TPM

RATING, RANKING, OR BOTH? A JOINT APPLICATION OF TWO PROBABILISTIC MODELS FOR THE MEASUREMENT OF VALUES

DEBORA DE CHIUSOLE
LUCA STEFANUTTI
UNIVERSITY OF PADOVA

Following the debate between the proponents of the ranking and rating methods for the measurement of values, this article considers the hypothesis that a person's responses in ranking and rating tasks are governed by the same unidimensional latent trait. The hypothesis was tested through a probabilistic modeling approach. The results of a joint application of the rating scale and ranking models indicated that this latent common trait exists and seems to have a role in molding the relationship between the latent variable and the observable responses in both ranking and rating formats.

Key words: Rank data; Rating scales; Measurement of values; Ranking model; Rating scale model.

Correspondence concerning this article should be addressed to Debora de Chiusole, Dipartimento di Psicologia Applicata, Università di Padova, Via Venezia 8, 35131 Padova (PD), Italy. E-mail: debora.dechiusole@studenti.unipd.it

INTRODUCTION

In the literature on the measurement of individual preferences and values, a thirty-year old debate has been raging between the proponents of the ranking method and those of the rating method (Alwin & Krosnick, 1985; Barylko-Pikielna et al., 2004; Baumgartnet & Steenkamp, 2001; Grimm & Church, 1999; Jackson & Alwin, 1980; Johnson, Sallis, & Hovell, 1999; Krosnick & Alwin, 1988; Maio, Bell, & Esses, 1996; Munson & McInyre, 1979; Ovadia, 2004; Rokeach, 1973; Russell & Gray, 1994; Tourangeau, Rips, & Rasinski, 2000; Van Herk & Van de Velden, 2007).

In the ranking method, the items belonging to a given set are ordered according to some criterion (e.g., preference or importance). Subjects are required to examine all items and to assess their position relative to the others, so that one item is considered the most important, another one the second most important, and so on. In the rating method, instead, participants are free to assess each item independently of the others, by scoring it on an *n*-point scale like, for example, a Likert scale. Each of the two methods has advantages and disadvantages.

Rating scales have become increasingly popular because of their ease of use. They are easy to understand and less cognitively demanding than ranking. There is no need to compare the items with one another, as different items can be given exactly the same score. All of this allows a respondent to complete the task up to three times faster than a ranking one (Munson & McInyre, 1979). On the other hand, rating scales are not free of criticism. For instance, it has been found that a rating task can decrease the motivation to discriminate between items (Alwin &

Krosnick, 1985). Furthermore, in extreme situations, all items may receive the same score (Maio et al., 1996). An obvious consequence of this is low variability in the data, which makes data analysis problematic (Jackson & Alwin, 1980; Ovadia, 2004). Another drawback of rating scales is their sensitivity to respondents' biases such as the tendency to score all items high or vice versa (Feather, 1973; Baumgartnet & Steenkamp, 2001; Tourangerau et al., 2000).

In contrast to rating scales, ranking allows to obtain higher quality data (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; Rokeach, 1973) in the sense that the cognitive effort and the level of attention required to complete the ranking task force respondents to better discriminate among the items (Barylko-Pikielna et al., 2004; Villanueva, Petenate, & da Silva, 2005) and to resolve the response tendencies issue. However, these advantages have a negative side. On the one hand, the task requires a higher cognitive effort and longer time to be completed (Rokeach, 1973). On the other hand, people may be led to overestimate the real differences between items (Maio et al., 1996).

In the context of this debate, an important theoretical issue was raised by Ovadia (2004) concerning the cognitive organization of values. A different set of assumptions about the nature and structure of the value system underlies each of the two methods. When a researcher chooses the ranking method, he (often implicitly) adheres to the assumption that values are organized in a hierarchy, where the position occupied by an element strictly depends upon the position of other elements. For instance, two distinct elements must always occupy different positions.

Instead, in the rating approach, a system is assumed where the values are independent of one another, so that the position of a value does not necessarily have an effect on the others (Gergen, 1991). Values would thus receive an absolute evaluation, with the consequence that some of them may happen to have the same score.

Despite appearances, Ovadia (2004) argued that these two sets of assumptions are not necessarily incompatible. The underlying organization of values might well be reflected by both a hierarchical rank order, and a rating system where each value receives a certain amount of total importance.

A methodological implication of this theoretical consideration regards the possibility to develop a mixed approach integrating the information obtained by the two methods and, at the same time, retaining the respective advantages. In the present article, this question is tackled through a probabilistic modeling approach, where the location of the items along a single unidimensional latent variable governs the responses of a person in both the rating and ranking tasks.

In particular, in our study we considered the hypothesis of the existence of a latent trait, where item importance is a parameter measured on an interval scale, and this parameter is unique, up to permissible transformations, across the ranking and rating methods. If the psychological construct underlying a given set of items (e.g., a linear order on the set of items) is to be the same, irrespectively of the method used to collect the data, the assessment of hypotheses like this is of fundamental importance.

If this hypothesis should turn out to be false, any difference at the individual level would necessarily reflect a difference (in the form of a bias) due to the two methods at the population level. In such a circumstance, one would obtained two methods that are inconsistent with one another, providing, for example, conflicting information about which is the true psychological construct underlying a given set of items.

Therefore our investigation is on a level that comes before the level where single individuals are assessed and it is, in a sense, a prerequisite for it. If the two methods favored distinct and incompatible representations of the underlying psychological construct, any further investigation of the differences between ranking and rating at the individual level would make little sense.

THE PROBABILISTIC MODELS

Both rating scale and rank data are quite often analyzed according to a probabilistic modeling approach. A clear benefit of this perspective is that it allows a researcher to make rigorous (i.e., formal) assumptions about the essential nature of the unobservable process that generated the data, and to test them against a suitable empirical dataset. Thus, we adopted the perspective that the process underlying each of the two types of data could be captured by a distinct set of assumptions. Most importantly, our request is that the two sets of assumptions should not be disjoint, such requirement will give rise to common and different assumptions.

Concerning the common assumptions, for both types of observable data, the existence of a unidimensional latent trait was postulated, along which a specific characteristic of the items (the same for both types of data) was measured. Because in our application the items are values, the degree of importance of a value will be the measured characteristic. Regarding the different assumptions, instead, the essential one is a constraint on the rank data, which does not apply to rating scale data. Such constraint imposes a linear order on the rank data,[1] so that there are no two equivalent items in the choice set.

Unidimensional latent trait modeling of rating scale data is a frequent choice based on item response theory (Andrich, 1978; Masters, 1982). Besides unidimensionality of the latent trait, Andrich's Rating Scale Model (RSM) displays a number of remarkable properties, among which: (a) person and item characteristics are measured on the same latent trait and the measurement scale is interval; (b) person and item characteristics are separable (the obtained measures are *specifically objective*); (c) all items have equal discrimination; (d) each item is characterized by a given number of thresholds, this number is the same for all items and the threshold dispersion about their arithmetic mean is the same for all items. Formally, given any two items i and j, $\tau_{ki} - \mu_i = \tau_{kj} - \mu_j$ for all scale points k, where μ_i and μ_j are the arithmetic means of items i and j respectively. An additional restriction that sometimes is introduced is that the thresholds τ_{ki} of an item form a monotonic sequence. In our study this restriction was not considered.

According to the RSM, the probability that a person v scores item i with $0 \leq x < n$ on an n-point rating scale is a function of a location β_v of person v along the latent trait and the k-th threshold τ_{ki} of item i on the same continuous dimension. For $x > 0$ it corresponds to

$$P(X_{vi} = x) = \frac{exp\left[\sum_{k=1}^{x}(\beta_v - \tau_{ki})\right]}{1 + \sum_{y=1}^{n-1} exp\left[\sum_{k=1}^{y}(\beta_v - \tau_{ki})\right]}, \quad (1)$$

where X_{vi} is a random variable whose realizations are integer values in the set $\{0,\ldots,n-1\}$, whereas for $x = 0$ the probability is:

$$P(X_{vi}=x)=\frac{1}{1+\sum_{y=1}^{n-1}exp\left[\sum_{k=1}^{y}(\beta_v-\tau_{ki})\right]}, \qquad (2)$$

where the β_v and τ_{ki} parameters turn out to be unique up to positive linear transformations with a common multiplicative constant, namely, they have interval scale properties with a common unit of measurement.

The arithmetic mean μ_i of all threshold parameters of an item *i* is taken as the location of the item itself along the latent trait. Concerning item location, it has to be noticed that when the *n*-point rating scale is meant to range from 0 = *totally unimportant* to *n* – 1 = *totally important* (as it is the case of our application), the more positive the item location, the less important the item. This is because the threshold parameters τ are taken in the negative form in Equations (1) and (2). A straightforward way to obtain a direct relationship between the *n*-point rating scale and the item location parameters is to replace the threshold parameters τ_{ki} with new parameters $\tau'_{ki}=-\tau_{ki}$ for each item *i*. In the sequel, we refer with μ'_i to the arithmetic mean of the thresholds τ'_{ki}. In our study this type of transformation was applied.

One further comment concerns the person parameters β_v. In our application only the item parameters are considered, while the β_v are regarded as nuisance parameters representing subjects' possible biases.

As far as rank data are concerned, a rather natural choice is the Plackett-Luce ranking model (Luce, 1959; Plackett, 1975). For an overview see, for example, Marden (1995). This model is applicable whenever individuals are asked to rank the elements of a finite set *X* according to some specific characteristic (e.g., importance, satisfaction, preference, etc.). A fundamental assumption of this model is that the specific characteristic of the elements in *X* is measured on a ratio scale with an arbitrary unit of measurement. Accordingly, each item $i \in X$ is represented by a non-negative real-valued parameter $v(i)$ that provides a measure of the specific characteristic for that item (i.e., a degree of importance) along the ratio scale.

In a ranking task the response pattern of an individual turns out to be a permutation $r=(r_1,r_2,\ldots,r_m)$ of the elements in *X*, where, for $k=1,2,\ldots,m$, $r_k=i$ if item $i \in X$ occupies position *k* in the permutation. Let *R* be a random variable whose realizations are all possible permutations of the set *X*. According to the ranking model, for the chain rule of conditional probabilities, the probability of observing permutation $(r_1,r_2,\ldots,r_m)$ is given by

$$P[R=(r_1,r_2,\ldots,r_m)]=P(r_1)P(r_2\mid r_1)\cdots P(r_m\mid r_1,r_2,\ldots,r_{m-1})$$

and each conditional probability has the form

$$P(r_k\mid r_1,r_2,\ldots,r_{k-1})=\frac{v(r_k)}{v(r_k)+v(r_{k+1})+\cdots+v(r_m)}.$$

So that, for $r=(r_1,r_2,\ldots,r_m)$, the model equation takes on the following form

$$P(R=r)=\prod_{k=1}^{m}\frac{v(r_k)}{v(r_k)+v(r_{k+1})+\cdots+v(r_m)}.$$

A usual re-parameterization of the model consists of obtaining new parameters π_i of the items through the bijective transformation $\pi_i=ln\,v(i)$. This transformation has, in particular, the

effect of changing the level of the scale from ratio to interval. More precisely, the parameters π_i are measured on an interval scale with common unit of measurement and origin, and the scale is unique up to the choice of an origin. A common choice is to fix the origin by constraining it to be equal to the mean of the item locations along the scale.

An important point should be clarified concerning how the data were coded in the ranking and rating tasks. In the ranking data the code 1 means *maximum importance* (i.e., chosen as first) while in the rating data the smallest code, that is 0, means *minimum importance*. On the other hand, as far as the parameters of the two models are concerned, the more positive the parameter $v(i)$ in the ranking model or μ'_i in the rating scale model, the more important item *i*.

The interesting point is now a fundamental assumption that the two models have in common: the item locations are measured on a continuous latent trait, and the scale level is the interval one for both models (we are considering the re-parameterized ranking model). This observation puts the basis for a comparison of the rating and ranking tasks from a probabilistic modeling perspective. The primitive idea is that if the two tasks allow us to measure the same aspect, the scale underlying each of them is exactly the same, up to permissible transformations. In other words, one should come up with the same measure for an item, regardless of which of the two tasks is chosen. One way to test this hypothesis would be to jointly fit the two models by constraining the item parameters to be the same across the RSM and the ranking model. This however would, at least, imply the derivation of a suitable estimation procedure, which is not trivial and, anyway, goes beyond the purposes of this article.

We opted, instead, for a less elegant, but still informative, route. Even when the parameters of the two models are estimated independently, the corresponding latent variables should display a high correlation, suggesting the existence of a common factor. If so, such common factor would represent, for example, in a within-subject design, the unique scale governing both the ranking and the rating tasks. Therefore, if θ_i^{RK} is the location of item *i* in the ranking model, and θ_i^{RT} is the parameter of the same item in the rating scale model, a simple common factor model would be

$$\theta_i^{RK} = \lambda^{RK} V_i + s^{RK}$$
$$\theta_i^{RT} = \lambda^{RT} V_i + s^{RT}$$

where: V_i is the location of item *i* on the common factor; λ^{RK} and λ^{RT} are, respectively, the factor loadings of the ranking latent variable and the rating latent variable; s^{RK} and s^{RT} are, respectively, specific factors of the two latent variables. In the present context the residual variance of each of the two specific factors s^{RK} and s^{RT} can be interpreted as method-specific variance. In the next section it will be seen how this common factor model has been applied to a slightly more complex design.

METHODS

Materials and Participants

The data set used in this study consists of the responses to a questionnaire provided by 93 Italian students (66% females, 34% males; mean age = 23.6, *SD* = 3.91) attending the graduate courses in Psychology (4th year) at the University of Padua.

The questionnaire used in the present research is an Italian version of the popular List of Values (LOV; Kahle, 1983). This scale can be applied in both rating and ranking versions (see, e.g., Chryssohoidis & Krystallis, 2005; Grunert, Grunert, & Beatty, 1989; Kamakura & Novak, 1992). The LOV scale is composed of nine items (1. Sense of belonging; 2. Excitement; 3. Warm relationships with others; 4. Self-fulfilment; 5. Being well-respected; 6. Fun and enjoinment of life; 7. Security; 8. Self-respect; 9. A sense of accomplishment) which, in the rating version, are assessed on a nine-point scale (1 = *very unimportant*, 9 = *very important*), and in the ranking version, are ordered from the most important (1) to the least important (9).

Procedures

In comparing ranking and rating, within-subject designs are rarely used. As pointed out by Van Herk & Van de Velden (2007) it is often unknown how the same subject responds to both tasks. For this reason, we opted for a mixed design, which include a between-subject and a within-subject analysis.

Each participant was randomly assigned to one of four different conditions: ranking-first (23 participants) and rating-first (25 participants) for the within-subject design; ranking-only (23 participants) and rating-only (22 participants) for the between-subjects design.

In the ranking-first condition, the same participant responded to a ranking version of the questionnaire, followed by a rating version. Hence, two separate datasets were obtained (referred to as RK_1 and RT_1 in the sequel). In the rating-first condition, the order of presentation of the two versions was inverted. In this case, the two datasets were RT_2 and RK_2. In the ranking-only (RK_0) and rating-only (RT_0) conditions, participants had just one version of the questionnaire.

The rating scale model and the ranking model were fitted independently to each of the six different datasets, thus obtaining six different sets of item parameter estimates. In order to verify any influence of one method on the other, the item parameter estimates of the within-subject conditions were compared with one another. In order to verify the existence of a unique scale underlying the six sets of parameters, a factor analysis model with a common single factor was hypothesized (Figure 1) and applied to them.

In Figure 1, V is the latent common factor (the *value system*) which explains the other latent variables, represented by circles. Factor loadings between the common factor V and each of the other six latent variables are represented by λ. For each of the six variables, ψ represents the proportion of variance not explained by the factor V and thus the variance of the specific factor.

Results

To obtain the item parameter estimates for the rating scale model, the RUMM software (Andrich, Sheridan, Lyne, & Luo, 2000) was used, while for the ranking model, MATLAB functions, programmed by one of the authors (LS) were applied. As a formal test of the goodness-of-fit of the models, Pearson's Chi-square statistic was used. For the ranking model, due to the

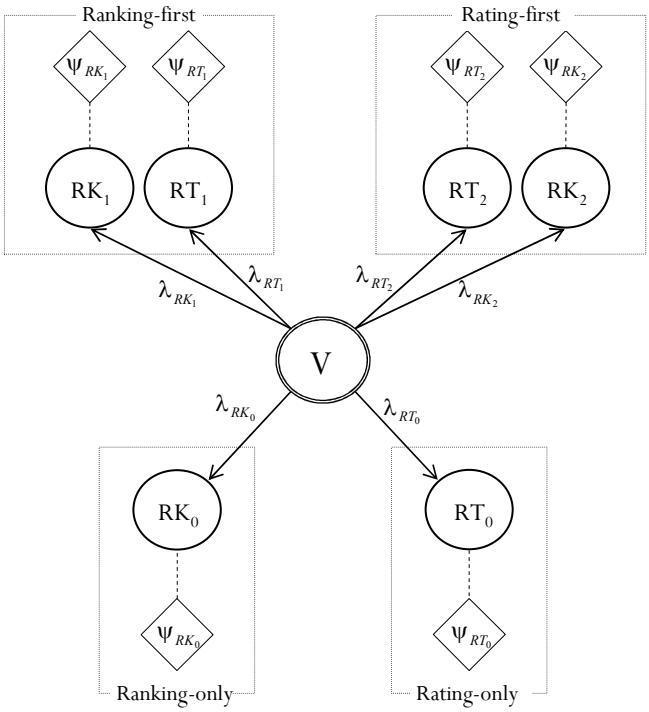

FIGURE 1
Graphical representation of the hypothesized single factor model.

sparseness of the data matrix, the *p*-value of the Chi-square statistic was estimated by parametric bootstrap (Efron, 1979; Efron & Tibshirani, 1991, 1993) with a total of 1,000 replications.

The *p*-value obtained by parametric bootstrap for the ranking models was: .404 for RK_1, .071 for RK_2, and .192 for RK_0. With a first type error probability fixed at .1, the fit of RK_2 was not really satisfactory. The reason of this result is not quite clear, however, by an inspection of the item parameter estimates displayed in Table 1 for the condition RK_2 two items (1 and 2) are found to have exactly the same location. This lack of discrimination between items 1 and 2 might have some influence on the model fit.

On the other hand, the Chi-square statistics for the three rating scale models were: $\chi^2_{18} = 15.0$ ($p = .663$) for RT_1 and $\chi^2_{18} = 15.6$ ($p = .619$) for RT_2, $\chi^2_{18} = 14.9$ ($p = .666$) for RT_0.

The item parameter estimates obtained in the within-subjects design are shown in Table 1, along with the respective standard errors.

If in Table 1 the items are ordered from the one having the highest to the one having the lowest parameter value, in the ranking-first condition, the position of the less important items (1, 2, 6, 7, and 5), as well as that of the most important one (3), is the same in both RK_1 and RT_1. Instead, for items occupying central positions (items 4, 8, and 9) there is more disagreement. Thus, it seems that the two methods perform equally well in discriminating between the extreme items (the most important and the least important). Conversely, they seem to behave differently in ordering items having a moderate importance. This observation is in line with the results obtained by Van Herk & Van de Valden (2007).

TABLE 1
Item parameters estimates (v for ranking and μ' for rating) obtained in the within-subjects design along with the corresponding standard errors (*se*)

	RK_1		RT_1		RT_2		RK_2	
Items	v	*se*	μ'	*se*	μ'	*se*	v	*se*
1	–1.291	0.321	–0.852	0.122	–0.710	0.152	–1.060	0.324
2	–0.837	0.280	–0.586	0.127	–0.889	0.141	–1.060	0.327
3	1.422	0.250	0.994	0.239	1.380	0.383	2.230	0.357
4	0.446	0.242	0.365	0.179	0.337	0.245	0.642	0.259
5	0.180	0.245	0.004	0.154	0.408	0.252	0.287	0.266
6	–0.815	0.280	–0.473	0.131	–0.200	0.193	–0.890	0.314
7	–0.555	0.263	–0.465	0.131	–0.437	0.173	–0.780	0.287
8	0.856	0.240	0.323	0.176	0.178	0.228	0.448	0.264
9	0.594	0.229	0.692	0.208	–0.066	0.205	0.182	0.266

Similar conclusions cannot be drawn, however, for the rating-first condition. Indeed, only the most important item (number 3) occupies the same position in RK_2 and in RT_2. Another interesting observation is that in RK_2 the parameter estimates of the two less important items (1 and 2) are exactly the same (which is rather unexpected in a ranking task).

On the whole, these results suggest that in both within-subject conditions participants are strongly influenced by the task which comes first. In particular, when rating comes before ranking, the discrimination among the items in the ranking part seems to become less reliable. This conclusion is motivated by the observation that the agreement between the ranking and the rating tasks about the order of the items, is weaker in the rating-first condition.

To summarize, all these aspects can be read in light of the influence of a method over the other. The ranking task could facilitate a subject to accomplish the rating task in a way that the answers to the two parts would be quite similar. The same conclusion cannot be drawn when rating comes before ranking, as the former seems to hinder the latter.

Comparing RK_0 and RT_0 parameters (Table 2), only the last important value (item 3) occupies the same position, while there is systematic disagreement concerning the position of all other values.

The results obtained in this condition are not much different from those obtained in the rating-first situation. However, this is a between-subjects condition, and the general expectation would be to find, in this design, more disagreement between ranking and rating than in any of the within-subject conditions.

Despite differences, ranking and rating turn out to have strong and positive correlations in all conditions: .97 for the item parameters of the ranking-first, .96 for the rating-first, and .92 for the between-subjects condition. The model hypothesized in Figure 1, a single common factor explaining all six latent variables, seems thus reasonable. As expected, the factor analysis model applied to the item parameters was strongly accepted ($\chi^2_9 = 13.7$, p = .13). On the other hand, models with more than a single common factor turned out to be unidentifiable. The model parameters are shown in Table 3.

TABLE 2
Item parameters estimates (v for ranking and μ' for rating) obtained in the between-subjects design along with the corresponding standard errors (*se*)

	RK_0		RT_0	
Item	v	*se*	μ'	*se*
1	–0.953	0.304	–1.130	0.125
2	–1.082	0.309	–0.974	0.140
3	1.758	0.280	1.339	0.393
4	0.179	0.249	0.711	0.310
5	0.248	0.250	0.371	0.276
6	–0.589	0.276	–0.314	0.218
7	–0.542	0.278	–0.386	0.211
8	0.479	0.246	–0.144	0.233
9	0.502	0.248	0.527	0.291

TABLE 3
Parameter estimates of the single factor model

	RK_1	RK_2	RK_0	RT_1	RT_2	RT_0
λ	0.992	0.963	0.981	0.933	0.953	0.951
ψ	0.015	0.074	0.038	0.129	0.092	0.095

Factor loadings λ are comprised between .93 and .99, indicating that all the variables included in the model are strongly correlated with the latent variable V, and that ranking and rating measure the same latent trait in an appropriate way.

However, again, a difference between ranking and rating emerges: the factor loadings of the ranking variables (RK_1, RK_2 and RK_0) are all higher than those of the rating ones (RT_1, RT_2, and RT_0). At the same time, the specific factor variances ψ of the rating variables are systematically higher than those of the ranking ones. According to what already observed, altogether these results indicate that ranking is slightly more representative of the latent values system V than rating.

To conclude, the overall procedure followed up to here allowed to use ranking and rating in a conjoint manner to produce a unique measurement scale for the nine values of the LOV questionnaire (Table 4).

TABLE 4
Item parameter estimates in the single factor model

Item	1	2	3	4	5	6	7	8	9
V	–1.147	–1.146	1.945	0.404	0.286	–0.694	–0.649	0.501	0.501

DISCUSSION

This article assesses the hypothesis that a person's responses in ranking and rating tasks are governed by the same latent trait. In order to test this hypothesis, we considered the ongoing debate, mentioned at the beginning of the paper, from a different point of view.

Rather than focusing on the respective advantages and disadvantages of the two methods, we wondered if a particular type of data analysis could allow us to move toward a new perspective. For this reason, we choose a probabilistic modeling approach, which provides precise assumptions about the process underlying each of the ranking and rating tasks.

The joint analysis of the rating and rank data was performed by an application of two exemplary models: Andrich's rating scale model and the Plackett-Luce ranking model. Each of the two models, when applied to the corresponding data, allows to overcome a number of problems that typically arise in a traditional statistical setting (e.g., the use of mean and variance statistics at an ordinal scale level, response biases in rating, the dependence of the items in ranking, etc.). To begin with, although the observed data are measured on an ordinal scale, both models provide item parameters which are measured on an interval scale. This happens because, in both models, the parameters are unique up to positive linear transformations with a common multiplicative constant, a sufficient condition for concluding that they are measured on interval scales with a common unit of measurement. Therefore, for instance, both mean and variance are meaningful statistics — that is, invariant up to permissible transformations (see, e.g., Luce, Krantz, Suppes, & Tversky, 1990; Stevens, 1946) — for the model's parameters. Moreover, the dependence among the items in a ranking task is often seen as an undesirable aspect to be removed from the analysis (e.g., Jackson & Alwin, 1980; Ovadia, 2004). In the ranking model, instead, this dependence is an explicit assumption which captures the very essential nature of the task, thus conferring more ecological validity to the data analysis. Furthermore, the property of the rating scale model to separate person parameters from item parameters, allows to control the effect of respondents' possible response biases. Finally, the two models share the important common assumption that item locations are measured on a one-dimensional and continuous latent trait.

The two models were applied in a mixed within- and between-subjects design, with the aim of answering the following two questions: (1) concerning ranking and rating, is there any influence to one task on the other when both of them are accomplished by the same subject? (2) do ranking and rating allow to measure the same latent variable?

Concerning question (1), a systematic comparison between the item parameter estimates in the ranking- and rating-first conditions reveals that there is an influence in both directions but the effect is different. If ranking comes before rating, reliability of the subjects' responses seems better than in the opposite condition and, also, the discrimination among the items in the rating task is improved. It thus seems that ranking acts as a facilitator for rating. On the practical side, this result suggests that, when a subject performs both tasks, ranking should come before rating.

As for question (2), the factor analysis on a single common factor confirms the hypothesis that the responses in both tasks are governed by the same latent trait. In agreement with what Ovadia (2004) argued, this suggests that the same underlying value system is used to accomplish both tasks, which seem to have a role in molding the relationship between latent variable and observable responses. In doing so, ranking seems however to better represent the underlying value system.

Now a question concerning the ability of ranking and rating to represent the latent variable arises. Actually, it cannot be excluded that in other application fields (e.g., quality assessment, marketing research, etc.) or with other types of questionnaires, the differences in representation ability of the two formats may become even more pronounced. It seems that there is room for further investigation along this direction, and this could be helpful in deciding whether, in a specific field, one format is more appropriate than the other.

An additional remark should be made regarding the comparison of individual differences. Our work is restricted to a comparison between the two methods in terms of item properties, that is properties that are invariant at the level of the whole population. Nonetheless, the next step might well be to provide a comparison in terms of individual differences. In this direction, for instance, individual differences are expressed in the rating scale model, as biases that are measured for each person along a unidimensional latent trait. There are also multidimensional approaches in both IRT (Reckase, 2009) and Classical Test Theory (see, e.g., Davison, Kim, & Close, 2009).

Less trivial is the case of ranking data: because of the constraints imposed by the ranking task to the items, individual biases cannot occur. A possible way to examine individual differences when ranking data are concerned could be to consider latent classes of individuals and to obtain item parameter estimates for every single class, which would be regarded as a subpopulation (see, e.g., De Carlo & Luthar, 2000). This route, however, will not be pursued in the present article.

Some limitations of the present study have to be pointed out. First of all, analyses were performed with a limited sample size. It would be worthwhile to extend this research to a larger data set. In the second place, as already stated, it has to be seen how much the results obtained here can be generalized to other application fields. Finally, the possibility of arriving at a joint estimation of the two models by constraining the item parameters to be the same across the rating scale and the ranking models, seems appealing and could be the subject of future research.

NOTE

1. We recall that a linear order on a set A is a binary relation for A which happens to be reflexive, transitive, connected, and antisymmetric.

REFERENCES

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly*, *49*, 535-552.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D., Sheridan, B., Lyne, A., & Luo, G. (2000). *RUMM: A Windows-based item analysis program employing Rasch unidimensional measurement models*. Perth, WA: Murdoch University.

Barylko-Pikielna, N., Matuszewska, I., Jeruszka, M., Kozlowska, K., Brzozowska, A., & Roszkowski, W. (2004). Discriminability and sensory preferences in elderly. *Food Quality and Preferences*, *15*, 167-175.

Baumgartnet, H., & Steenkamp, J. B. E. M. (2001). Response style in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*, 143-156.

Chryssohoidis, G. M., & Krystallis, A. (2005). Organic consumers' personal values research: Testing and validating the list of values (LOV) scale and implementing a value-based segmentation task. *Food Quality and Preference*, *16*, 585-599.

Davison, M. L., Kim, S. E., & Close, C. (2009). Factor analytic modeling of within person variation in score profiles. *Multivariate Behavioral Research*, *44*, 668-687.

De Carlo, L. T., & Luthar, S. S. (2000). Analysis and class validation of a measure of parental values perceived by early adolescents: An application of a latent class model for rankings. *Educational and Psychological Measurement*, *60*, 578-591.
Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26.
Efron, B., & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science*, *253*, 390-395.
Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.
Feather, N. T. (1973). The measurement of values: Effect of different assessment procedures. *Australian Journal of Psychology*, *29*, 221-231.
Gergen, K. J. (1991). *The saturated self: Dilemmas of identity in contemporary life*. New York: Basic Books.
Grimm, S. D., & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, *33*, 414-441.
Grunert, K. G., Grunert, S. C., & Beatty, S. E. (1989). Cross-cultural research on consumer values. *Marketing and Research Today*, *17*, 30-39.
Jackson, D. J., & Alwin, D. F. (1980). The factor analysis of ipsative measures. *Sociological Methods and Research*, *9*, 218-238.
Johnson, M. F., Sallis, J. F., & Hovell, M. F. (1999). Comparison of rated and ranked health and lifestyle values. *American Journal of Health Behaviour*, *23*, 356-367.
Kahle, L. R. (1983). *Social values and social change: Adaptation to life in America*. New York: Praeger.
Kamakura, W. A., & Novak, T. P. (1992). Value-system segmentation: Exploring the meaning of LOV. *Journal of Consumer Research*, *19*, 119-132.
Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Rating, rankings and the measurement of values. *Public Opinion Quarterly*, *52*, 526-538.
Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.
Luce, D. R., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement: Representation, axiomatization, and invariance* (Vol. III). San Diego, CA: Academic Press.
Maio, G. R., Bell, D. W., & Esses, V. M. (1996). Ambivalence and persuasion: The processing of messages about immigrant groups. *Journal of Experimental Social Psychology*, *32*, 513-536.
Marden, J. I. (1995). *Analyzing and modeling rank data*. London: Chapman and Hall.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.
Munson, J. M., & McInyre, S. H. (1979). Developing practical procedures for the measurement of personal values in cross-cultural marketing. *Journal of Marketing Research*, *16*, 48-52.
Ovadia, S. (2004). Ratings and rankings: Reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology*, *7*, 403-414.
Plackett, R. L. (1975). The analysis of permutations. *Applied Statistics*, *24*, 193-202.
Reckase, M. D. (2009). *Multidimensional item response theory*. Heidelberg, Germany: Springer.
Rokeach, M. (1973). *The nature of human values*. New York: Free Press.
Russell, P. A., & Gray, C. D. (1994). Ranking or rating? Some data and their implications for the measurement of evaluative response. *British Journal of Psychology*, *85*, 79-92.
Stevens, S. S. (1946). On the theory of scales measurement. *Science*, *103*, 667-680.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of survey response*. Cambridge, UK: Cambridge University Press.
Van Herk, H., Van de Velden, M. (2007). Insight into the relative merits of rating and ranking in a cross-national context using three-way correspondence analysis. *Food Quality and Preference*, *18*, 1096-1105.
Villanueva, N. D. M., Petenate, A. J., & da Silva, M. A. A. P. (2005). Performance of the hybrid hedonic scale as compared to the traditional hedonic, self-adjusting and ranking scales. *Food Quality and Preference*, *9*, 413-419.