2007

How constrained a response: a comparison of binary, ordinal and metric answer formats

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

Bettina Grun
Vienna University of Technology, Austria, bettina@uow.edu.au

Publication Details

This article was originally published as: Dolnicar, S & Grun, B, How constrained a response: a comparison of binary, ordinal and metric answer formats, Journal of Retailing and Consumer Services, 2007, 14 (2), 108-122.

How constrained a response: a comparison of binary, ordinal and metric answer formats

Abstract
A question is the main measurement instrument in the social sciences. Yet no conclusive results exist with respect to the suitability of alternative answer formats for typical constructs studied in this field. Furthermore, no prior studies have used real answers from respondents to investigate differences in responses as a consequence of answer formats, typically assuming the way in which individuals translate their responses from one answer format to another. In this study we make a first step towards filling the above two gaps. We investigate answer format effects for two different constructs (attitudes, intentions) and three formats (binary, ordinal, metric) using a repeat measurement design. Results indicate that formats lead to the same managerial interpretations with the same reliability; differ in speed and perceived speed while being perceived as equally simple, pleasant, and useful to express feelings. Binary and metric answer formats are found to represent interesting alternatives to the predominantly used ordinal format, especially when speed of completion or the availability of metric data for analyses is essential.

Keywords
Answer format, response scale, binary, ordinal, metric, Likert

Disciplines
Business | Social and Behavioral Sciences

Publication Details
This article was originally published as: Dolnicar, S & Grun, B, How constrained a response: a comparison of binary, ordinal and metric answer formats, Journal of Retailing and Consumer Services, 2007, 14 (2), 108-122.

How constrained a response:

A comparison of binary, ordinal and metric answer formats

Sara Dolnicar*

School of Management & Marketing

marketing research innovation centre (mric)

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.

How constrained a response:

A comparison of binary, ordinal and metric answer formats

Abstract

A *question* is the main measurement instrument in the social sciences. Yet no conclusive results exist with respect to the suitability of alternative answer formats for typical constructs studied in this field. Furthermore, no prior studies have used real answers from respondents to investigate differences in responses as a consequence of answer formats, typically assuming the way in which individuals translate their responses from one answer format to another.

In this study we make a first step towards filling the above two gaps. We investigate answer format effects for two different constructs (attitudes, intentions) and three formats (binary, ordinal, metric) using a repeat measurement design.

Results indicate that formats lead to the same managerial interpretations with the same reliability; differ in speed and perceived speed while being perceived as equally simple, pleasant, and useful to express feelings. Binary and metric answer formats are found to represent interesting alternatives to the predominantly used ordinal format, especially when speed of completion or the availability of metric data for analyses is essential.

Keywords: Answer format, response scale, binary, ordinal, metric, Likert

INTRODUCTION

The success of marketing activities depends on the quality of market research that has been undertaken to understand market mechanisms. Market research results are only as good as the data that is collected from the market. A major decision that can influence data quality in survey research, the most popular form of market research, is the choice of the answer format. Presently, ordinal answer format dominates commercial and academic marketing research (Van der Eijk, 2001). In the academic field this is best illustrated by reviewing recent publications in the leading marketing journals. In recent issues of the Journal of Marketing Research (42(2)), the Journal of Consumer Research (32(1)) and the Journal of Marketing (69(3)), 21 articles reported empirical findings based on consumer responses. Of these 21, 18 (86 percent) chose an ordinal scale as answer format. The dominance of the ordinal answer format is surprising given the methodological problems associated with it, which will be reviewed later in this manuscript. One possible reason for the popularity of ordinal answer formats is that data is simpler to enter as opposed to asking respondents to make a cross on a line, which requires measurement of every single response to determine the precise metric value. By assuming that the construct underlying the ordinal scale is metric and asking respondents to use an ordinal scale, data entry becomes simpler and quicker while it is assumed that it is justifiable that data is treated as metric.

Two major changes in the marketing research environment make it worthwhile to reinvestigate answer formats: the increased used of web-based surveying, and the high level of saturation of consumers with market research. The fact that web-based surveying is quickly replacing conventional paper-and-pencil surveys makes is feasible to collect metric data by offering respondents a scroll bar. Saturation of consumers with market research puts pressure on questionnaire development. The shorter and simpler the questionnaire the higher the probability

that potential respondents will agree to participate, thus potentially reducing response bias and possibly fatigue effects. In view of these changes, the dominance and continuous popularity of ordinal scales becomes questionable. Metric answer formats are preferable with respect to the data properties and binary formats can speed up the data collection process. However, recommending substitution of ordinal answer formats with metric and binary answer formats is only legitimate under certain conditions which will be discussed below.

Due to the importance of answer formats in the marketing research process, numerous comparisons of answer format effects have been undertaken in the past, which guide this reinvestigation. A number of distinct streams of research have developed using different criteria for the evaluation of the "optimality" of an ordinal 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 published so far have, however, adopted a within respondent repeated measurement approach to answer format comparison, thus implicitly assuming to know the respondents' transformations from one scale to another[1]. The present work makes one step towards filling this gap by using within respondent repeated measurement 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 measured.

[1] 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.

Clearly, in such comparisons of answer formats it is not known which the true answer or the true managerial interpretation is. We deal with this problem by requiring the formats which are referred to as alternative formats (metric and binary) to not perform significantly worse than ordinal answer formats along these criteria. Should this be the case, the different advantages of alternative formats would warrant selective substitution as it indicates that the systematic errors due to the specific answer format are different.

The present study makes the following contributions: whether or not there are differences in noise levels (content-unrelated error) between alternative answer formats is studied by (1) analyzing the way in which respondents transform their answers from one scale to another. Managerial interpretations derived from different answer formats are evaluated by comparing (2a) differences in the mean answers of respondents and (2b) differences in structural equivalence between answer formats. (3) Reliability values based on repeated measurements on different answer formats are compared and generalizability coefficients are determined for the specific objects of measurement. The burden of respondents is established by investigating differences in (4a) duration, (4b) perceived duration of the questionnaire, (4c) perceived complexity, and (4d) the perceived ability to express feelings.

The results of this study have major implications for marketing research: if the same managerial conclusions result, respondents are not burdened more, results are equally valid and reliable and do not contain more noise, binary or metric scales offer marketing researchers some distinct advantages and should be seriously considered as alternatives to the typically used ordinal answer format.

PRIOR WORK

The most comprehensive discussion of ordinal answer formats was provided 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-rater 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.

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). A number of studies, however, conclude the opposite (Symonds, 1924; Nunnally, 1967; Jones, 1968; Oaster, 1989; Finn, 1972; Ramsay, 1973).

Controversy also results 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 to 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) formats. 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 studies (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."

While the issue of optimal answer formats has clearly attracted attention from social scientists in the past, two significant gaps can be identified in the body of prior work: (1) scale comparisons were typically undertaken with artificial data or by using actual responses based on one answer format only which were then artificially transformed to another scale level. In doing so, these studies assumed to know a priori in which way respondents would translate an ordinal response to a binary response. The present study attempts to fill this gap by collecting actual responses on three different answer formats by respondents in a repeat measurement design. Furthermore, (2) prior work was limited to one – seemingly randomly chosen - construct under study, although it is plausible to assume that different answer formats would be more or less suitable when used to measure specific constructs. The present study includes two constructs which are typically studied in marketing research: behavioral intentions and attitudes.

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 because 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. The assignment of questionnaire versions to tutorials was random, the assignment of students to tutorials, however, was not. The bias that could be expected using a student sample is that they are more highly educated and their cognitive capabilities may be better than this would be the case in the general population. Findings can therefore not be generalized beyond the student population. However, the findings derived from the student sample are indicative of mechanisms that may be at work and could be replicated in other populations of interest to researchers using surveys as the instrument of data collection.

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. They were informed that they would be recognizing the survey in the following two

weeks, but that their response to all three questionnaires was crucial to enable us to investigate differences in their responses across different kinds of questionnaires. The limitation of this approach is that students were vaguely aware of the aim of the study and may have changed their response behavior as a consequence of this knowledge. However, experiences from prior survey studies with students indicated that this approach was more likely to keep students motivated in participating than any attempt to surprise them in the consecutive weeks, as students inevitably ask why they have the same questionnaire.

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 and Van Liere, 1978, 1984; Dunlap, Van Liere, Mertig and Jones, 2000). The New Ecological Paradigm Scale in its long (and later shortened) version has been validated and revalidated later by the original authors (Dunlap et al., 2000) and has been extensively used in studies of environmental behavior to assess different aspects of environmental concern. 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 balanced seven-point scale all seven scale points had numbers from 1 to 7 and only

the endpoints were verbally anchored as “Strongly disagree’ and “Strongly agree”. The seven point scale was chosen because of the recommendation by Cox (1980) resulting from an extensive review of prior work comparing different formats of ordinal answer options. 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. It should be noted at this point that – although it is typically assumed that the construct studied is represented in a consumer’s mind as a metric construct – we do not know which metric best represents the opinions sought from the respondents. While the metric scale may be most desirable from the perspective of data analysis, it may well be that the constructs studied are not in fact represented in a metric way in the consumer’s minds.

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”. It should be mentioned that it is possible that student respondents did not undertake many of those activities at the point of being surveyed because, for instance, the parents clean the driveway. However, the questions were hypothetical by very

nature. It is therefore expected that students responded in a scenario-evaluation manner rather than they answered based on past experience, given that recycled water is presently not available to these respondents. The results for those items which require students to undertake them personally (drinking, showering) did not differ from those that other household members could be doing.

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. The five point format was chosen because it was different from all the answer formats used in the different versions of the questionnaire and because it helped the respondents to separate between the task of completing the questionnaire and the task of evaluating the questionnaire.

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, 2005).

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. This is based on the assumption that strong agreement is captured equally on the different scales, whereas slight agreement can be expressed by a smaller value in the ordinal and metric scale, but the same value as for strong agreement is assigned on the binary scale. This transformation was chosen because it is assumed that it suitably minimizes the differences in the estimation of mean values for the different answer formats. It is also important to note that - due to the within respondent repeated measurement 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.

1 (Noise) Mappings between the answer formats

The within respondent repeated measurement design enables the estimation of mapping functions between different answer formats. It is assumed that individual mappings can be described by a binomial logit model for metric to binary and for ordinal to binary, while a proportional odds-model (McCullagh, 1980) is assumed for the mapping from metric to ordinal. The proportional odds assumption signifies that the odds ratio of cumulative probabilities is independent of the scale category and depends only on the difference between the covariate

values, i.e. an increase in the observed metric value increases the cumulative probabilities for each scale category of the ordinal scale. While for a multinomial logit model different parameters are estimated for each category and covariate, the proportional odd-model only has different parameters for each category for the intercept and the same parameters for all categories for the other covariates.

It is unlikely that the mapping functions are the same for all respondents. To account for heterogeneity in the respondents' mapping functions finite mixture models are fitted which provide a model-based approach to include unobserved heterogeneity (Wedel and Kamakura, 2001). Finite mixtures of logit models are therefore fitted using the binary responses as dependent variables and the metric and the ordinal answers as independent variables, respectively, and mixtures of proportional odds-models for the relationship between ordinal and metric. The finite mixture model is given by

$$H(y \mid x, \Theta) = \sum_{s=1}^{S} \pi_s F(y \mid x, \beta_s)$$

where S are the number of segments, π_s are the segment sizes which are nonnegative and sum to one and β_s are the parameters for segment s and the component distribution F. Θ is the vector of all parameters which determine the mixture model. The y are the dependent variables and x are the covariates. The posterior probabilities are the probabilities to be from a certain segment given the observation *(x, y)*.

The use of finite mixture models in contrast to a mixed-model approach allows to partition the respondents into segments where the respondents in each segment have similar mapping functions. In order to account for possible variation in the mapping functions due to the construct

under study the mappings are allowed to vary for the two constructs while the segment membership is fixed.

The 2-segment solution for the mapping of metric responses to binary responses is shown in Figure 1 (left). For this Figure we clustered the observations with respect to the a-posteriori probabilities given the mixture model and depicted the relationship between the independent and dependent variable in the model for each of the derived segments separately. The top half of Figure 1 depicts aggregated binary answers given for each segment and construct. In the bottom half of Figure 1 the relationship between binary answer and metric answer is illustrated for each segment. The choice of two segments is supported by the Bayesian information criterion (BIC) which is in general used for model selection because it allows to decide on a trade-off between model fit and model complexity. 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. In addition the prediction of the binary answers based on the metric is better for the behavioral intentions. 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 1 ----------

In Figure 1 (center) 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 for the behavioral intentions, because all categories except for the seventh are mapped to “no”. For the first segment it can be again seen that the prediction of the binary answer based on the ordinal is better for the behavioral intentions. This accordance between the binary-metric and binary-ordinal solutions is confirmed by a comparison of the segment memberships: 90 percent of the respondents are assigned to the same segment based both on the binary-metric and the binary-ordinal model.

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 1 (right). Segment 1 with 8 percent of the respondents contains the students who appear to give random answers. Such respondents, if identifiable, should be eliminated from the data set. However, given that the identification was only possible due to the repeated measurements, it is unlikely that they would have been identified in a typical single wave survey. Segment 2 with 74 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 the middle points on the NEP scale and levels two and six for the behavioral intentions on the ordinal scale. For this segment prediction of the endpoints of the ordinal scale given the metric answer is better than for segment 2. A cross-tabulation of the answering behavior segmentation with the sequence of confrontation with the answer formats reveals no association (p-value=0.72), i.e. there is no indication that segment three is an artifact of the difference in the sequence of answer formats.

The estimated mixtures of mapping functions were 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 marketing researchers and managers is limited, given that repeated measures are not typically undertaken. However, if a repeat design would be affordable for the mere purpose of identifying respondents who do not complete the questionnaire carefully, such respondents should be eliminated form the data prior to analysis.

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.

Please note that the above conclusion is based on the assumption that respondents have used the metric scale in a metric way, thus acknowledging that each unit along the scale is of equal distance.

2 Managerial interpretation

The analyses for managerial interpretation focus on the items, i.e. the likelihood of use of recycled water for the different purposes and the agreement levels with the different NEP statements. For this purpose the mean values of the answers on the different answer formats are

compared for each item and factor analysis is applied to investigate the relationship between the different items and assess if some items can be combined to form a single factor. This means that with respect to these analyses the items are the objects of measurement.

2a (Managerial interpretation) 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.

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

The mean answers for all three formats are very similar. For behavioral intentions only one single item ("washing clothes") demonstrates differences of more than 0.15 in absolute values: 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 < 0.05 for ordinal versus metric. Between binary and metric the

interaction is significant with p-value < 0.05 for the question on “balance of nature”, “washing clothes”, “disastrous consequences” and “severely abusing” while this is the case between binary and ordinal for the question on “balance of nature” and “washing clothes”. This signifies that the average binary answers differ from the metric and ordinal answers only for a small number of questions (4 respectively 2 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.

2b (Managerial interpretation) Structural equivalence

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. The factor analysis method chosen is principal component analysis applied to the correlation matrix of the answers on each of the different answer formats followed by *varimax* rotation. The sample sizes are rather small for this kind of analysis. In general the recommended minimum necessary sample sizes depend on different influence factors, such as the number of factors, the variables-to-factor ratio or the level of communality. Mundfrom, Shaw and Lu Ke (2005) perform a simulation study and conclude that for a variables-to-factor ratio of 7 the minimum sample size never exceeds 85 for good-level agreement. Given these results it can be assumed that factor analysis gives good results for the behavioural intentions as this ratio is equal to 6.5, whereas it suggests that for the analysis of the NEP scale the number of respondents might be rather low as this ratio is only 4. To the authors' knowledge PCA with factor rotation is one of the most popular methods for exploratory factor analysis. In recent issues in JRCS for example eight papers refer to the use of factor analysis. Five of these papers apply as method for exploratory factor analysis PCA with factor rotation, while one gives no details about the EFA method, one uses SEM and the third applies CFA with SEM after an exploratory step using qualitative research.

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 58.3 for the ordinal, 57.5 for the binary and 65.4 for the metric scale. The two factors[2] which result after *varimax* rotation with Kaiser normalization of

[2] After rotation the components are in the following referred to as factors.

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.

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

In Table 3 the factor loadings of the NEP scale are given. The questions are sorted in ascending order with respect to the loadings of the first factor for the ordinal answer format. For the ordinal and metric scale the factor structure reflects pro-environmental versus not pro-environmental statements. 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. The NEP scale is a validated instrument where only a shortened version was used in this survey. The factor analysis does not reflect the factor structures that have been found for the NEP scale in other contexts[3] but factors are clearly associated with having an environmentally friendly or unfriendly attitude towards nature.

[3] Please note that the original scale was not developed using factor analysis. Many subsequent studies have factor analysed their data and come to different conclusions about the factor structure.

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 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. It should be noted, however, that the attitudinal items contained questions which indicated a pro-environmental and items which indicated a non pro-environmental attitude, whereas all items in the set of behaviors were inherently non pro-environmental.

3 Reliability and generalizability

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 congruence of the answers is 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 were either excluded or assigned either to “yes” or to “no” for both the ordinal and the metric scale. The overall agreement using this approach is found to be quite high amounting to 80 percent across all scale comparisons and aggregated over all three collapsing strategies. The overall agreement is higher for the behavioral intentions with 83 percent than the NEP scale with 75 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 if the answers on the middle point are assigned to “yes” or “no”. Omitting the middle categories increases the agreement between the ordinal scale and any of the two others. The assumption that the percentage of agreement is the same is rejected using a test for equal proportions for the omitting strategy while it is not rejected for the assignment to “yes” or “no”. 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 unless the middle category is omitted for the ordinal scale.

The reliability analysis only takes different occasions confounded with difference in answer format into account as possible source of variation. This lack of consideration of other sources of variation in classical measurement theory is criticized and generalizability theory has been

proposed to overcome these shortcomings and improve marketing measures (Finn and Kayande, 1997). In a generalizability study the variance components are estimated for all sources of variances and then generalizability coefficients can be determined for the given objects of measurement. The sources of variance in our study are respondents, constructs, items and answer formats confounded with occasions. The constructs are assumed to be a fixed factor and therefore separate analyses are conducted for them. Answer format can be assumed to be a random factor as well as a fixed factor. The variance components are determined using linear mixed-effects models fitted with restricted maximum likelihood estimation (REML; Pinheiro and Bates, 2000). The variance component estimates are given in Table 7 for the data where format is used as random factor and for a separate analysis for each answer format. Answer format together with the interaction accounts for 5.6% of the variance in the behavioural intentions data set and for 9.4% for the NEP data set.

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

With respect to our analyses and managerial interpretations the items are the objects of measurements. The generalizability coefficient (GC) for the aggregated data for relative decisions is given by:

$$GC = \frac{\sigma^2_{Items}}{\sigma^2_{Items} + \frac{\sigma^2_{Items:Respondents}}{n} + \frac{\sigma^2_{Items:Format}}{f} + \frac{\sigma^2_{Error}}{n \cdot f}},$$

where n is the number of respondents and f the number of formats. For the separate analysis the generalizability coefficient is given by:

$$GC = \frac{\sigma^2_{Items}}{\sigma^2_{Items} + \frac{\sigma^2_{Error}}{n}}.$$

The GC for the aggregated data are 0.99 for the behavioral intentions and 0.93 for the NEP scale. The separate analysis gives a GC of 0.99 for the 7-point and metric scale and 0.98 for the binary scale for the behavioral intentions and 0.96, 0.94 and 0.95 for the 7-point, binary and metric scale for the NEP data set. The generalizability coefficients are above the recommended level of 0.9 and this indicates that the items' scores can be generalized over respondents for each answer format and also over respondents and answer formats. These results also confirm the conclusions drawn with respect to managerial interpretations.

From a managerial perspective similar conclusions can be derived for marketing 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.

4a-d Burden on respondents

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). The eliminated six students who indicated to take less than zero (two respondents) or longer than 20 minutes (4 respondents) must have misprinted the hour of either the starting or the finishing time as it was impossible for students to take more than 20 minutes, because the questionnaires

were collected earlier than that. 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 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. The influence of repetition and answer format is evaluated using ANOVA and the results are given in Table 8.

---------- Table 8 ----------

As can be seen, repetition and answer format have a significant influence on the duration of filling in the questionnaire. 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.

For evaluating the perception of the scale in dependence of repetition and answer format a multivariate ANOVA (MANOVA) was conducted using perceived simplicity, pleasantness, quickness and ability to express feelings as dependent variable as MANOVA allows assessing group differences simultaneously for multiple dependent variables. Both independent variables have a significant influence on the overall perception of the scale as indicated by a Pillai-Bartlett test (Repetition: Pillai=0.15, p-value< 0.01; Format: Pillai=0.11, p-value=0.01). In order to assess which of the four items are differently evaluated given repetition and format, separate ANOVAs are made for each of the four dependent variables and the results are given in Table 8. Repetition has a significant influence for p=0.05 for simple, pleasant and quick. The p-value for ability to express the feelings is rather small and might be only insignificant due to lack of power for a sample size of 60. For the answer formats perceived quickness is significant: the binary answer format is perceived as significantly quicker than the ordinal, while the ordinal and metric are perceived as equally quick. Furthermore, the p-value for simple is rather small and might only be insignificant due to lack of power. However, the answer formats are equally perceived with respect to pleasantness and ability to express feelings.

The findings on the user-friendliness of questionnaires have major implications for marketing research practice: if indeed respondents perceive binary scales to be as pleasant and simple as ordinal scales 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 within respondent repeated measurement 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 mappings between answer formats, managerial interpretation, reliability and burden on the respondents. The within respondent repeated measurement design extended past work in the area which typically compared independent samples. This 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 of the mappings between the different answer formats while allowing for heterogeneity between the respondents reveals that the answers on the metric and ordinal 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 endangers 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 exploratory factor analysis were applied. Regardless of the answer format the main conclusions drawn are the same. Consequently it appears that marketing 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.

---------- Table 9 ----------

The findings from all analyses reported in this study are summarized in Table 9. In conclusion, it seems that with regard to behavioral intentions marketing 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 present study could further be extended by including ordinal scales with different numbers of answer options as well as labeled and non-labeled answer options. Another interesting direction for future work would be to investigate the effect of answer format familiarity with the evaluation of user-friendliness. Furthermore, research into the way in which different constructs typically studied in marketing are represented in consumers' minds would be of interest as the results from this study clearly indicate that answer format recommendations cannot be made independent of the construct under study.

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. Due to the dependence of fieldwork on class attendance, no balanced design for the rotation of the answer formats could be achieved which might confound the results of the analysis where the influence of repetition and answer format was assessed. A replication study with an improved sample should be conducted in future. Furthermore, other constructs that are typically measured in the marketing research context should be included to determine whether the findings for behavioral intentions and attitudes are generalizable. In addition the analyses focused on items being the objects of measurement. In

market research and retailing possible other objects of measurements are consumers, brands or stores. As the results might be different if the object of measurement is changed it is not possible to generalize the results to situations where the objects of measurement are different. Further analyses would be necessary to investigate this. Finally, another important area of future work is the study of whether respondents actually use metric response scales as metric in nature, meaning that they are fully aware of the fact that each unit on the scale (or turn on a knob) is of precisely equal distance.

ACKNOWLEDGEMENTS

To be added after review.

REFERENCES

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

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

Chang, L. (1994). 'A Psychometric Evaluation of Four-point and Six-point Likert-type Scales in Relation to Reliability and Validity.' Applied Psychological Measurement 18: 205–215.

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

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., 1978. The „new environmental paradigm“: A proposed measurement instrument and preliminary results. Journal of Enviornmental Education 9, 10-19.

Dunlap, R. E., Van Liere, K. D., 1984. Commitment to the dominant social paradigm and concern for environmental quality. Social Science Quarterly 65, 1013-1028.

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

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.

Finn, A., and U. Kayande (1997). 'Reliability Assessment and Optimization of Marketing Measurement.' Journal of Marketing Research 34(2), 262–276.

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.

Hancock, G. R. and A. J. Klockars (1991). 'The Effect of Scale Manipulations on Validity: Targeting Frequency Rating Scales for Anticipated Performance Levels.' Applied Ergonomics 22 (3), 147–154.

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.

Loken, B., P. Pirie, K.A. Virnig, R.L. Hinkle, and C.T. Salmon (1987). 'The Use of 0–10 Scales in Telephone Surveys.' Journal of the Market Research Society 29(3), 353–362.

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

Mundfrom, D.J., Shaw, D.G., Lu Ke, T., 2005. Minimum Sample Size Recommendations for Conducting Factor Analyses. International Journal of Testing 5(2), 159-168.

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.

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.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-Plus. Springer.

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, 2005. 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.

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

Wedel, M., Kamakura, W.A., 2001. Market Segmentation – Conceptual and Methodological Foundations. Kluwer Academic Publishers, 2nd edition.

TABLES AND FIGURES

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

		Both constructs	Behavioral Intentions	New Ecological Paradigm
Test-Retest Reliability	Ordinal vs. Binary	0.66	0.71	0.57
	Ordinal vs. Metric	0.74	0.78	0.63
	Binary vs. Metric	0.63	0.71	0.48
Agreement omitting middle category	Ordinal vs. Binary	0.84	0.86	0.80
	Ordinal vs. Metric	0.84	0.86	0.79
	Binary vs. Metric	0.78	0.82	0.72
	χ^2	15.34	6.26	8.62
	p-value	<0.01	<0.01	<0.01
Agreement collapsing middle category to "no"	Ordinal vs. Binary	0.79	0.83	0.74
	Ordinal vs. Metric	0.80	0.83	0.74
	Binary vs. Metric	0.78	0.82	0.72
	χ^2	1.03	0.65	0.44
	p-value	0.60	0.72	0.80
Agreement collapsing middle category to "yes"	Ordinal vs. Binary	0.80	0.83	0.75
	Ordinal vs. Metric	0.79	0.83	0.74
	Binary vs. Metric	0.78	0.82	0.72
	χ^2	0.95	0.17	1.16
	p-value	0.62	0.92	0.56

Table 7

Estimates of Variance Components

		Behavioral Intentions		New Ecological Paradigm	
		Variance Comp.	Percent (%)	Variance Comp.	Percent (%)
Aggregated	Formats	0.000	0.0	0.000	0.0
	Items	0.089	47.2	0.028	18.9
	Respondents	0.021	11.0	0.004	2.9
	Formats:Items	0.002	2.2	0.003	2.2
	Formats:Respondents	0.009	7.2	0.011	7.2
	Items:Respondents	0.028	14.6	0.042	29.1
	Error	0.040	21.5	0.058	39.6
7-point scale	Items	0.084	53.7	0.025	25.9
	Respondents	0.040	12.7	0.007	7.2
	Error	0.053	33.6	0.066	66.9
Binary	Items	0.109	42.2	0.052	20.3
	Respondents	0.048	18.5	0.016	6.3
	Error	0.102	39.3	0.187	73.4
Metric	Items	0.079	52.6	0.017	18.8
	Respondents	0.022	14.5	0.021	23.3
	Error	0.066	32.9	0.053	57.9

Table 8

ANOVA of the linear models for the logarithmised duration and the perception of the scales

		Sum of squares	Degrees of Freedom	F-value	p-value
log(Duration)	Repetition	8.00	2	23.29	< 0.01
	Format	1.87	2	5.45	< 0.01
	Residuals	28.98	169		
Simple	Repetition	4.90	2	3.82	0.02
	Format	3.27	2	2.55	0.08
	Residuals	112.38	175		
Pleasant	Repetition	8.40	2	6.69	< 0.01
	Format	1.89	2	1.50	0.23
	Residuals	109.91	175		
Quick	Repetition	10.43	2	7.02	< 0.01
	Format	5.34	2	3.59	0.03
	Residuals	130.03	175		
Feelings	Repetition	3.34	2	2.82	0.06
	Format	1.22	2	1.03	0.36
	Residuals	103.76	175		

Table 9

Summary of findings

Criterion	**Result** Answer format
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.
Differences in average values	Results of all three answer formats do not differ significantly if only mean values are of interest.
Construct equivalence	Factor analytic results indicate the same underlying structure across all answer formats.
Reliability / agreement	Scales render equally high levels of agreement.
Time required for completion	Binary format is quicker to complete.
Perceived speed	Binary format is perceived as quicker to complete.
Perceived simplicity	No difference between scales.
Perceived pleasantness	No difference between scales.
Perceived ability to express feelings	No difference between scales.

Figure 1

Mappings between two answer formats