

*Faculty of Commerce*

*Faculty of Commerce - Papers*

---

*University of Wollongong*

*Year 2007*

---

Question stability in brand image  
measurement - Comparing alternative  
answer formats and accounting for  
heterogeneity in descriptive models.

S. Dolnicar\*

B. Grun†

\*University of Wollongong, sarad@uow.edu.au

†Vienna University of Technology, Austria

This article originally published as Dolnicar, S and Grun, B, Question stability in brand image measurement: Comparing alternative answer formats and accounting for heterogeneity in descriptive models, *Australasian Marketing Journal*, 15(2), 2007, 26-41.

This paper is posted at Research Online.

<http://ro.uow.edu.au/commpapers/413>

## **Question stability in brand image measurement**

Comparing alternative answer formats and accounting for heterogeneity in  
descriptive models

**Sara Dolnicar\***

Marketing Research Innovation Centre (MRIC)

School of Management and Marketing

University of Wollongong, Wollongong, NSW 2522, Australia

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

[sara\\_dolnicar@uow.edu.au](mailto:sara_dolnicar@uow.edu.au)

**Bettina Grün\***

Marketing Research Innovation Centre (MRIC)

Department of Statistics and Probability Theory

Vienna University of Technology

Wiedner Hauptstrasse 8–10, A-1040 Vienna, Austria

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

[bettina.gruen@ci.tuwien.ac.at](mailto:bettina.gruen@ci.tuwien.ac.at)

\*Authors listed in alphabetical order.

### **Acknowledgements**

This research was supported by the Australian Research Council (through grants DP0557257 and LX0559628) and the Austrian Science Foundation (through grant P17382).

### **Autobiographical notes**

Sara Dolnicar holds a position as Professor of Marketing in the School of Management & Marketing at the University of Wollongong and is currently serving as the Associate Dean (research) of the Faculty of Commerce. She finished her degrees in Psychology and Business Administration at the University of Vienna and the Vienna University of Business Administration, respectively. She was awarded her PhD in Business Administration 1996. Her primary research interests are in market segmentation, marketing research methodology and tourism.

Bettina Grün is a Research Fellow at the Department of Statistics and Probability Theory of the Vienna University of Technology. She finished her degree in Applied Mathematics at the Vienna University of Technology and received a PhD in Applied Mathematics in 2006. Her main research interests are finite mixture modelling, statistical computing and quantitative methods in marketing research and tourism.

## **Executive summary**

Organisations are interested in how consumers perceive their brand. Consequently, many organisations regularly conduct brand image surveys. In such surveys respondents are asked to state with which brands they associate a list of attributes.

A number of researchers have recently warned managers that brand images are not stable. If brand images are not stable that means that a respondent who, for instance, states that McDonalds is expensive when surveyed for the first time does not express this same belief when asked a second time, even if no advertising or other intervention occurred during the two measurements that may explain a change of beliefs. Unstable brand images indicate that either consumers do not have a clear idea of a brand or do not associate it strongly with certain attributes. Both would be highly concerning results for brand managers in any organisation which would have to lead to seriously questioning marketing action aimed at brand development.

Recent research into brand image stability has suggested that the answer format used in the brand image surveys may be a reason for low brand image stability. Furthermore, recent work assumes that all brands-attribute associations are equally stable. This implies, for instance, that it is not possible that only a small group of consumers holds a very strong belief that McDonalds is expensive. This is not plausible given that consumer heterogeneity is widely acknowledged and target markets frequently form the basis of most brand managers marketing activities.

In this paper we investigate the extent to which answer formats used in brand image studies affect the stability of brand-attribute associations and we propose a model which accounts for consumer heterogeneity.

The results are of major importance for managers who rely on empirical brand image data. The study demonstrates (1) that brand images are more stable than previously reported and that brand image data therefore represents a valid basis for the development of marketing activities, (2) that all answer formats lead to equally stability levels, (3) that heterogeneity exists in brand-attribute associations, thus making it possible for managers to design customized strategies for different attributes, and (4) that the reliability of brand image data depends strongly on how the brand image survey is designed. Most importantly, brand images should be measured among consumers for whom the product category is meaningful.

## Question stability in brand image measurement

### Comparing alternative answer formats and accounting for heterogeneity in descriptive models

#### Abstract

High quality image data on how consumers perceive brands is essential to make good brand management decisions. Prior studies reveal that brand images are not very reliable, as they are typically measured in industry, which might be due to the answer format typically used (Rungie et al., 2005). The practical implication is that brand image data — as currently collected in consumer surveys — is not a valid source of market information. We challenge this implication.

Using three measures of stability we test whether the binary answer format produces image data less reliable than alternative formats. We investigate whether the aggregate descriptive model of brand image stability proposed by Rungie et al. can be improved by accounting for heterogeneity.

Results indicate that, compared to alternative formats, binary answer formats lead to equal stability levels, and most brand-attribute associations are stable. Unstable associations typically fail to describe adequately the brands under study.

Practical implications include that binary brand-attribute associations can be used safely to measure brand images. Also, practitioners can get guidance about required brand management measures by discriminating between stable and unstable brand-attribute associations. A model that helps managers classify brand-attribute associations into stable or unstable is proposed in the article.

**Deleted:** can use a descriptive model of brand-attribute associations that accounts for heterogeneity to

**Deleted:** e

**Deleted:** which call for different brand management measures.

#### Key words

brand image stability, brand image stability, answer formats, questionnaire design, finite mixture models, unobserved heterogeneity

## 1. Introduction

Brand image is defined and measured as a “set of associations which a brand has acquired for an individual” (Joyce, 1963, p. 45) and as “brand associations in consumer memory” (Keller, 1993). Strategic marketing decisions, such as positioning and segmentation, are typically based on market information obtained through consumer surveys. Brand-based industries use key market information from brand image survey data to determine how consumers perceive their brands. Because strategic decisions, and consequently expensive marketing actions, are based on information contained in brand image data sets, these must be of the highest quality.

Several studies over the past decade have questioned the quality of brand image data resulting from typical brand image surveys. These mainly criticise brand image data for its instability — if respondents are asked repeatedly to state brand-attribute associations, they do not reproduce the results of the first measurement very well in the second measurement. For a brand-attribute association to be stable for one particular respondent, the respondent would have to express agreement with the association in all repeated measurements. For instance, if a respondent states that McDonalds is expensive when asked for the first time, stability means that he or she would also say that McDonald is expensive when resurveyed.

Castleberry et al. (1994) use response levels (RL) to indicate the proportion of respondents assigning an attribute to a brand, and repeat rates (RR) to indicate the proportion of respondents assigning an attribute to a brand multiple times out of those who initially made this brand-attribute association. While response levels are stable at the aggregate level, answers are very unstable at the individual level, averaging at repeat rates of about 50 per cent (Castleberry et al., 1994). “Error of measurement” (p. 161) may explain this low level of stability. Dall’Olmo Riley et al. (1997) provide additional empirical support for the findings of Castleberry et al., with average repeat rates ranging from 40 to 60 per cent. They propose a simple model, in which RR and RL are linearly related by a constant of 20, to describe the relationship between RL and RR at the aggregate level across all brands and attributes measured. The model notation states that the constant of 20 is a percentage because both RL and RR are percentage values by definition.

Deleted: instable

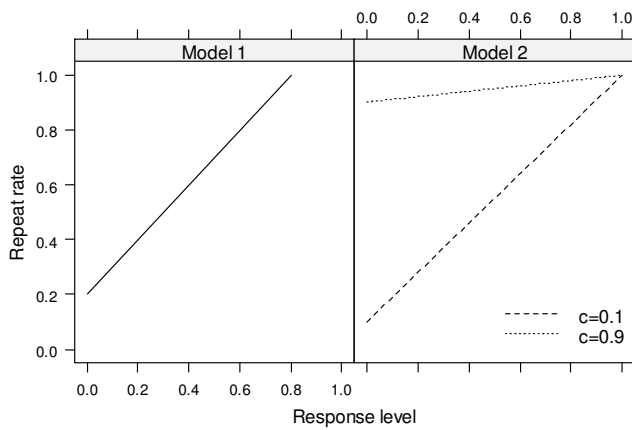
$$RR = RL + 20\%$$

Model 1

The practical interpretation of Model 1 is illustrated by discussing three kinds of brand-attribute associations along the linear function: (1) those held by a high proportion of consumers in a stable manner (see the top right-hand corner of Model 1 in Figure 1); (2) those held by a small proportion of consumers in an unstable manner (bottom left-hand corner); and (3) those held by a subset of consumers which are of medium stability (the middle area). Model 1 does not allow for a subset of consumers which has stable beliefs about a brand-attribute association. It does not account for consumer heterogeneity, although targeting specific sub-groups of the marketplace, and thus harvesting knowledge about consumer heterogeneity, is fundamental to brand-based industry marketing.

Deleted: are

**Figure 1: Graphical Illustration of Models 1 and 2**



Almost a decade after Castleberry et al. (1994), Rungie et al. (2005) reinvestigated brand image stability, and empirically demonstrated, over several data sets, the instability of brand images when measured in a binary way. This result throws the stability of binary answer formats into question, although the authors state explicitly that “a similar lack of reliability may exist for attitude questions in a Likert format” (Rungie et al., 2005, p. 317). “Reliability” is a broader term than “stability”, because it includes both test–retest reliability and internal consistency. “Stability” is therefore preferred, because it refers to the relation of two repeat

measurements of the same individual to each other, thus capturing only the test–retest reliability component. Rungie et al. (2005) propose an improved model that describes the aggregate relationship of RR and RL in which the coefficient  $c$  is referred to as “reliability”:

**Deleted:** That particular answer format is a possible cause of brand image instability. We

$$RR = c + (1 - c) RL$$

Model 2

The coefficient  $c$  subsumes the total variation in consumers’ responses from the first to the second survey wave. Specifically,  $c$  contains: (1) actual attitudinal change that may have occurred; (2) instability due to unstable brand images/insecurity about attribute-brand associations; (3) lack of stability of the answer format used; and (4) other possible measurement errors.

**Deleted:** instable

Model 2 can represent the subset of consumers with highly stable brand-attribute associations by  $c$ , and the subset of consumers with random (with probability RL) brand-attribute associations by  $(1-c)$ . The model thus accounts for consumer heterogeneity in stability, but assumes that all brand-attribute associations elicit the same stability in consumers. If  $c$  is low (see the bottom line in Model 2 in Figure 1), then Model 2 is similar to Model 1; if  $c$  is high (the top line in Model 2 in Figure 1), Model 2 postulates high stability levels for all brand-attribute associations. This limitation of Model 2 might not be realistic, because it cannot describe a market situation where two types of brand-attribute associations exist for low RLs: stable and unstable beliefs.

**Deleted:** instable

In summary, a brand manager studying the problem of brand image stability could be led to believe that binary measurement is not a good choice for brand image studies and that the stability of specific brand-attribute associations is the same for all consumers. Our study contributes to the area of brand image measurement research in questioning the above two managerial conclusions;

**Deleted:** two ways

**Research Objective 1** tests the hypothesis that the binary answer format does not cause a lack of stability in brand image data.

**Deleted:** Question

**Deleted:** proposed by Rungie et al. (2005):

**Research Objective 2** comparatively tests the two competing models that relate RR to RL with respect to how well they describe empirical brand image data sets at the aggregate level.

**Deleted:** Question

We propose an extended model ([Model 3](#)) which accounts for heterogeneity, and consequently overcomes the limitations of both models.

Research Question 1 is relevant to both researchers and practitioners: most contemporary brand image studies are conducted using binary data. If the binary answer format is responsible for low levels of stability, the validity of most current brand image studies conducted by organisations is highly doubtful. We may need to develop new answer formats to improve the validity of brand image studies. Our study offers guidance for both researchers and practitioners about how better to measure brand image.

The industry requires an improved model of describing brand-attribute associations from brand image surveys which accounts for both stable and [unstable](#) brand associations at low RL levels (Research Question 2). From a theoretical perspective, not all brand-attribute associations lack stability (lack of stability was concluded by Castleberry et al., 1994; Dall’Olmo Riley et al., 1997; Rungie et al., 2005). In practice, the ability to distinguish between stable and [unstable](#) brand-attributes at low RL levels allows brand managers to select suitable marketing actions for each case. Stable brand images at low RL levels indicate a market segment with a very stable brand perception — and represent segments very suitable for marketing action. [Unstable](#) attributes at low RL levels are either not particularly good brand descriptors, or alternatively, past advertising campaigns were not successful, so future campaigns should specifically target such attributes to increase consumer awareness, as well as strengthen the brand association.

Deleted: instable

Deleted: instable

Deleted: Instable

## 2. Data Collection

The study involved collecting data from university students, who were approached in compulsory tutorials over two consecutive weeks. Student IDs were used to match the two subsequent responses. The participants were asked to state their associations of six fast food chain brands with 11 attributes.

The fast food brands and attributes were selected in a multi-stage qualitative pre-study which aimed at identifying a product category that is relevant to the student population as well as known brands in the product category and attributes that are used by students to describe brands within the product category. In Stage 1 students were asked to list product categories they were interested in, then to complete a short questionnaire, which included the product

categories derived from Stage 1. The study asked participants to list as many brand names as they knew for each product category. The fast food product category emerged as most relevant to the majority of the student population based on frequency statistics. Other categories were highly relevant to a subset of the student population only and were therefore not suitable for our study. For instance, beer was one of the first product categories mentioned, but few students could list brand names, and entire segments (for example, Asian students) could not list a single brand. Finally, students were asked to state attributes of fast food brands in a separate, written short survey which asked them to list attributes of fast food chains, and attributes of a particular fast food chain (named in the questionnaire). This process ensured that the full range of attributes was collected. The highest frequency attributes were included in the final survey.

This study aims to understand the mechanism of how people respond to brand image questions. This mechanism is expected to be universal to all consumers, provided that basic principles of questionnaire design are ensured and that the product category they are asked to evaluate is meaningful to them. Consequently, our research aims can legitimately be investigated using a sub-sample of consumers — here, students in a large undergraduate subject. The process of selecting the product category, brands and attributes ensures the relevance of the brand image task to the population under study.

This data collection method should provide data of greater validity than the commercially collected data sets Rungie et al. used, because the product category, brands and attributes were specifically chosen to be meaningful to the population under study. This compares to commercial brand image studies, where consumers are asked to evaluate several product categories, brands and attributes, some of which may not be meaningful to them at all.

We used five alternative answer formats: 1) a six-point multi-category answer format with a fully verbalised subversion (that is, all categories are labelled); 2) a six-point multi-category answer format, with a subversion anchoring the endpoints only; 3) a five-point multi-category answer format with a fully verbalised subversion; 4) a five-point multi-category answer format, with a subversion anchoring the endpoints only; and 5) a full binary answer format where respondents had to choose between “yes” and “no”. The full binary format is not identical to the binary format (free choice) used in the data sets of Rungie et al. (2005). Our binary format forced respondents to answer each brand-attribute combination by either agreeing or disagreeing; whereas the free-choice format gives respondents more flexibility in

Deleted: one

Deleted: one

naming only selected brand-attribute combinations. To ensure consistency across experimental test conditions, all our answer formats were forced-choice formats. Using a “pick any” format would have favoured students confronted with the binary version of the questionnaire, because they would not have been forced to choose one answer option. This could potentially lead to interaction effects not separable from the binary answer format effect itself.

Hughes (1969) recommends the use of forced-choice formats in survey situations where respondents are aware of the attribute objects. This condition is met in our study, because the product category, brands and attributes were specifically chosen as relevant to the student population. Hughes also demonstrates the biases possible in forced-choice data, that is, the tendency to use the middle category when an uneven number of answer options is provided, and the tendency not to answer a question if an even number of options is provided. We tested whether the number of unanswered questions (missing data) was significantly higher for the even-answer format for both the endpoint anchored and the fully verbalised formats. We concluded that this was not the case, and therefore we consequently assume that no forced-choice bias was in our data set.

We included five-point and six-point scales because we hypothesised that allowing for a neutral option would affect responses — more respondents might have chosen the middle answer category, which could increase the stability of the brand image measurement over time.

In total, 272 students completed both questionnaires. Of those, 57 (21 per cent) used the six-point fully verbalised answer format, 55 (20 per cent) the six-point endpoints anchored format, 49 (18 per cent) the five-point fully verbalised, 60 (22 per cent) the five-point endpoints anchored and 51 (19 per cent) the binary answer format. Although we did not achieve identical numbers of respondents for each condition, the number of respondents do not differ significantly between the different answer formats, as indicated by a chi-squared test ( $\chi^2=1.46$ ,  $df=4$ ,  $p\text{-value}=.83$ ).

Respondents completed their task very conscientiously, and only 1.6 per cent of responses were missing. Because the analysis is based on either binary or category-specific associations (see Section 3 below) missing data enter computations as non-associations and do not have to be imputed. This approach leads to more conservative results because response levels are

reduced. However, given the small proportion of missing data, the effect is negligible.

### 3. Method

The comparison of stability is based on the RR and RL measures from Castleberry et al. (1994), which have so far been used exclusively for binary answer formats. For our comparison of alternative answer formats, including of multi-category formats, modification of the way RR and RL are computed is necessary. Three alternative approaches are possible:

- To determine which of the multi-category answer formats indicate that the respondents identified an association between a brand and an attribute, binarise the responses accordingly, and use the resulting RL and RR measures (“agreement stability”). This approach will work in favour of multi-category answer formats because slight variations in responses (for example, from answer option 1 to answer option 2 on the scale) will not be penalised by the stability measure.
- The RR measure can be redefined as indicating only an identical response in both waves on the exact same point on the answer format. This is the stricter measure, because any variation is interpreted as instability (“response category stability”).
- If the model for the assessment of reliability proposed by Rungie et al. (2005) holds for all answer formats after binarisation, the coefficient  $c$  can be used comparatively to assess stability of alternative answer formats. The estimated coefficient  $c$  is hence a reliability measure derived using the same binarized data which is used to analyse agreement stability.

Formatted: Bullets and Numbering

In our comparative study we implemented all three approaches. For the “agreement stability” approach we split the balanced answer formats along the agreement and disagreement dimension. For instance, the six-point answer formats were split into three agreement levels, set equal to a “yes” response in the binary format and into three disagreement levels set equal to a “no” response. For unbalanced answer formats, the midpoint was counted as non-agreement, because it was labelled in the fully verbalised version as “neither agree nor disagree”. This is consistent with the interpretation of typical free-choice answers in brand image measurement where the lack of a clear answer is not assumed to indicate agreement.

For the “response category stability” approach only identical responses in the two survey waves were counted as a reliable answer. A respondent had to use option 2 on a five-point multi-category answer format both in wave 1 and 2 for their response to be deemed reliable.

#### 4. Results

##### 4.1 Agreement stability

The response level (RL) across all answer formats averaged 52 per cent, with a standard deviation of 31. The mean repeat rate (RR) was 74 per cent across all answer formats, with a standard deviation of 24. Both these values are significantly higher than those reported in Rungie et al. (2005), which average an RL of 28 per cent and an RR of 49 per cent over eight data sets. This is not unexpected, because Rungie et al. use several data sets, most of which include several product categories. The respondents in those studies were presented with a large number of questions for assessment, and not all the product categories would have been relevant to them. Fatigue effects are known to affect data quality (Johnson et al., 1990), and shown to reduce stability in the brand image measurement context in the past (Dolnicar and Heindler, 2003).

Table 1 includes all key indicators used in Rungie et al. for all answer formats compared in our study: RL, RL2, RR and  $\rho$ . RL2 is the response level of the second wave and  $\rho$  is the proportion of all brand-attribute combinations which were agreed to in both waves. We refer, as do Rungie et al., to  $\rho$  as the “double positive rate”.

Deleted: ,

Deleted: which

**Table 1: Average positive response level, positive repeat rate and double positive rate  
using the criterion of agreement stability**

		Response level (RL)	Response level for wave 2 (RL2)	Repeat rate (RR)	Double positive rate ( $\rho$ )
		Mean (standard deviation)	Mean (standard deviation)	Mean (standard deviation)	Mean (standard deviation)
5-point	Fully verbalised	45% (31)	45% (31)	70% (26)	37% (29)
	Endpoint anchored	43% (30)	41% (26)	68% (22)	33% (27)
6-point	Fully verbalised	61% (31)	61% (29)	77% (22)	52% (32)
	Endpoint anchored	58% (31)	59% (29)	75% (24)	49% (32)
Binary		52% (31)	52% (30)	78% (23)	46% (31)
<b>Pooled data</b>		<b>52% (31)</b>	<b>51% (30)</b>	<b>74% (24)</b>	<b>44% (31)</b>

Before undertaking comparisons across answer formats, we compared aggregate response levels for both waves (see columns 1 and 2 in Table 1). A test for equality of proportions confirms no significant difference between the waves. We therefore assume that no major structural effects (such as fatigue effect on the side of the respondents) affected the responses in the second survey wave.

An analysis of variance indicates significant differences between the answer formats for all measures (see Table 2).

**Table 2: Analysis of variance**

	F	df <sub>1</sub>	df <sub>2</sub>	p-value
RL	4.41	4	325	.002
RR	2.62	4	319	.035
$\rho$	4.96	4	325	<.001

These significant differences are caused by the five-point scales, as indicated by pair-wise *t*-

tests, because only these comparisons have significant p-values for a significance level of five per cent. The five-point scales' performance is worse than the other answer formats, with an average repeat rate of 70 per cent for the fully verbalised and 68 per cent for the endpoint-anchored alternatives. The double positive rate is as low as one-third. By classifying the midpoint as an answer which does not indicate agreement, the RL will likely be lower for scales with a midpoint than scales without a midpoint, where those respondents who would have ticked the midpoint are forced to decide between agreeing and disagreeing. No significant differences exist between the two alternative forms of the six-point scale and the binary scale, indicating that agreement and disagreement are captured in an equally reliable manner. These findings confirm results of previous answer format comparisons (Dolnicar, 2003; Dolnicar et al. 2004; Dolnicar and Grün, 2007).

## 4.2 Response category stability

In addition to the measure of agreement stability we computed the stability of respondents in responding with the precisely same answer category to the brand-attribute association questions. Table 3 shows these results.

**Table 3: Average positive response level, positive repeat rate, double positive rate using  
the criterion of response category stability.**

Deleted: .

Answer format	Category	Response level (RL)	Repeat rate (RR)	Double positive ( $\rho$ )
		Mean*	Mean*	Mean*
5-point, fully verbalised	Strongly disagree	7%	39%	4%
	Disagree	19%	50%	12%
	Neither agree nor disagree	27%	60%	17%
	Agree	31%	62%	23%
	Strongly agree	14%	43%	8%
	<b>All</b>	<b>20%</b>	<b>52%</b>	<b>13%</b>
5-point, endpoint anchored	Strongly disagree	14%	58%	9%
		16%	42%	7%
		26%	56%	16%
		20%	44%	10%
	Strongly agree	22%	61%	16%
	<b>All</b>	<b>20%</b>	<b>52%</b>	<b>12%</b>
6-point, fully verbalised	Strongly disagree	7%	49%	4%
	Disagree	16%	45%	9%
	Mildly disagree	14%	36%	6%
	Mildly agree	21%	45%	10%
	Agree	26%	53%	16%
	Strongly agree	14%	46%	8%
<b>All</b>	<b>17%</b>	<b>45%</b>	<b>9%</b>	
6-point, endpoint anchored	Strongly disagree	11%	50%	6%
		14%	36%	6%
		17%	37%	7%
		22%	46%	11%
		20%	51%	11%
	Strongly agree	16%	59%	10%
<b>All</b>	<b>16%</b>	<b>46%</b>	<b>9%</b>	
binary	No	45%	76%	39%
	Yes	52%	78%	46%
	<b>All</b>	<b>49%</b>	<b>77%</b>	<b>42%</b>

\* In order to increase readability, standard deviations are not included in the table. The standard deviations for

RL values ranged from 10 to 31, for RR values from 16 to 34 and for the double positive values from 5 to 31.

Two interesting findings emerge from this computation. Binary answer formats reach the highest level of RR across all answer categories, followed by the five-point and six-point scale. However, the answer patterns within the categories differ significantly between the endpoint-anchored and the fully verbalised scale alternatives (both for the five-point and the six-point version), with the endpoint-anchored version attracting more responses to the endpoints than the fully verbalised alternative. Generally, the endpoints of the multi-category answer formats achieve surprisingly high RR levels, given the relatively low initial RL levels. This is plausible because people using the endpoints probably have a clear, and consequently stable, perception of that particular brand-attribute association.

### 4.3 Coefficient $c$

In order to assess whether coefficient  $c$  can be used as a stability measure for our answer format comparison, we first established the validity of the two models proposed in prior work on brand image stability (Models 1 and 2 discussed in the introduction).

The validity of the two models proposed was tested by computing ordinary least squares regressions using  $\rho$ , the probability of two agreement answers in both waves as the dependent variable, and RL and squared RL, as independent variables. Model 1 implies the following relationship between  $\rho$  and the RL:

$$\rho = \text{RL}^2 + 0.2 \text{RL}$$

Because both RL and  $\rho$  in the formula above enter as probabilities rather than percentage values, the constant of 20 per cent rescaled accordingly (to 0.2). According to Model 2, it is given by:

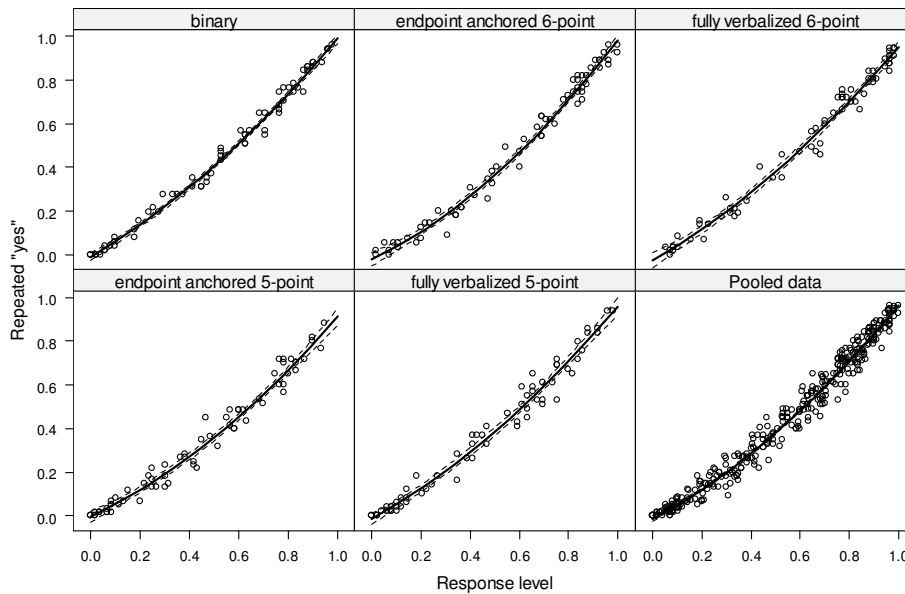
$$\rho = (1-c) \text{RL}^2 + c \text{RL}$$

Table 4 shows the results. Figure 2 depicts the data and the fitted regressions for all five answer formats and the pooled data. The full lines indicate the estimated mean values; the dashed lines indicate the estimated 95 per cent confidence interval for the mean values.

**Table 4: Empirical ordinary least squares estimates of the regression of  $\rho$  on RL and  $RL^2$**

<b>Answer format</b>	<b>N</b>	<b>R<sup>2</sup></b>	<b>F</b>	<b>Constant (t)</b>	<b>1<sup>st</sup> degree coefficient (t)</b>	<b>2<sup>nd</sup> degree coefficient (t)</b>	<b>Total</b>
5-point, fully verbalised	66	.978	1407	-.013 (t=-1.061)	.630 (t=8.909)	.343 (t=4.512)	.972
5-point, endpoint anchored	66	.977	1330	-.004 (t=-.319)	.540 (t=8.312)	.375 (t=5.320)	.914
6-point, fully verbalised	66	.983	1843	-.025 (t=-1.441)	.642 (t=8.095)	.335 (t=4.622)	.976
6-point, endpoint anchored	66	.985	2123	-.020 (t=-1.274)	.539 (t=7.630)	.464 (t=7.022)	1.003
Binary	66	.991	3303	-.008 (t=-.850)	.663 (t=13.956)	.334 (t=6.980)	.998
Pooled data	330	.981	8562	-.012 (t=-1.870)	.588 (t=19.191)	.388 (t=12.796)	.976

**Figure 2: Empirical relationships between response levels and probability  $\rho$  of a repeated positive answer**



The first model assumption, which is identical for both Model 1 and 2, is that the intercept is zero. The regression results show that this is the case for the empirical data investigated: the column “Constant (t)” in Table 4 contains all the estimated intercepts. All are very small and, as the respective t-values indicate, insignificant. Consequently, we can confirm that the assumption of a zero intercept holds, consistent with both Model 1 and 2.

Deleted: supporting

With respect to the coefficients, Model 1 assumes that the value of the first-order coefficient is 0.2, and the value of the second-order coefficient is one. Contrarily, Model 2 implies that the sum of the first-order and second-order coefficients is 1. The regression results provided in Table 4 contradict both Model 1 assumptions: that of a 0.2 first-order coefficient (which empirically ranges from 0.539 to 0.663) and that of a second-order coefficient equal to 1 (which empirically ranges from 0.334 to 0.464). However, Model 2 assumptions are supported, with the sum of first- and second-order coefficients at very close to 1 for all answer formats.

Model 2 describes the data resulting from alternative answer formats better than Model 1. Consequently, coefficient  $c$  can be used to compare the stability of the alternative answer formats in our experimental design. The differences in coefficient  $c$  are less ambiguous in our comparison than the absolute value of coefficient  $c$  presented in Rungie et al., given that a possible attitudinal change towards fast food chains, as well as insecurity about brand attribute evaluations, are held constant across all conditions. We therefore assume that the difference in coefficient  $c$  in our experiment captures the differences in stability of answer formats only.

Table 5 provides coefficient  $c$  estimates for each answer format and the pooled data. They are estimated under the assumption that Model 2 is valid; that is, they are fitted under the restriction that the intercept and the coefficient of the RL sum to 1 when modelling the linear relationship between RL and RR. Results indicate that the stability of brand images derived from alternative answer formats is very similar, with coefficient  $c$  values ranging from .407 for the five-point multi-category answer format to .487 for the binary answer format. These results reflect the findings based on agreement stability. Given that the coefficient  $c$  is a derived measure using data which was binarized using an agreement/disagreement split and hence is based on data which was also used for the analysis of agreement stability, the similarity of results is not surprising.

- Deleted: T
- Deleted: based on
- Deleted: , as is the
- Deleted: measure. Consequently

**Table 5: Estimated reliabilities  $c$  for each answer format**

Answer format		Estimate of the stability $c$	Standard error
5-point	Fully verbalised	.407	.033
	Endpoint anchored	.440	.034
6-point	Fully verbalised	.419	.031
	Endpoint anchored	.413	.036
Binary		.487	.034
Pooled data		.433	.015

An analysis of variance indicates that the separate regression models for each answer format do not fit the data significantly better than the pooled regression ( $F=.904$ ,  $df_1=4$ ,  $df_2=319$ ,  $p\text{-value}=.462$ ). Based on the coefficient  $c$  as a stability measure, we conclude that all answer formats are equally reliable (or unreliable) in capturing brand images.

#### 4.4 Accounting for heterogeneity

Having used coefficient  $c$  for comparative assessment, we investigate how the brand image data can be better described, and whether accounting for heterogeneity improves the model.

Figure 3 depicts the regression lines derived from Model 2, as well as all the data points. If respondents assigned brand-attribute associations randomly, all data points would be located along the main diagonal. Because the population under study (students) is aware of the product category of fast food restaurants as well as the brands and attributes, we would expect that responses to be more stable over repeat measurements than the random model indicates. Consequently, we expect data points to be located above the main diagonal. This is clearly the case. Therefore the variability of RR is higher at lower levels of RL.

**Figure 3: Repeat rate versus response level with the linear relationship implied by Model 2**

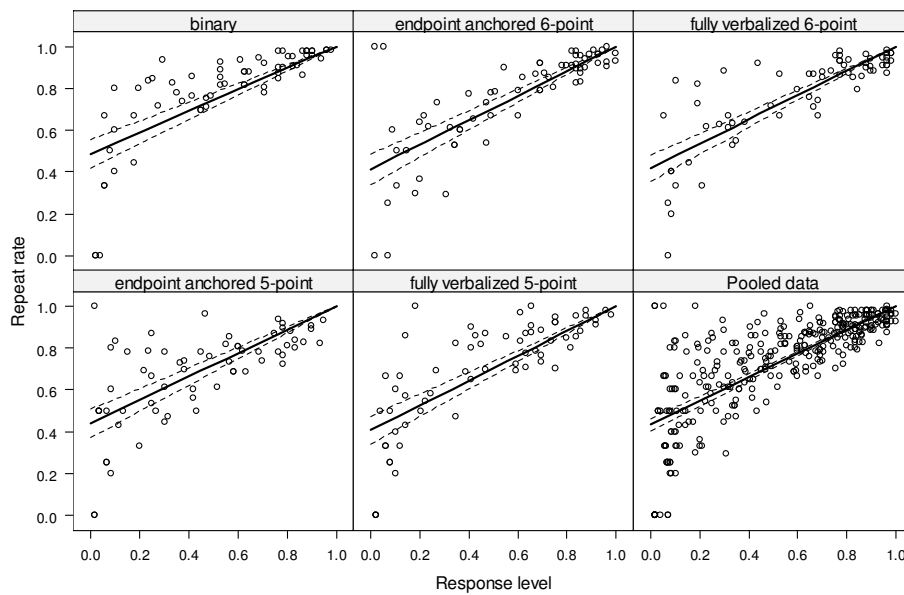


Figure 3 shows that the regression lines do not fit the data well. A substantial number of data points lie far away from the regression line. These data points are of particular interest, because they are in contradiction with the proposed models to describe the RL–RR relationship. Of specific interest is the case where RL is low, but RR is high. This means that only few respondents assign an attribute to a brand, but those who do are consistent in their assessment. In order to identify which brand-attribute associations in our data set demonstrate this pattern, we selected all brand-attribute associations for which RR was higher than RL + 40 (representing a constant twice as high as postulated by Model 1). Across all answer formats one brand attribute association complied with this criterion: “McDonald’s” with “expensive”. This indicates that there is a sub-segment (indicated by a low RL) which consistently states that McDonald’s is expensive. The attribute that most frequently occurs in the selected subset of brand-attribute combinations is “disgusting”. For McDonald’s and KFC, sub-segments of students exist who repeatedly evaluate these brands as disgusting (across four and three answer formats). Figure 3 shows, and the examples discussed illustrate, that the better of the two alternative models (Model 2) does not fit the data well, because it systematically underestimates the stability of brand-attribute associations.

Deleted: ¶

Deleted: ¶

We propose to relax the assumption of Model 2, that the coefficient  $c$  is constant across all brand-attribute associations, thus assuming the same level of stability across all associations. This leads to the formulation of an alternative model (Model 3), in which coefficient  $c$  consists of multiple coefficients which describe subsets of brand-attribute associations (thus accounting for heterogeneity; see Grün et al., 2007). However, in this model the assumption that the sum of the coefficients is equal to 1 is not relaxed. We tested Model 3 by fitting finite mixtures of regressions with two components. The components of Model 3 are restricted to have equal variances and to contain at least 10 per cent of the observations. The EM algorithm (Dempster et al., 1977) is used to fit the models and obtain the maximum likelihood estimates. The mixture regression model is given by:

$$H(RR, \Theta) = \sum_{k=1}^K \pi_k N(RR; \mu_k(RL), \sigma^2) \quad \text{Model 3}$$

— where  $H(\cdot)$  is the mixture distribution,  $\Theta$  is the vector of all parameters of the mixture

distribution and  $N(y; \mu, \sigma^2)$  is the Normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The component specific mean is determined by:

$$\mu_k(RL) = c_k + (1 - c_k)RL$$

— where  $c_k$  is the component specific coefficient  $c$  measuring stability. The number of components is given by  $k$ . Each class/component is of size  $\pi_k$  and the component sizes have to fulfil the following constraints:

$$\pi_k \geq 0.1 \quad \forall k = 1, \dots, K \quad \wedge \quad \sum_{k=1}^K \pi_k = 1.$$

Model 3 (Figure 4) outperforms Model 2 (Figure 3) with respect to both the AIC and BIC criteria (see Table 6), indicating that it is better to account for heterogeneity instead of assuming a constant stability coefficient.

Table 6: AIC and BIC values for Model 2 and Model 3

		<u>BIC</u>		<u>AIC</u>	
<u>Answer format</u>		<u>Model 2</u>	<u>Model 3</u>	<u>Model 2</u>	<u>Model 3</u>
<u>5-point</u>	<u>Fully verbalised</u>	<u>-42.4</u>	<u>-50.5</u>	<u>-46.7</u>	<u>-59.1</u>
	<u>Endpoint anchored</u>	<u>-38.6</u>	<u>-43.9</u>	<u>-42.9</u>	<u>-52.6</u>
<u>6-point</u>	<u>Fully verbalised</u>	<u>-77.6</u>	<u>-94.4</u>	<u>-82.0</u>	<u>-103.2</u>
	<u>Endpoint anchored</u>	<u>-52.7</u>	<u>-64.9</u>	<u>-57.1</u>	<u>-73.6</u>
<u>Binary</u>		<u>-53.5</u>	<u>-86.4</u>	<u>-57.9</u>	<u>-95.1</u>
<u>Pooled data</u>		<u>-285.3</u>	<u>-377.9</u>	<u>-292.9</u>	<u>-393.1</u>

Figure 4 shows the fitted regression lines of Model 3 for each component. The observations are plotted in different symbols, according to the assignment to one of the two components with respect to their maximum *a posteriori* probability. Observations assigned to the smaller component are depicted using crosses and for the observations assigned to the larger component triangles are used. . The full lines indicate the estimated mean values for each of the components separately and the dashed lines indicate the estimated 95 per cent confidence interval for these mean values. Table 7 shows the corresponding estimated coefficients *c* and approximate standard errors for each of the components, as well as the relative size of the components.

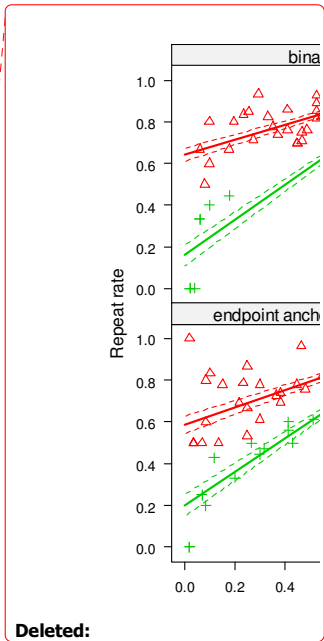
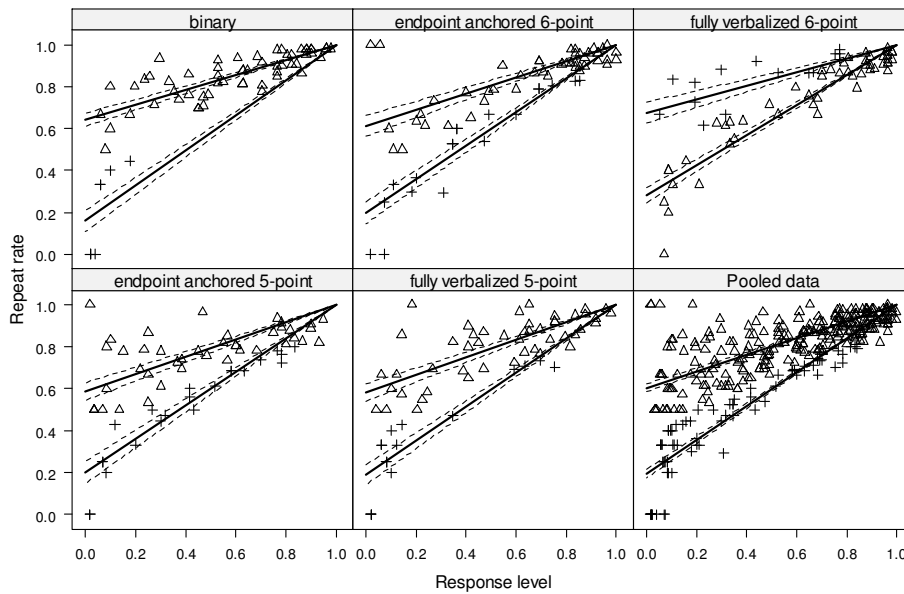
Deleted: /colours

Deleted: 6

**Table 7:** Estimated coefficients  $c$  for each component of the mixture model and each answer format

Answer format		Component 1			Component 2		
		Estimated coefficient	Standard error	Relative size	Estimated coefficient	Standard error	Relative size
		$c$			$c$		
5-point	Fully verbalised	.582	.031	.642	.189	.034	.358
	Endpoint anchored	.589	.030	.601	.200	.039	.399
6-point	Fully verbalised	.281	.026	.642	.676	.036	.358
	Endpoint anchored	.612	.036	.556	.198	.037	.444
Binary		.643	.024	.830	.159	.034	.170
Pooled data		.605	.013	.640	.196	.016	.360

**Figure 4:** Repeat rate versus response level with the linear relationship implied by Model 2 allowing for heterogeneity of the stability coefficients



Deleted:

The assumption that the sum of the coefficients is equal to 1 is tested by comparing the BIC of the two-segment solution with unrestricted coefficients in the linear model of the components to the BIC of the two-segment solution, where the coefficients are restricted to sum to 1. BIC values are better for the restricted model for all answer formats except the endpoint-anchored five-point scale and the pooled data. A likelihood ratio test comparing the two models for each answer format supports this finding (binary: p-value=.10; six-point endpoint anchored: p-value=.06 and fully verbalised: p-value=.11; five-point endpoint anchored: p-value=.01 and fully verbalised: p-value=.28; pooled data: p-value<.01).

Based on the above computations, we conclude that Model 3 outperforms Model 2 in describing the data, and that the model assumptions underlying Model 3 hold.

The practical implications of Model 3 are that brand-attribute associations are not all equally stable. Some associations, represented by the top components in Figure 4, are reproduced in a stable manner by consumers across the entire range of RL. The associations within this group that are characterised by both high RLs and high RRs represent brand image dimensions which are perceived by a large proportion of consumers in a stable way. One example from our empirical data set is the brand-attribute association “McDonald’s” and “yummy”, with an average RL across all answer formats of 64, and an average RR of 94. Practically, this means that a large group of consumers (64 per cent of respondents) perceive McDonald’s to be yummy, and they do so in a highly stable manner, in 94 per cent of cases. If such stable beliefs in the majority of the population are in line with brand management aims, mass marketing campaigns could be used to reinforce the belief. If they are negative, mass marketing campaigns would be necessary to counteract such beliefs among the majority of consumers.

Associations with low RLs and high RRs indicate the existence of sub-segments of consumers holding very firm views about specific brand-attribute associations which they do not share with the majority of consumers. One such example is “McDonald’s” and “expensive”, as discussed above. A second example is “Pizza Hut” and “healthy”: five per cent of respondents stated in the first survey wave that they perceived Pizza Hut to be healthy. Eighty per cent of these respondents repeated this evaluation in the second survey wave, indicating that they really do perceive Pizza Hut as being healthy, in a stable manner. It may be attractive for brand managers to target such sub-segments specifically to reinforce their positive brand

beliefs or counteract negative brand beliefs.

Other associations are not stable, for example, “McDonald’s” and “healthy”. A small proportion of respondents associated McDonald’s with healthy in the first survey wave, but only in 13 per cent of cases was this association repeated in the second wave survey. This indicates either: (1) advertising campaigns were not successful, and people consequently do not have stable brand-attribute associations; or (2) respondents might have been presented with irrelevant brand-attribute combinations. The brand management implication for the first case is that improved campaigns have to be developed. In the second case, increased qualitative work is necessary before the quantitative fieldwork for the brand image study is conducted, in order to ensure that only attributes that are relevant descriptors of brands in the eyes of consumers are included.

The practical implications of Model 3 are very plausible, because some brand-attribute combinations (such as “Subway” and “healthy”) are stored in consumers’ minds to a greater degree than others (such as “Subway” and “spicy”). This could be due to good promotion that has achieved a strong association between a brand name and an attribute in consumers’ memories. Alternatively, it could be due to brand-attribute combinations that can be easily guessed by respondents.

The results in Table 7 show that in all but one case (six-point fully verbalised) the more reliable brand-attribute combinations (those with a higher coefficient  $c$ ) represent the larger of the two groups of brand-attribute combinations. For example, in the binary case, 83 per cent of brand-attribute combinations have a coefficient  $c$  of .643, and only 17 per cent have a lower stability of .159. We therefore conclude that Model 2 underestimates the stability of brand-attribute associations.

Deleted: 6

In order to analyse if there is an association between the components and the different brands and attributes, a multinomial logit model, with the posterior probabilities as dependent variables, and brands and attributes as independent variables, is estimated for each answer format. However, no significant relationships were detected at a significance level of five per cent. This indicates that there is not a single brand or attribute which is less reliable; however, this might be the case for specific brand-attribute associations. An inspection of the brand-attribute associations assigned to the component with the smaller stability reveals that eight brand-attribute associations are never assigned to this component by each of the models:

“yummy” and Burger King, KFC, McDonald’s and Pizza Hut, “cheap” and McDonald’s and Subway, “expensive” and McDonald’s, and “fast” and Pizza Hut. No brand-attributes association was assigned to the less-reliable component for each model, but “spicy” and Burger King was in the less-reliable component for each answer format except for the endpoint-anchored six-point scale.

## **5. Conclusions**

Past research questions the stability of empirical brand image data sets. One suggested explanation for low stability is the answer format used in brand image surveys. Because brand image surveys typically use binary answer formats, the implicit conclusion from past research is that the binary answer format may not be suitable for brand image surveys because it lacks stability.

One aim of the present study was to test whether alternative answer formats perform better (if they led to more reliable results). A data set including five different answer formats were used for the empirical investigation. Results indicate that the answer format does not significantly affect brand image stability.

We also compared two alternative models of describing brand image data proposed over the past decade. The model proposed by Rungie et al. (2005) outperforms the simpler model proposed earlier by Dall’Olmo Riley et al. (1997). We extended the Rungie model by taking heterogeneity into account and demonstrate further improvement in the description of brand image data. Accounting for heterogeneity reveals that the majority of brand-attribute associations demonstrate higher levels of stability than previously reported.

These results are of major importance for research and practitioners who rely on empirical brand image data as the basis of their knowledge development or branding strategy. The results of this study demonstrate — contrary to some prior work — relatively high levels of stability of brand images, meaning that brand image data is a valid source of information for management action. The study also demonstrates that the reliability of brand image data is highly dependent on the set-up of the brand image study. In particular, it is important that brand images are measured among consumers for whom the respective product categories are

meaningful, rather than across a random sample of the entire population.

The study provides empirical evidence that the binary answer format is suitable for measuring brand image data in a reliable manner, a notion recently questioned by Rungie et al. Finally, the model proposed to describe empirical brand image data enables brand managers to identify different kinds of brand-attribute associations in the marketplace which require different brand marketing actions to be taken.

Our study also raises several new questions, which offer interesting future investigations. The current study is limited to forced-choice formats. A direct comparison of free and forced choice binary answer formats might be undertaken to ensure that the free-choice (“pick any”) format is not causing higher levels of instability. Our research design in measuring brand images included extensive qualitative work and was customised to a specific segment of the market. Consequently, the questions were relevant to most respondents. Our questionnaire was easy to understand and short, so fatigue effects would not have occurred in data collection. Yet, despite these almost perfect measurement conditions, the brand image associations were not perfectly reproduced in consecutive weeks. This requires two kinds of follow-up investigations: repeat measure studies that control for intervening variables, such as advertising exposure, media reports and so on, which may explain changes in brand-attribute associations; and alternative hypotheses about instability resulting from survey research theory. For example, “satisficing” (Krosnick et al., 1996) could explain lack of stability.

Satisficing means that people minimize the effort needed to make a decision, as long as the outcome is acceptable. In the survey context satisficing means that respondents do not go through the stepwise process of properly responding to survey questions (interpreting meaning, searching memory, integrating information into a summary judgement and responding). Instead they either go through this process superficially (weak satisficing) or they omit the retrieval and judgement steps.

**Deleted:** due to respondents constructing responses while completing the questionnaire.

Finally, replication studies should be conducted to test whether our results are replicated for different subsets of consumers and different competitive sets of brands in different product categories.

## References

- Castleberry, S.B., Barnard, N.R., Barwise, T.P., Ehrenberg, A.S.C., Dall'Olmo Riley, F., 1994. Individual attitude variations over time. *Journal of Marketing Management* 10 (1–3), 153–162.
- Dall'Olmo Riley, F., Ehrenberg, A.S.C., Casleberry, S.B., Barwise, T.P., Barnard, N.R., 1997. The variability of attitudinal repeat-rates. *International Journal of Research in Marketing* 14 (5), 437–45.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B39*, 1–38.
- Dolnicar, S., 2003. Simplifying three-way questionnaires — Do the advantages of binary answer categories compensate for the loss of information? CD Proceedings of the Australian and New Zealand Marketing Academy Conference (ANZMAC).
- 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 (2), 108–122.
- 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.
- Dolnicar, S., Heindler, M., 2004. If you don't need to know, don't ask! Does questionnaire length dilute the stability of brand images? Proceedings of the 33rd Annual Conference of the European Marketing Academy, CD version.
- Grün, B., Dolnicar, S., Rossiter, J., 2007. Extending Rungie et al.'s model of brand image stability to account for heterogeneity. CD Proceedings of the European Marketing Academy Conference.
- Hughes, G.D., 1969. Some confounding effects of forced-choice scales. *Journal of Marketing Research* 6, 223–226.
- 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, 35–43.

- Joyce, T., 1963. Techniques of brand image measurement. In: *New Developments in Research*. Market Research Society, London, pp. 45–63.
- Keller, K., 1993. Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing* 57, 1–22.
- Krosnick, J.A., Narayan, S., Smith, W.R., 1996. Satisficing in surveys: initial evidence. *New Directions for Evaluation* 70 (Summer), 29–44.
- Rungie, C., Laurent, G., Dall’Olmo Riley, F., Morrison, D.G., Roy, T., 2005. Measuring and modelling the (limited) reliability of free choice attitude questions. *International Journal of Research in Marketing* 22 (3), 309–318.