2011

Quick, simple and reliable: forced binary survey questions

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

Bettina Grün
WU Wirtschaftsuniversität Wien, Austria

Friedrich Leisch
Ludwig-Maximilians-Universität München, Germany

Recommended Citation

Dolnicar, Sara; Grün, Bettina; and Leisch, Friedrich: Quick, simple and reliable: forced binary survey questions, International Journal of Market Research: 53(2) 2011, 231-252.
https://ro.uow.edu.au/commpapers/782

Quick, simple and reliable: forced binary survey questions

Abstract

Consumers are increasingly saturated by market research which leads to decreasing response rates and an increased danger of response bias. Market researchers thus face the challenge of recruiting respondents, increasing response rates and reducing respondent fatigue by making questionnaires as short and pleasant as possible. One way of achieving this is to replace traditionally used ordinal multi-category answer formats (such as Likert scales) with forced binary scales. This proposition is only attractive if it indeed shortens the survey time while not compromising the quality of managerial insights from the data.

This study investigates these conditions. Results from a repeat-measurement design indicate that managerial interpretations do not differ substantially between the two answer formats, responses are equally reliable, and that the binary format is quicker and perceived as less complex.

Keywords

answer format, multi-category, ordinal, binary, Likert scale

Disciplines

Business | Social and Behavioral Sciences

Publication Details

This article was originally published as, Dolnicar, S, Grün, B, and Leisch, F, Quick, simple and reliable : forced binary survey questions, International Journal of Market Research, 53(2), 2011, 231-252. Original item available here.

Quick, Simple and Reliable: Forced Binary Survey Questions

Sara Dolnicar*

Professor
Institute for Innovation in Business and Social Research
University of Wollongong
Northfields Ave, 2521 Wollongong, Australia
Telephone: (61 2) 4221 3862
Email: sara_dolnicar@uow.edu.au

Bettina Grün*

Research Fellow
Institute for Statistics and Mathematics,
WU Wirtschaftsuniversität Wien
Augasse 2-6, A-1090 Vienna, Austria
Telephone: (43 1) 31336 5032,
Email: Bettina.Gruen@wu.ac.at

Friedrich Leisch*

Professor
Department of Statistics
Ludwig-Maximilians-Universität München
Ludwigstrasse 33, 80539 München, Germany
Telephone (49 89) 2180 3165,
Email: Friedrich.Leisch@stat.uni-muenchen.de

* Authors listed in alphabetical order.

Quick, Simple and Reliable: Forced Binary Survey Questions

ABSTRACT

Consumers are increasingly saturated by market research which leads to decreasing response rates and an increased danger of response bias. Market researchers thus face the challenge of recruiting respondents, increasing response rates and reducing respondent fatigue by making questionnaires as short and pleasant as possible. One way of achieving this is to replace traditionally used ordinal multi-category answer formats (such as Likert scales) with forced binary scales. This proposition is only attractive if it indeed shortens the survey time while not compromising the quality of managerial insights from the data.

This study investigates these conditions. Results from a repeat-measurement design indicate that managerial interpretations do not differ substantially between the two answer formats, responses are equally reliable, and that the binary format is quicker and perceived as less complex.

Formatted: Indent: First line: 0 pt

Keywords: answer format, multi-category, ordinal, binary, Likert scale

Biographies

Sara Dolnicar, PhD, is a Professor of Marketing at the University of Wollongong (Australia) and the Director of the Institute for Innovation in Business and Social Research. Sara's research focuses on measurement and methodology in marketing research.

Bettina Grün, PhD, is a research fellow in the Institute for Statistics and Mathematics at the Wirtschaftsuniversität Wien (Austria) and an Associate Member of the Institute for Innovation in Business and Social Research. Her research interests include finite mixture modelling, statistical computing, and quantitative methods in marketing research.

Friedrich Leisch, PhD, is a Professor in the Department of Statistics at the University of Munich and an Associate Member of the Institute for Innovation in Business and Social Research at the University of Wollongong. His research interests include clustering, finite mixture modelling, statistical computing, and quantitative methods in marketing research.

1. INTRODUCTION

The ordinal multi-category answer format dominates commercial and academic marketing research (Lietz, 2010; Van der Eijk, 2001). Yet, a number of good reasons exist for selectively substituting multi-category scales with alternative answer formats: first, negative effects on data quality are known to occur when surveys are too long. When questionnaires become very time-consuming and tedious, respondents may not answer properly at later stages of the questionnaire or may stop completing the questionnaire half way through, at the expense of both data quality and field work efforts (Johnson et al., 1990; Drolet & Morrison, 2001; for a comprehensive list of negative effects see Vriens et al., 2001). While this may not be problematic in the case of short opinion polls, many market research tasks require a large set of questions to be included. One such example is brand image measurement (Driesener & Romaniuk, 2006; Dolnicar & Rossiter, 2008), where managers need to know how consumers view their brand and competitors' brands along numerous attributes. This increases the number of questions by the number of attributes included for each new brand to be evaluated. Ways of making this kind of questionnaires easier, faster or more pleasant for respondents are therefore urgently needed.

Second, people are increasingly reluctant to volunteer their time to participate in market research. A decrease in response rates has been noted by numerous authors (e.g., Hardie & Kosomitis, 2005; Bednell & Shaw, 2003). Hardie and Kosomitis have furthermore investigated the main reasons for respondents to refuse participation in market research. The length of the interview emerged as the second most important consideration, with short questionnaires increasing participation likelihood. Only the survey topic was considered more influential. Furthermore, questionnaire length is also crucial in attracting respondents with lower probabilities of participating in market research. Among those who refused, the first reason given was that they were too busy; the second reason was that the survey was too long. Hardie and Kosomitis conclude that measures need to be taken to stop the trend of decreasing response rates, both through questionnaire design and active advertising of market research: "*interviews need to be sold*" (2005, p.1). These findings support the need for simpler, faster and less burdensome questioning procedures.

A recent comparison of a binary and a Likert-scale version of a standardised health survey led to the conclusion that replacing the original multi-category answer options with binary options did not decrease validity or the component structure of the test instrument (Grassi et al., 2007), but significantly reduced the time required to complete the questions, thus making it more suitable for administration in the clinical setting.

We therefore hypothesise that binary answer formats are easier and quicker for respondents and they therefore represent an attractive alternative to ordinal multi-category formats. However, before a recommendation in favour of the binary format can be justified, the drawbacks of data collection with binary data have to be evaluated. It may well be that the results derived from binary answer formats will differ from those obtained from ordinal multi-category formats, be less reliable, or perceived as more difficult because respondents are more familiar with multi-category scales. These possibilities will be investigated in the present study. For binary answer format to represents an attractive alternative, the following conditions should be met: (1) typical managerial interpretations based on positioning analyses of the binary data should not differ from those derived from ordinal multi-category data; and (2) binary format should not be more burdensome to respondents.

If it were proven that the binary answer format saves respondents' time, it might then be preferable to the ordinal multi-category answer format. As discussed above, shorter questionnaires are likely to increase data quality due to both a decrease in respondent fatigue (Johnson et al., 1990) and a reduction in the non-response rate (Hardie & Kosomitis 2005). In addition, personnel costs for interviewers or survey administrators can be reduced if respondents are invited to participate in a survey personally. Based on cost information from an Australian permission-based online company the increase in cost for doubling the questionnaire length between 5 and 30 minutes lies at about 30%. A reduction of questionnaire length thus leads to significant savings in data collection costs.

We investigate the above questions in the brand image measurement context using a repeat measure design. In so doing we follow the recommendations by Hauser and Koppelmann (1979) not to use a sample that favours any of the two alternatives, to use analytic techniques that are representative of analyses used by marketing researchers, and to use criteria for comparison that are relevant to marketing academics and practitioners, namely (1) reliability; (2) differences in managerial interpretations; (3) answering speed; and (4) answering ease.

Findings from this study are relevant to academics and practitioners conducting market research: if binary questions are quicker, perceived as easier, equally reliable, and the managerial implications derived do not substantially differ, the binary answer format should be used more. Also, the binary format does not require questionable assumptions about the nature of the data.

2. PRIOR WORK

The optimal number of response options in questionnaires has been extensively studied, although little has been published recently. A number of distinct streams of research have developed using different criteria for the evaluation of the "optimality" of an answer format.

2.1 Reliability

A large number of researchers have chosen reliability or validity as criterion. Overall, the area is dominated by studies demonstrating that the number of answer categories is not related to reliability (Bendig, 1954; Jacoby & Matell, 1971; Komorita, 1963; Komorita & Graham, 1965; Matell & Jacoby, 1971; Peabody, 1962; Preston & Colman, 2000; Remington et al. , 1979), although there are studies which conclude the opposite (Finn, 1972; Nunnally, 1967; Oaster, 1989; Ramsay, 1973; Symonds, 1924).

Methodologically, these studies are either simulated or empirical by nature. Empirical studies typically use small sample sizes and each respondent is exposed to one format only. Where respondents were questioned twice, this was for the purpose of evaluating test-retest reliability, which was the most frequently used reliability measure besides Cronbach's alpha (Cronbach, 1951). Most studies compare respondents who used different answer formats, or re-categorize multi-category data into binary data and compare the two.

Consequently, prior work is limited in the following ways: (1) studies including a range of constructs and a range of answer formats have led to a wide variety of findings; (2) re-categorizing data to binary format (wrongly) assumes knowledge about the correct cut-off points; (3) sample size were small; and (4) Cronbach's alpha is heavily relied on as a quality indicator, despite being challenged as a valid measure (Rossiter, 2002).

2.2 Validity

Findings also contradict each other with respect to answer format validity. Matell and Jacoby (1971), Jacoby and Matell (1971) and Preston and Colman (2000) correlate the aggregate score for each respondent with an independently obtained overall rating of the construct and find no significant difference between answer formats. Using an external criterion, Chang (1994) shows that 4-point and 6-point formats are equally valid. He also warns that using an internal measure would not account for higher method variance of the 6-point format.

In contrast, Loken et al. (1987) and Hancock and Klockars (1991) conclude that more response options increase validity and discriminating power. The generalisability of these findings is limited by: (1) the limited variety of constructs measured; (2) the use of different validity measures; and (3) the fact that the external criteria are generally derived from the same respondents' evaluation rather than from independent sources.

Explicit recommendations to use a binary format were made by Peabody (1962), Komorita and Graham (1965), Matell and Jacoby (1971) and Jacoby and Matell (1971). A general warning about rating scales with multiple answer options came from Albaum et al. (2006) who raise concerns about increased central tendency error due to confounding the direction (e.g. agree or not agree, yes or no) and intensity (how much, to which extent) dimensions.

2.3 Interpretability

An alternative approach is to take an interpretational perspective and compare factor analysis results. While Martin et al. (1974) and Percy (1976) re-categorized empirical data to achieve this, Green and Rao (1970) used artificial data and were thus able to compare the results to the true specifications. Martin et al. (1974) compare factor analysis results from ordinal multi-category and binary data (both collapsed from responses collected on a metric format) concluding that results do not differ between two and nine answer categories. The same approach was taken by Percy (1976) who constructed binary data from a multi-category data set resulting from a 5-point Likert scale. He interprets the findings graphically and computes a quantitative compliance measure (Tucker coefficients) to compare the managerial insights, concluding that the two solutions are nearly equal. Green and Rao (1970) determined the number of optimal scale points based on a numerical simulation; they compared how well the inter-product relations were recovered. They suggest using at least a 6-point response format and at least eight attributes.

These studies are of direct value to managers because they focus on differences in final interpretation, which determines marketing decisions and action. However, findings are typically derived from artificially constructed data sets rather than repeat measurements. This assumes (wrongly) that the cut-off levels for re-categorizing answer formats are known, thus ignoring that respondents might have different, heterogeneous transformations from one answer format to another. This is the case because if people are not asked to respond using two different answer formats, and instead it is assumed that the researcher knows at which point in a multi-point scale people move from “yes” to ”no” on a binary scale, the data is not based on people’s true translations. Therefore any differences in such translations between people are not accounted for.

2.4 User-friendliness

To the best of our knowledge, only three studies investigated respondents’ subjective preferences: Jones (1968) finds preference for multiple categories and Preston and Colman (2000) conclude that people can better express their feelings using more response options. By contrast, perceived higher speed of questionnaire completion is associated with fewer answer options. In a repeat measurement study Dolnicar and Grün (2007) find that respondents perceived the binary, 7-point and metric format as equally pleasant, while seeing the binary format as quicker to complete.

The most comprehensive review was published by Cox (1980) who concluded that – while a democratic vote for the best number of response alternatives would be seven –

additional research is needed to replicate prior findings. Specifically, he believes that the issues of response error and bias have been insufficiently studied and that "*Surprisingly little is known about the process of psychological judgment*." (p. 419).

In the context of brand image measurement, Rungie et al. (2005) have discussed the issue of alternative answer formats in the context of brand image stability. This study is an extension of prior work by Dall'Olmo Riley et al. (1997), which concludes that brand images are instable, with an average repeat rate of about 50%. While Dall'Olmo Riley et al. (1997) argue that low repeat rates are due to brand image instability, Rungie et al. (2005) suggest the reason may lie in the lack of reliability of binary questions, noting, however, that ordinal questions may not perform better. Dolnicar and Rossiter (2008), based on an empirical investigation of brand image stability, recommended measures to improve reliability: surveying users of the product category only, using short questionnaires, including a "Don't know" option, and not instructing respondents to guess.

Finally, Driesener and Romaniuk (2006) compared the performance of a binary pick-any format, a rating format and a ranking format, and conclude that the three formats are "virtually interchangeable" (p. 681). They emphasize, however, that the binary format was substantially quicker, taking only about half the time of the other answer formats to complete. These experimental findings are in direct contradiction with the recommendations made in a recent review by Lietz (2010) that Likert scales with between 5 and 8 response options should be used and binary formats should be avoided.

3. METHODOLOGY

Respondents were exposed to two versions of the same questionnaire at two points in time: a binary version and an ordinal multi-category version one week later. They evaluated six fast food chains using eleven attributes, which represents a simple task for a brand image study. Brand names and attributes were derived from an exploratory study phase prior to the survey. The question wording (binary version) was as follows:

> *There are 11 attributes that you can agree or disagree with for each brand. Please enter into each one of the cells "1" if the attribute applies to the brand and "0" if the attribute does not apply to the brand. E.g. if you put a 1 into the first cell, you are expressing that "McDonalds is yummy!"*

In the multi-category version respondents chose from the following options: "perfectly applies"(1), "applies well"(2), "applies a little bit"(3), "does rather not apply"(4), "does not describe well"(5), "does absolutely not describe"(6). Respondents were also asked whether they perceived the questionnaire as "easy to answer", "ok", or "difficult to answer" and whether they felt it was "long" or "short". Furthermore, respondents were asked to note the starting and finishing time, which gave the actual duration.

Respondents were Australian university students enrolled in a compulsory third year subject. We do not expect this to negatively affect the validity of results, because the research question can be investigated legitimately for a subset of the population, especially if the product category is of relevance to the subgroup. Only respondents who completed both questionnaires were included in the final sample (n=148). The limitation of the data is that the order of presentation of the two versions was not rotated. However, ex post cross-tabulations indicate that no "binarization" of responses occurred.

A follow-up study was conducted using the same items but presenting identical answer formats twice in a row. This was necessary to assess the reliability of the two answer formats. This survey resulted in 46 usable samples for the binary and 35 samples for the multi-category format.

All computations and graphics have been produced using the R statistical software package (R Development Core Team, 2009).

4. RESULTS

4.1 Congruence, reliability and stability

If answers on a 6-point multi-category scale are compared to binary answers, the naïve expectation is that the respondents who ticked one of 1-3 on the multi-category scale indicated “yes” on the binary scale and those who ticked one of 4-6 indicated “no”. This would be in line with the semantic description of the multi-category scale. Re-categorizing multi-categorial responses into binary responses using the midpoint as splitting point is standard practise (e.g. in Jacoby & Matell, 1971; Percy, 1976). We test whether the assumption that respondents translate from ordinal to binary responses using the midpoint as cut-off point is legitimate and whether the prediction of binary answers via cut-off points is equally good over all fast food chains and attributes.

The multi-category information was used to identify empirical cut-off points for each respondent. This procedure assumes that there is a unique cut-off point for each respondent, which is consistent across attributes and brands, but accounts for differences in respondents’ response styles which are know to be constant (Cronbach, 1950). The cut-off level for each respondent then serves as an assignment rule prescribing which multi-category values become a 0 and which become a 1 on the binary scale. This is determined by minimising the prediction error. Table 1 shows the distribution of the resulting individual cut-off values.

TABLE 1 Optimal cutoff-values

Number of respondents	*Never*	*1*	*1-2*	*1-3*	*1-4*	*1-5*	*Always*
Frequency	1	2	31	88	19	6	1
Percent	1	1	21	59	13	4	1

Using a 6-point scale means that seven cut-off points are possible. The cells, labelled "Never" and "Always" in Table 1 represent respondents where 1 is never or always predicted, respectively. The cut-off point with label "1-2" signifies that 1 is predicted for the multi-category answers 1 and 2. The middle of the scale (cut-off value "1-3") renders optimal predictions for 88 respondents (59%) thus representing the most frequent optimal cut-off level. However, for 41% of respondents the relationship between the multi-category and the binary answers is not represented correctly by using the midpoint as the splitting criterion. For 31 respondents the optimal cut-off point is "1-2", meaning that "applies a little bit" on a multi-category scales translates into a disagreement on the binary scale.

After identifying the individual cut-off levels for each respondent, individual-level scale transformation predictions are made and analysed. The better the binary data can be predicted from the multi-category answers, the higher the extent of congruence between the two answer formats. The congruence of the answer formats was operationalised by computing a precision value, which is defined as follows:

$$Precision\ of\ "yes" = \frac{correctly\ predicted\ "yes"}{total\ number\ of\ "yes"} \qquad (1)$$

and the precision of "no" is analogously defined.

The overall precision levels are 0.76 for "no" and 0.89 for "yes" responses. Table 2 provides precision values for all fast food chains and attributes.

TABLE 2 Test-Retest Reliability for Fast Food Brands and Attributes

	Fast food chain					
	Burger King	*KFC*	*McDonalds*	*Pizza Hut*	*Red Rooster*	*Subway*
No	0.78	0.77	0.85	0.73	0.72	0.72
Yes	0.91	0.89	0.93	0.87	0.86	0.85
	Fast food chain attributes					
	cheap	*convenient*	*disgusting*	*expensive*	*fast*	*fattening*
No	0.62	0.46	0.79	0.76	0.55	0.59
Yes	0.93	0.92	0.74	0.74	0.93	0.94
	greasy	*healthy*	*spicy*	*tasty*	*yummy*	
No	0.68	0.93	0.87	0.68	0.71	
Yes	0.92	0.80	0.58	0.89	0.89	

It can be seen that, generally, agreement with brand-attribute combinations demonstrates a slightly higher precision level than disagreement. For instance, in case of the characteristic “convenient”, the difference between “yes” predictions and “no” predictions is 0.46. The only exceptions are the attributes “spicy” and “healthy” where “no” responses can be predicted better and the attributes “disgusting” and “expensive” where the precision levels are nearly equal. One possible explanation – which would be interesting to investigate in a follow up study – is that attributes for which the respondent has a very well formed opinion trigger more consistent use of the full multi-category scale range, consequently enabling better prediction of binary values. This would mean that multi-category answers only contain additional valuable information for well pre-selected object attributes, shifting the market researcher’s responsibility to extensive exploratory work in attribute selection, as already recommended decades ago by Joyce (1963), Myers and Alpert (1968), Alpert (1971) and Wilkie and Weinreich (1972). The attribute “healthy”, for instance, seems to be associated with Subway only and leads to high precision values in prediction. The difference in precision for “cheap” and “expensive” could be caused by the different connotations of these attributes, where “expensive” is more price-related whilst “cheap” bears additional associations about quality .

In general, however, differences in precision levels indicate that respondents do not differentiate between multi-category levels equally well for all attributes. They can better distinguish between the multi-category options if they have a clearer opinion on the attribute-brand association. But as these opinions can be agreeing or disagreeing with brand-attribute combinations, it depends on the specific brand-attribute association as to which part of the scale is exploited. Hence, it is impossible to evaluate a priori which multi-category options add information to the binary scale and which do not.

Test-retest reliability can also be computed as the relative number of respondents where agreement is predicted the second time given that they have indicated “yes” the first time. This measure is used to assess in/stability of brand images and is referred to as Repeat Rate (Dall’Olmo Riley et al., 1997; Rungie et al., 2005). Note that neither the term *reliability* nor the term *stability* describe precisely the reasons for low repeat rates, because they capture two effects: (1) instability of images, which could be due to either certain attributes not being associated with certain brands or respondents’ exposure to advertisements between survey waves; and (2) unreliability of the answer format, which is a measurement issue.

We computed the repeat rate using the individual cut-off levels for the multi-category data. Resulting repeat rate levels are 88.8% for respondents who ticked “yes” the first time and, given the individual cut-off levels, have also indicated agreement the second time. If a general cut-off point at the middle of the multi-category scale is used the repeat rate value amounts to 89.2% of the “yes” answers at the first measurement. Individual cut-off points which minimise the prediction error gave slightly worse results for the repeat rate than the general cut-off point. This rather surprising result is due to the fact that respondents use the positive part of the scale more frequently for the ordinal multi-category than for the binary questionnaire.

The repeat rate values for both answer formats are very high. Values typically reported average at 49% (Rungie et al., 2005). This result could be interpreted as high compliance between the two answer formats (which supports our hypothesis that binary formats could safely substitute multi-category formats). It is fair to note, however, that this is due in part to the study design: respondents were asked about a product category relevant to them, only well known brands were included and all attributes were selected to represent relevant descriptors. Other studies on brand image stability typically include multiple product categories, brands and attributes and report pooled results. It is thus likely that not all product categories, brands and attributes were known and/or relevant to the respondents. Empirical evidence for the effect of the above measurement factors on repeat rates has been reported by Dolnicar and Rossiter (2008) and Dolnicar and Heindler (2004). It is thus plausible to assume that the fast food study reached significantly higher levels of repeat rates than reported by Dall'Olmo Riley et al. (1997) and Rungie et al. (2005) due to its design. Furthermore, previous studies on brand image stability often used free choice format while our binary format forces respondents to choose one option.

Finally, we conducted a test-retest comparison of reliabilities using data collected in a follow-up survey in which respondents used the same answer format twice. This resulted in an repeat rate of 88.9% for the binary data set as well as for the ordinal data set where we used a general cut-off point at the middle of the scale. A pair wise comparison of the repeat rates using a proportion test indicates that there is no significant difference between the repeat rate of the binary-binary and the binary-ordinal survey (χ^2=0.022, df=1, p-value=0.88) as well as the binary-ordinal and the ordinal-ordinal survey (χ^2=0.018, df=1, p-value=0.89).

4.2 Differences in mean values

Table 3 contains the mean values for both answer formats. Values have been transformed to the [-1, 1] interval to allow direct comparison.

TABLE 3 Mean Values of the Answers for each Answer Format

	McDonalds		*KFC*		*Pizza Hut*		*Burger King*		*Subway*		*Red Rooster*	
	B	*M*	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*	*B*	*M*
yummy	0.35	0.23	0.49	0.29	0.64	0.50	0.20	0.27	0.62	0.58	0.14	0.06
fattening	0.99	1.00	0.97	0.99	0.90	0.81	0.87	0.94	-0.56	-0.15	0.74	0.67
greasy	0.87	0.76	0.97	0.94	0.78	0.65	0.85	0.71	-0.66	-0.39	0.68	0.58
fast	0.99	0.99	0.69	0.74	-0.19	0.15	0.83	0.75	0.51	0.52	0.48	0.53
cheap	0.72	0.67	0.33	0.36	0.08	0.22	0.53	0.49	0.21	0.25	0.39	0.38
tasty	0.20	0.12	0.39	0.26	0.59	0.46	0.24	0.24	0.63	0.58	0.13	0.18
expensive	-0.81	-0.42	-0.38	-0.13	-0.29	0.02	-0.64	-0.21	-0.37	-0.05	-0.44	-0.16
healthy	-1.00	-0.99	-0.99	-0.97	-0.91	-0.74	-0.93	-0.92	0.46	0.42	-0.79	-0.65
disgusting	-0.33	-0.18	-0.51	-0.20	-0.70	-0.39	-0.47	-0.29	-0.76	-0.58	-0.38	-0.14
convenient	0.94	0.97	0.71	0.72	0.43	0.53	0.72	0.75	0.68	0.61	0.47	0.56
spicy	-0.96	-0.87	-0.07	-0.20	-0.54	-0.54	-0.94	-0.79	-0.69	-0.49	-0.66	-0.50

Only few differences can be found. The binary means tend to deviate more strongly from zero, but the pattern of applicable and non-applicable attributes is the same. In particular the perception of Subway, which is quite distinct from other fast food chains, is captured well by both answer formats. The average absolute difference between the mean values of the binary and the multi-category data amounts to 6.2% of the scale range.

4.3 Differences in managerial interpretations: Positioning analysis

Brand image data is usually collected to conduct positioning analysis, the results of which are used for strategic or operational marketing decision making. Therefore, it is important to undertake a comparison from a managerial perspective. For this purpose, we constructed perceptual maps using principal component analysis (PCA) separately for the two answer formats (after unfolding them to two-mode by row wise stacking the chains times attributes matrices). PCA is a data reduction technique providing a projection of data to lower dimensional space by identifying sub-spaces with high variability. Polychoric correlations are used and the principal components are determined using an eigenvalue decomposition. Polychoric correlations are, in general, used for ordinal variables under the assumption that the ordinal variables dissect continuous latent variables that follow a Gaussian distribution. Results are compared from an interpretational point of view and using Tucker coefficients of congruence (Harman 1964), a measure indicating the similarity of factor analysis results.

For both answer formats the Kaiser criterion (Kaiser, 1960) suggested to retain three factors with variance explained amounting to 74% (binary) and 63% (multi-category), respectively. Varimax rotation was used to improve interpretability of factors. Table 4 contains the loadings after *varimax* rotation. From a managerial perspective, the three factors are nearly equivalent for the two answer formats and could be named “health”, “convenience” and “taste”. The biggest difference is for the attribute “spicy” in the first factor (0.17). The average absolute difference is 0.05.

For both answer formats each attribute except for “spicy” loads on exactly one factor for the 3-factor solutions. This indicates that the three factors combine the contribution of nearly all attributes and are complementary, reflecting different latent traits. After rotation, each factor consists of at least three attributes, which all contribute to a similar extent but which is not outstandingly high for any attribute.

TABLE 4 Loadings of the 3-Factor Solutions after Varimax Rotation for both Answer Formats

	Factor 1		*Factor 2*		*Factor 3*	
	B	*M*	*B*	*M*	*B*	*M*
yummy	0.02	0.01	0.02	-0.02	-0.59	-0.60
fattening	0.60	0.62	-0.01	0.00	-0.07	-0.08
greasy	0.56	0.58	0.03	-0.02	0.04	0.04
fast	-0.04	0.07	0.48	0.50	0.02	-0.08
cheap	0.03	-0.06	0.55	0.58	0.08	0.01
tasty	0.01	-0.02	0.02	-0.02	-0.58	-0.59
expensive	0.04	0.14	-0.53	-0.51	-0.03	-0.13
healthy	-0.55	-0.48	0.00	-0.05	-0.07	-0.10
disgusting	0.05	0.06	0.02	0.02	0.51	0.45
convenient	0.05	0.16	0.43	0.35	-0.18	-0.23
spicy	0.13	-0.03	-0.11	-0.15	-0.12	-0.08

While the interpretative comparison is by very nature subjective, the Tucker coefficient of congruence (CC) offers an objective measure

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

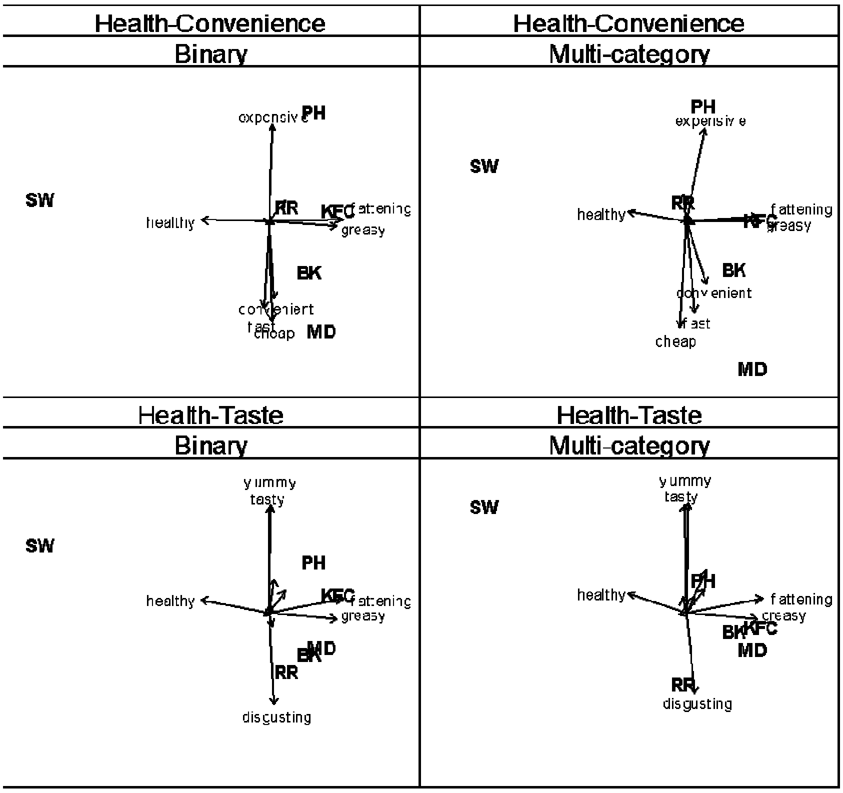
where b_{jp} is the jp^th^ element in the binary loadings matrix, m_{jq} the jq^th^ element of the multi-category loadings matrix and n the number of attributes. The Tucker coefficients lie between -1 and 1 and measure the similarity between two factors on a factor-to-factor basis.

The Tucker coefficients for the PCA solution after *varimax* rotation are 0.963 for the first, 0.992 for the second, and 0.984 for the third factor. These values are close to one, indicating substantial congruence between the two factor matrices.

Figure 1 shows the perceptual maps for the PCA solution after *varimax* rotation, with the first rotated factor being plotted against the second and the first against the third. For simplicity of presentation, the names of attributes with small absolute loadings are omitted.

Brand positions are determined by aggregating perceptions over all respondents. From an interpretational point of view, both perceptual maps lead to similar conclusions: Subway (SW) claims the most distinct position in the perceptual maps. Respondents perceive Subway as healthy and yummy/tasty, but recognise that Subway is not a cheap snack solution. McDonalds (MD) and Burger King (BK) are located close to each other, indicating a competitive relation, although all attributes describing the two chains are perceived to apply more to McDonalds. Both are perceived as cheap, convenient and fast. KFC is seen to be greasy and neither cheap, nor expensive. Red Rooster (RR) is perceived as disgusting. Typical attributes of fast food chains (e.g. cheap, fast or convenient) are not associated with it. Pizza Hut (PH) is perceived as expensive, but yummy and tasty.

FIGURE 1 Perceptual Maps



Based on these maps it is unlikely that completely different strategic marketing action plans would emerge. For instance, from the point of view of Subway, the competitive position should be maintained and the only interesting positioning modification might be to improve the perception regarding the price. From the point of view of Pizza Hut, a major repositioning might be required, shifting away from being seen as expensive without providing the typical fast food benefits (fast service, convenience).

The relationship between the individual factor scores is analysed by regressing the binary scores on the ordinal scores of each respondent-brand combination for each factor. The coefficient of the binary scores are highly significant as indicated by a *t*-test (t_1=27.4, t_2=21.5, t_3=24.9, all p-values<0.001). The R^2 values are equal to 0.49, 0.37 and 0.44, respectively. The congruence between the results of the individual analysis is not as high as for the aggregate analysis. However, as brand image measurement is known to be rather unstable for a given individual, the correspondence between the factor scores for these individuals is surprisingly high.

It can be concluded that positioning results are nearly identical for both answer formats.

4.4 Duration of the questionnaire

Questionnaire length is relevant from three perspectives: shorter surveys lead to less fatigue and consequently better data quality, they reduce fieldwork cost, and are likely to increase response rates, thus reducing response bias. In sum: the shorter the better, management insights being equal.

In our study, respondents needed four minutes (standard deviation 1.7) on average to complete the binary and 50% (6 minutes, standard deviation 1.9) longer for the ordinal multi-category questionnaire. A paired sample *t*-test (t=-11.81, p-value<0.001) indicates that this difference is highly significant, leading to the conclusion that it took respondents significantly longer to complete the questionnaire using multi-category answer options.

Nevertheless, it is possible that respondents felt more comfortable using the traditional multi-category scale and may not have perceived the multi-category questionnaire version as longer. To investigate this possibility, respondents were asked about their perceptions. A cross-tabulation of the paired observations shows that 91 respondents (63%) thought that both questionnaires were short, 21 (14%) thought both were long, 9 (6%) thought the binary was long and the multi-category short and 24 (17%) said the binary was short and the multi-category long. The difference in perception is tested using McNemar's test of symmetry. The association is significant (χ^2=5.94, df=1, p-value=0.015), leading to the conclusions that the binary questionnaire is perceived as significantly shorter.

4.5 Perception of the difficulty of the questionnaire

Respondents were asked to state how difficult they perceived the questionnaires to be. The more difficult the questionnaire, the more likely that some respondents will not be able to carefully complete the task as intended. In our study the questionnaires were identical except for the answer format. Any difference in perceived difficulty can thus be attributed to the answer format.

A significant differences in perceived difficulty is found (Fisher's exact test for count data, p-value<0.001). Table 5 provides the frequency distribution. As can be seen, more respondents perceived the binary questionnaire was easy (around 70%) than the multi-category version (around 40%). This is even more remarkable because respondents were confronted with the multi-category version in the second week, and thus had already experienced it.

This result is in contrast with findings reported in Preston and Colman (2000), where a 6-point format was perceived by respondents as significantly easier than a binary format. This contradiction may be due to the difference in questionnaires; our questionnaire used the same answer format for all questions within one version of the questionnaire, whereas Preston and Colman (2000) used different answer formats for different questions in their survey.

TABLE 5 Difficulty of Answering the Questionnaire

	Frequency (multi-category)	*Frequency (binary)*	*Percent (multi-category)*	*Percent (binary)*
easy	66	101	45	70
ok	78	43	53	30
difficult	3	1	2	1
Total	147	146	100	100

Results show that respondents perceived the binary questionnaire as significantly easier than the multi-category version.

5. CONCLUSIONS

This study investigated whether the binary answer format represents an attractive alternative to the ordinal multi-category format for market researchers. This would be the case if:

(1) the binary format led to equally reliable results;

(2) managerial interpretations based on typical managerial analyses would not differ;

(3) the binary format saved respondent time; and

(4) the binary format was perceived as simpler.

The study findings support the above points empirically for a repeat measure student sample containing evaluations of fast food brands through attribute association statements: binary questions were answered more quickly, perceived as less difficult, were equally reliable and led to the same managerial implications using positioning analysis.

One characteristic of ordinal multi-category formats that can never be substituted by a binary alternative is the way they are able to capture intermediate shades of respondent opinion. Typically, however, those intermediate shades are not exploited in data analysis. Consequently, if intermediate shades are not of interest (in the form of, for instance, a frequency count of how many respondents are "moderately satisfied"), the present study leads to the conclusion that the binary format outperforms ordinal multi-category formats with respect to survey efficiency, without generating different results from a typical positioning analysis point of view. The managerial recommendation resulting from this research is to selectively substitute the ordinal multi-category answer format with binary questions.

There is at least one instance, however, where such a substitution cannot not be recommended: if inter-individual change of brand image (e.g. due to advertising) needs to be evaluated. For an aggregate evaluation binary format is sufficient, because it would be expected that the proportion of agreements relating to the advertised attributes would increase. But if it is essential to understand changes occurring within one individual, then multi-category formats are logically required.

We include in Figure 2 a proposed classification which summarizes under which circumstances binary answer formats are preferable, under which circumstances they should be considered as an alternative, and under which circumstances they cannot be used. The two key dimensions we propose should be used for this assessment are: (1) importance of keeping the survey short and, consequently, fatigue effects to a minimum; and (2) whether multiple-answer options are logically (not methodologically or psychometrically) needed. For example, as illustrated above, binary measures are logically not suitable for tracking changes in attitudes at the individual level over time. Binary measures, however, are preferable in cases where duration of the survey is critical and there is no logical need for multi-category options. The value of using the binary answer format should also be carefully assessed in all other cases where there is no logical requirement for offering respondents multiple response categories.

FIGURE 2 Summary recommendation chart

		Reduction of duration / fatigue	
		Not important	Important
Multiple options logically required	No	The value of using the binary answer format instead of a conventional multi-category format should be assessed	Binary answer format preferable
	Yes	Multi-category answer format preferable	Multi-category answer format necessary

The present study has some limitations. Although we did not detect a "binarization of multi-category responses" as a consequence of presenting the binary version to respondents first, it would have been preferable to split the sample and present each version to half of the sample first. Also, our research design did not permit testing for comparative predictive validity. This would require behavioural data related to fast food restaurant usage. Finally, our study is limited to a specific product category and a sub-section of the population which heavily uses this product category. Optimally, however, a number of product categories and the respective users of those product categories should be included to empirically ensure generalisability of results. Currently generalisability is argued theoretically, because there is no logical reason why the selected sub-segment would differ systematically from other sub-sections of the population or the population as a whole with respect to answer format use. We do not expect that the fact that the students who participated in the sample were marketing students affects the results negatively, because the study investigated effects of answer formats rather than the absolute or relative brand image positions of brands.

Future work could support selective substitution of the ordinal multi-category answer format with the binary answer format by investigating systematically whether factors such as familiarity with the object of study (brand, product), or the usage frequency, have significant impact on the suitability of alternative answer formats.

ACKNOWLEDGEMENT

To be added after blind review.

REFERENCES

Albaum, G., Roster, C., Yu, J. H. & Rodgers, R. D. (2006). Simple rating scale formats – exploring extreme responses. *International Journal of Market Research*, *49*(5), 633-650.

Alpert, M. I. (1971). Identification of determinant attributes: a comparison of methods. *Journal of Marketing Research*, *8*(2), 184-191.

Bednell, D. H. B. & Shaw, M. (2003). Changing response rates in Australian market research. *Australasian Journal of Market Research, 11*(1), 31-41.

Bendig, A. W. (1954). Reliability and the number of rating scale categories. *Journal of Applied Psychology*, *38*(1), 38-40.

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*(3), 205-215.

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*(1), 3-31.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *6*, 297-334.

Dall'Olmo Riley, F., Ehrenberg, A. S. C., Castleberry, S. B., Barwise, T. P. & Barnard, N. R. (1997). The variability of attitudinal repeat-rates. *International Journal of Research in Marketing*, *14*(5), 437-450.

Dolnicar, S. & Grün, B. (2007). How constrained a response: a comparison of binary, ordinal and metric answer formats. *Journal of Retailing and Consumer Services*, *14*, 108-122.

Dolnicar, S. & Heindler, M. (2004). If you don't need to know, don't ask! Does questionnaire length dilute the stability of brand images?. *CD Proceedings of the 33rd EMAC Conference*. Murcia, Spain: European Marketing Academy.

Dolnicar, S. & Rossiter, J. R. (2008). The Low Stability of Brand-Attribute Associations is Partly Due to Measurement Factors. *International Journal of Research in Marketing*, *25*(2), 104-108.

Driesener, C. & Romaniuk, J. (2006). Comparing methods of brand image measurement. *International Journal of Market Research*, *48*(6), 681-698.

Drolet, A. L. & Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research*, *3*(3), 196-204.

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.

Grassi, M., Nucera, A., Zanolin, E., Omenaas, E., Anto, J. M. & Leynaert, B. (2007). Performance comparison of Likert and Binary formats of SF-36 version 1.6 across ECRHS II adult populations. *Value in Health*, *10*(6), 478-488.

Green, P. E. & Rao, V. R. (1970). Rating scales and information recovery - how many scales and response categories to use? *Journal of Marketing*, *34*(3), 33-39.

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

Hardie T. & Kosomitis, N. (2005). The F word in market research: high impact public relations in the 21st century. *CD Proceedings from the AMSRS Conference 2005: Impact*. Sydney: Australian Market & Social Research Society.

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

Hauser, J. R. & Koppelman, F. S. (1979). Alternative perceptual mapping techniques: relative accuracy and usefulness. *Journal of Marketing Research*, *16*(4), 495-506.

Jacoby, J. & Matell, M. S. (1971). Three-point Likert scales are good enough. *Journal of Marketing Research*, *8*(4), 495-500.

Johnson, M. D., Lehmann, D. R. & Horne, D. R. (1990). The effects of fatigue on judgments of interproduct similarity. *International Journal of Research in Marketing*, *7*(1), 35-43.

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*. San Francisco, CA.: American Psychological Association.

Joyce, T. (1963). *Techniques of brand image measurement. New developments in research*. London: Market Research Society.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151.

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.

Lietz, P. (2010). Research into questionnaire design – a summary of the literature. *International Journal of Market Research*, *52*(2), 249-272.

Loken, B., Pirie, P., Virnig, K. A., Hinkle, R. L. & Salmon, C.T. (1987). The use of 0-10 scales in telephone surveys. *Journal of the Market Research Society*, *29*(3), 353-362.

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

Myers, J. H. & Alpert, M. I. (1968). Determinant buying attitudes: meaning and measurement. *Journal of Marketing*, *32*(4), 13-20.

Nunnally, J. C. (1967). *Psychometric Theory.* New York: McGraw-Hill.

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, *68*, 549-550.

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

Percy, L. (1976). An argument in support of ordinary factor analysis of dichotomous variables. In B. B. Anderson (Ed.), *Advances in Consumer Research* (Volume III). Ann Arbor, MI: Association for Consumer Research.

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), 1-15.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, *37*, 513-532.

R Development Core Team (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*(4), 305-335.

Rungie, C., Laurent, G., Dall'Olmo Riley, F., Morrison, D. G. & Roy, T. (2005). Measuring and modeling the (limited) reliability of free choice attitude questions. *International Journal of Research in Marketing*, *22*(3), 309-318.

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*(3), 325-341.

Vriens M., Wedel M. & Sandor Z. (2001). Split-questionnaire designs: a new tool in survey design and panel management. *Marketing Research*, *13*(1), 14-19.

Wilkie W. L. & Weinreich R. P. (1972). Effects of the number and type of attributes included in an attitude model: more is not better. *Proceedings of the Third Annual Conference of the ACR*. Association for Consumer Research. 325-240.