On the optimal number of scale points in graded paired comparisons

De Beuckelaer, Alain ; Toonen, Stef ; Davidov, Eldad

DOI: https://doi.org/10.1007/s11135-012-9695-2

On the optimal number of scale points in graded paired comparisons

Alain De Beuckelaer[1], Stef Toonen[2], Eldad Davidov[3]

[1]Ghent University, Ghent, Belgium / Renmin University of China, P.R. China / Radboud University Nijmegen, The Netherlands – [2]Radboud University Nijmegen, The Netherlands – [3]University of Zurich, Zurich, Switzerland

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in the journal **Quality & Quantity** *following peer review. It was first published online in this journal on March 23, 2012. The definitive publisher-authenticated version is available online at:*

http://www.springerlink.com/content/f318162764233013/

or under

doi:10.1007/s11135-012-9695-2

RESEARCH NOTE

On the optimal number of scale points in graded paired comparisons

Abstract In market research, it is common practice to measure individuals' brand or product preference through graded paired comparisons (GPCs). One important decision concerns the (odd) number of scale points (e.g., five, seven, nine, or eleven) that has to be assigned to either brands or products in each pair. Using data from an experiment with 122 students, we assessed the extent to which GPCs with a higher number of scale points (requiring more cognitive effort) really outperform GPCs with a smaller number of scale points (requiring less cognitive effort). Our data analyses have shown that one may reduce the (odd) number of scale points from eleven to nine or seven, depending on what minor compromises one is willing to make. The detailed psychometric results presented in this paper are useful to applied researchers as they help them in making well-informed decisions on the number of scale points in a GPC task.

Keywords Graded / ordered / constant sum paired comparisons - Preference measurement - Measurement validity

1. Introduction

The method of paired comparison is one of the pillars of market and social research (Dawes 2007; Peterson, Brown, McCollum, Bell, Birjulin and Clarke 1996). It enables making multiple comparative judgments as to which of two choice alternatives (e.g., brands, products) is preferred by an individual consumer. The simplified paired comparison, as introduced by the psychologist Thurstone (1927), instructs individuals to choose between two choice alternatives by distributing one point. For example, an individual may indicate that product A is preferred (=1) over product B (=0) in a given pair. As indicated by elementary combinatorics, the number of paired comparisons in a paired comparison task can be calculated as follows: for k choice alternatives the number of paired comparisons each participant has to make is $\binom{k}{2}$. So, for five choice alternatives the number of paired comparisons to be made equals $\frac{5!}{2!\,3!} = 10$.

A more fine-grained assessment of consumers' preferences relies on a *graded* paired comparison (GPC, also referred to as *constant sum* paired comparison; see Day 1965; Netzer and Srinivasan 2011; Netzer, Toubia, Bradlow, Dahan, Evgeniou, Feinberg, Feit, Hui, Johnson, Liechty, Orlin and Rao 2008; Oishi, Hahn, Schimmack, Radhakrishan, Dzokoto and Ahadi 2005; Scholz, Meissner and Decker, 2010). In a GPC, individual consumers are instructed to rate the direction and intensity of their preference by dividing a particular sum of points among the two alternatives in every pair. A constant sum equal to an odd number, for instance seven, results in a forced choice as seven cannot be subdivided in two equal (whole) numbers. Forcing individuals to make a choice may be reasonable as long as there are substantial differences in preference among the choice alternatives (see, for instance, Friedman and Amoo 1999). However, it is not clear what the optimal odd number of points to be divided between the two alternatives is[1]. In addition, when individuals do not prefer one choice alternative over the other in a given pair, and are forced to assign a larger number to (at best, an arbitrarily chosen) choice alternative, GPC data are likely to suffer from logically

[1] To date, no consensus exists as to what the optimal number of scale points is (Janhunen in press; Moors 2008). Some studies have suggested the optimal number of scale points to be seven: see Hofmans et al. 2007; Lozano et al. 2008; Weng 2004; few other studies suggested the optimal number of scale points to be higher than seven (see Alwin 1997; Scholz et al. 2010), whereas other studies have argued that using five scale points is just as good as using seven or more scale points (see Böcker 1988; Churchill and Peter 1984).

incoherent judgments, leading to so-called *transitivity errors* (De Beuckelaer, Kampen and Van Trijp in press; Vershuren and Arts 2004).

To illustrate the concept of a transitivity error, assume that an individual has no preference for either choice alternative A and B, but assigns the larger number of preference points (PP) to choice alternative B, for instance PP(B)=4 and PP(A)=3; in numerical terms: PP(B) > PP(A). The same individual also assigns the larger number to choice alternative C [PP(C)=4] when compared to choice alternative B [PP(B)=3], but is also in reality indifferent when choosing between C and B. In this case, it is plausible that when choice alternative A has to be evaluated against choice alternative C, choice alternative A will receive the highest number of points [e.g., PP(A)=4]. Obviously, this assignment is logically incoherent with the assignment of points in the other two pairs [i.e., the inequality PP(A) > PP(C) does not match with the two inequalities PP(B) > PP(A) and PP(C) > PP(B)]. To be logically consistent, choice alternative B should have received the larger number of points.

In the case of k different stimuli (corresponding to k different brands that are compared), a total of $\binom{k}{3}$ different transitivity errors can be made. As a consequence, making paired comparisons may become very complex, especially when the number of stimuli to be compared is large. The reader must realize that transitivity errors may also occur when respondents are *not* indifferent in terms of brand preference. Unfortunately, individuals' preference structures are not always logically consistent, and – even if an individual has a coherent preference structure – he/she may still provide preference data which are not completely in line with his/her preference structures.

A recent study (De Beuckelaer et al. in press), which examined the cross-national measurement validity of GPCs in a 14-country consumer study, showed that 27.8% of the respondents made at least one transitivity error. Nation-specific percentages in the study ranged from 10.2% in The Netherlands to 44.4% in the United Kingdom. This study also showed that transitivity errors were not related to the respondents' background statistics, such as gender, age, working status, social class, but do depend on the nation to which the respondent belongs. Obviously, the results obtained in this study may be partly an artifact of key characteristics of the research design, such as the number of choice alternatives (i.e., 5) and the total number of points to be assigned in each pair, namely, eleven[2].

[2] As a side note, we would like to mention that graded comparisons using eleven scale points have extensively been used in various tools for "market pretesting" which were designed

Eleven as the number of preference points to be assigned in each pair may be considered a high number[2] if one takes into account that relevant methodological studies have typically considered a scale length ranging between five and eleven (i.e., [5, 11]) (see, e.g., Alwin 1997; Bech et al. 2006; Bendig 1954; Böcker 1988; Churchill and Peter 1984; Friedman and Amoo 1999; Hofmans, Theuns and Mairesse 2007; Komorita and Graham 1965; Lozano, García-Cueto and Muniz 2008; McKelvie 1978; Preston and Colman 2000; Roth, Schroeder, Huang and Kristal 2008; Weng 2004). The high number eleven may be beneficial in that it allows (at least potentially) a high level of precision (when indicating one's preference), but – at the same time - it may put too high a cognitive burden on respondents, especially if (virtually) the same level of precision (accuracy) can be achieved with a lower number of scale points, such as nine, seven, or five (Moors 2008). As the methodological literature discussing effects of the number of scale points (in Likert-type rating scales; see works referred to above) is not conclusive as to what number of scale points is optimal in terms of precision, we wonder whether an (odd) number of scale points smaller than eleven may provide comparable levels of precision. If this were the case, GPCs with a reduced number of scale points may be preferable as they do not require the respondent to choose between that many combinations (e.g., eleven vs. zero; ten vs. one; nine vs. two; etc.), some of which do not increase precision of measurement. As such, one may save precious research time and decrease the probability of measurement error and intransitive choices because of the reduced computational burden on the interviewees. Based on data from an experiment, this study examines levels of precision and transitivity errors produced by all (odd) numbers of scale points situated in the interval [5, 11].

2. Method

To study the differential impact of a varying number of scale points, some key factors were controlled (i.e., fixed) in this study. The factors under control comprised: the nature of the choice alternatives (i.e., brands of bottled mineral water); the mode of data collection, namely an electronic survey (e-survey); the layout of the e-survey containing the paired comparison task; and the instructions preceding the paired comparison task. The initial sample used in this

---

using principles of preference scoring as introduced in the ASSESSOR model (Silk and Urban 1978; Urban 1993; Urban & Katz, 1983; Netzer and Srinivasan 2011).

research consisted of 139 students enrolled at a Dutch university. More detailed information on the choice alternatives and the sample used are provided in the following paragraphs.

2.1 Choice alternatives

In the experiment included in the e-survey, students were asked to make a forced choice between different 0.5-liter bottles of branded water (which were all offered at the same price). Bottled water is considered to be a low involvement product, meaning that consumers, most typically, do not spend a lot of time and effort to choose between alternatives. By working with a low involvement product rather than a high involvement product (e.g., a car), we can expect the probability of the individual being indifferent when choosing between multiple choice alternatives to be relatively high. As a consequence, the probability of making transitivity errors may be relatively high as well (see illustrative example provided in the Introduction). In total, five different brands of bottled water were included in this study: "Spa" (brand A), "Sourcy" (brand B), "Chaudfontaine" (brand C), "Evian" (brand D), and "Vittel" (brand E). Among the (manufacturer-owned) brands of bottled water available on the Dutch market, these five brands were found to score highest on brand awareness in a small-scale market research [N=30] among students which we conducted prior to this experiment (i.e., brand awareness scores were: 100% for Spa; 82% for Sourcy; 75% for Evian; 62% for Chaudfontaine; and 48% for Vittel). Prior to collecting paired comparison data in our study we asked respondents to check a box for each of the five brands they were familiar with. Of all respondents who provided data for final analysis (N=122), 97.5% were familiar with "Spa" (brand A); 96.7% were familiar with "Evian" (brand D); 92.6% with "Sourcy" (brand B); 91.0% with "Chaudfontaine" (brand C), and 87.7% with "Vittel" (brand E).

In every paired comparison to be made, the labels of five brands (e.g., "Spa") were depicted together with a picture of the packaging of the corresponding 0.5 liter bottle sold in stores. The instructions given to the respondents were formulated as follows: "Please consider all presented bottled water brands equal in price and content (0.5 L). Which brand do you prefer? Please distribute x preference point/points." (with x being replaced by one, five, seven, nine, or eleven).

2.2 Sample

One hundred thirty-nine students participated in this experiment. They were all enrolled as an undergraduate or graduate student in one of the universities in The Netherlands. We chose to work with students because of two main reasons. First, the population of students in the Netherlands is known to be relatively homogeneous in terms of sample composition (e.g., as far as age, income, stage in the life cycle, and consumption behavior is concerned) (see Van de Vijver and Leung 1997, p. 30). Second, by using a relatively homogeneous sample, the risk of obtaining biased experimental results (across experimental conditions with a varying number of scale points) is minimized. The mean age in the final sample (N=122) was 23 years and 7 months (SD=2 years and 2 months). Fifty percent of the participants in the final sample were female. Students were divided randomly into the four experimental conditions (five, seven, nine, or eleven scale points). Seventeen students were dropped from the analysis because they did not show up for the second round of the experiment, so that explains why the actual number of participants decreased to 122.

2.3 Experimental procedure and underlying rationale

As mentioned in the Introduction, Thurstone's approach to paired comparisons relies on indicating which choice alternative in each pair is the preferred one. Data from such a paired comparison task comprise a series of zeroes and ones; one if a 'specific choice alternative in a given pair' (i.e., representing the variable under study) is chosen, and zero if that specific choice alternative in that pair is not chosen. Given that simple (i.e., Thurstonian) paired comparisons are relatively easy to conduct (when compared to GPCs), the 'performance' (see section 'Performance measures and analysis approach' located further down in this research note) of simple paired comparisons may serve as a 'natural benchmark' against which the 'performance' of GPCs (with a varying number of scale points) is evaluated. To set this benchmark and to ensure that all students have the same 'experience' when conducting GPCs, they were all asked to first complete the simple paired comparison (simple PC) task. They were also informed that, in the near future, they would be invited again to participate in a second round to provide additional data. Participants were informed that after the second round two randomly chosen students would receive a 25 Euro gift certificate to purchase books. A unique student identifier allowed the authors to differentiate between those students who dropped out after the first round and those who continued to provide useful data for analysis in the second round.

As explained in the Introduction, ten paired comparisons are required to make a systematic comparison between the five brands of water. More specifically, the relevant list of cell entries may be taken from the upper triangular (symmetric) brand matrix, and includes the following ten cell entries: (brand A vs. brand B), (brand A vs. brand C), (brand A vs. brand D), (brand A vs. brand E), (brand B vs. brand C), (brand B vs. brand D), (brand B vs. brand E), (brand C vs. brand D), (brand C vs. brand E), and (brand D vs. brand E). The paired comparison procedures used in our study were implemented such that the order of brands in a pair was determined at random. For instance, about half of the respondents were offered the pair (brand A vs. brand B), whereas the other half of the respondents were offered the pair (brand B vs. brand A). By doing so, the effect of possible violations of the "symmetry assumption" (i.e., the assumption stating that the order in which both brands in a pair are presented does not affect the number of preference points assigned to each brand) are "leveled out". This means that none of the (five) brands will get an inflated (or deflated) number of preference points, at least not when summarizing preference data across all respondents completing a specific paired comparison task. In addition to the ten pairs offered to the respondent (with the brands in each pair listed in random order), an additional pair was offered to each respondent. The eleventh pair offered the same brands as the first pair, either in the same order or in reverse order (also determined at random). As this eleventh pair was added, it was possible to assess "consistency of preference scoring" across all respondents completing a particular paired comparison task (e.g., a GPC with five scale points). A simple (Pearson) correlation computed between the number of preference points assigned to any given brand across both (corresponding) pairs (i.e., the first and the eleventh) provides an adequate indicator of the extent to which scores have been assigned consistently in the paired comparison task.

Two weeks after the first round we invited the 139 students for the actual experiment, in which they had to repeat essentially the same paired comparison task, with the total number of preference points to be assigned in each pair fixed to either five, seven, nine, or eleven (random draw from these four possibilities). A time span of minimally two weeks between the first round (i.e., simple paired comparison task) and the second round (i.e., the GPC task) was long enough to prevent memory effects from having an impact on the data, and short enough to avoid shifts in brand preference which may occur over a (longer) period of time (see Langbroek and De Beuckelaer 2007). When necessary, we sent out multiple reminders to avoid a substantial level of dropout. A high percentage of students (87.8%; i.e., 122 out of 139) who provided data in the first round also provided data in the second round. For

respondents who did not drop out, the actual number of days between the first round and the second round was 22 days on average (*SD*=7.5). Two of the respondents who showed up for the second round were randomly selected and were asked per email to provide their private address to receive the 25 Euro gift certificate for books.

Data from such a GPC task comprise a series of (paired) variables, each of the variables in a pair representing the number of preference points given to a specific choice alternative (e.g., brand A, brand B, …, brand E) offered in a paired comparison. If participants entered the data correctly, the total sum computed over the two variables in each pair was equal to the fixed total (e.g., combination of zero / five; four / one; and three / two are possible when using a five-point scale). If participants did not enter the data correctly, the e-survey reported (if necessary, repetitively) an error, and instructed the participant to reenter the numbers reflecting his/her preferences so that the total of preference points assigned to both variables in a pair was equal to the fixed total. In other words, additivity errors (i.e., the sum of preference points given to both choice alternatives in a pair is different from the fixed total) were simply not present in the data. However, the data may contain transitivity errors. Table 1 presents a summary of the experimental design.

Insert Table 1 about here

2.4 Performance measures and analysis approach

In this study, we relied on several measures to assess the measurement validity of GPCs with varying numbers of scale points. As mentioned before, data obtained through a simple PC task served as a benchmark to assess the performance of GPCs. In the following, the measures used to perform the evaluation are presented. The results of our comparative analyses (across paired comparison approaches) are listed in detail in the Results section.

2.4.1 Difference in brand dominance between GPCs and simple PCs: NEBD and EBD[0/1]

GPCs evaluated in this study employed five, seven, nine, or eleven scale points (four conditions). In each condition, dominance in each pair is determined by the brand to which the largest number of preference points was assigned. All respondents providing GPC data (*N*=122) for one of these four conditions already provided valid data on brand dominance in each pair in a simple PC. So, two methods, namely a GPC and the simple PC, were used to

obtain brand dominance data from all respondents. As such, one may easily determine, for each respondent, the *number of pairs* (less than or equal to ten) in which the data obtained though both methods (i.e., the GPC and the simple PC) showed an inconsistency in terms of the dominant brand. This measure is indicated as the number of errors in brand dominance judgments, 'NEBD'. In addition, a second (derived) measure, referred to as the error in brand dominance judgments expressed as a dichotomy, 'EBD[0/1]', flags participants for whom NEBD exceeds zero. In other words, the data from respondents who scored 1 on EBD[0/1] contained at least one judgment error in brand dominance across all corresponding pairs included in a GPC and a simple PC.

The values computed for NEBD and EBD[0/1] may (only) be compared across different types of GPC tasks (i.e., GPC with five, seven, nine, or eleven scale points). The values for NEBD and EBD[0/1] as obtained for simple PC are, by definition, zero as the number of differences between an object (i.e., brand dominance in the simple PC) and itself is always zero (see also Table 2).

2.4.2 Intensity of brand preferences: IBP

As explained before, a simple paired comparison task does not allow the participant to score the intensity of preference for each brand within a pair. As such, the ratio between the number of times a given brand is chosen/preferred (for all five brands in our study this is a number between zero and four) and the number of times that brand was offered in a pair (always four) provides a natural benchmark tor assessing the intensity of brand preferences as indicated by a GPC task.

Instead of using data obtained through a simple paired comparison task (as was done already for the NEBD and EBD[0/1] measures) as a benchmark, we now used the data from the GPC to derive (or mimic) 'simple brand preference data'. In each *graded* pair, the brand which was assigned the larger number of preference points qualified as the preferred brand, and received a score of one. In contrast, the other, nonpreferred brand received a score of zero. Once such simple brand preference data was extracted from graded comparison data, the number of times a given brand is preferred (out of four times) was set as a benchmark to assess the intensity of preference scores assigned (to each brand) in the same GPC task.

In the GPC task, a functionally equivalent measure of brand preference (which we name BPG, brand preference in the GPC task) is calculated as the total number of preference points assigned to a brand (across all relevant pairs) divided by the maximum number of

points the brand could receive (which can be computed by multiplying the total number of preference points to be assigned in each pair – five, seven, nine, or eleven – and the number of pairs in which that particular brand was offered – four). By comparing brand preference measured by means of a GPC task with simple brand preference as computed from the same GPC task, the intensity of brand preference of a particular brand (over all other brands) was quantified. As a summary statistic, one may then average brand preference across all brands involved in the GPC task (for further discussion about the formula, see De Beuckelaer et al. in press[3]). More formally, it is stated that:

$$IBP = \left(\sum_{b=1,\ldots,K} \sqrt{(BPG_b - BPd_b)^2} \right) / K$$

with

b: an index ranging from 1 to *K* to indicate the number of brands to be graded in the paired comparison task (that is, *K*=5 in this study)

BPG_b: brand preference in the GPC task, that is, the total number of preference points assigned to brand b divided by the multiplication of the total number of preference points that it was assigned in each pair and the number of pairs in which brand b was offered

BPd_b: simple brand preference derived from the *same* GPC task, that is, the number of times brand b is chosen divided by the number of times that brand b was offered in a pair

The value of IBP may then be compared across GPC tasks with a varying number of preference points (five, seven, nine, or eleven). A larger value for IBP obtained for GPC task A in comparison to GPC task B signifies that in GPC task A brands were assigned preference points in a more intense manner than in GPC B.

2.4.3 Transitivity error indicator: NTE and TE[0/1]

[3] De Beuckelaer et al. (in press) refer, within the context of a cross-national study, to the extremity of one's responses rather than intensity of one's response.

As explained in the Introduction, transitivity errors reflect an incoherent assignment of (one or more) preference points across all pairs in a paired comparison task. In a GPC task which involves five brands (A, B, C, D, and E), one might produce up to ten possible transitivity errors (i.e., one in each set of two pairs: {(A, B), (B, C)}, {(A, B), (B, D)}, {(A, B), (B, E)}, {(A, C), (C, D)}, {(A, C), (C, E)}, {(A, D), (D, E)}, {(B, C),(C, D)}, {(B, C), (C,E)}, {(B, D), (D, E)}, and {(C, D), (D, E)}). The number of transitivity errors made (out of ten) was indicated by a measure labeled Number of Transitivity Errors, 'NTE'. In addition, a second (derived) measure, referred to as 'TE[0/1]' (transitivity error expressed as a dichotomy), flags respondents for which NTE exceeds zero. In other words, the data from respondents who scored 1 on TE[0/1] showed at least one transitivity error across the ten pairs.

Just as for all other measures presented above, it is worthwhile to compare the measures NTE and TE[0/1] across GPC tasks with a varying number of preference points.

2.4.4 Consistency in preference scoring: CON

As explained earlier in the Method section, the (Pearson) correlation between the number of preference points assigned to brand A (or brand B) across the first and the eleventh pair provides a measure of the extent to which respondents have scored consistently in the paired comparison task. This measure is labeled 'CON' and was also computed for all types of GPCs.

3. Results

The results of the analyses are summarized in Table 2.

Insert Table 2 about here

3.1 NEBD

Table 2 indicates that the number of judgment errors in brand dominance (NEBD), when compared to the reference (a simple PC), was highest for GPCs with the number of scale points equal to five.

3.2 EBD[0/1]

Table 2 also indicates a relatively high number of participants making at least one inconsistency in their judgment of the dominant brand (see EBD[0/1]) in each pair as measured in a GPC with five scale points, and in a simple PC task (i.e., the reference method). Exactly 60 percent of all respondents provided nonconsistent data for at least one paired comparison. This percentage was significantly ($p < 0.05$; binomial test of equal proportions) higher than the one observed for GPCs with a higher number of scale points (seven, nine, and eleven), where the corresponding percentages consistently dropped below 20.0%. Thus, a scale with length five is clearly disadvantageous, especially when compared to scales of a higher length. GPCs with seven, nine, or eleven scale points were found to produce brand dominance results which are largely consistent with simple PCs (EBD[0/1]=0.194, 0.152, and 0.152, respectively).

3.3 IBP

A higher intensity of brand preferences was found for the scales with five points than for simple PCs. Table 2 indicates that the 'gain' in preference intensity amounts to an increase of 0.150 preference points (on average) per brand. Table 2 also shows that by using seven scale points, one cannot further 'intensify' respondents' preference scores over and above the level of intensity that was already obtained with GPCs with five scale points (IBP=0.163 for seven scale points; the difference between 0.163 and 0.150 is not significant: p=0.258 as reported by an independent-samples t-test not assuming equal variances across groups). Similar results were obtained when using scales with nine or eleven points.

3.4 NTE

Table 2 presents also the number of transitivity errors for GPCs with various scale points. It turns out that the number of transitivity errors were highest for the seven scale points (0.355) and lowest for nine and eleven scale points.

3.5 TE[0/1]

As far as the percentage of respondents making at least one transitivity error is concerned, a somewhat lower percentage was reported for GPCs with five scale points when compared to simple PCs (i.e., 8.0% instead of 11.5%), but the difference is not statistically significant (binomial test of equal proportions results in p=0.611). GPCs with seven scale points were found to generate a somewhat higher percentage of respondents making a transitivity error (16.1%) when compared to both simple PCs and five scale points. These differences in

percentages are, however, not significant (p=0.483 and 0.361, respectively). The fraction of respondents making at least one transitivity error was lowest (3.0%) when nine or eleven scale points were used. This percentage (3.0%) was not significantly different from the corresponding percentage reported for simple PCs (i.e., 11.5%; p=0.145); it was, however, significantly different from the corresponding percentage reported for GPCs with seven scale points (i.e., 16.1%; p=0.074, p<.10).

3.6 CON

GPCs with five scale points were also found to produce rather consistent results regardless of the order in which brands are listed in a pair (a correlation of 0.853 was reported in the last column of Table 2). Seven scale points produced a lower consistency (0.756) but consistency was again higher for nine or eleven scale points (0.846 and 0.887, respectively).

3.7 Combined results

As mentioned before, we are interested in finding out whether or not GPCs with a larger number of scale points allow for a much higher gain in preference intensity. Due to a high level of inconsistency in brand dominance reported for graded comparisons with five scale points we recommend considering a higher number of scale points. When taking a closer look at the values reported in Table 2 for scales with nine and eleven points, one immediately notices that their performance is almost identical to one another, while outperforming the performance of scales with seven points in almost all aspects. The use of both nine and eleven scale points leads to: (a) a high level of consistency in brand performance (when compared to simple PCs) as reflected in the EBD[0/1] index; (b) virtually no transitivity errors; and (c) a high level of consistency in the CON index. Only preference intensity did not change substantially when increasing the number of scale points from five up to eleven (see Table 2).

Insert Tables 3a, 3b, 3c about here

To complement our univariate analyses reported in Table 2, we also performed some additional regression-type analyses. The 'count variables' NEBD and NTE were not considered as dependent variables as both the numbers of errors in brand dominance and transitivity errors were generally very small, that is 0.290 and 0.355 on average, respectively. Instead, we used the (derived) dichotomous variables EBD[0/1] and TE[0/1], and entered them as dependent variables in two separate logistic regression analyses. In addition, we also

performed an OLS regression analysis with IBP as the dependent variable. These three regression analyses are briefly summarized in Tables 3a, 3b, and 3c. We included gender (i.e., indicating females [1] instead of males [0]) as a control variable in all regression analyses to ensure the consistency in modeling across comparable studies (see De Beuckelaer et al., in press).

Tables 3a, 3b, and 3c confirm our earlier findings in that: (a) the fraction of respondents making transitivity errors was found to be slightly higher when using GPCs with seven scale points (see Table 3c); and (b) the inconsistency in brand dominance between GPCs with five scale points on the one hand and simple PCs on the other hand, is obvious from a highly significant (positive) effect reported in Table 3b. In addition, Table 3a also showed that, unlike male respondents, female respondents (i.e., half of the sample) had a tendency to provide somewhat 'flattened' preference scores, that is, preference scores that do not differ that much across the five brands of bottled water. This may indicate that, in contrast to male respondents, female respondents attach less importance to the specific brand of bottled water. A gender effect was not found in any other table (Tables 3b and 3c), implying that women do not differ from men when it comes to consistency in preference scoring. Comparable levels of consistency in preference scoring across gender were also reported by De Beuckelaer et al. (in press).

4. Conclusions

This study was the first to make a comparative assessment of the measurement validity of GPCs with a varying number of scale points. As indicated by the methodological literature on grading tasks (primarily using Likert-type rating scales), it is common practice to work with a number of scale points within the interval [5,11]. Within the context of GPCs (with forced choice), scales with eleven points have been systematically implemented in market pretesting tools / models (see Introduction), and were also used in a 14-country consumer study conducted by De Beuckelaer et al (in press). In this 14-country study, scales with eleven points were found to ensure adequate levels of measurement validity, both on a national and an international level. This study was undertaken to examine whether GPCs with nine, seven, and five (odd numbers only) points perform as well as GPCs with eleven scale points.

Results demonstrate that reducing the number of scale points below eleven is certainly justified, but our 'indicative results' (our sample was rather small) seem to point in the direction of an important 'trade-off'. As expected, using a high number of preference points (e.g., eleven) is beneficial in that the analytical operations to be performed, which require

increased cognitive effort, result in highly consistent scoring (e.g., low levels of transitivity errors as well as a consistent production of brand dominance data when compared to simple PCs). In this respect, one may be inclined to choose the maximal number of scale points included in this study, namely, eleven.

However, one may be willing to compromise slightly on performance, provided one doesn't mind working with a reduced number of scale points. Our data analysis showed that nine scale points performed equally well on all (psychometric) quality measures. So, in that sense, one would not really have to compromise on measurement validity if one decides to work with nine scale points. Our results also indicate that GPCs with seven scale points offer a reasonable alternative to GPCs with eleven and nine points. A further reduction to GPCs with five points is troublesome, especially as such GPCs were found to produce a relatively high level of inconsistency in brand dominance when compared to simple PCs. Such type of inconsistency is, without any doubt, worrisome for market researchers who use paired comparison data to elicit brand preferences in a given market.

Obviously, we must admit that we cannot fully exclude the possibility that our study results are partially dependent on unique characteristics of our rather small samples (i.e., the number of respondents providing data for a particular GPC is about 30). Even though respondents were assigned randomly to the groups completing a particular GPC, minor sampling fluctuations across experimental conditions may have influenced our results. Despite these possible shortcomings, our study has revealed some interesting patterns, such as: (a) over and above five points, one can hardy gain accuracy in terms of preference intensity, and (b) from a 'purist point of view', a high number of preference points (e.g., nine or eleven) is preferable as the higher number of preference points seems to (slightly) increase consistency in scoring behavior.

From a practical perspective, our study has shown that if one aims to reduce cognitive effort on the part of the respondents, one may consider using nine or seven scale points rather than using eleven scale points. However, this recommendation may apply only to The Netherlands where our research was conducted or to similar Western countries. In other countries with a different culture, individuals may attach a strong (unintended) meaning to one-digit numbers such as seven and nine (e.g., both seven and nine represent a lucky number in China), which may bias the number of preference scores assigned to choice alternatives. A two-digit number such as eleven is less likely to carry such unintended meaning (this is, at least, not the case in China). GPCs with eleven points have already demonstrated adequate measurement validity in an international context (De Beuckelaer et al. in press). At present,

researchers working in an international comparative setting may choose to ‘stay on the safe side’, and rely on GPCs with eleven points. Future research may examine the extent to which GPCs with scales containing nine and seven scale points (or with more than eleven points) also exhibit measurement validity across regional (e.g., European) or truly international samples.

Finally, our study, which was based on a well-balanced sample of female and male university students, provided further evidence to support the assumption that women and men (with similar levels of education) do not differ from one another in terms of scoring brand preferences in a consistent manner, that is, across paired comparison methods (e.g., GPC and simple PCs; see Table 2b) as well as within (graded) paired comparison methods (see Table 3c).

References

Alwin, D.F.: Feeling thermometers versus 7-point scales. Which are better? Sociological Methods & Research **25**, 318-340 (1997)

Bech, M., Gyrd-Hansen, D., Kjaer, T., Lauridsen, J., Sorensen, S.: Graded pairs comparison. Does strength of preference matter? Analysis of preferences for specialized nurse home visits for pain management. Health Economics **16**, 513-529 (2006)

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

Böcker, F.: Scale forms and their impact on ratings' reliability and validity. Journal of Business Research **17**, 15-26 (1988)

Churchill, G.A., Peter, J.P.: Research design effects on the reliability of rating scales: a meta analysis. Journal of Marketing Research **21**, 360-375 (1984)

Day, R.L.: Systematic paired comparisons in preference analysis. Journal of Marketing Research **2**, 406-412 (1965)

Dawes, J.: Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. International Journal of Market Research **50**, 61-77 (2007)

De Beuckelaer, A., Kampen, J.K, Van Trijp , H.C.M.: An empirical assessment of the cross-national measurement validity of graded paired comparisons. Quality & Quantity (in press)

Friedman, H. H., Amoo, T.: Rating the rating scales. Journal of Marketing Management **9**, 144-123 (1999)

Hofmans, J., Theuns, P., Mairesse, O.: Impact of the number of response categories on linearity and sensitivity of self-anchoring scales. A functional measurement approach. Methodology **3**, 160-169 (2007)

Janhunen, K.: A comparison of Likert-type rating and visually-aided rating in a simple moral judgment experiment. Quality and Quantity (in press)

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

Langbroek, I., De Beuckelaer, A.: Between-method convergent validity of four data collection methods in quantitative means-end-chain research. Food Quality and Preference **18**, 13-25 (2007)

Lozano, L.M., García-Cueto, E., Muñiz, J.: Effect of the number of response categories on the reliability and validity of rating scales. Methodology **4**, 73-79 (2008)

McKelvie, S.J.: Graphic rating scales: how many categories? British Journal of Psychology **69**, 185-202 (1978)

Moors, G.: Exploring the effect of a middle response category on response style in attitude measurement. Quality and Quantity **42**, 779-794 (2008)

Netzer, O., Toubia, O., Bradlow, E.T., Dahan, E., Evgeniou, T., Feinberg, F.M., Feit, E.M., Hui, S.K., Johnson, J., Liechty, J.C., Orlin, J.B., Rao, V.R.: Beyond conjoint analysis: advances in preference measurement. Marketing Letters **19**, 337-354 (2008)

Netzer, O., Srinivasan, V.: Adaptive self-explication of multiattribute preferences. Journal of Marketing Research **158**, 140-156 (2011)

Oishi, S., Hahn, J., Schimmack, U., Radhakrishan, P., Dzokoto, V., Ahadi, S.: The measurement of values across cultures: a pairwise comparison approach. Journal of Research in Personality **39**, 299-305 (2005)

Peterson, G.L., Brown, T.C., McCollum, D.W., Bell, P.A., Birjulin, A.A., Clarke, A.: Moral responsibility effects in valuation of WTA for public and private goods by the method of paired comparison. In: Adamowicz, W.L., Boxall, P.C., Luckert, M.K.,

Phillips, W.E., White, W. (eds.) Forestry, Economics and the Environment, pp. 134-159. Cab International, Wallingford, UK (1996)
Preston, C.C., Colman, A.M.: Optimal number of response categories in rating scales: reliability, validity, discriminating power and respondent preferences. Acta Psychologica **104**, 1-15 (2000)
Roth, A.V., Schroeder, R.G., Huang, X., Kristal, M.M.: Handbook of Metrics for Research in Operations Management: Multi-item Measurement Scales and Objective Items. Sage, Thousand Oaks, CA (2008)
Scholz, S.W., Meissner, M., Decker, R.: Measuring consumer preferences for complex products: a compositional approach based on paired comparisons. Journal of Marketing Research **47**, 685-698 (2010)
Silk, A., Urban, G.: Pre-test market evaluation of new packaged goods: a model and measurement methodology. Journal of Marketing Research **13**, 171-191 (1978)
Thurstone, L.L.: A law of comparative judgment. Psychological Review **34**,273-286 (1927)
Urban, G.L.: Pretest market forecasting In: Eliashberg, J., Lilien, G. (eds.) Handbook in Operations Research and Management Science, Vol. 5: Marketing, pp. 315-348. Elsevier Science Publishers B.V. North Holland, New York, NY (1993)
Urban, G.L., Katz, M.: Pre-test market models: validation and managerial implications. Journal of Marketing Research **20**, 221-234 (1983)
Van de Vijver, F.J.R., Leung, K.: Methods and Data Analysis for Cross-cultural Research. Sage, London, U.K. (1997)
Verschuren, P., Arts, B.: Quantifying influence in complex decision making by means of paired comparisons. Quality and Quantity **38**, 495-516 (2004)
Weng, L.: Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educational and Psychological Measurement **54**, 956-972 (2004)

Table 1 Experimental design

Round	Groups involved	Group size (*n*)	Number of preference points	Condition
First round: simple, Thurstonian PC	All groups	122[a]	One	Condition 1 (benchmark)
Second round: GPC	Group 1	25	Five	Condition 2
	Group 2	31	Seven	Condition 3
	Group 3	33	Nine	Condition 4
	Group 4	33	Eleven	Condition 5

Note. [a] only participants who also provided data for the second round are used for statistical analysis (i.e., n=122 instead of n=139). PC = Paired comparison; GPC = Graded paired comparison.

Table 2 Descriptive results

Type of paired comparisons	NEBD	EBD[0/1]	IBP (average per brand)	NTE	TE[0/1]	CON
Simple PC (Thurstonian): *n*=122	0.000 (0.000)	0.000 (0.000)			.115 (.320)	NA (non metric)
GPC						
Five scale points	1.240 (1.508)	0.600 (0.500)	0.150 (0.039)	0.080 (0.277)	0.080 (0.277)	0.853
Seven scale points	0.290 (0.693)	0.194 (0.402)	0.163 (0.445)	0.355 (0.877)	0.161 (0.374)	0.756
Nine scale points	0.182 (0.465)	0.152 (0.364)	0.165 (0.062)	0.030 (0.174)	0.030 (0.174)	0.846
Eleven scale points	0.182 (0.465)	0.152 (0.364)	0.152 (0.050)	0.030 (0.174)	0.030 (0.174)	0.887

Note. For number of participants in each group in the GPC conditions, see Table 1; Numbers displayed represent mean scores for count variables or fractions. Standard deviations are in parentheses. PC = Paired comparison; GPC = Graded paired comparison; NEBD = Number of judgment errors in brand dominance; EBD[0/1] = Fraction of respondents making at least one judgment error in brand dominance; IBP = Intensity of brand preference; NTE = Number of transitivity errors; TE[0/1] = Fraction of respondents making at least one transitivity error; CON = Pearson correlation indicating consistency in preference scoring; NA (nonmetric) = not applicable (nonmetric variable; Cramer's *V*=.919*).*

Table 3a OLS regression: Dependent variable IBP (n=122, GPC data)

	Unstandardized coefficient	Standard error	p-value
Constant	0.162[a]	0.011	0.000***
Number of scale points:			
Seven	0.012	0.013	0.369
Nine	0.013	0.013	0.339
Eleven	0.003	0.013	0.821
Female	-0.028	0.009	0.011**
R^2=.070			

Note. [a] reference is five scale points. IBP = Intensity of brand preferences.

**$p < 0.05$; ***$p < 0.001$.

Table 3b Logistic regression: Dependent variable EBD[0/1] (*n*=122, GPC data)

	Unstandardized coefficient	Standard error	Wald	*p*-value
Constant	-1.168[a]	0.852	1.882	0.170
Number of scale points:				
Five	2.122	0.636	11.130	0.001***
Seven	0.264	0.668	0.156	0.693
Nine	-0.054	0.691	0.006	0.937
Female	-0.359	0.462	0.604	0.437
Pseudo R^2 (Nagelkerke R^2) = .209				

Note. [a] reference is eleven scale points. EBD[0/1] = Fraction of respondents making at least one judgment error in brand dominance. ***$p < 0.001$

Table 3c Logistic regression: Dependent variable TE[0/1] (*n*=244, all data)

	Unstandardized coefficient	Standard error	Wald	*p*-value
Constant	-4.294[a]	1.272	11.395	0.001***
Scale points:				
One	1.465	1.056	1.924	0.165
Five	1.054	1.257	0.703	0.402
Seven	1.871	1.130	2.742	0.098*
Nine	0.076	1.439	0.003	0.958
Female	0.508	0.454	1.252	0.263
Pseudo R^2 (Nagelkerke R^2) = .065				

Note. [a] reference is eleven scale points. TE[0/1] = Fraction of respondents making at least one transitivity error. *$p < 0.10$; ***$p < 0.001$.