Qual Quant (2018) 52:1227–1239
https://doi.org/10.1007/s11135-017-0514-7

Comparison of four common data collection techniques to elicit preferences

Pasquale Anselmi[1] · **Luigi Fabbris**[2] · **Maria Cristiana Martini**[3] · **Egidio Robusto**[1]

Published online: 11 May 2017

Abstract We compare four common data collection techniques to elicit preferences: the rating of items, the ranking of items, the partitioning of a given amount of points among items, and a reduced form of the technique for comparing items in pairs. University students were randomly assigned a questionnaire employing one of the four techniques. All questionnaires incorporated the same collection of items. The data collected with the four techniques were converted into analogous preference matrices, and analyzed with the Bradley–Terry model. The techniques were evaluated with respect to the fit to the model, the precision and reliability of the item estimates, and the consistency among the produced item sequences. The rating, ranking and budget partitioning techniques performed similarly, whereas the reduced pair comparisons technique performed a little worse. The item sequence produced by the rating technique was very close to the sequence obtained averaging over the three other techniques.

Keywords Rating · Ranking · Budget partitioning · Paired comparisons · Bradley–Terry model

1 Introduction

In preference analysis, the competition between data collection techniques that may enable either ranking or scoring a finite set of interrelated items (or stimuli) has drawn the attention of social scientists for decades (see, among others, Coombs 1964; Elrod,

✉ Pasquale Anselmi
pasquale.anselmi@unipd.it

[1] Department FISPPA, University of Padua, Via Venezia, 8, 35131 Padua, Italy

[2] Department of Statistics, University of Padua, Via Cesare Battisti, 241, 35121 Padua, Italy

[3] Department of Communication and Economics, University of Modena and Reggio Emilia, Viale Allegri, 9, 42121 Reggio Emilia, Italy

Louviere, and Davey 1992; Guttman 1946; Hauser and Rao 2004; Huber 1963; Jech 1989; Louviere, Hensher, and Swait 2003; Luce 1959; McFadden 1986; Takane 1982; Thurstone 1927; Torgerson 1958).

In the present study, we analyze and compare the functioning of four common data collection techniques that may enable researchers to elicit people's preferences. We assume that, in order to collect preference data, a battery of p ($p > 1$) items is administered to a sample of n ($n \geq 1$) units using a homogeneous scale. Under the assumption that both the "choice set" of items and the responding units are randomly selected from the respective universes of all possible sets, comparisons among the data collection techniques are viable.

The following techniques have been taken into account in the present study:

1. The *rating* technique consists of presenting the items and asking respondents to assign each item i ($i = 1,\ldots, p$) a rate according to a common measurement scale.
2. The *ranking* technique consists of presenting all the items at once and asking respondents to simultaneously order them from the most to the least relevant according to a given construct.
3. The technique of *budget* (or *amount*) *partitioning* consists of giving a fixed amount of points to each respondent, for instance 100, and asking him or her to partition the points over the p items according to a given criterion.
4. The technique of *paired comparisons* consists of administering $p(p - 1)/2$ distinct pairs of items and asking respondents to choose the preferable item in each pair according to a prescribed criterion. If p is large, the technique is not viable even if visual devices and/or computer-assisted systems are in use. The reduced pair comparisons technique proposed by Fabbris (2013) has been used in the present study. This technique involves a certain ordering of the choice units and the submission, in a hierarchical fashion, of the $p/2$ pairs of items, then the submission of the $p/4$ pairs of items preferred at the first choice level, and so on until the most preferred item is sorted out.

Not all of these preference elicitation techniques are equally viable in all survey contexts. Whereas the rating technique and the reduced pair comparison do not require visual aids, the ranking and the budget partitioning techniques can return very unreliable responses if they are administered without visual aids and more than 3–4 items have to be ranked (Bradburn et al. 2004). This implies that the ranking and the budget partitioning techniques with a higher number of items can only be used in self-administered questionnaires, either computer-assisted or paper-and-pencil, and cannot be adopted in telephone surveys or personal interviews, unless visual aids are made available to respondents.

A further point to be stressed is that the techniques differ in the effort that is required to respondents (e.g., time needed to complete the task, cognitive effort required, prior knowledge). Aloysius et al. (2006) argued, in fact, that any potential normative superiority of a preference elicitation technique must be balanced against its potentially adverse effects on user acceptance. The authors found that pairwise comparisons generate stronger decisional conflicts in respondents, are more effortful and, overall, less desirable to use than absolute measurements (i.e., ratings).

Not all the preference elicitation techniques allow the occurrence of ties. According to Krosnick and Alwin (1988), ranking should be preferred to the rating technique in the field of value surveys because the latter technique tends to provide high and undifferentiated scores; nevertheless, when ties are removed from the data, the rating and ranking techniques provide similar results. Conversely, ranking forces people to make distinctions that

they would not otherwise make (Alwin and Krosnick 1985), and the validity of ranking is lessened by these unimportant and/or inconsequential distinctions between similarly regarded items (Maio et al. 1996).

In the present study, each respondent was assigned a questionnaire employing one of the four data collection techniques. All questionnaires incorporated the same collection of items. For the differences between the results of the questionnaires to be attributable only to data collection technique, the following devices were used:

- The data collection setting was the same for all techniques;
- Respondents were randomly assigned to one of the four techniques;
- The data collected with the four techniques were converted into analogous preference matrices, and analyzed within a common methodological framework.

The preference matrix obtained for each of the four techniques was analyzed through the Bradley–Terry model (Bradley and Terry 1952). The four techniques were evaluated with respect to: (a) the fit to the model, (b) the precision and reliability of the estimated item measures, and (c) the consistency among the produced sequences of items.

The remainder of the paper is organised as follows. Section 2 describes the data collection procedure, the generation of the preference matrices, and the analysis of these matrices with the Bradley–Terry model. Section 3 presents the main results of the analyses. Sections 4 and 5 review the results and provide some conclusive argumentation.

2 Method

2.1 The sample

A total of 282 university students ($M_{age} = 22.85$, $SD = 2.83$; 70% were females) participated in the study on a voluntary basis. Seventy-three percent were attending a bachelor degree program, 17% a master degree program, and 10% a single-cycle degree program.

2.2 Procedure

The data were collected at the students' secretariat of a major Italian university over four weeks. The sample of students was selected in a systematic manner, picking one student from every ten and asking him or her to anonymously self-administer an electronic questionnaire (computer assisted self-administered interviewing [CASI]) that was accessible from two local PCs. The students were asked to express their opinion about which, from a list of 12 services, the university should prioritarily invest on (see "Appendix"). A fixed order was used for presenting the items in the rating, ranking, and budget partitioning techniques, whereas two different hierarchical systems of comparisons were arranged to apply the reduced pair comparisons technique. A seven-point response scale, with an anchor point at each extreme (maximum and minimum), was employed for rating the items. An amount of 100 points was used in the budget partitioning technique.

Any sampled student was randomly assigned a questionnaire, which evaluated the 12 services through one of the four techniques. Since the data collection setting and the 12 items were common to all students, we can assume that differences in the produced item sequences only depend on the data collection technique (Takane 1989; Tversky and Russo 1969). Although non-responses were rare, the sample sizes differed because the experiment was articulated to attain various aims. The sample sizes were $n_1 = 94$, $n_2 = 49$, $n_3 = 47$

and $n_4 = 92$ for the rating, ranking, budget partitioning and paired comparison samples, respectively. The data are available on request.

2.3 Data preparation

The data collected with the four techniques were converted into analogous preference matrices. The possibility of observing ties differs across the four techniques. Ties were not allowed with the ranking and paired comparison techniques. Ties were unavoidable with the rating technique since the number of points in the rating scale was lower than the number of items, whereas ties were possible with the budget partitioning technique. At the first stage, ties in rating and budget partitioning data were ignored. Hence, for each respondent h ($h = 1, \ldots, n_1; n_2; n_3; n_4$), a strict preference relation can be stated (David 1988)

$$\begin{aligned} x_{ijh} &= 1 \; if \; A_i > A_j \quad (i, j = 1, \ldots, n; \; h = 1, \ldots, n(t), t = 1, \ldots, 4) \\ &= 0 \; if \; A_i < A_j \\ &= missing \; if \; A_i = A_j \end{aligned}$$

where A_i and A_j denote the i-th and j-th items, respectively, and $A_i > A_j$ indicates that A_i is preferred to, or dominates A_j. In this way, the ranking, rating and budget partitioning outcomes emulate the paired comparison outcomes. In the reduced pair comparisons application, the item winning a direct comparison at a certain level was forced as the winner against all items implicitly contrasted in previous matches.

Under multiple judgements, the preferences can be represented by the proportion of subjects in the population who choose stimulus A_i over stimulus A_j. The preferences can be ordered in a preference matrix $\mathbf{P} = \{\pi_{ij} \; (i, \; j = 1, \; \ldots, \; p)\}$. The maximum likelihood estimate of π_{ij} is: $p_{ij} = \Sigma_h \, x_{ijh}/n_{ij}$, where n_{ij} ($n_{ij} > 0$) is the number of non-tied comparisons between items A_i and A_j (Fienberg and Larntz 1976). This makes the skew-symmetric feature of matrix $\mathbf{P}$: $\pi_{ij} = 1 - \pi_{ji}$ evident for all is and js. The values on the main diagonal are null, $\pi_{ii} = 0$.

2.4 Data analyses

The Bradley–Terry model (Bradley and Terry 1952) was applied to the analysis of the $\mathbf{P}$ matrices obtained for each of the four techniques. It is a logit model for paired evaluations, and takes on the form: $\log(\Pi_{ij}/\Pi_{ji}) = \beta_i - \beta_j$, where Π_{ij} denotes the probability that i is preferred to j in the population, and β_i is the parameter expressing the location of item i on the latent trait under consideration. In the basic form of the model, $\Pi_{ij} + \Pi_{ji} = 1$ for all pairs, that is, ties are not allowed. Π_{ij} equals 1/2 when $\beta_i = \beta_j$, and exceeds 1/2 when $\beta_i > \beta_j$.

The Bradley–Terry model was estimated through the computer program FACETS 3.70.0 (Linacre 2012). The analysis provided an estimate (β) of the location of each item on the latent trait, and a measure of the precision (SE) of the estimate itself. The purpose of the present work is to compare the functioning of the four data collection techniques rather than to understand the actual students' preferences for the 12 services. The focus will be on the fit of the four techniques to the Bradley–Terry model, the precision and reliability of the item estimates, and the consistency among the produced item sequences.

The *mean-square fit statistic* evaluates the fit of the data to the Bradley–Terry model. Under the null hypothesis, mean-squares are χ^2 statistics divided by their degrees of freedom, with an expected value of 1. Values greater than 1 indicate underfit to the model

(i.e., the data are less predictable than the model expects), whereas values smaller than 1 indicate overfit (i.e., the data are more predictable than the model expects; Linacre 2002, 2012). For instance, a mean-square of 1.4 indicates that there is 40% more randomness in the data than modelled, whereas a mean-square of 0.6 indicates a 40% deficiency in the randomness expected by the model. Underfit is more deleterious for measurement purposes than overfit. There are two types of mean-square fit statistics: Outfit and Infit. Outfit is based on the conventional χ^2 statistic. Infit is based on the χ^2 statistic with each observation weighted by the variance of its expected value. Outfit is influenced more by the choices made between two contrasted items with different priority levels, whereas Infit is influenced more by the choices made between two contrasted items with similar priority levels. Outfit and Infit indices are computed for each of the 12 items.

Separation reliability (R) and separation (G) provide information about the reliability of item measures (Fisher 1992; Smith 2001). R represents the proportion of variance that is not due to measurement error. It is computed as the ratio of true variance SA^2 to observed variance SD^2 ($R = SA^2/SD^2$, where $SA^2 = SD^2 - \sum_{i=1}^{n} SE_i^2$). R ranges from 0 to 1. The closer the value of R is to 1, the greater the probability that differences among the measures express actual differences among the items. R is non-linear (e.g., an improvement from .6 to .7 is not twice an improvement from .9 to .95) and suffers from "ceiling effects" (i.e., R cannot be greater than 1).

The index G overcomes these two limits: it is on a ratio scale and ranges from 0 to infinity. G is computed as $SA/RMSE$ (where $RMSE = (\sum_{i=1}^{n} SE_i^2)^{1/2}$). G compares the "true" spread of the item measures with the size of their measurement error (Fisher 1992). The greater the value of G, the more the spread of the items on the latent trait expresses true differences among them. The indices R and G reported in the present paper are corrected for possible misfit of the data to the Bradley–Terry model (for details, see Linacre 2012) and represent lower boundary values for the reliability of item measures.

The four techniques were evaluated with respect to the fit of the respective preference matrices to the Bradley–Terry model and to the reliability of the item measures resulting from them. In addition, Pearson's correlation coefficients between item measures were computed to investigate whether the different techniques allowed for the elicitation of an analogous item sequence. This provides evidence of a possible convergent validity of the techniques.

The item estimates obtained for each of the four techniques were compared with the average item estimates of the three complementary techniques. For instance, the item measures estimated for the rating were compared with the average of the item measures estimated for ranking, budget partitioning, and reduced pair comparisons. The SE of the average measure of each item was computed by the Delta theorem (Bollen 1989). The statistic $z_{in(t)} = (\beta_{in(t)} - \overline{\beta}_{ic}) \Big/ (SE_{in(t)}^2 + SE_{ic}^2)^{1/2}$ allows for testing the significance of the difference between the measure of item i produced by applying the technique $n(t)$ ($\beta_{in(t)}$) and the average ($\overline{\beta}_{ic}$) of the measures of item i computed with the three complementary techniques ($SE_{in(t)}^2$ and SE_{ic}^2 are the standard errors of $\beta_{in(t)}$ and $\overline{\beta}_{ic}$, respectively).

3 Results

The Bradley–Terry model assumes one-dimensionality of the measures. For each of the four techniques, the first eigenvalue computed on the preference matrix was very close to its maximum positive value of 6.5 (eigenvalues of 6.19, 6.04, 5.98, 5.54 for ranking, rating,

budget partitioning, and reduced pair comparison, respectively). Thus, the measures are substantially one-dimensional.

Table 1 shows the β parameter estimates of the 12 items, together with their *SE*s and mean-square fit statistics. Larger values of β indicate greater priority levels.

3.1 Mean-square fit statistics

The only unpredictable results concern item [J] in the reduced pair comparisons technique. The Infit of item [J] is 1.66: If, in a pair, item [J] is contrasted with an item of a similar priority level, it is difficult to predict whether item [J] will be preferred or not.

For rating, ranking and budget partitioning, the fit statistics of all the items are smaller than 1: This means that, when two items are compared to each other, the preferred one is too predictable. The overfit observed for these three techniques is due to the deductive procedure that was used for emulating paired comparison data starting from rates, ranks or budget partitions: The assumption was made that, if two specific items had been contrasted in a pair, the item that received the greater rank, rate or budget point would have been preferred.

Results of all 66 distinct pairs of items were inferred in the rating, ranking and budget partitioning techniques. In the reduced pair comparisons, only the results of 20 distinct pairs of items were available. This may explain the lower predictability of the data produced with the reduced pair comparisons technique, compared with those of the other three techniques.

3.2 Reliability of item measures

For each of the four preference elicitation techniques, Fig. 1 depicts the estimates of the location of the items on the latent trait. The reduced pair comparisons technique produced the largest (2.80) range of measures (i.e., the difference between the top and the bottom value), followed by budget partitioning (2.31), rating (1.90) and ranking (1.61). *SE*s of the estimates were a bit larger in the former two techniques, compared with the latter two (see Table 1). When taking the measurement error of the estimates into account, the ranking, budget partitioning and reduced pair comparisons techniques performed similarly, whereas the rating technique performed a little better. For the former three techniques, the spread of measures was five times greater than the measurement error ($G = 5.48$, 5.65, and 5.80, for ranking, budget partitioning, and reduced pair comparisons, respectively), and it was seven times greater for the latter ($G = 7.32$).

The number of respondents affects the size of *SE*s and, therefore, the value of G. For excluding the sample size effect on *SE*s, 10 different samples of 48 respondents were randomly extracted from both the rating data and the reduced pair comparisons data, and new analyses were run on them. The mean *SE* across samples and items was .11 for the rating technique (the *SE* is of the same order of magnitude as the *SE* of ranking and budget partitioning) and .27 for the reduced pair comparisons technique. The value of G observed on the rating data (the mean across the 10 samples is 5.29) is of the same order of magnitude as that observed on ranking and budget partitioning, whereas that observed on the reduced pair comparisons is smaller (2.78). If the number of respondents was the same, rating, ranking and budget partitioning techniques performed similarly with respect to precision (*SE*) and reliability of estimated item measures, whereas the reduced pair comparisons exhibited a worse performance.

Table 1 Parameters β, SEs and mean-square fit statistics of the items for each of the four techniques. Larger values of β indicate greater priority levels

	Rating ($n = 94$)				Ranking ($n = 49$)				Budget partitioning ($n = 47$)				Reduced pair comparisons ($n = 92$)			
Item	β	*SE*	Infit	Outfit	β	*SE*	Infit	Outfit	*B*	*SE*	Infit	Outfit	β	*SE*	Infit	Outfit
A	−0.28	0.08	0.12	0.13	0.07	0.09	0.36	0.37	0.16	0.11	0.14	0.15	0.04	0.12	1.03	0.80
B	−0.21	0.08	0.25	0.25	0.02	0.09	0.21	0.22	0.26	0.11	0.22	0.22	0.11	0.13	0.21	0.31
C	0.84	0.09	0.13	0.12	0.36	0.09	0.14	0.14	0.87	0.11	0.23	0.24	1.25	0.12	0.37	0.47
D	0.63	0.08	0.40	0.44	0.59	0.09	0.40	0.38	0.49	0.11	0.25	0.28	0.75	0.12	0.44	0.48
E	−0.64	0.08	0.23	0.26	−1.02	0.10	0.42	0.43	−0.46	0.12	0.23	0.21	−1.55	0.15	0.93	0.32
F	0.32	0.08	0.15	0.14	0.38	0.09	0.31	0.30	0.14	0.11	0.25	0.27	−0.97	0.14	0.98	0.75
G	0.70	0.08	0.22	0.22	0.60	0.09	0.28	0.28	0.40	0.11	0.07	0.08	0.49	0.12	0.81	0.73
H	−0.38	0.08	0.16	0.17	−0.47	0.09	0.23	0.22	−0.92	0.12	0.14	0.14	−0.41	0.14	1.01	0.63
I	0.48	0.08	0.35	0.33	0.43	0.09	0.49	0.50	0.09	0.11	0.15	0.20	0.10	0.12	0.94	0.64
J	−0.63	0.08	0.30	0.33	−0.16	0.09	0.33	0.33	−0.67	0.12	0.37	0.43	−0.77	0.15	1.66	1.23
K	−1.06	0.09	0.18	0.19	−0.83	0.10	0.30	0.31	−1.44	0.14	0.59	0.61	−0.93	0.16	0.19	0.21
L	0.53	0.08	0.18	0.18	−0.04	0.09	0.20	0.20	0.58	0.11	0.33	0.40	0.29	0.11	0.88	0.69
β range	1.90				1.61				2.31				2.80			
R	0.98				0.97				0.97				0.97			
G	7.32				5.48				5.65				5.80			
RMSE		0.08				0.09				0.11				0.13		

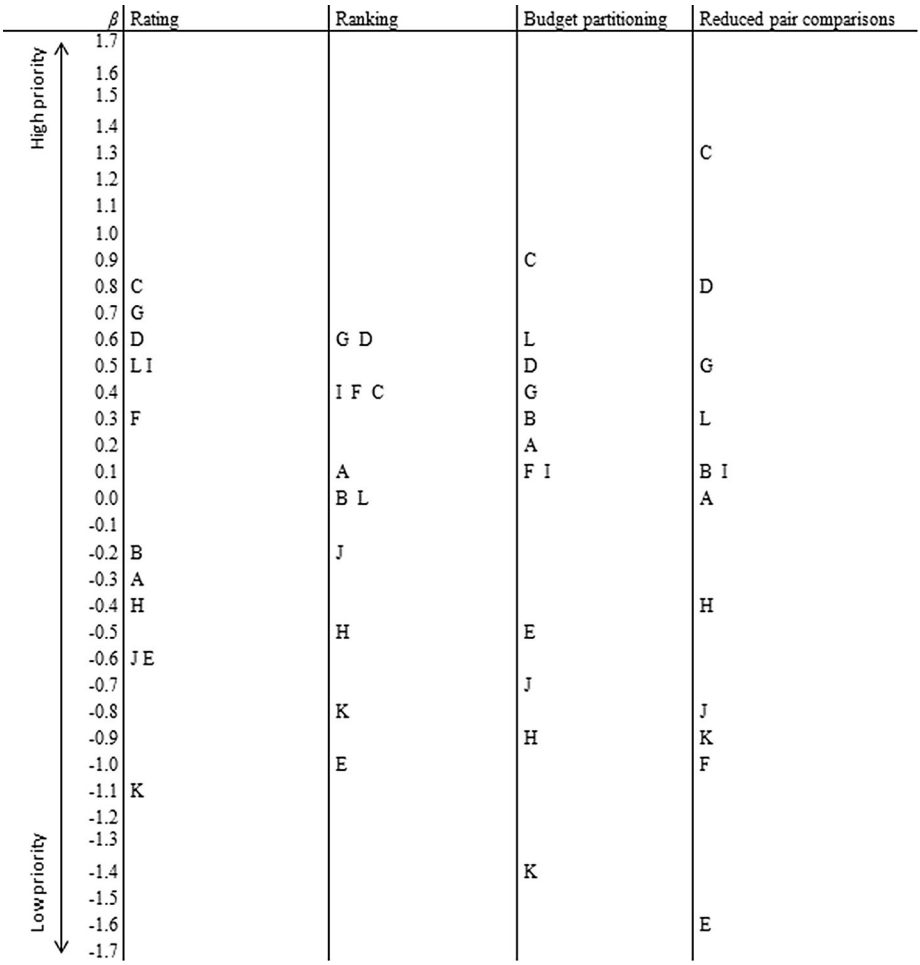

Fig. 1 Location of the items on the latent trait of the priority for each of the four techniques. Larger values of β indicate greater priority levels

3.3 Comparisons across measures

Table 2 shows the correlation coefficients between the item measures resulting from the four techniques. The weakest correlation was observed between the reduced pair comparisons and the ranking techniques (.74), whereas the strongest correlation was observed between the budget partitioning and the rating techniques (.88). Even though all correlation coefficients were high, they were not as high as one would have expected since the data collected with the four techniques were converted into analogous preference matrices and analyzed with a common model.

There are differences among the item sequences produced by the four techniques (see Fig. 1). In the budget partitioning and reduced pair comparisons techniques, item [C] turned out to be by far the item with the highest priority, whereas it shared the highest

Table 2 Correlations between the item measures resulting from the four techniques

	Rating	Ranking	Budget partitioning
Ranking	.86		
Budget partitioning	.88	.78	
Reduced pair comparisons	.78	.74	.75

priority level with item [G] ($\beta_C = .84$, $\beta_G = .70$, $SE_C = .09$, $SE_G = .08$; $z = (\beta_C - \beta_G)/(SE_C - SE_G)^{1/2} = 1.10$, $p = .22$) in the rating technique. The ranking technique put item [C] in fifth position, although its priority level did not significantly differ from that of the item in first position ($\beta_C = .36$, $\beta_G = .60$, $SE_C = SE_G = .09$; $z = -1.89$, $p = .07$). The rating, budget partitioning and reduced pair comparisons techniques agreed in identifying the four most preferred items, even if with some order differences, but differed in their capacity to differentiate among their relative positions. That is, whereas the reduced pair comparisons technique highlighted the relative differences among the top items, the rating technique smoothed things over and the budget partitioning technique fell somewhere in-between. The ranking technique did not fully agree with the three other techniques in identifying the four most preferred items.

Item [K] was by far the item with the lowest priority in the rating and the budget partitioning techniques, whereas it shared the lowest priority level with item [E] ($\beta_K = -.83$, $\beta_E = -1.02$, $SE_K = SE_E = .10$; $z = -1.34$, $p = .16$) in the ranking technique. Item [K] is far from the last item in the reduced pair comparisons technique ($\beta_K = -.93$, $\beta_E = -1.55$, $SE_K = .16$, $SE_E = .15$; $z = 2.83$, $p < .01$). The last four items were the same for the four techniques, again with some differences in the orderings.

It is noted in passing that analogous results are obtained if the number of students presented with rating or reduced pair comparisons was equalized to the number of students presented with ranking or budget partitioning. There was a strong agreement among the item measures estimated on the 10 samples of 48 respondents extracted from the rating data (Robinson's coefficient of agreement $A = .94$, $t(11) = 5.33$, $p < .001$), as well as among the item measures estimated on the 10 samples extracted from the reduced pair comparisons data ($A = .78$, $t(11) = 3.23$, $p < .01$). The mean correlation between the item measures estimated on each of the 10 samples and those estimated on the full data set was .84 for the rating technique, and .97 for the reduced pair comparisons technique.

3.4 Comparing the measures estimated for each technique with the average measures

In order to control for their different spread on the latent trait, the item measures produced by the four techniques were standardized. None of the standard values was statistically significant (p-values $\geq .09$); thus, none of the item measures resulting from any technique differed from the average item measure computed on the three other techniques.

The largest correlation was observed between the item measures produced by the rating technique and the average of the measures of three other techniques (.92). The smallest correlation (.80) was observed between the item measures resulting from the reduced pair comparisons technique and the average item measures computed on the three other techniques. This suggests that using the rating technique alone allows the definition of an

item sequence that is very close to the sequence that could be obtained by using a collection of the other three techniques.

3.5 Considering ties in rating and budget data

The rating data contained 1,892 ties, representing 30.5% of the data. In the budget data, the number of ties was 828 and represented 26.69% of the data. In order to consider also information about the ties, an extension of the Bradley–Terry model (see, e.g., Agresti 2007) was run on the rating and budget data including the ties.

Considering ties in the rating data, the estimates of the item measures became more precise (average $SE = .01$ instead of .08), and covered a smaller area on the latent trait (range $= .92$ instead of 1.90). This is an expected result because ties reduce the differences among items. Moreover, the fit statistics of all the items came close to the expected value of 1. The correlation coefficient between the item measures that were obtained with and without ties was .99. Analogous results were observed with the budget partitioning data (average $SE = .01$ instead of .12; range $= 1.49$ instead of 2.31; correlation coefficient $= .99$).

4 Discussion

The rating, ranking and budget partitioning techniques performed similarly with respect to precision (*SE*) and reliability of the estimated item measures. The reduced pair comparisons technique exhibited a worse performance since, by construction, it implies the analysis of a smaller number of pairs. Moreover, the rating technique turned out to be more consistent with the overall results produced by the three other techniques.

There are some differences in the item sequences produced by the four techniques. Not even the first and the last positions are shared among all techniques. The analysis of how item locations are determined in each approach gives some hints on the expected trustfulness of each technique. The reduced pair comparisons technique is likely to be the most trustful technique at the top end of the ordering, since the winner of the whole set of comparisons is directly compared to the winners of all the other comparisons. Conversely, this technique appears to be particularly unreliable on the bottom of the ordering, since the items on this end of the continuum are compared with only a few items.

Instead, the budget partitioning and the rating techniques appear to be the most reliable at the bottom end. As far as the rating approach is concerned, when all items to be rated are important (e.g., they pertain to values or prospective services), many respondents could take the easy way out of rating most of the items as highly important with no discernment (Feather 1973; Krosnick and Alwin 1988); therefore, the lowest end of the ordering might end up being more reliable than the top end. Some literature (Bech et al. 2007; Huber et al. 1993) reported the ranking technique to be reliable at the two ends of the continuum. This has been attributed to the respondents' tendency to define first the top and bottom positions, which are the easiest to choose, and to adjust then the intermediate positions, which require more effort. Our results concerning the ranking technique do not provide support to the aforementioned statements. The item that the ranking technique placed at the top of the continuum differed from that placed by the three other techniques. The item that the ranking technique placed at the bottom of the continuum was the same item placed by the reduced pair comparisons technique—which cannot be taken to be particularly reliable on

this end of the continuum—whereas it differed from the item placed by the rating and the budget partitioning techniques.

Another relevant difference among the four techniques is their capacity to differentiate the relative position of each item. The ranking method produced the lowest differentiation, which is in contrast with the findings of previous studies (e.g., Alwin and Krosnick 1985; Krosnick and Alwin 1988); the reason for this low differentiation may be due to the number of items to rank, that is too high to guarantee precision and differentiation. In contrast, the reduced pair comparisons technique gave rise to a greater amount of differentiation, especially at the two sides of the continuum. This is partly due to the strategy used to derive implicit preferences from the observed ones. Once an item had "won" the whole set of comparisons, it was taken as the winner of all comparisons in which it was indirectly involved. Once an item had "lost" the first comparison, it was taken as the loser in a number of further comparisons ('transitivity principle'). The rating and the budget partitioning techniques produced a differentiation that was somewhere in-between those produced by the ranking and the reduced pair comparisons techniques. It is worth noting that the differentiation of the rating technique also depends on the number of points in the rating scale, whereas that of the budget partitioning technique also depends on the amount of points to be partitioned.

As pointed out in the introduction, not all the preference elicitation techniques are equally viable in all survey contexts (Aloysius et al. 2006; Alwin and Krosnick 1985; Bradburn et al. 2004; Krosnick and Alwin 1988; Maio et al. 1996). When the number of items at hand and the adopted survey technique allow for the use of more than one preference elicitation method, one solution could be to choose the most efficient, that is the one that needs the smallest number of respondents to produce precise and reliable estimates. From a different perspective, the technique could be chosen that returns a given level of information with a lower respondent burden. Although the reduced pair comparisons technique implied a lower burden to respondents, it was not able to return the same level of information of three other techniques. In the present experiment, this partly depends on the fact that only two alternative sets of initial comparisons were randomly assigned to respondents. A complete randomization of the initial set of comparisons might, at least to some extent, ride over the risk of bias due to particular initial pairings.

The adoption of the Bradley–Terry model for analyzing the data collected with different techniques represents both a strength and a limitation of the present study: A strength, because the performance of the techniques is compared within a common methodological framework, and a limitation because the Bradley–Terry model may not represent the optimal model for all the considered techniques. The rating and ranking data were also analyzed by using the rating scale model (Andrich 1978). The item measures obtained with this model correlated .99 with those obtained with the Bradley–Terry model, and this can be taken as an indicator of the appropriateness of the Bradley–Terry model for the analysis of rating and ranking data.

5 Conclusions

In the present work, the data collected with four preference elicitation techniques were converted into a comparable structure (the preference matrix) and analyzed with a common formal model. Both the number of techniques that have been compared and the use of the same model for analyzing them are not common in the literature.

One of the main problems when trying to assess the validity of alternative preference elicitation techniques is the lack of a "true" item sequence to be used as a benchmark. A possible development of the present study could be to analyze a set of items with a "theoretical" ranking (e.g., illegal behaviors ranked according to their carried penalties), and to address construct validity by verifying if the item sequence produced by the various techniques resembles the theoretical item sequence.

Appendix: List of the university services submitted to evaluation

A. Counselling high school graduates for university choices, links with high school
B. Counselling for course changing (during courses)
C. Toward labour guidance (after graduation)
D. Organizational support of studies (reducing costs, improving canteen and accommodation, etc.)
E. Amusement, socialization, culture, sports and other types of relationships with guest town
F. Information exchange for and among students (call centre, internet, help desks, etc.)
G. Economic support for deserving students
H. Larger classrooms and (possibly self-managed) rooms for besides-study activities
I. Teaching materials (lecture notes, online textbooks, library accessibility, etc.) to improve study efficacy
J. Learning supporting activities (internships, study groups, summer schools, language courses)
K. Individual or group tutorship to improve learning
L. Increasing opportunities to study or work abroad (Erasmus, Leonardo, etc.)

References

Agresti, A.: An Introduction to Categorical Data Analysis, 2nd edn. Wiley, Hoboken (2007)
Aloysius, J.A., Fred, D.D., Darryl, D.W., Taylor, A.R., Kottemann, J.E.: User acceptance of multi-criteria decision support systems: the impact of preference elicitation techniques. Eur. J. Oper. Res. **169**, 273–285 (2006)
Alwin, D.F., Krosnick, J.A.: The measurement of values in surveys: a comparison of ratings and rankings. Public Opin. Q. **49**, 535–552 (1985)
Andrich, D.: A rating formulation for ordered response categories. Psychometrika **43**, 561–573 (1978)
Bech, M., Gyrd-Hansen, D., Kjær, T., Lauridsen, J.T., Sørensen, J.: Graded pairs comparison—does strength of preference matter? Analysis of preferences for specialised nurse home visits for pain management. Health Econ. **16**, 513–529 (2007)
Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York (1989)
Bradburn, N.M., Sudman, S., Wansink, B.: Asking Questions: The Definitive Guide to Questionnaire Design—For Market Research, Political Polls, and Social and Health Questionnairs, Revised edn. Jossey Bass, San Francisco (2004)
Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: the method of paired comparisons. Biometrika **39**, 324–345 (1952)
Coombs, C.H.: A Theory of Data. Wiley, Oxford (1964)
David, H.A.: The Method of Paired Comparisons, 2nd edn. Chapman and Hall, London (1988)
Elrod, T., Louviere, J.J., Davey, K.S.: An empirical comparison of ratings-based and choice-based conjoint models. J. Mark. Res. **29**, 368–377 (1992)
Fabbris, L.: Measurement scales for scoring or ranking sets of interrelated items. In: Davino, C., Fabbris, L. (eds.) Survey Data Collection and Integration, pp. 21–44. Springer, Heidelberg (2013)

Feather, N.T.: The measurement of values: effects of different assessment procedures. Aust. J. Psychol. **25**, 221–231 (1973)
Fienberg, S.E., Larntz, K.: Log linear representation for paired and multiple comparisons models. Biometrika **63**, 245–254 (1976)
Fisher Jr., W.P.: Reliability, separation, strata statistics. Rasch Meas. Trans. **6**, 238 (1992)
Guttman, L.: An approach for quantifying paired-comparisons and rank order. Ann. Math. Stat. **17**, 143–163 (1946)
Hauser, J.R., Rao, V.: Conjoint analysis, related modeling, and applications. In: Wind, Y., Green, P.E. (eds.) Marketing Research and Modeling: Progress and Prospects: A Tribute to Paul E. Green, pp. 141–158. Springer, New York (2004)
Huber, P.J.: Pairwise comparison and ranking: optimum properties of the row sum procedure. Ann. Math. Stat. **34**, 511–520 (1963)
Huber, J., Wittink, D.R., Fiedler, J.A., Miller, R.: The effectiveness of alternative preference elicitation procedures in predicting choice. J. Mark. Res. **30**, 105–114 (1993)
Jech, T.: A quantitative theory of preferences: some results on transition functions. Soc. Choice Welf **6**, 301–314 (1989)
Krosnick, J.A., Alwin, D.F.: A test of the form-resistant correlation hypothesis: ratings, rankings, and the measurement of values. Public Opin. Q. **52**, 526–538 (1988)
Linacre, J.M.: What do infit and outfit, mean-square and standardized mean? Rasch Meas. Trans. **16**, 878 (2002)
Linacre, J.M.: Facets Computer Program for Many-Facet Rasch Measurement, Version 3.70.0. Winsteps.-com, Beaverton (2012)
Louviere, J.J., Hensher, D.A., Swait, J.D.: Stated Choice Methods. Analysis and Application. Cambridge University Press, Cambridge (2003)
Luce, R.D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
Maio, G.R., Roese, N.J., Seligman, C., Katz, A.: Rankings, ratings, and the measurement of values: evidence for the superior validity of ratings. Basic Appl. Soc. Psychol. **18**, 171–181 (1996)
McFadden, D.: The choice theory approach to market research. Mark. Sci. **5**, 275–297 (1986)
Smith Jr., E.V.: Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. J. Appl. Meas. **2**, 281–311 (2001)
Takane, Y.: Maximum likelihood additivity analysis. Psychometrika **17**, 225–240 (1982)
Takane, Y.: Analysis of covariance structures and probabilistic binary choice data. In: de Soete, G., Feger, H., Klauer, K.C. (eds.) New Developments in Psychological Choice Modeling, pp. 139–160. North Holland, Amsterdam (1989)
Thurstone, L.L.: A law of comparative judgment. Psychol. Rev. **34**, 281–299 (1927)
Torgerson, W.S.: Theory and Methods of Scaling. Wiley, New York (1958)
Tversky, A., Russo, J.E.: Substitutability and similarity in binary choices. J. Math. Psychol. **6**, 1–12 (1969)