking Questions. Princeton, NJ: Princeton University Press.

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.

Peterson, C.R., Semmel, A., von Baeyer, C., Abramson, L.Y., Metalsky, B.I., Seligman, M.E.P. (1982). The Attributional Style Questionnaire. Cognitive Therapy and Research 6 (3), 287-300.

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

R Development Core Team, 2004. 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.

Roster, C.A., Rogers, R., Albaum, G., in press. A Cross-Cultural/National Study of Respondents' Use of Extreme Categories. Journal of Cross-Cultural Psychology.

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

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 de Vijver, F. J. R., Poortinga, Y. H., 2002. Structural Equivalence in Multilevel Research. Journal of Cross-Cultural Psychology 33(2), 141-156.

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

Van Herk, H., Poortinga, Y.H., Verhallen, T.M.M., 2004. Response Styles in Rating Scales - Evidence of Method Bias in Data From Six EU Countries. Journal of Cross-Cultural Psychology 35(3), 346-360.

Watkins, D., Cheung, S., 1995. Culture, Gender and Response Bias. Journal of Cross-Cultural Psychology 26(5), 490-504.

Watson, D., 1992. Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale: An Application to Class Consciousness. Sociological Methods and Research 21(1), 52-88.

Welkenhuysen-Gybels, J., Billiet, J., Cambre, B., 2003. Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items. Journal of Cross-Cultural Psychology 34(6), 702-722.

TABLES

Table 1

Mean answers for each answer format

	New Ecological Paradigm – Attitude			
	Disastrous consequences	Balance of nature	Severely abusing	Lead to ecological catastrophe
Ordinal	0.69	0.69	0.68	0.67
Binary	0.79	0.88	0.75	0.61
Metric	0.64	0.68	0.59	0.61
	Exaggerated ecological crisis	Right to modify	Meant to rule	Animals exist to be used
Ordinal	0.47	0.45	0.33	0.33
Binary	0.50	0.33	0.34	0.26
Metric	0.48	0.44	0.37	0.34

	Behavioral Intentions				
	Watering the garden	Toilet flushing	Washing the car	Washing the house	Fish pond
Ordinal	0.90	0.88	0.84	0.78	0.60
Binary	0.92	0.90	0.80	0.88	0.67
Metric	0.87	0.87	0.84	0.80	0.63
	Watering of vegetables	Air conditioning	Washing clothes	Swimming pool	Showering
Ordinal	0.57	0.56	0.50	0.37	0.21
Binary	0.68	0.67	0.36	0.25	0.17
Metric	0.60	0.65	0.53	0.37	0.23
	Taking a bath	Cooking	Drinking		
Ordinal	0.17	0.16	0.08		
Binary	0.17	0.09	0.03		
Metric	0.23	0.20	0.14		

Table 2

Type-II ANOVA for answers with respect to question and answer format

	Sum of squares	Degrees of Freedom	F-value	p-value
Ordinal versus Binary				
Question	170.35	20	70.78	<0.001
Format	0.03	1	0.23	0.63
Question:Format	4.66	20	1.94	0.01
Residuals	295.07	2452		
Ordinal versus Metric				
Question	131.27	20	88.48	<0.001
Format	0.05	1	0.61	0.44
Question:Format	1.09	20	0.74	0.79
Residuals	183.02	2467		
Binary versus Metric				
Question	156.83	20	64.19	<0.001
Format	0.002	1	0.01	0.90
Question:Format	5.91	20	2.42	<0.001
Residuals	298.91	2447		

Table 3

Two principal components after *varimax* rotation for each answer format for the NEP scale

		Animals exist to be used	Meant to rule	Exaggerated ecological crisis	Right to modify
Factor 1	Ordinal	0.07	0.04	-0.05	-0.06
	Binary	-0.05	-0.08	0.32	0.03
	Metric	0.07	-0.06	-0.07	0.06
Factor 2	Ordinal	0.48	0.52	0.43	0.54
	Binary	0.55	0.57	0.26	0.44
	Metric	0.51	0.52	0.45	0.51
		Lead to ecological crisis	Balance of nature	Disastrous consequences	Severly abusing
Factor 1	Ordinal	-0.42	-0.50	-0.52	-0.54
	Binary	-0.53	-0.07	-0.47	-0.62
	Metric	-0.46	-0.51	-0.53	-0.48
Factor 2	Ordinal	0.02	-0.08	-0.01	0.07
	Binary	0.09	-0.29	-0.07	0.05
	Metric	0.03	-0.09	-0.01	0.07

Table 4

Two principal components after *varimax* rotation for each answer format for behavioral intentions

		Drinking	Cooking	Taking a bath	Showering	Washing clothes
Factor 1	Ordinal	0.27	0.03	0.01	-0.01	-0.10
	Binary	0.07	0.03	0.02	0.00	-0.14
	Metric	0.20	0.10	0.04	0.00	-0.11
Factor 2	Ordinal	-0.34	-0.42	-0.45	-0.46	-0.30
	Binary	-0.39	-0.36	-0.49	-0.47	-0.30
	Metric	-0.37	-0.38	-0.40	-0.42	-0.35

		Swimming pool	Air conditioning	Watering of vegetables	Fish pond	Washing the house
Factor 1	Ordinal	-0.12	-0.24	-0.29	-0.30	-0.39
	Binary	-0.05	-0.30	-0.26	-0.35	-0.43
	Metric	-0.20	-0.17	-0.28	-0.29	-0.41
Factor 2	Ordinal	-0.32	-0.21	-0.09	-0.10	-0.06
	Binary	-0.37	-0.12	-0.04	-0.04	0.06
	Metric	-0.31	-0.30	-0.15	-0.12	-0.03

		Washing the car	Toilet flushing	Watering the garden
Factor 1	Ordinal	-0.40	-0.41	-0.44
	Binary	-0.35	-0.44	-0.43
	Metric	-0.44	-0.40	-0.43
Factor 2	Ordinal	-0.03	0.05	0.20
	Binary	-0.04	0.07	0.08
	Metric	0.02	0.08	0.17

Table 5

Tucker's coefficients of concordance between the rotated principal components

	New Ecological Paradigm			Behavioral Intentions		
	Ordinal	Ordinal	Binary	Ordinal	Ordinal	Binary
	Binary	Metric	Metric	Binary	Metric	Metric
Comp. 1	0.81	0.98	0.81	0.97	0.99	0.96
Comp. 2	0.95	1.00	0.95	0.97	0.99	0.96

Table 6

Test-retest reliability and agreement between the different answer formats for the complete questionnaire and for the two constructs separately

	Test-Retest Reliability			Agreement		
	Ordinal	Ordinal	Binary	Ordinal	Ordinal	Binary
	Binary	Metric	Metric	Binary	Metric	Metric
Both constructs	0.66	0.74	0.63	0.79	0.80	0.78
Behavioral Intentions	0.71	0.78	0.71	0.83	0.83	0.82
New Ecological Paradigm	0.57	0.63	0.48	0.74	0.74	0.72

Table 7

Estimated coefficients and standard deviations for the linear mixed-effects model with the logarithmised duration and the proportional-odds models for the perception of the scales

		Duration		Simple		Pleasant		Quick		Feelings	
Variables		Coef.	Std. Err	Coef.	Std. Err	Coef.	Std. Err	Coef.	Std. Err	Coef.	Std. Err
Format	Binary	-0.30**	0.08	-0.48	0.44	-0.58	0.43	-1.12*	0.44	0.51	0.43
	Metric	-0.04	0.08	0.55	0.44	0.20	0.45	-0.16	0.43	0.02	0.44
Repetition	2	-0.31**	0.08	-0.09	0.43	-0.24	0.43	-1.20**	0.44	0.01	0.43
	3	-0.60**	0.08	-1.18**	0.45	-1.39**	0.45	-1.60**	0.45	-0.53	0.45
χ^2-Test	p-value			0.52		0.16		0.14		0.79	

*p-value < 0.05; **p-value < 0.01

Table 8

Summary of findings

Criterion	Result	Construct-dependence
Susceptibility to response styles	Empirical support for prior findings that ordinal and metric scales are susceptible to response styles.	Yes, ordinal and metric more susceptible.
Individual level transformations between answer formats	Mappings between binary and the other two formats can be achieved in a reliable manner, ordinal and metric mappings suffer from the impact of response styles on the transformations.	No, when mapped to binary. Yes, when metric mapped to ordinal.
Differences in average values	Results of all three answer formats do not differ significantly if only mean values are of interest.	No
Construct equivalence	Factor analytic results indicate the same underlying structure across all answer formats.	Yes, behavioral intentions show higher equivalence values.
Reliability / agreement	Scales render equally high levels of agreement.	Yes, behavioral intentions show higher agreement levels.
Time required for completion	Binary format is quicker to complete.	-
Perceived speed	Binary format is perceived as quicker to complete.	-
Perceived simplicity	Binary significantly simpler than metric. No difference between ordinal and the other two answer formats.	-
Perceived pleasantness	No difference between scales.	-
Perceived ability to express feelings	No difference between scales.	-

FIGURES

Figure 1

Answering patterns of the segments of the fitted finite mixture for the ordinal answer format with respect to the two different constructs

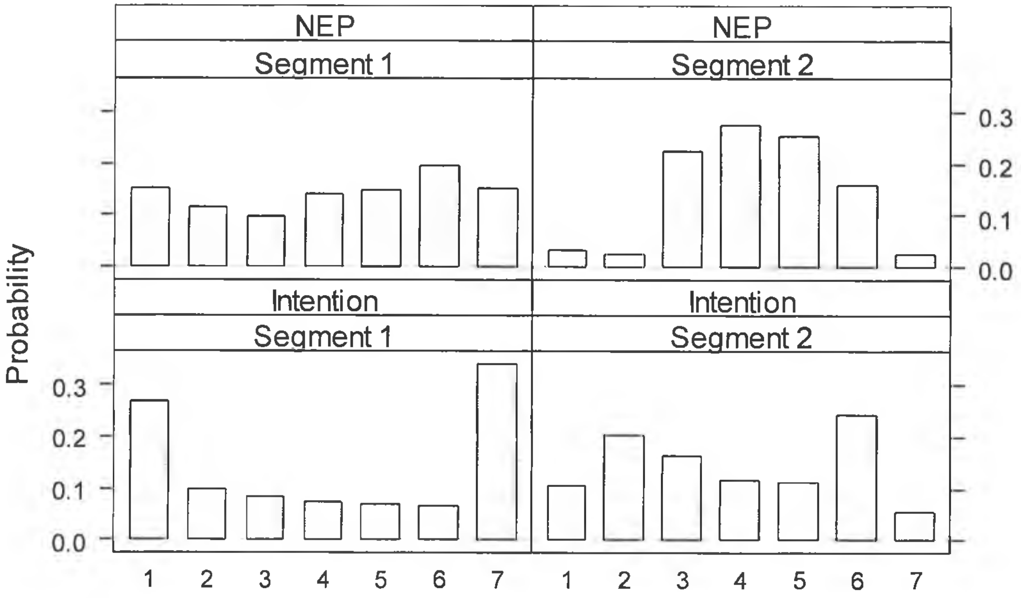


Figure 2

Answering patterns of the segments of the fitted finite mixture for the binary answer format with respect to the two different constructs

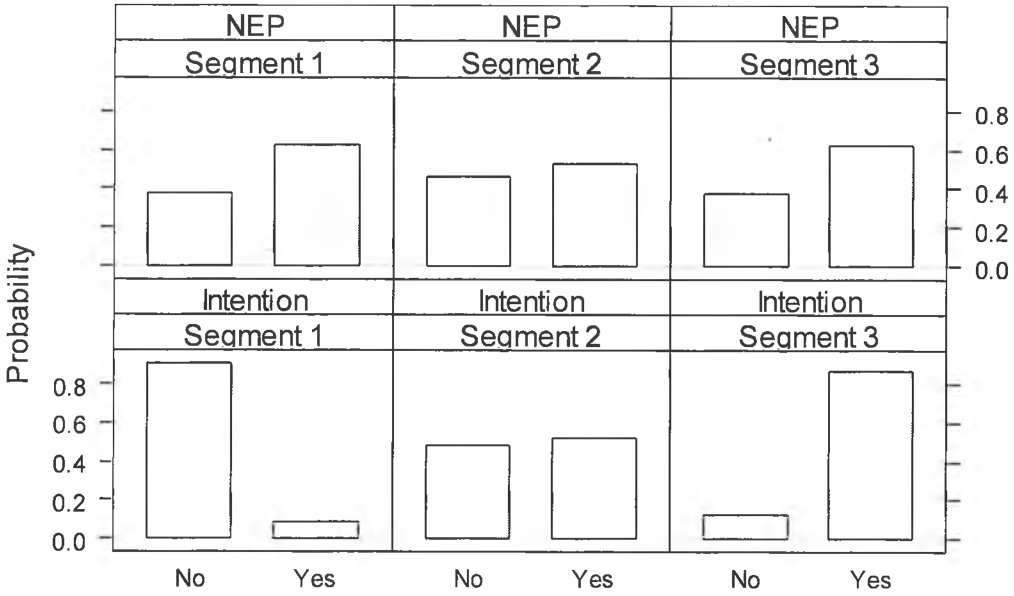


Figure 3

Answering patterns of the segments of the fitted finite mixture for the metric answer format with respect to the two different constructs

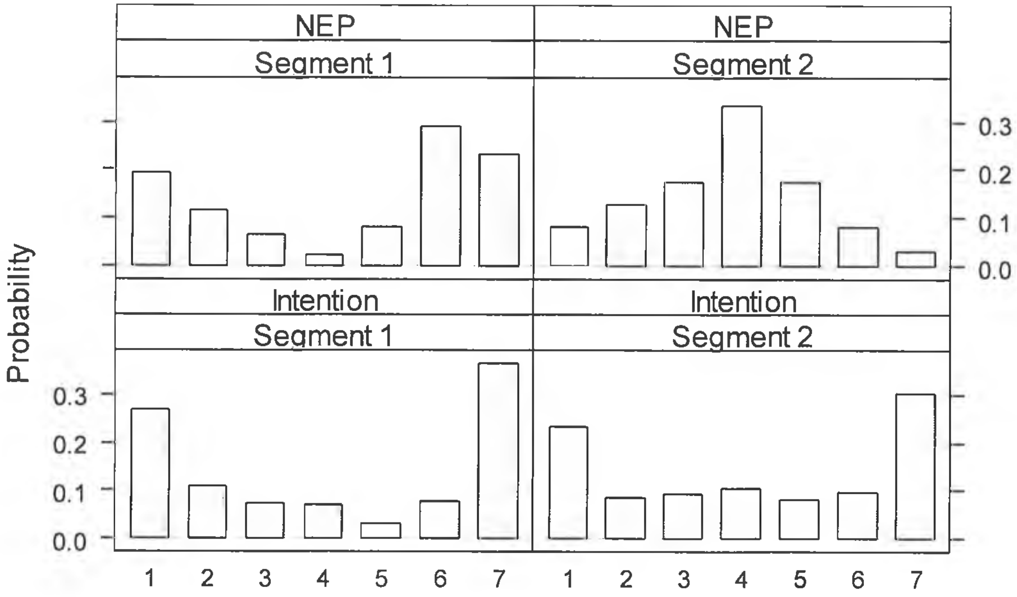


Figure 4

Mappings from the metric to the binary scale

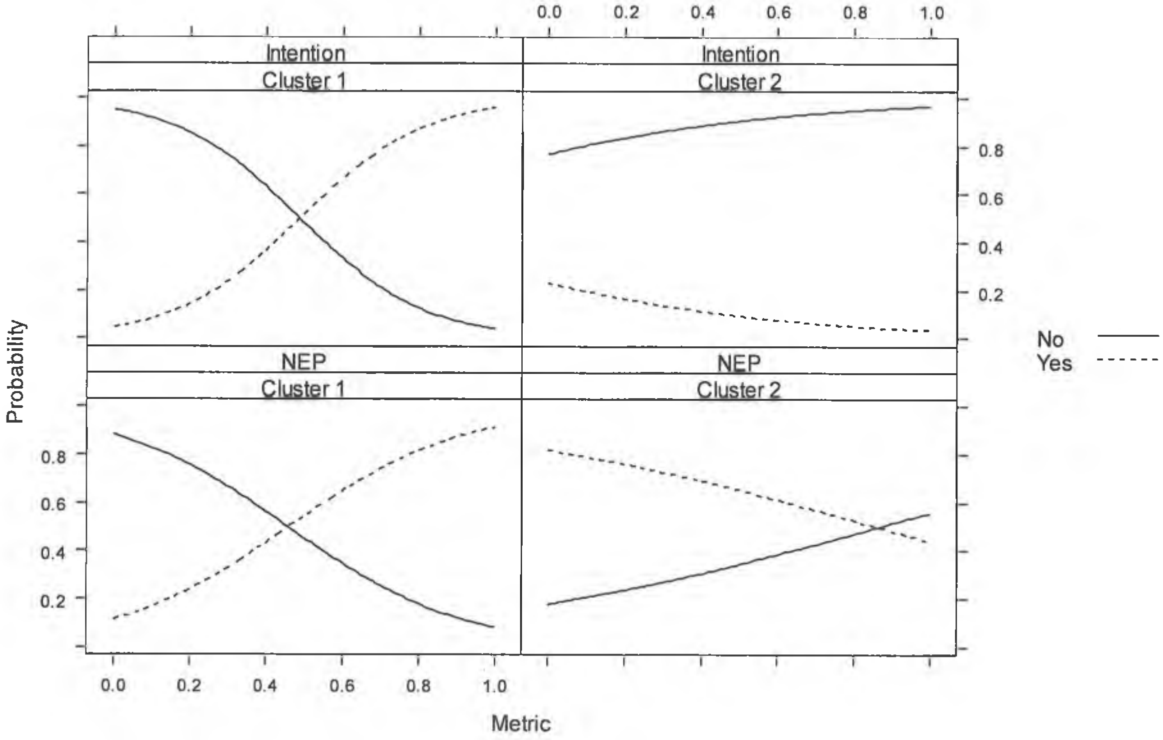


Figure 5

Mappings from the ordinal to the binary scale

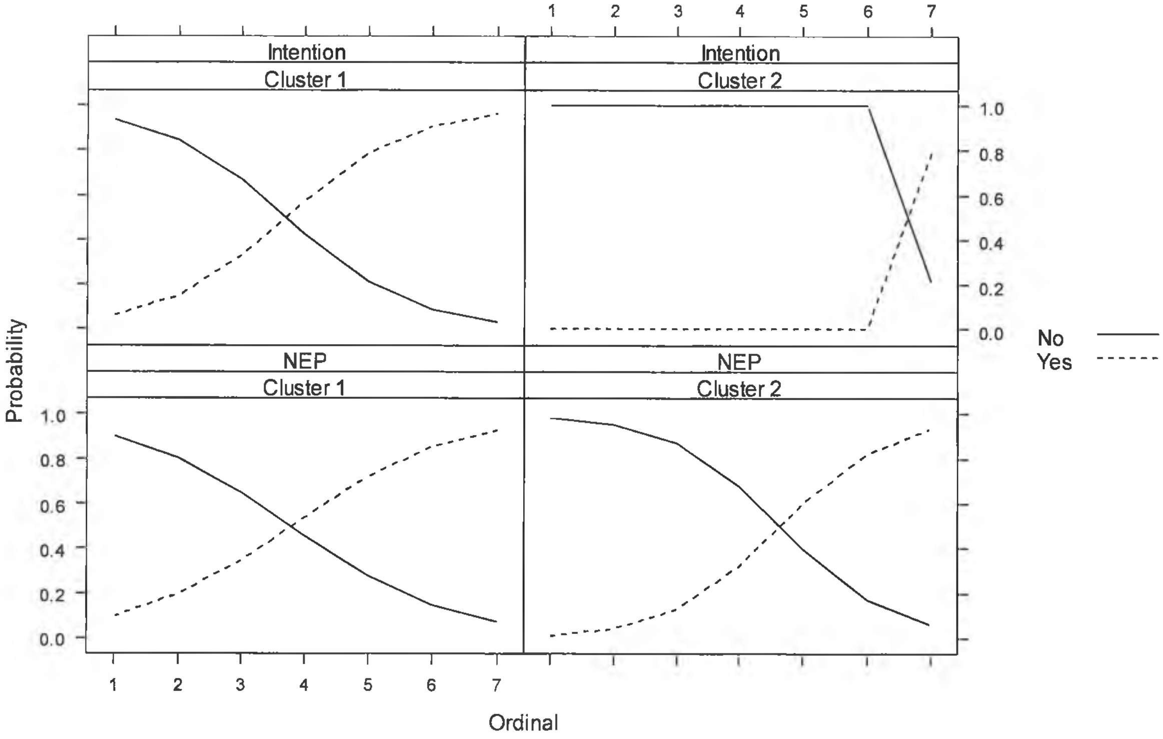
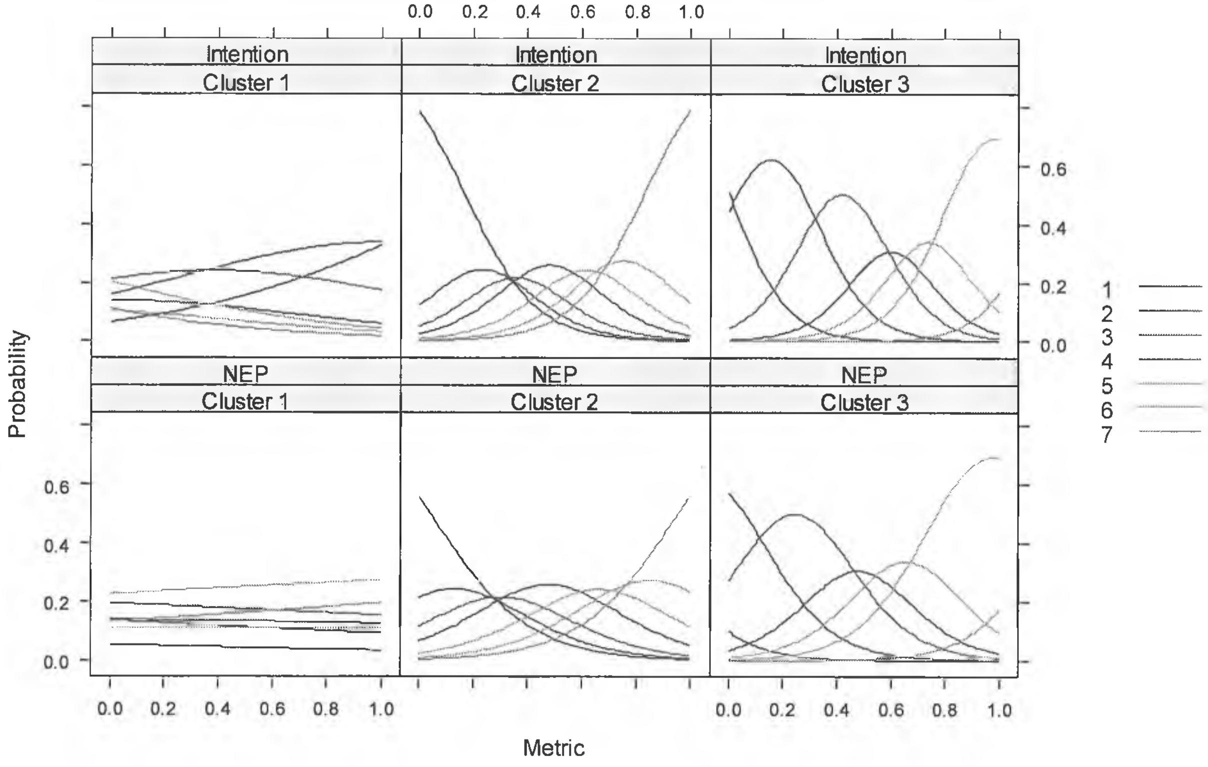


Figure 6

Mappings from the metric to the ordinal scale


Intention
Cluster 1
Intention
Cluster 2
Intention
Cluster 3
NEP
Cluster 1
NEP
Cluster 2
NEP
Cluster 3
Probability
Metric