Journal of Business Research 66 (2013) 1298–1306

Contents lists available at SciVerse ScienceDirect

Journal of Business Research

# “Translating” between survey answer formats☆

Sara Dolnicar [a,*], Bettina Grün [b,1]

[a] Institute for Innovation in Business and Social Research, University of Wollongong, Wollongong, NSW 2522, Australia
[b] Department of Applied Statistics, Johannes Kepler Universität Linz, 4040 Linz, Austria

ARTICLE INFO

*Article history:*
Received 1 April 2011
Received in revised form 1 October 2011
Accepted 1 November 2011
Available online 2 March 2012

*Keywords:*
Questionnaire design
Survey research
Answer formats
Likert

ABSTRACT

Survey research remains the most popular source of market knowledge, yet researchers have not yet established one consistent technique for measuring responses. Some market research companies offer respondents two answer options; others five or seven. Some answer formats use middle points on the answer scales, others do not. Some formats verbalize all answer options, some only the endpoints. The wide variety of answer formats that market research companies and academic researchers use makes comparing results across studies virtually impossible. This study offers guidance for market researchers by presenting empirical translations for the answer formats they most commonly use, thus enabling easier comparisons of results.

## 1. Introduction

Organizations heavily use survey research to learn about consumer behavior, preferences, and perceptions. While repeat surveys by the same organization using the same market research company typically use the same answer format, this does not occur in studies that different organizations, market research companies, or academic researchers conduct, which makes comparing results across different studies virtually impossible.

A good example of this problem occurs in research into the stated acceptance for recycled water. Researchers first conducted studies in this area in the early 1970s, and continue to conduct them internationally. Two Australian examples illustrate the point well. They were both published in 2006 and refer to the same geographic region, yet report acceptance levels for drinking recycled water of 11% and 47% respectively; a difference that suggests that how the questions are asked, and what answer options are offered, significantly affect results. Hurlimann (2006), who reports the higher acceptance level, asked respondents how happy they would be using recycled water, and offered a ten-point scale ranging from *not at all happy to use recycled water* to *extremely happy to use recycled water*. The authors added responses with the value of six or more on the ten-point scale to determine the 47% acceptance level. Dolnicar and Schäfer (2006) report the lower acceptance level of 11%. They asked respondents in that study a scenario question and offered five fully verbalized answer options; the 11% acceptance level represents the respondents who selected the *very likely* answer option.

The consequences of such measurement inconsistencies and the absence of guidance on how to compare results across studies are that recycled water usage studies have produced many heterogeneous and incompatible numbers, instead of making definitive contributions to the body of knowledge. Such dissimilar results appear in many contexts, because no strategies are available for comparing survey results that employ different answer formats. The lack of tools to compare results effectively weakens our ability to draw valid conclusions and develop a body of knowledge in certain research areas.

The present study addresses the problem of heterogeneous and incompatible survey results by offering empirical translations that support comparisons of results across studies, regardless of the answer formats employed. The tools that this study generates should be particularly useful to market researchers, academic researchers, and users of market research studies. Specifically, this study provides translations that allow practitioners to compare: the forced-choice full binary answer format against other answer formats in common use; answer formats with middle points against answer formats without middle points; Likert-type and bipolar answer formats; and answer formats with fully verbalized options against endpoint-labeled answer formats.

In offering empirical translations to compare results from different survey methodologies, this study contributes to the theoretical understanding of answer formats in survey research, and is of direct practical value to market researchers, academic researchers and users of market research results.

This study does not determine a single, most-valid answer format. Rather, it accepts that different studies use different answer formats, and the consequent virtual impossibility of comparing results across

☆ This research was funded by the Australian Research Council under the Discovery and Linkage International Grant Schemes and by the Austrian Science Fund (FWF) under Elise-Richter grant V170-N18. Authors are listed in alphabetical order.

* Corresponding author. Tel.: +61 2 4221 3862; fax: +61 2 4221 4210.
*E-mail addresses:* sarad@uow.edu.au (S. Dolnicar), Bettina.Gruen@jku.at (B. Grün).
[1] Tel.: +43 732 2468 6829; fax: +43 732 2468 6800.

0148-2963 © 2012 Elsevier Inc. Open access under CC BY-NC-ND license.
doi:10.1016/j.jbusres.2012.02.029

studies. This paper is the first to provide guidance for translating different answer formats onto one another. Such guidance is important when comparing findings across studies, or comparing results over time in longitudinal studies, because researchers often encounter dissimilar answer formats. In addition, researchers frequently binarize multi-categorical data using the middle point to split respondents. This study demonstrates that such binarization does not actually match the internal translation process of respondents, which leads to invalid data transformations before data analysis even starts. The presented translations address problems associated with changed or different answer formats, and validity in the binarization of multi-categorical data.

The context of brand image measurement limits empirical investigation in the present study; traditionally, the free-choice binary or pick any/n answer format dominates commercial research (such as in brand tracking studies). According to Rossiter (2011, p. 75), brand-attribute beliefs, which brand image studies measure, are the single most common construct measured in marketing research. Also, interactions often occur between the construct under study and the answer format; and therefore, results may deviate somewhat for other constructs under study (Dolnicar & Grün, 2007a, 2009).

### *1.1. Prior work*

Prior work that relates to this study resides in two areas. First are studies that seek the best answer formats. Second are studies that attempt to translate between answer formats. The research debate over the best answer format is as old as survey research itself. Authors tend to (rather passionately) take one of two positions: either they propose that binary measures are sufficient (Bendig, 1954; Dolnicar & Grün, 2007a, 2007b; Dolnicar, Grün, & Leisch, 2011; Komorita & Graham, 1965; Martin, Fruchter, & Mathis, 1974; Matell & Jacoby, 1971a, 1971b; Schutz & Rucker, 1975), or they tend to reject absolutely binary measures and instead use multi-category answer formats. Within the latter group, views differ regarding the optimal number of answer options, with recommendations ranging from five (Boote, 1981; Jenkins & Taber, 1977; Lissitz & Green, 1975; Remmers & Ewart, 1941), to six (Finn, 1972; Green & Rao, 1970), to seven (Cicchetti, Showalter, & Tyrer, 1985; Miller, 1956; Oaster, 1989; Symonds, 1924) and nine (Hancock & Klockars, 1991), and 18 or more (Champney & Marshall, 1939; Garner, 1960). The key argument between these opposing groups is whether additional answer options add precision to the measurement, or merely capture noise (such as response styles).

Garner (1960, p. 352) is representative of the opinion of multi-category proponents: “information transmission cannot be lost by increasing the number of rating categories. Therefore, it is better to err on the side of having too many categories than to err by having too few.” Peabody (1962, p. 73) characterizes the position of binary measure proponents: differences in responses using multi-category answer formats “primarily represent response sets, and only to a secondary degree actual differences in intensity.” This group believes that response sets represent contamination of data, rather than additional information. Avoiding response bias, according to Rossiter (2002, 2011), is a key requirement for any measure to be content valid, and content validity is the ultimate quality criterion for measures in the social sciences.

The body of literature on answer formats does not lead to any firm conclusion about what is ultimately the best answer format. This vagueness is attributable to how past researchers have conducted studies in a range of different contexts, using a range of different evaluation criteria for answer formats, and with many variations in how they word answer options or present them to respondents. Despite the significant body of research comparing answer formats, no work has yet been conducted comparing different formats of binary measures (e.g., pick any/n compared to forced full binary).

Only a very small number of studies are available that relate to translating responses from one answer format to another. Haley and Case (1979) provide the first study of this kind, evaluating 13 commonly used scales in brand image measurement with respect to answer patterns, measured content, concurrent validity, and discrimination between brands. They conclude that forced-choice answer formats, as well as answer formats with fully verbalized answer options, perform better. Hui and Triandis (1989) compare responses from five- and ten-point answer formats for Hispanic and non-Hispanic respondents. However, their research design, which is not longitudinal, does not permit mapping across answer formats. The chart they provide shows frequencies of use for each answer option for both formats, and indicates that more answer options reduce extreme response styles.

Dolnicar and Grün (2007a) and Dolnicar et al. (2011) examine transformations between a limited number of answer formats. Dolnicar and Grün (2007a) scrutinize measures of two different constructs (behavioral intentions and attitudes), employing a repeat measurement design on three different answer formats (full binary, metric and ordinal seven-point); while Dolnicar et al. (2011) investigate the mappings between a full binary and an ordinal six-point answer format.

## 2. Data and method

The experiment used a permission-based internet panel that asked respondents representative of the Australian adult population to complete two brand image questionnaires with an approximate two-week break between measurements. Both questionnaire versions were identical, except for the answer format. This design enabled the derivation of individual-level translations, because the collected data allowed mapping of how each respondent answered from one answer format to another. Any variation between the two measurements was not caused by inter-individual differences or changes in brand perception, because the time between measurements was short, and no changes in advertising campaigns or the marketplace occurred that could have changed respondents' brand evaluations.

Brand image measurements are not perfectly stable, even under unchanged market conditions or when the same answer format is used (Dolnicar & Grün, 2007b; Dolnicar & Rossiter, 2008; Rungie, Laurent, Dall'Olmo Riley, Morrison, & Roy, 2005). Therefore, also the present study will capture some of this instability. However, a reduction of this effect was achieved by following the measurement recommendations of Dolnicar and Rossiter (2008). Also, any variations due to instability in brand image measurement should affect all experimental conditions equally, with no bias toward any of the answer formats. In addition, base instability levels are reported for repeat measurements on the same answer format.

Respondents assessed two brands: McDonald's (very well known among Australians) and Red Rooster (less well known). The five attributes presented to respondents were *yummy*, *fast*, *cheap*, *healthy*, and *convenient*. These attributes were derived from a prior, extensive, qualitative study where interview respondents were asked about the relevant characteristics of fast food brands. Each item identified through the qualitative study was viewed by respondents as relevant to consumers, easy to understand, and formulated in consumer language.

The affirmative binary format is better known as the pick any/n format. Respondents were given a list of attributes and asked to select those that applied to a given brand. If they did not wish to assign an attribute to a brand then they were asked not to select the attribute. The full binary format version of the questions required respondents to state whether or not they believed that each of the listed attributes applied to any given brand. As with the affirmative binary format version, the information available in the data set was binary, but respondents had to think about every single brand-attribute combination. Presumably, this approach should lead to a greater number of positive association responses than the affirmative binary format, which allows respondents to easily evade a response by indicating non-association (e.g., if they are fatigued or not motivated in the first place). The versions that used the Likert five verbal and Likert five endpoints answer formats (Likert, 1932) offered respondents five answer options including a neutral middle point. For the Likert five verbal version, all answer options came with a verbal description; whereas for the Likert five endpoints version, only the endpoints had a verbal description (e.g., *strongly agree* and *strongly disagree*). The Likert four verbal answer format version was the same as the Likert five verbal one, except that it had no middle point. This is not the answer format that Likert (1932) originally recommended, but is a variation thereof, used here to assess respondents' changes in response when no middle point is available.

The unipolar four verbal answer format version offered respondents four answer options, all of which came with a verbal description. Respondents were asked to evaluate which option an attribute applied to, with options ranging from, for example, *not at all* to *extremely*. The bipolar seven verbal answer format version offered respondents choices ranging between the positive and negative extremes of the attribute under study: three to the right and three to the left of the neutral midpoint. Respondents were asked to state which of the seven labeled options applied — either the neutral, one of the three degrees of positive, or one of the three degrees of negative. The bipolar seven endpoints answer format version was identical to the bipolar seven verbal answer format, except that only the endpoints have labels. The bipolar six verbal and bipolar six endpoints answer formats were the same as the bipolar seven-point answer formats, but without a middle point. Fig. 1 includes examples of all answer formats.

Table 1 shows the experimental design and the sample sizes for all answer formats included in the study (total n = 2609). Some conditions included two measurements using identical answer formats to enable the calculation of base level instability (control groups); while others exposed respondents to two different answer formats to enable the translation of responses. Respondents were randomly assigned to one experimental condition.

## 3. Results

### 3.1. Translations from the full binary answer format

The first analysis gives translations from the full binary answer format to all other verbally labeled answer formats included in the experiment. The full binary answer format is translated onto itself using control group data (row 1 of Table 1) in order to assess the base instability level. Fig. 2 shows the results: the top row indicates how respondents who used a *yes* answer in the first measurement (on a full binary answer format) responded in the second measurement; and the second measurement was either full binary (column 1), affirmative binary (column 2), Likert four verbal (column 3), Likert five verbal (column 4), unipolar four verbal (column 5), bipolar six verbal (column 6), or bipolar seven verbal (column 7). The bottom row shows how respondents translated the *no* responses in the first measurement on a full binary answer format onto other answer formats in the second measurement. The bar heights in the figure indicate the percentage of answers for each answer option.

For example, as reported in column 3 of Fig. 2, only 22% of respondents who said *yes* in the first measurement selected *strongly agree* in the second measurement using the Likert four verbal answer format; while 68% selected *agree*. Of those who said *no* in the first measurement, 16% selected *strongly disagree* and 59% selected *disagree*.

Affirmative binary (pick any/n):

McDonald's is yummy □

Full binary:

McDonald's is yummy □Yes □No

Likert 5 verbal: McDonald's is

| | Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree |
|---|---|---|---|---|---|
| Yummy | □ | □ | □ | □ | □ |

Likert 5 endpoints: McDonald's is

| | Strongly agree (+2) | +1 | 0 | -1 | Strongly disagree (-2) |
|---|---|---|---|---|---|
| Yummy | □ | □ | □ | □ | □ |

Likert 4 verbal: McDonald's is

| | Strongly agree | Agree | Disagree | Strongly disagree |
|---|---|---|---|---|
| Yummy | □ | □ | □ | □ |

Unipolar 4 verbal: McDonald's is

| | Not at all | Slightly | Quite | Extremely |
|---|---|---|---|---|
| Yummy | □ | □ | □ | □ |

Bipolar 7 verbal: McDonald's is

| | Very | Rather | Slightly | Neither / nor | Slightly | Rather | Very | |
|---|---|---|---|---|---|---|---|---|
| Yummy | □ | □ | □ | □ | □ | □ | □ | Yuk |

Bipolar 7 endpoints: McDonald's is

| | Very +3 | +2 | +1 | 0 | -1 | -2 | Very -3 | |
|---|---|---|---|---|---|---|---|---|
| Yummy | □ | □ | □ | □ | □ | □ | □ | Yuk |

Bipolar 6 verbal: McDonald's is

| | Very | Rather | Slightly | Slightly | Rather | Very | |
|---|---|---|---|---|---|---|---|
| Yummy | □ | □ | □ | □ | □ | □ | Yuk |

**Fig. 1.** Answer format examples (note: only one item is provided in the example).

**Table 1**
Experimental design and sample sizes.

| First measurement | Second measurement | Sample size |
|---|---|---|
| Full binary | Full binary | 203 |
| Full binary | Affirmative binary | 101 |
| Full binary | Likert 5 verbal | 95 |
| Full binary | Likert 4 verbal | 101 |
| Full binary | Bipolar 7 verbal | 99 |
| Full binary | Bipolar 6 verbal | 94 |
| Full binary | Unipolar 4 verbal | 83 |
| Likert 4 verbal | Likert 5 verbal | 103 |
| Bipolar 6 verbal | Bipolar 7 verbal | 100 |
| Likert 4 verbal | Bipolar 6 verbal | 101 |
| Likert 4 verbal | Likert 4 verbal | 208 |
| Bipolar 6 verbal | Bipolar 6 verbal | 202 |
| Likert 5 verbal | Bipolar 7 verbal | 95 |
| Likert 5 endpoints | Likert 5 verbal | 101 |
| Likert 5 endpoints | Likert 5 endpoints | 206 |
| Likert 5 verbal | Likert 5 verbal | 207 |
| Bipolar 7 endpoints | Bipolar 7 verbal | 103 |
| Bipolar 7 endpoints | Bipolar 7 endpoints | 203 |
| Bipolar 7 verbal | Bipolar 7 verbal | 204 |

Fig. 2 supports the following conclusions from the translations.

### 3.1.1. Base level instability

The base level instability for the full binary format version of the questions was approximately 15%. This percentage reflects the proportion of respondents who changed answers between two consecutive measurements where the answer format was identical (see column 1).

### 3.1.2. Asymmetrical use of affirmative binary

Respondents used the affirmative binary answer format version of the questions asymmetrically. They tended to tick *yes* less often if they did not have to choose between a *yes* and a *no* option, compared to when they had to choose (as in the full binary format). As illustrated in the empirical map in column 2, only 63% of *yes* answers on the full binary answer format retained *yes* answers on an affirmative binary answer format; this represents a discrepancy much higher than the base level instability of 14%.

On the other hand, 92% of *no* answers remained *no* answers. For the practitioner, this suggests that *yes* is a stronger statement on an affirmative binary answer format than on a full binary answer format; and that a *no* answer may not necessarily be interpreted as a negative, but instead may capture respondents' evasive behavior.

### 3.1.3. Likert four verbal captures yes answers

The Likert four verbal answer format captured *yes* responses on the full binary answer format very well. The total of *strongly agree* (22%) and *agree* (68%) responses was almost identical to the *yes* responses on the full binary answer format. Deviation is well in the range of base instability (the sum of the 2% and 8% totals are only slightly smaller than 14%). The results are similar for a *no* response in the first measurement. The translation from full binary format to the Likert four verbal format is quite consistent, making practical comparisons of results reported on these answer formats relatively uncomplicated. Interestingly, however, the majority of *yes* answers translate to the more conservative *agree* option, not to *strongly agree*.

### 3.1.4. Middle points hamper translation

The introduction of a middle point in the Likert five verbal answer format makes translating results from full binary to Likert five verbal answer formats less straightforward. As seen in column 4, 21% of *yes* responses and 36% of *no* responses shift to the *neither agree nor disagree* option. Consequently, only 73% of original *yes* respondents remain positive on the Likert five verbal format (the sum of 55% and 18%); while only 52% of the original *no* respondents remain negative (the sum of 14% and 38%). This means that empirical results derived from a Likert five verbal answer format tend to underreport agreement in comparison to both full binary and Likert four verbal results.

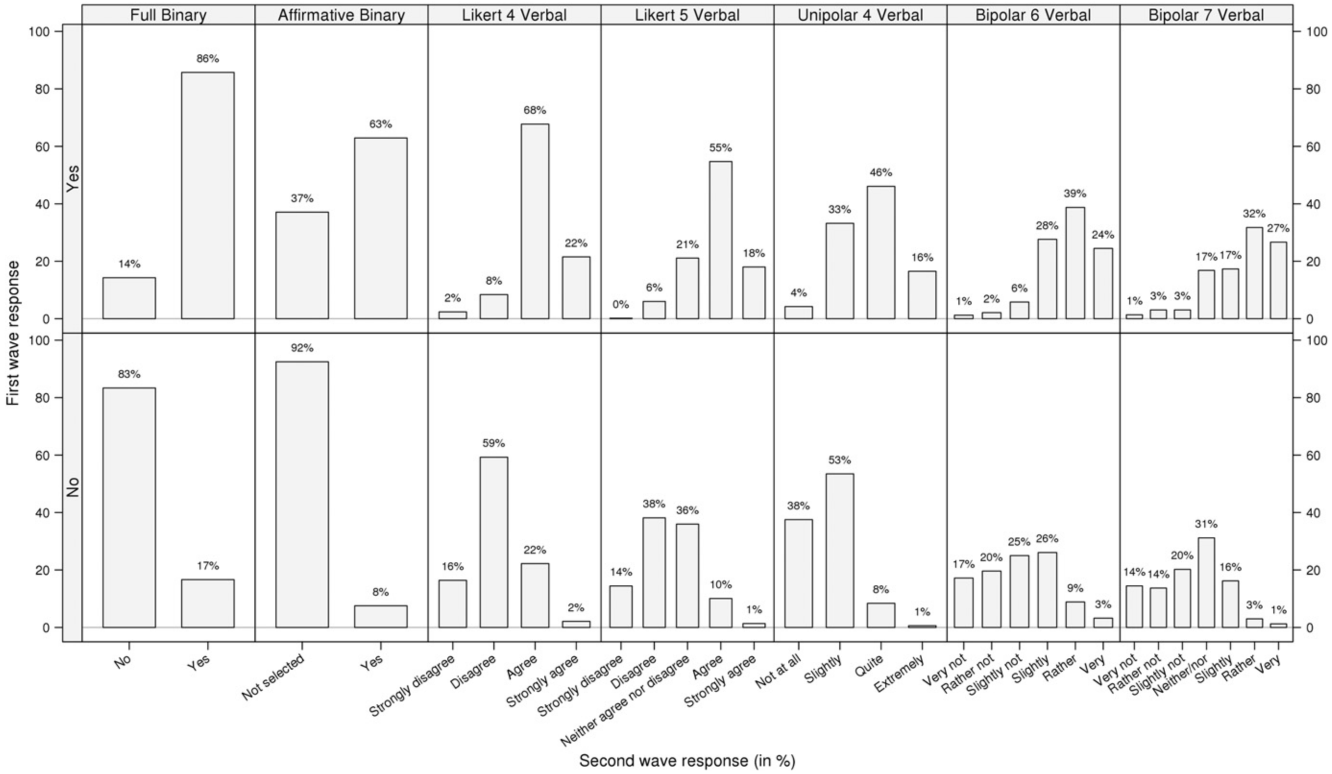

**Fig. 2.** Translation from a full binary onto different scales.

The translation of full binary responses to the unipolar four verbal answer format (column 5) indicates that people are able to validly translate positive responses (the three positive answer options of the unipolar four verbal answer format — namely *slightly*, *quite*, and *extremely* — captured 96% of the original *yes* answers). However, this is not the case for the negative responses: 53% of *no* responses moved to *slightly*, and only 38% selected *not at all*. This means that — at least in the context of brand image measurement — the unipolar four verbal answer format is strongly biased toward positive responses.

Upon translation of the full binary responses to a bipolar six verbal answer format, the positive agreement is relatively high. Of those who answer *yes* on the full binary answer format, 91% also selected one of the three positive answer options on the bipolar six verbal format. The agreement of negative responses is not as high: only 62% of people who responded with a *no* on the full binary format respond with one of the three negative options provided by the bipolar six verbal format. The practical implication is that results from bipolar six verbal answer formats are likely to have a positive bias as opposed to simple *yes*/*no* formats.

Finally, the translation of the full binary responses onto the bipolar seven verbal format, which contains a middle point, led to conclusions similar to the Likert five verbal format. The *neither*/*nor* option attracted a substantial amount of responses, reducing the positive agreement to 76% (17% who responded that the attribute 'slightly' applied to the brand, plus 32% who responded that the attribute 'rather' applied, and 27% who responded that the attribute applied 'very much'), and reducing the negative agreement to only 48%.

Table 2 summarizes the positive, negative and total agreement between the answer formats that respondents mapped against one another in the first study.

In sum, these results indicate that quite substantial deviations in responses occur, depending on the answer format offered in a survey. The translations that this study report also uncover some systematic deviations. Affirmative binary answer formats are prone to evasion, and therefore must always be expected to lead to lower agreement levels than forced binary answer options. For all other answer formats, positive agreement tends to be higher than negative agreement; and answer formats with midpoints deflect positive and negative responses toward the neutral middle point. The empirical translations in Fig. 2 may guide the comparison of results from empirical studies using different answer formats.

### 3.2. Translations from answer formats without a middle point to answer formats with a middle point

We compared the Likert four verbal containing no midpoint and the Likert five verbal containing a midpoint (Fig. 3 shows the translations on the left), and the bipolar six verbal containing no midpoint and the bipolar seven verbal containing a midpoint (on the right).

The following key conclusions follow from these translations.

The translations of the Likert four verbal format containing no midpoint, to the Likert five verbal format with a midpoint, indicate that *strongly agree* responses are seldom redirected to the *neither agree nor disagree* option; although only 52% of respondents repeatedly selected the *strongly agree* option. For all other original answer options, the switch to the midpoint option is quite substantial: 27% moved from the agree option to the midpoint, 42% moved from the disagree option to the midpoint and, most surprisingly, 18% moved from the strongly disagree option to the midpoint. The practical conclusion from these results is that including a midpoint offers a convenient answer option to respondents who do not strongly agree with a brand attribute association. The high proportion of *strongly disagree* responses redirected to the midpoint demonstrates the implausibility of suggesting that respondents who are genuinely unsure of an answer randomly choose any other option when no midpoint is available. Based on these results, omitting the midpoint option appears preferable — at least in the brand image measurement context — if the choice offered is between four or five-point answer scales.

The translation of the bipolar six verbal format with no midpoint against the version with a midpoint, shows a different picture: only a few of the respondents switched from the extremes to the middle (6% for *very* and 2% for *very not*). The movement from both slightly options (negative and positive) is symmetric, with approximately one-third switching to the midpoint option. Asymmetry is only evident in the original *rather* responses, where only 10% switched to the midpoint on the positive side; whereas 18% did so on the negative side. The substantial overall movement to the middle option (20%) apparently makes answer formats with midpoint options — particularly if they are longer scales — unattractive in the brand image measurement context.

**Table 2**
Percentage of repeat answers for positive and negative associations, and the aggregate results.

| | Positive (percent) | Negative (percent) | Both (percent) |
|---|---|---|---|
| Full binary | 86 | 83 | 85 |
| Affirmative Binary | 63 | 92 | 75 |
| Likert 4 verbal | 89 | 76 | 84 |
| Likert 5 verbal | 73 | 53 | 65 |
| Unipolar 4 verbal | 96 | 38 | 72 |
| Bipolar 6 verbal | 91 | 62 | 79 |
| Bipolar 7 verbal | 76 | 48 | 65 |

### 3.3. Translations from Likert-type answer formats to bipolar scales

Fig. 4 illustrates the translations of Likert four verbal against bipolar six verbal (on the left), and Likert five verbal against bipolar seven verbal (on the right). In order to interpret Fig. 4 correctly, we calculated the base level instability for each of the answer formats. As for the full binary translations, the base level instability indicates the percentage of respondents who did not select the same answer option twice in a row when presented with the same answer format.

The base instabilities are at 29% for Likert four verbal, 35% for Likert five verbal, 52% for bipolar six verbal, and 53% for bipolar seven verbal. Even if we take into account apparent stability due to random guessing (as indicated by Schmittlein, 1984), base instability grows with the number of answer options offered; therefore, full binary formats offer the highest level of stability over all other formats. These differences in stability themselves have major practical implications. While most users of multi-category answer formats argue that they want more than two answer options to capture finer levels of agreement, the price for this precision is low reliability, which raises fundamental questions regarding the validity of multi-category measures.

The following key insights result from this analysis. The translation from Likert four verbal to bipolar six verbal answer options is generally quite consistent with expectations: respondents divided responses within the two most extreme options in the four-point answer format version into the four most extreme options (two positive and two negative). In the case of negative responses, the two most extreme negative options contain 74% of all original *strongly disagree* responses, and in the case of positive responses, the two most extreme positive options contain 84% of the original *strongly agree* responses. The same effect occurs for the two middle options of the four-point answer format. The only surprising translation result is that 29% of those who originally selected *disagree* in the four-point format selected *slightly* on the positive side in the six-point format, thus effectively switching from a negative to a positive brand attribute association.

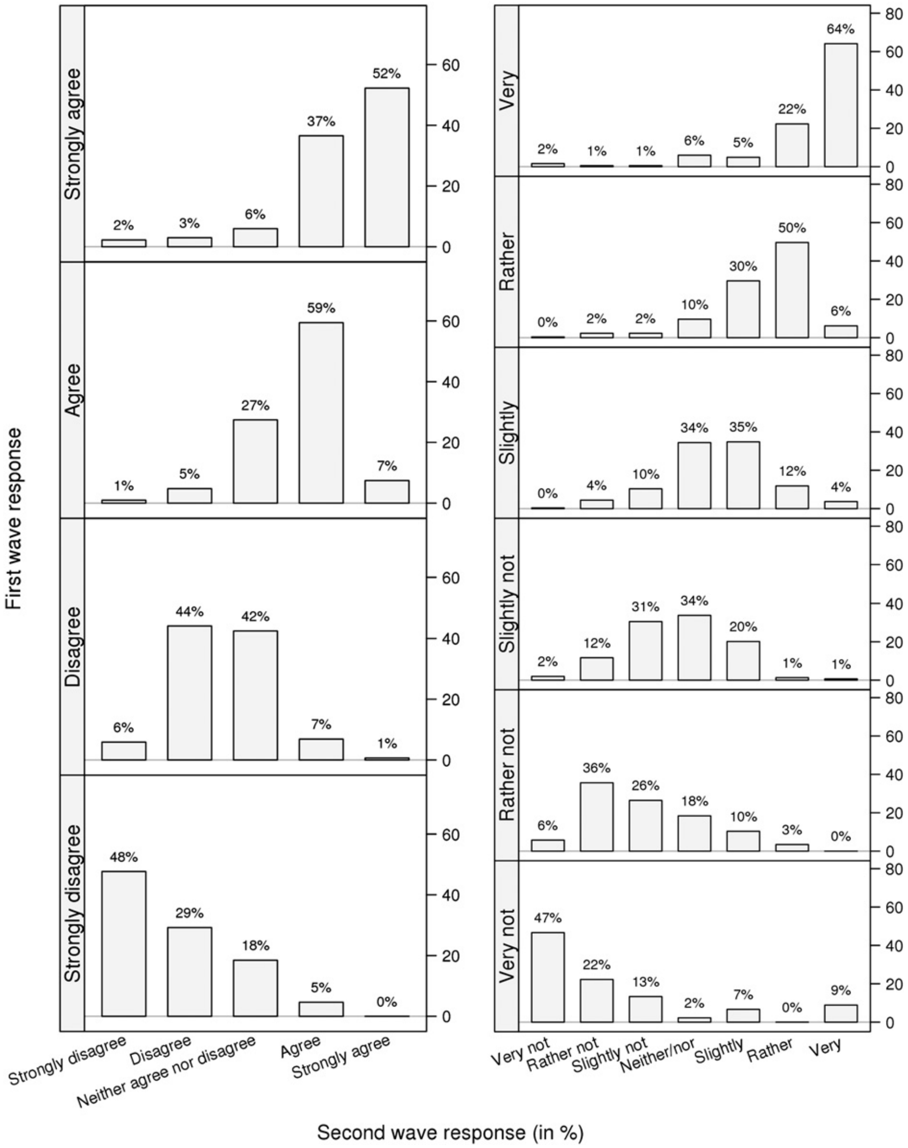

**Fig. 3.** Translation from the Likert 4 verbal to the Likert 5 verbal (on the left). Translation from the bipolar 6 verbal to the bipolar 7 verbal (on the right).

The translation from Likert five verbal to bipolar seven verbal answer options led to similar results: the extreme options in the seven-point answer format capture 92% of the original *strongly agree* responses and 79% of the original *strongly disagree* responses. Switching over to the positive side occurs again: 16% of *disagree* responses moved to the *slightly* positive option. In addition, a substantial movement occurred with respect to the original *neither agree nor disagree* response.

### *3.4. Translations from versions with endpoint labeled to fully labeled answer options*

We mapped Likert five verbal and bipolar seven verbal against Likert five endpoints and bipolar seven endpoints, respectively, and thereby calculated base level instability at 35% for Likert five verbal, 53% for bipolar seven verbal, 46% for Likert five endpoints and 52% for bipolar seven endpoints.

The present study includes an analysis of the number of endpoint responses. If only the endpoints have verbal labels, and if verbal labeling acts as a pointer for respondents, then we might expect that more respondents should select endpoints. This assumption is supported empirically: in this study, only 20% used the endpoints for Likert five verbal, compared to 27% for the Likert five endpoints version ($\chi^2 = 69$, df = 1, p-value < 0.001); and only 19% used the endpoints for the bipolar seven verbal answer format, compared to 21% for the bipolar seven endpoints version ($\chi^2 = 7.5$, df = 1, p-value = 0.006). These differences are significant for both answer formats.

Fig. 5 shows the resulting translations. Overall, the switching behavior from a fully verbalized answer format to an endpoint labeled answer format amounts to 42% for five-point formats and 54% for seven-point formats. These results indicate that the level of switching between the seven-point formats is practically identical to the level of switching that occurs when respondents face the same answer formats twice (the test of proportions for the two base instability levels and the switching rate indicates that they are not statistically significant, with $\chi^2 = 1.5$, df = 2, p = 0.477).

The following key insights are gained from these translations:

Approximately one-third of the respondents whom we first presented with a Likert five endpoint format and later with a Likert five verbal format moved from strongly agree and strongly disagree to agree and disagree, respectively (Fig. 5 on the left).

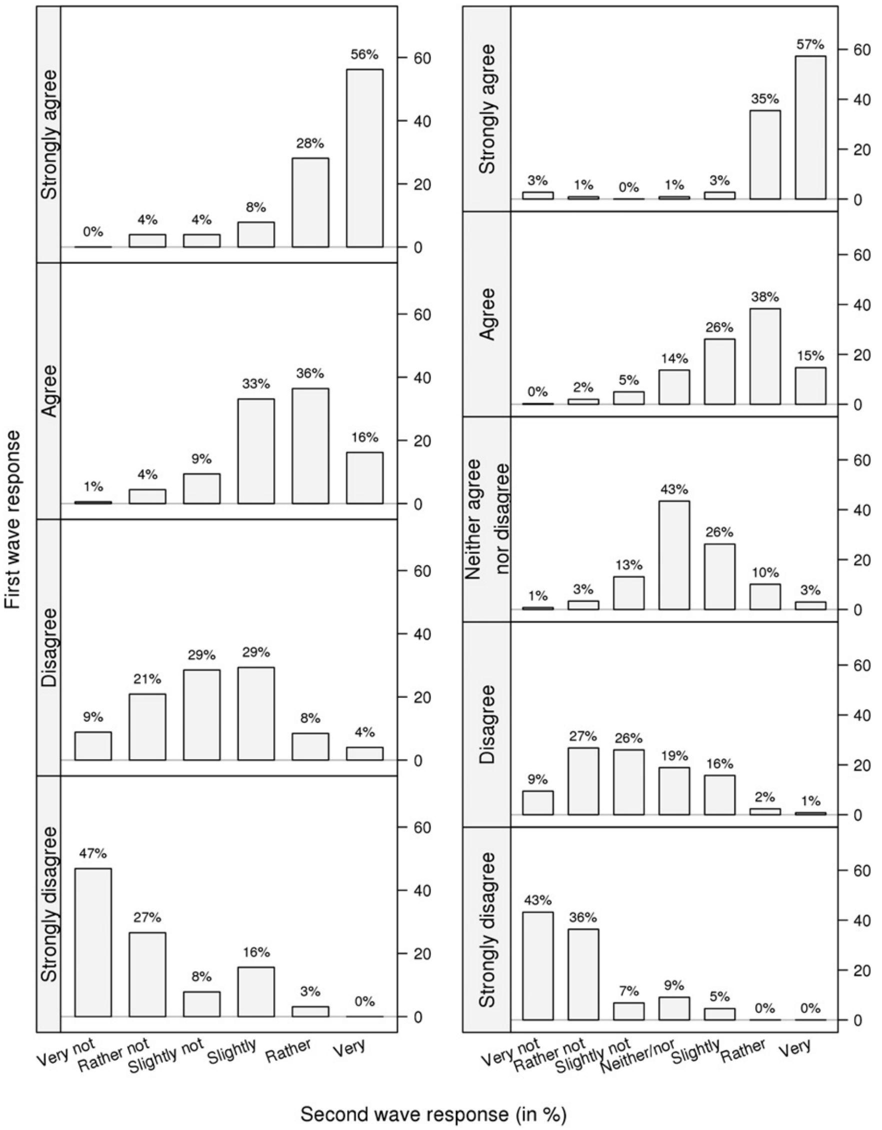

**Fig. 4.** Translation from the Likert 4 verbal to the bipolar 6 verbal (on the left). Translation from the Likert 5 verbal to the bipolar 7 verbal (on the right).

However, among the respondents who originally selected agree or disagree, only a few moved to strongly agree (8%) or strongly disagree (13%). These results provide empirical support for the previously expressed assumption that endpoint labeled formats stimulate extreme responses.

Fig. 5 on the right shows the translation from bipolar seven endpoints to bipolar seven verbal. The tendency remains the same as described for Likert five; the only difference being that the level of switching was generally higher — a finding in line with this answer format's higher base instability rate.

## 4. Conclusions

The aim of this study is to provide empirical translations of different survey answer formats to facilitate the comparison of findings across studies. Several fundamental behaviors related to answer formats are observed through the experiment conducted with a total of 2609 respondents.

The full binary answer format has a very low level of base instability (14%) compared to answer formats with higher numbers of answer options (29% for Likert four verbal, 35% for Likert five verbal, 46% for Likert five endpoints, 52% for bipolar six verbal, 53% for bipolar seven verbal, and 52% for bipolar seven endpoints). Consequently, the switching patterns observed in answer formats with a higher number of answer options are more difficult to interpret because the instability of responses and switching are heavily confounded. This is a relevant finding that questions the validity of multi-category formats for surveys that measure brand image.

Two design features of answer formats appear to reduce the general level of agreement. First, the non-forced nature of an answer format, which the finding that the affirmative binary format leads to systematically lower agreement levels than all other answer formats tested, illustrates. Second, the inclusion of a neutral midpoint, which appears to stimulate evasion behavior. One design factor leads to an increase in agreement level: the unipolar answer format. In theory, researchers should only use unipolar formats if the construct under study (or the attribute in a brand image investigation) is in fact unipolar. However, this is not always the case in empirical studies; hence, market researchers should be aware that unipolar answer formats that offer multiple agreement options, but only one disagreement option, will generally increase the stated level of agreement. Finally, a substantial increase in extreme responses occurs if only the endpoints of an answer format have labels.

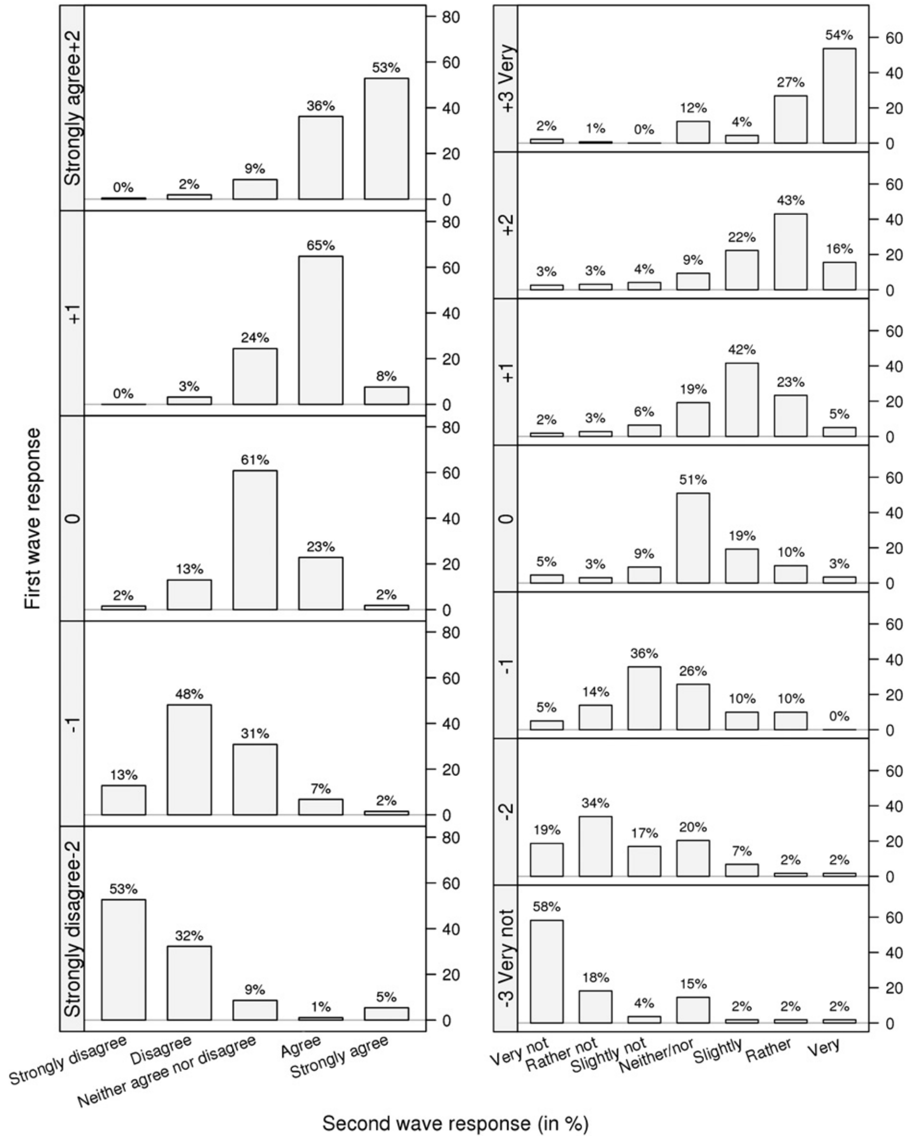

**Fig. 5.** Translation from Likert 5 endpoints to Likert 5 verbal (on the left). Translation from bipolar 7 endpoints to bipolar 7 verbal (on the right).

The primary contribution of these findings is the knowledge regarding the effects of answer format choice in empirical marketing research; and they provide strategies for comparing the results of different answer formats to each other. The secondary contribution is to increase our understanding about what behaviors related to answer formats have implications for researchers when selecting answer formats for survey research. For example, commonly used seven-point multi-category answer formats (as recommended by Cox, 1980) suffer from a very high base level instability, and, rather than providing a more detailed response, may actually capture a lot more noise, thus making the measurement less valid overall than a simple full binary answer format.

The conclusions from this study cannot be generalized beyond the context of brand image measurement, but we expect that replication studies in other contexts will find that the same base tendencies apply for each investigated answer format. All translations in the present study are based on one particular order of exposure for the two answer formats under study. Future studies should consider randomizing the order of exposure. Furthermore, all translations assume homogeneity among respondents. However, various sub-segments of respondents who use different translation functions may in fact be present.

## References

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

Boote, A. S. (1981). Reliability testing of psychographic scales: Five-point or seven-point? Anchored or labeled? *Journal of Advertising Research*, *21*, 53–60.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology*, *23*, 323–331.

Cicchetti, D. V., Showalter, D., & Tyrer, P. (1985). The effect of number of rating scale categories upon levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, *9*, 31–46.

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

Dolnicar, S., & Grün, B. (2007a). 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. (2007b). Question stability in brand image measurement — Comparing alternative answer formats and accounting for heterogeneity in descriptive models. *Australasian Marketing Journal*, *15*(2), 26–41.

Dolnicar, S., & Grün, B. (2009). Does one size fit all? The suitability of answer formats for different constructs measured. *Australasian Marketing Journal*, *17*(1), 58–64.

Dolnicar, S., Grün, B., & Leisch, F. (2011). Quick, simple and reliable: Forced binary survey questions. *International Journal of Market Research*, *53*(2), 231–252.

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.

Dolnicar, S., & Schäfer, A. I. (2006). Public perception of desalinated versus recycled water in Australia. *Proceedings of the 2006 AWWA Desalination Symposium (CD)*.

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

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review*, *67*(6), 343–352.

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

Haley, R. I., & Case, P. B. (1979). Testing thirteen attitude scales for agreement and brand discrimination. *Journal of Marketing*, *43*(4), 20–32.

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

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309.

Hurlimann, A. (2006). Melbourne office worker attitudes to recycled water use. *Water (Journal of the Australian Water Association)*, *33*(7), 58–65.

Jenkins, G. D. J., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*, *62*(4), 392–398.

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

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, *60*(1), 10–13.

Martin, W. S. M., 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. (1971a). Communication and research notes. *Journal of Marketing Research*, *8*(4), 495–500.

Matell, M. S., & Jacoby, J. (1971b). Is there an optimal number of alternatives for Likert scale items? Study 1: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657–674.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

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.

Remmers, H. H., & Ewart, E. (1941). Reliability of multiple-choice measuring instruments as a function of the Spearman–Brown prophecy formula. *Journal of Educational Psychology*, *32*(1), 61–66.

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

Rossiter, J. R. (2011). *Measurement for the social sciences — The C-OAR-SE method and why it must replace psychometrics.* New York: Springer.

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.

Schmittlein, D. C. (1984). Assessing validity and test–retest reliability for "Pick K of N" data. *Marketing Science*, *3*(1), 23–40.

Schutz, H. G., & Rucker, M. H. (1975). A comparison of variable configurations across scale lengths: An empirical study. *Educational and Psychological Measurement*, *35*, 319–324.

Symonds, P. M. (1924). On the loss of reliability in rating due to coarseness of the scale. *Journal of Experimental Psychology*, *7*, 456–461.