Survey Research Methods (2016)
Vol. 10, No. 1, pp. 15-33
doi:10.18148/srm/2016.v10i1.6209

© *European Survey Research Association*
ISSN 1864-3361
http://www.surveymethods.org

Two of a kind. Similarities between ranking and rating data in measuring work values.

Guy Moors
Tilburg University

Ingrid Vriens
Tilburg University

John P.T.M. Gelissen
Tilburg University

Jeroen K. Vermunt
Tilburg University

The key research question asked in this research is to what extent the respondents' answers to ranking a set of items is mirrored in the response pattern when using rating questions. For example: Do respondents who prefer intrinsic over extrinsic work values in a ranking questionnaire also rate intrinsic values higher than extrinsic values when ratings are used? We adopt a modified version of the form-resistant hypothesis, arguing that each questionnaire mode yields unique features that prevent it from establishing a perfect match between both modes. By adopting a unified latent class model that allows identifying latent class profiles that share a particular preference structure in both question modes, we show that a large portion of respondents tend to identify similar preferences structures in work values regardless of the questionnaire mode used. At the same time the within-subjects design we use is able to answer questions regarding how non-differentiators in a rating assignment react to a ranking assignment in which non-differentiation is excluded by design. Our findings are important since – contrary to popular belief – ranking and ratings do produce results that are more similar than often thought. The practical relevance of our study for secondary data analysts is that our approach provides them with a tool to identify relative preference structures in a given dataset that was asked by rating questions and hence not directly designed to reveal such preferences.

Keywords: Rating and Ranking questions, Survey methodology, Measuring attitudes and values, Latent class analysis, Questionnaire modes

1 Introduction

In survey research the overwhelming mode of asking opinion questions makes use of ratings. Ratings involve respondents indicating the level of agreement, satisfaction or importance with statements. Rankings, on the other hand, are much more rarely used. Rankings imply a respondent to list his or her priorities in a given set of items rather than indicating a level of importance or agreement. In the context of values research it has been debated whether the concept of values reflects absolute evaluations of an individual's values or rather expressing a relative preference of a particular value over others. The absolute evaluation perspective follows from Kluckhohn's idea that values are "conceptions of the desirable" (Parsons & Shils, 1962, p. 405), while the relative preference perspective follows from Rokeach's vision that "a value is an enduring belief that a specific mode of conduct or end-state of existence is personally preferable to an opposite or converse mode" (Rokeach, 1973, p. 5). Measuring values from Kluckhohn's conceptualization implies using ratings, whereas proponents of Rokeach's definition of values prefer rankings. Hence, from a conceptual point of view it is suggested that ratings and rankings would fundamentally measure different things. Admittedly, Rokeach's and Kluckhohn's discussion regarding the meaning of values is ancient but still highly relevant in the field of values research (de Chiusole & Stefanutti, 2011; Klein, Dülmer, Ohr, Quandt, & Rosar, 2004; McCarty & Shrum, 2000; Ovadia, 2004; Van Herk & Van de Velden, 2007).

Regardless of this theoretical view, there has been research that focused on comparing the two questionnaire modes with both proponents for the rating method (Braithwaite & Law, 1985; Maio, Roese, Seligman, & Katz, 1996; Munson & McIntyre, 1979) as well as for the ranking method (de Chiusole & Stefanutti, 2011; Harzing et al., 2009; Krosnick & Alwin, 1988; Miethe, 1985; Van Herk & Van de Velden, 2007). With Jacoby (2011) we believe that, although not stated explicitly in most literature, there is a consensus that the ranking approach is better than the rating approach because the ranking approach is more in accordance with the fundamental idea of the structure of individual values. How-

Contact information: Guy Moors, Tilburg University, School of Social and Behavioral Sciences Department of Methodology and Statistics (Guy.Moors@uvt.nl)

ever, in practice rankings are rarely used mainly for pragmatic reasons since common and acknowledged statistical methods of measurement are not straightforwardly applicable with ranking data. Applied researchers are more familiar with rating questions and hence do not always feel the urge to adopt rankings even if the concept of values refers to this type of measurement.

Most of the studies that compared the rating and ranking methods used a between-subjects split-ballot design. This means that different respondents were randomly assigned to either the rating or ranking method and that these two groups were compared with each other. Undoubtedly very valuable insights are obtained from such an approach. However, an essential question of whether respondents react similarly or differently to ranking versus rating assignments remains unanswered. This central question is: are there (groups of) respondents that react in a similar way to a set of items regardless whether it is asked by means of ratings or rankings? This is the central topic of our research that can most convincingly be answered by adopting an adequate within-subjects design alongside a split-ballot design.

There are a few previous studies that also used the within-subjects design for measuring values using the rating versus the ranking approach (de Chiusole & Stefanutti, 2011; Maio et al., 1996; Moore, 1975; Ovadia, 2004; Van Herk & Van de Velden, 2007). Compared to the design we implemented in this research we observe three disadvantages with respect to these studies. First, the rating and ranking method was applied to questions in the same questionnaire on one time-point only. Therefore, the results of the second question can be influenced by the first question because of recognition of the question. When the ranking task was shown before the rating task, Moore (1975) found that the responses to the rating question were consistently lower. de Chiusole and Stefanutti (2011) found evidence for an improved discrimination in the rating task and a better reliability for both methods, when the ranking preceded the rating task compared to the opposite order. Both these studies demonstrate that responses to a question format are affected by the preceding format used on the same set of items. In this study both question formats are asked on two separate occasions and as such we avoid this crossover effect within one measurement. A second disadvantage with previous within-subjects studies is that they could only compare what happened if the same respondents got a different measurement method at both measurement occasions. None of the previous studies included the same measurement method twice. Measuring the same method twice provides more information on comparing response consistencies. More specifically, how consistent do respondents answer to the same set of items when question format changes compared to when the same format is used on each occasion? Third and finally, with few exceptions (de Chiusole & Stefanutti, 2011; Moore, 1975) these studies did not vary in the ordering of the rating and ranking items. The order in which items are presented in a ranking assignment can have an effect on the choices respondents make. Primacy and recency effects might bias the measurement and make comparison with ratings more difficult to establish (Becker, 1954; Campbell & J., 1950; Fuchs, 2005; Klein et al., 2004; Krosnick, 1992; Krosnick & Alwin, 1988; McClendon, 1986, 1991; Schuman & Presser, 1996; Stern, Dillman, & Smyth, 2007). Ratings on the other hand are vulnerable to non-differentiation (Moore, 1975; Rankin & Grube, 1980). In fact, it was this issue of non-differentiation that initiated Alwin and Krosnick's research (1985) on the form-resistance hypothesis. Controlling for question format specific response biases is hence crucial to any comparison.

In this paper we will overcome the problems of previous within-subjects studies by showing results of a within-subjects comparison of the rating and ranking method by having all four possible combinations (rank-rank, rank-rate, rate-rank, rate-rate) tested on two measurement occasions with two months in between. These design features return in the presentation of results (see Table 4). A novelty of our approach compared to previous research is that we use a latent class choice modeling approach that allows us to distinguish between clusters of cases that share a common preferences pattern in the ranking as well as the rating measurement. Mode specific biases such as primacy effects, in the case of the ranking assignment, and non-differentiation, in the case of the rating assignment are simultaneously modeled. The major benefit of this approach is that it allows identifying latent classes or clusters of respondents that respond similarly to both the ranking and rating task while at the same time defining classes that reveal different ways of responding across modes. Previous research adjusted the ranking data in such a way that established methods for analyzing rating data, i. e. confirmatory factor analyses and structural equation modeling, are applicable. The work of Alwin and Krosnick (1985) is exemplary for this approach. Our approach does exactly the opposite: Rating data are modeled in such a way that the analysis shows relative preferences of particular items compared to others rather than general agreement. A second difference is that we define latent classes rather than latent factors, which is a distinction similar to cluster versus dimensional approach respectively. It is exactly this combination of modeling choices with defining latent classes that reveals clear similarities in response patterns across the two measurement methods that have previously been left unidentified.

In what follows we first take a closer look at the evidence on comparing ratings with rankings from the literature. Then we present the method and our approach in an intuitive way so that even scholars who are not familiar with latent class modeling can appreciate the benefits of our approach. Having some basic notion on logit modeling should be sufficient to understand the method. After describing the setup of our data collection we elaborate on the consecutive analysis indicating how they contribute to researching similarities and differences between ranking and ratings. The two subsamples that received the same format in each method serve as a comparative basis for the subsamples that differed in task on two occasions. Furthermore the former subsamples allow to more formally test for what is known in the literature as testing for measurement invariance (Meredith, 1993). The logic of our series of analyses will become clear as we progress through presenting our approaches and results.

2 Rating versus Ranking

In this research we focus on the issue of work values in which respondents need to either rate or rank a list of items that they consider to be of importance in work. The usual distinction made is between intrinsic and extrinsic work values (Elizur, 1984; Elizur, Borg, Hunt, & Beck, 1991; Furnham, Petrides, Tsaousis, Pappas, & Garrod, 2005; Super, 1962) sometimes complemented with a social dimension (Elizur, 1984; Elizur et al., 1991; Furnham et al., 2005; Kalleberg, 1977; Knoop, 1994; Ros, Schwartz, & Surkiss, 1999; Super, 1962). There are other examples of social concepts that are similar in how a distinction is made between two or more aspects (e. g. intrinsic versus extrinsic) of a global concept (work values) for instance: Inglehart's materialistic versus post-materialistic political values orientations (1977, 1990); Kohn's intrinsic versus extrinsic parental values (1977); Rotter's internal versus external locus of control (1966) – to name some of the classics in the field. All these concepts share one thing: they refer to different – often assumed opposite – aspects of an overarching concept. It is within this context that the question regarding (dis)similarities between ratings and rankings is particular relevant.

Methodological differences between the two measurement methods play an important role in the rating-ranking controversy. The methodological benefits of the rating approach are that rating questions are easy to administer, less time-consuming, can be administered over the telephone, allow identical scoring of items and that they are easier to statistically analyze (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; McCarty & Shrum, 2000; Munson & McIntyre, 1979). A main disadvantage of the rating approach is that it is susceptible of response biases like agreement response style (ARS: tendency to always agree with every item irrespective of the item content) and non-differentiation (tendency to not really differentiate between the items irrespective of the item content) (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988). These response biases may be the consequence of satisficing behavior, which Krosnick and Alwin (1987) define as looking for the first acceptable answer instead of going for the optimal solution. This satisficing behavior leads to a reduced quality of the data.

Contrary to ratings rankings are under appreciated in survey research mainly because of certain disadvantages that are associated with it. Ranking of items is a more cognitive demanding task for the respondents compared to the rating approach, more time-consuming, and less easy to statistically analyze because of the ipsativity of the data (Alwin & Krosnick, 1985). Ipsativity means that the ranking of the items is dependent on one another and therefore traditional statistical techniques are flawed (Jackson & Alwin, 1980). Task difficulty in a ranking assignment may lead to satisficing since choices may be made arbitrarily (Maio et al., 1996). However, previous research on comparing ratings and rankings has shown that the ranking approach gives higher quality and more informative data, higher test-retest and cross-sectional reliability, higher validity of the factor structure, higher discriminate validity and higher correlation validity (Krosnick, 2000; Munson & McIntyre, 1979; Reynolds & Jolly, 1980). Furthermore, since respondents are being forced to discriminate between items satisficing behavior in the form of non-differentiation or acquiescence is excluded by design. All of these advantages of ranking compensate for the major drawbacks of using ratings.

A comparison of rating and ranking methods in previous research showed only limited comparability in measurement between the two methods. Both Maio et al. (1996) and McCarty and Shrum (1997) found that the results of the rating and ranking approach were similar within participants that freely differentiated using the rating approach. Krosnick and Alwin (1988) were able to solve part of the rating-ranking discrepancy by accounting for the level of non-differentiation in ratings and adjusting for ipsativity in the ranking assignment. Other researchers found that the two methods perform equally well in differentiating between extreme items, but the items that are of moderate importance behave different using the two approaches (de Chiusole & Stefanutti, 2011; Van Herk & Van de Velden, 2007). What all these studies have in common is that in the end they indicate that the ranking assignment somewhat arbitrarily forces the conceptually opposite aspects – such as intrinsic versus extrinsic orientation – to be bipolar on a single dimension whereas the rating assignment defines the two aspects as separate – although often negatively related – dimensions. The contribution of our study to the literature is that we take a different look at the same issue that sheds a new light on the alleged bipolarity of two aspects of work values related items. We will show that in both ratings and rankings distinct classes of respondents can be found that clearly assign greater preference to one type of work values over the other and vice versa. We will also show that respondents do this consistently across both methods. As argued before, previous research primarily used a between-subjects design whereas our study includes a within-subjects design as well. Different from previous research in which the ranking data are adjusted in such a way that the methods used with rating data are applicable, we adjust the rating data in such a way that the specific methods to deal with choice data can be applied in a similar way as they are used to model ranking data. The inspiration of this perspective is provided to us from consumer research (Magidson & Vermunt, 2006). As with survey research in consumer research rating questions are the predominant method of data collection. Typical research questions in this field refer to what kind of brand is preferred by which segments of the population (Moors, 2010). Finding an adequate answer to this kind of questions is important since new products are developed toward a targeted population. One major problem with consumer data is that an overall liking tends to dominate the response pattern of respondents when ratings are used (Magidson & Vermunt, 2006). For instance, tasting different brands of cakes and rating their tastefulness is– for most consumer subjects – a pleasant experience skewing the average rating towards positive overall evaluations. The same logic applies to work values: work can be regarded as of crucial importance in the life of (most) people. From this perspective it is far more difficult to find aspects of work as not being valuable than it is to indicate that they are important. As a result scores on rating questions regarding work values tend to be skewed towards positive scale points as well. It is important to not misinterpret the meaning of this "overall liking" or "overall importance" that dominates the response pattern. We do not suggest this reflects a response bias. A tendency towards overall liking or importance is only a response bias if it is independent of the true content that is measured. This is definitely not the case with expressing an overall liking in tasting goods, nor with feeling that work is generally important and by consequence also its different aspects.

There have been attempts to use within-case "centering" as a solution to eliminate the overall response tendency in a set of rating items (Cattell, 1944; Cunningham, Cunningham, & Green, 1977). This involves subtracting the within-case mean score in a set of items from each observed score of each item and analysing these transformed data. This approach has been criticized from a statistical point of view since it creates ipsative data (Cheung, 2006; Cheung & Chan, 2002; Dunlap & Cornwell, 1994). This means that data on different items are not observed independently of each other. More specifically, within-case centering implies that the sum of all items scores in the set is fixed to the constant value of zero. Most statistical models require independent data though and hence are not applicable in a straightforward manner. A model that overcomes the shortcomings of within-case centering has been proposed by Magidson and Vermunt (2006) who demonstrated the usefulness of a latent class ordinal regression model with random intercept in identifying latent class segments in a population that differ in their preference structure of tasting crackers. Moors (2010) has demonstrated that this approach works well whenever a researcher's aim is to construct a latent class typology of respondents with survey data on locus of control, gender roles and civil morality. This model reflects methods developed to model sequential choice processes (Böckenholt, 2002; Croon, 1989; Kamakura, Wedel, & Agrawal, 1994; Vermunt & Magidson, 2005). Sequential choice modelling implies the analysis of ranking data in which a first choice is made out of K alternatives and each consecutive choice as a choice made out of K minus the alternative in the previous step. This model hence requires data to be ipsative. In the following sections we elaborate on these methods used in our research. To the best of our knowledge, this research is the first attempt to compare rating and ranking questions using methods developed to analyse ipsative (ranking) or ipsatized (rating) data and compare its outcome in a within-subjects design. We do not adopt this approach for the sole sake of its "novelty" but because it does allow us to identify segments in a sample whose work values preferences are similar regardless whether ratings or rankings are used. In what follows we explain the method in some detail, describe our data and the sequence of analyses we conducted to investigate (dis)similarity in work values preferences across measurement mode as well as the stability in preference structures both within and between modes. The logic of this sequential analysis will be explained in the process of presenting the setup of each part of the research.

3 Latent Class Choice Modeling of Ranking and Rating Data

Lazarsfeld (1950) was the first to introduce latent class analysis as a tool to build typologies based on dichotomous observed variables and Goodman (1974) extended it for polytomous manifest variables. Current software development (e. g. Mplus, Latent Gold, lEM) has made the method accessible to applied researchers. Most readers thus probably have some intuitive understanding of the classical latent class model. Probably the best way of giving latent class analysis an intuitive meaning is by reference to cluster analysis. The principal aim of latent class as well as cluster analysis is to identify classes or clusters of cases that are similar in the manifest variables. The current research makes use of the generalized framework that latent class analysis has provided to deal with choice data that are typically provided with a ranking assignment, i. e. the latent class choice model for ranking data. Furthermore, by adopting a latent class regression model with random intercept, choice preferences in a rating assignment can also be revealed. In this section we elaborate on these two models and explain how the within-subjects comparison is modeled.

3.1 Latent Class Choice Model for Ranking Data

The model used for the ranking data in the current study is the Latent Class Choice (LCC) model. This model is based on the work of McFadden (1986) and makes it possible to model the actual choice process (Croon, 1989; Vermunt & Magidson, 2005). We use a partial ranking approach in which respondents needed to rank their top 3 most important work values and the least important one out of j items. Let item a_1 be the item that was chosen as the most important one, a_2 as the second most important, a_3 as the third most important and $a_{(-1)}$ as the least important item selected by a respondent. Making the assumption that the successive choices are made independently of one another, the probability of this response pattern (a_1, a_2, a_3, $a_{(-1)}$) equals:

$$P(a_1, a_2, a_3, a_{(-1)}) = P(a_1) \cdot P(a_2 | a_1) \cdot P(a_3 | a_1 a_2) \cdot P(a_{(-1)} | a_1 a_2 a_3) \quad (1)$$

This means that the probability of the response pattern is a product of the probability of selecting item a_1 out of the full list of j items, times the probability of selecting item a_2 out of $j-1$ items given that item a_1 was already chosen, times the probability of selecting item a_3 out of the remaining $j-2$ items given that items a_1 and a_2 were already selected, times the probability of selecting $a_{(-1)}$ as the least favorite item out of the remaining $j-3$ items given that items a_1, a_2 and a_3 were chosen already. Next, we follow the random utility model in which we are able to estimate a utility μ_{a_j} for each item. A higher utility for one item in comparison with another item means that this item has a higher ranking (Allison & Christakis, 1994). Using a logit model to determine the response pattern shown above, the equation becomes:

$$P(a_1, a_2, a_3, a_{(-1)}) = \frac{\exp(\mu_{a_1})}{\sum_T \exp(\mu_{a_t})} \cdot \frac{\exp(\mu_{a_2})}{\sum_S \exp(\mu_{a_s})} \cdot \frac{\exp(\mu_{a_3})}{\sum_R \exp(\mu_{a_r})} \cdot \frac{\exp\left(-\mu_{a_{(-1)}}\right)}{\sum_Q \exp\left(-\mu_{a_q}\right)} \quad (2)$$

The value μ_{a_j} is the degree to which item a_j is being preferred over all other items by a respondent. T equals the full set of items, S is the remaining set of $j-1$ items (minus the alternative chosen first), R is the remaining set items minus the alternatives selected first and second, and Q is the item set minus the items ranked as top 3 most important items. The item that was chosen as the least favorite one ($a_{(-1)}$) is negatively related to the utility of the item. This was made possible by including scale weights which could have a value of +1 when an item was chosen as the top 3 most important versus -1 when an item was chosen as the least important one. Taking the exponent of μ_{a_j}, the odds is determined that an item is being chosen out of a set of possible alternatives.

In the current application we are interested in applying a latent class analysis in which respondents are being clustered that have a similar value preference structure. Thus, each group (latent class) of respondents has its own value for the utilities. Using the LCC model, different utilities can be estimated for different latent classes (Magidson, Eagle, & Vermunt, 2003; McFadden & Train, 2000). Equation 2 needs to be slightly changed to account for the differences between the latent classes and becomes:

$$P(a_1, a_2, a_3, a_{(-1)} | X = c) = \frac{\exp(\mu_{a_1 c})}{\sum_T \exp(\mu_{a_t c})} \cdot \frac{\exp(\mu_{a_2 c})}{\sum_S \exp(\mu_{a_s c})} \cdot \frac{\exp(\mu_{a_3 c})}{\sum_R \exp(\mu_{a_r c})} \cdot \frac{\exp\left(-\mu_{a_{(-1)} c}\right)}{\sum_Q \exp\left(-\mu_{a_q c}\right)} \quad (3)$$

in which X is the discrete latent variable and c is a particular latent class. The higher the value of μ_{ac}, the higher the probability that a respondent belonging to latent class c selects alternative a as one of the most important items.

In the current study we will model the utilities based on the following formula:

$$\mu_{ac} = \alpha_a + \beta_{ac} \quad (4)$$

(see also: Moors & Vermunt, 2007). Effect coding is used for identification purposes, and therefore intercept parameter α_a can be seen as the average utility of item a and slope parameter β_a as the deviation from the average utility for respondents belonging to latent class c. A positive β_{ac} value means that respondents belonging to latent class c have a higher probability than average of choosing item a as one of the most important items. Since the β_{ac} values are estimated relative to the average utility, the sum of all β_{ac} values within a latent class equals zero.

Last, we are also interested in the presence of a response order effect. A response order effect is present when items that are shown as one of the first or last alternatives in the list of items have a higher probability of being chosen than one of the more important items, irrespective of the actual content. In this research we present two alternative orderings of items in a split-ballot design (see later). Since the placement of the items is the same for the respondents in each subsample, the response order effect is also forced to be the same for all respondents. This means that it is an alternative-specific trait and modeled as such as an attribute of alternative. Equation 4 needs to be extended to be able to model the response order effect and becomes:

$$\mu_{acz} = \alpha_a + \beta_{ac} + \beta_z z \quad (5)$$

Let z be the response order effect indicator (takes on the value 1 for the items presented first or last in the list, and 0 otherwise) and β_z the effect of this attribute of choice. Thus, when a response order effect is present, it can be accounted for by adding β_z to the utility of the items that may be affected by a response order effect.

3.2 Latent Class Regression Model with Random Intercept for Rating Data

The main interest in the current study is to be able to compare the results from ranking data with the results from rating data. Therefore, a model is chosen for the rating data that allows controlling for the overall agreement level and estimate latent classes that differ in their relative ratings of particular items compared to other items in the set. This model is called the latent class regression model with random intercept. The inclusion of a random intercept in this regression model makes it possible to control for the overall level of agreement or importance (Magidson & Vermunt, 2006; Moors, 2010). Specifically, with the random intercept the average agreement across rating items is modeled as it varies across respondents. The latent class regression coefficients will then indicate relative – as opposed to absolute – differences in importance between the items. In this research we are particularly interested in the relative preference information because this information is similar to the relative preferences obtained by using the ranking method.

As indicated before the latent class regression model with random intercept is a model-based alternative to within-case centering (Magidson & Vermunt, 2006; Moors, 2010). The benefit of using the model-based approach is that the original ordinal measurement level of the rating data is being maintained (Magidson & Vermunt, 2006) and it suits the analysis of ipsatized data.

Let Y_{ij} be the rating of respondent i of item j and let m be the discrete values of the rating Y_{ij}. Since the rating is a discrete (ordinal) response variable, an adjacent-category logit model is being defined as follows:

$$\log\left(\frac{P\left(Y_{ij}=m|c\right)}{P\left(Y_{ij}=m-1|c\right)}\right)=\alpha_{im}+\beta_{cj}=\alpha_m+\lambda F_i+\beta_{cj} \quad (6)$$

This is a regression model for the logit of giving rating m instead of $m-1$ for item j conditional on belonging to latent class c. α_{im} is the intercept which is allowed to differ over individuals and is a function of the intercept's expected value (α_m) and a continuous factor (F_i) which is normally distributed and has a factor loading equal to λ. β_{cj} is the effect of item j for latent class c. For the identification of the parameters effect coding is used, which leads to a sum of zero for the α_m parameters over the possible ratings and to a sum of zero for the β_{cj} parameters over items and classes. A positive value for β_{cj} indicates that respondents belonging to latent class c value an item as more important than average. Thus, α_{im} accounts for the overall importance/agreement level and β_{cj} gives an indication of the relative preference of an item in comparison with the average importance level.

Last, it is also possible to control for a response order effect in rating items. Again, the response order effect is modeled as an attribute of choice, which is choice-specific meaning that it has the same effect over all individuals. Extending equation 6 to account for a response order effect, the formula becomes:

$$\begin{aligned}\log\left(\frac{P\left(Y_{ij}=m|c,z\right)}{P\left(Y_{ij}=m-1|c,z\right)}\right)&=\alpha_{im}+\beta_{cj}+\beta_z z_j\\&=\alpha_m+\lambda F_i+\beta_{cj}+\beta_z z_j \quad (7)\end{aligned}$$

The z_j parameter indicates whether items were presented first or last in the item list and β_z is the effect of this attribute on the respondents' ratings. This term is only needed when a response order effect is found to be present. A requirement for identication of the order effect is that (at least) two randomly assigned subsamples receive alternative orderings of the set of items.

3.3 Comparing Latent Class Assignments

In both models it is possible to assign respondents to particular classes based on their posterior membership probabilities. These probabilities then are the input for subsequent analyses in which the association between repeated measurements is investigated. We make use of a recently developed approach to adequately estimate associations in a three-step design (Bakk, Tekle, & Vermunt, 2013; Vermunt, 2010). These three steps include: (1) estimating a measurement model (as presented in section 2.1 and 2.2); then (2) obtaining class assignments and adding these as new variables to the dataset; and then (3) estimating associations between the class memberships using these class assignments. It has been shown that outcomes from the latter analysis may lead to severely downward-biased estimates of the associations (Bolck, Croon, & Hagenaars, 2004). In this research we make use of the correction method as proposed by Vermunt (2010). In the current study proportional assignment will be used as classification method, which means that respondents are treated as belonging to each of the latent classes with a weight equal to the posterior membership probability. The adjustment method that is used is the maximum likelihood (ML) method which is the preferred option for most situations (Vermunt & Magidson, 2013).

Assume that X is the latent variable, c is a particular latent class and y is a particular response pattern. The posterior class membership probabilities can be estimated using the following formula:

$$P(X=c|Y=y)=\frac{P(X=c)\,P(Y=y|X=c)}{P(Y=y)} \quad (8)$$

This means that the probability of belonging to a certain latent class conditional on a respondent's response pattern can be calculated by multiplying the latent class proportions $P(X=c)$ with the class-specific response probabilities $P(Y=y|X=c)$ and then dividing this multiplication by the probability of having a certain response pattern

$P(Y = y)$. The proportional assignment to each of the latent classes d equals the posterior membership probabilities $P(W = d|Y = y) = P(X = c|Y = y)$. The proportional assignment values are used in step 3 of the stepwise approach in which we investigate the associations between measurements across occasions. This is the main interest of our study. We want to know the consistency in results when alternative measurement methods (ratings versus rankings) are presented to the respondents. Results from the same method subsamples will serve as a comparative basis. In the next section we present our between- and within-subjects design in detail.

4 Design

To collect our data, we made use of the LISS (Longitudinal Internet Studies for the Social Sciences) panel administered by CentERdata. This panel is a probability-based internet panel that participates in monthly internet surveys. The LISS panel is based on a true probability sample of households drawn from the population register in the Netherlands in 2007. Households that did not have the materials to participate, like a computer or internet access, were provided with these materials. The questionnaire used in the current study was implemented in a small experiment in the summer of 2012. Since a between- and within-subjects design was used, we had two time-points at which the questionnaire was administered. The first measurement took place in June and July and the second measurement in September and October. The time between the two measurements was at least two months for all respondents. The first questionnaire was sent to 7425 panel members, aged between 16 and 92, of which 5899 responded (response rate of 79.4%). For the second measurement the questionnaire was distributed among 5697 of these respondents. 5492 of them filled in the questionnaire (response rate of 96.4%).

Since we are comparing rating and ranking methods, the sample was a priori randomly divided into subsamples. This division led to a subsample of 1675 respondents who received the ranking questionnaire twice (subsample 1), 1035 who received first the ranking and then the rating questionnaire (subsample 2), 1104 who received first the rating then the ranking questionnaire (subsample 3), and 1678 respondents that received the rating questionnaire twice (subsample 4). One panel member for the rank-rank condition was excluded because this respondent did not completely fill in the questionnaire and one panel member for the rank-rate condition was excluded from the subsample because this respondent did not respond at the first measurement occasion.

To measure work values, a survey question from the European Values Study (EVS) 2008 was used in which respondents needed to indicate the importance of 17 job aspects. The items given to the respondents (see Table 1) were similar to items used in previous work values research (Elizur et al., 1991; Furnham et al., 2005; Knoop, 1994; Ros et al., 1999). The question from the EVS was transformed for the current application into a rating task and a partial ranking task. For the rating task a 5-point scale was used with only labels for the endpoints. The rating questionnaire was set up in such a way that the items had to be rated from top to bottom. Altering an answer to an item was not possible after a respondent rated the next item. In the ranking task, respondents were asked to indicate their top 3 most important items and the item that was least important to them personally out of the full list of items. Once an item was chosen as the most important one and the respondent went to the next page, which contained the next question, the chosen item was dropped out of the list of possible items to select. This means that each item could be chosen only once. Also, respondents were able to choose only one item in each of the ranking tasks. See the bottom part of Table 1 for the rating and ranking question formats that were used.

To be able to detect a response order effect in both ranking and rating data, different orderings of the questionnaire in a split-ballot experiment were needed. Respondents were randomly assigned to either version A or version B of the questionnaire. In version A the items were shown to the respondents in the same order as the items are ordered in Table 1 (see also the numbers that are placed in front of the item names). In version B of the questionnaire the item set was split in half (see the dotted line in Table 1) and then the order of the items was reversed for each half (see also the numbering behind each item in Table 1).This approach differs from previous studies, in which the items are shown in a simply reversed order. The main reason why items from the middle of the list (version A) are presented at the beginning or end of the alternative list (version B) is that it makes it possible to research primacy or recency response order effects in case they would occur at the same time. With simple reversed ordering this would not be possible.

5 Results

5.1 Preliminary Analyses

Before presenting the main results of our study we briefly summarize the results from preliminary analyses. The preliminary analyses involved: (a) investigating whether response order effects need to be taken into account, and (b) deciding on the number of latent classes in each measurement mode. Detailed information on these model selection procedures have been reported in previous studies that analyzed the first wave (Vriens, Moors, Gelissen, & Vermunt, 2015) and repeated with data from the second wave (Vriens, 2015, pp. 45-68). We found evidence of primacy effects in the ranking assignment that affected the measurement of work values. This was not the case with rating data. Hence, primacy is accounted for in the measurement model using

Table 1
Questionnaire design

Ordering of job aspect items in two experimental conditions		
Version A		Version B
(1)	Good pay	(9)
(2)	Pleasant people to work with	(8)
(3)	Not too much pressure	(7)
(4)	Good job security	(6)
(5)	Good hours	(5)
(6)	An opportunity to use initiative	(4)
(7)	A useful job for society	(3)
(8)	Generous holidays	(2)
(9)	Meeting people	(1)
(10)	A job in which you feel you can achieve something	(17)
(11)	A responsible job	(16)
(12)	A job that is interesting	(15)
(13)	A job that meets one´s abilities	(14)
(14)	Learning new skills	(13)
(15)	Family friendly	(12)
(16)	Have a say in important decisions	(11)
(17)	People treated equally at the workplace	(10)
Question format ranking		
(a)	Here are some aspects of a job that people say are important. The question is which of these you personally think is the most important in a job?	
(b)	Of the remaining aspects of a job, which one do you consider next most important?	
(c)	Of the remaining aspects of a job, which one do you then consider next most important?	
(d)	And which one of the remaining aspects do you consider least important of all?	
Question format rating		
	Here are some aspects of a job that people say are important: How important is each of these to you personally?	
	1 “Very unimportant” 2 3 4 5 “Very important”	

ranking data. We also found that a three-class model represented the data adequately in the ranking assignment whereas a four-class model is preferred with rating data. This choice depended on methodological criteria (fit statistics) as well as theoretical interpretation of the results for each model. The results of these models are in accordance with work values literature in which two main types of work values are being distinguished, namely intrinsic and extrinsic work values, sometimes complemented with a third social work values dimension (Elizur, 1984; Elizur et al., 1991; Furnham et al., 2005; Kalleberg, 1977; Knoop, 1994; Ros et al., 1999; Super, 1962). The latter is not always observed consistently in the literature. The extra latent class for the rating data consists of the non-differentiating respondents. In what follows we start by comparing the parameter estimates for similar latent classes found at each time-point and for each measurement method. Then we will show the results of investigating the association between the proportional latent class assignments to each of the latent classes at the two measurement occasions. All the analyses reported in the next sections are estimated using Latent Gold version 5.1.

5.2 Latent Class Comparisons

The results for the latent class segment analyses are shown in Table 2 and 3. Table 2 includes the findings on the first (T1) and second (T2) ranking measurements. In Table 3 we present findings on the two waves of rating questions. The final two columns in each of these tables contrast the effect sizes of the intrinsic versus extrinsic latent class found in both the ranking and rating data. To interpret the results the following characteristics of the estimates need to be kept in mind:

1. Column wise the parameter estimates sum to zero. Positive values indicate higher than average preference for the items in the given set of items when rankings are used. When ratings are used positive values indicate a higher than average rating relative to the overall rating of items, the latter which is measured by the random intercept. Negative values mean the opposite.

2. If one wants to assign a meaning to the different latent classes, one should compare results row wise. An item may be ranked or rated highly (positive parameter estimates) in each latent class but with different magnitude across classes. For example: in the first analysis (ranking 3-Class model T1) "Meeting one's abilities" has higher than average (positive) rankings in each latent class but clearly highest in the first "intrinsic" latent class (b = 1.763) and least in the second "extrinsic" latent class (b = 0.255). "Pleasant people to work with" is also an item with positive estimates across latent classes but is highest on the second "extrinsic" latent class (b = 2.051) and lowest on the first "intrinsic" latent class (b = 0.892). Although both items have higher than average preferences, "meeting one's abilities" contributes to identifying a more intrinsically oriented latent class whereas "pleasant people to work with" contributes to defining the second latent class as extrinsically oriented.

3. To facilitate interpretation we regrouped items into three categories. The top 7 items are linked to intrinsic work values, the bottom 6 items refer to extrinsic work values and the remaining 4 items in the middle differ in meaning depending on the analysis. This regrouping is based on our empirical findings but is at the same time consistent with theoretical conceptualization.

4. Each of the analyses includes different subsamples. There were four subsamples coinciding with the four test conditions: "rank-rank" (subsample 1), "rank-rate" (subsample 2), "rate-rank" (subsample 3), and "rate-rate" (subsample 4). Respondents were randomly assigned to one of these four test conditions.

Reading the table it can be seen that the intrinsic and extrinsic work values class is consistently observed across measurement method (rankings and ratings) and across occasions (first and second wave). The third latent class in the ranking assignment can be linked to social work values and is observed consistently in both first and second measurement. In the rating assignment the four items grouped in the middle do not define a particular latent class, although the third class seems to put greater emphasize on the "people" items. The fourth and last class identifies a class of respondents that reveal little differentiation (small parameter estimates and few significant differences) in their ratings and thus can be regarded as non-differentiators.

Having a closer look at what items have their highest relative parameter estimate across classes thus reveals the meaning that can be given to each latent class. The first latent class present in both the rating and ranking method is the intrinsic work values class. Tables 2 and 3 show that the items "a job that meets one's abilities", "a responsible job", "a job that is interesting", "a job in which you can achieve something", "have a say in important decisions", "an opportunity to use initiative" and "learning new skills" all have the highest probability to be preferred by the respondents belonging to this class. The item "have a say in important decisions" shows a slightly deviant result for the rating approach at time-point 1 since the parameter value associated with the fourth "non-differentiation" latent class (−0.192) is marginally higher than with the first latent class (−0.219). The difference is too small to interpret this particular result as contradiction, especially not since its value on the intrinsic class is in contrast with the very low value on the extrinsic class. Hence it is safe to conclude that the overall pattern of the ranking and rating method in identifying an intrinsic work values class is quite similar.

Although tempting, it is dangerous to compare the magnitude of the estimated parameters across occasions for two reasons. First of all results from first and second measurement refer to different samples and second it is not formally tested whether observed differences are meaningful from a statistical point of view. For instance, we observe that effect parameters tend to be (slightly) higher for most of the intrinsic items on the second occasion. Furthermore class sizes also differ from first to second measurement which might merely be an artifact of differences in measurement model. We tested measurement invariance in the case of subgroups 1 and 4 that had repeated measures at two waves and found that the differences between first and second measurement are statistically not significant. Furthermore, differences in class sizes reduced when the same measurement model was applied on first and second wave. These additional analyses provide evidence that measurement did not depend on measurement occasion.

The second latent class is labeled as the extrinsic work values class. The items "generous holidays", "pleasant people to work with", "good pay", "good job security", "good hours" and "not too much pressure" are all the most preferred by respondents belonging to this latent class. All of these items refer to benefits beyond the content of the job itself. One could think of the item "pleasant people to work with" as symbolizing a social aspect as well, but in all four analyses its highest observed effect parameter is linked to the extrinsic class. Keep in mind that overall it is a very popular item, but most popular amongst the extrinsically motivated. Having

Table 2
Comparison of the parameter estimates for the work values classes for the ranking and rating latent class models (Ranking, 3-Class model)

	T1			T2			(I) - (E)	
	Intrinsic (I)	Extrinsic (E)	Social	Intrinsic (I)	Extrinsic (E)	Social	T1	T2
Meeting abilities	**1.763***	0.255*	0.908*	**2.144***	0.474*	0.869*	1.508	1.670
Responsible job	**0.333***	−0.887*	−0.573*	**0.675***	−0.959*	−0.791*	1.220	1.634
Interesting	**1.210***	0.130	−0.612*	**1.033***	0.364*	−0.264*	1.080	0.669
Achieve something	**0.548***	−0.514*	−0.162	**0.814***	−0.504*	−0.594*	1.062	1.318
Have a say	**−0.557***	−1.578*	−0.955*	**−0.103**	−1.259*	−1.272*	1.021	1.156
Use initiative	**0.082**	−0.379*	−0.157	**0.233***	−0.381*	−0.131	0.461	0.614
Learn new skills	**−0.184***	−0.458*	−0.267*	**−0.126**	−0.426*	−0.239*	0.274	0.300
Useful for society	0.098	−1.137*	**0.659***	−0.100	−0.987*	**0.842***	1.235	0.887
Meeting people	0.352*	−0.275*	**1.163***	0.345*	−0.361*	**1.446***	0.627	0.706
People equally treated	0.208*	0.367*	**1.229***	0.066	0.316*	**1.041***	−0.159	−0.250
Family friendly	−1.568*	−1.178*	**0.097**	−1.638*	−0.929*	**−0.429***	−0.390	−0.709
Holidays	−1.302*	**−0.884***	−1.207*	−1.731*	**−0.803***	−1.245*	−0.418	−0.928
Pleasant people	0.892*	**2.051***	1.310*	0.874*	**1.952***	1.773*	−1.159	−1.078
Pay	1.139*	**2.279***	−1.023*	0.947*	**2.164***	−0.777*	−1.140	−1.217
Job security	−0.392*	**0.944***	−0.859*	−0.470*	**0.788***	−0.432*	−1.336	−1.258
Good hours	−0.810*	**1.112***	0.310*	−0.844*	**0.849***	0.184	−1.922	−1.693
No pressure	−1.812*	**0.152**	0.140	−2.119*	−0.298*	**0.019**	−1.964	−1.821
Class size (proportion)	0.402	0.409	0.189	0.306	0.443	0.251		
Subsamples		(1) + (2)			(1) + (3)			

Values in bold indicate for each item the highest preference value over all latent classes in each model
* $p < 0.05$

Table 3

Comparison of the parameter estimates for the work values classes for the ranking and rating latent class models (Rating, 4-Class model)

	T1				T2				(I) - (E)	
	Intrinsic (I)	Extrinsic (E)	People	Non-diff	Intrinsic (I)	Extrinsic (E)	People	Non-diff	T1	T2
Meeting abilities	**1.279***	−0.136*	0.529*	0.349*	**1.459***	−0.109*	1.104*	0.134	1.415	1.568
Responsible job	**0.286***	−1.357*	−0.753*	−0.239*	**0.369***	−1.371*	−0.318*	−0.379*	1.643	1.740
Interesting	**1.168***	−0.331*	0.264*	0.252*	**1.203***	−0.428*	0.854*	−0.030	1.499	1.631
Achieve something	**0.332***	−0.771*	−0.226*	0.032	**0.470***	−0.696*	0.067	−0.196*	1.103	1.166
Have a say	−0.219*	−1.228*	−1.066*	**−0.192***	**−0.064**	−1.203*	−0.679*	−0.329*	1.009	1.139
Use initiative	**0.660***	−0.180*	0.266*	0.135*	**0.667***	−0.287*	0.438*	−0.002	0.840	0.954
Learn new skills	**0.437***	−0.397*	0.036	0.169*	**0.355***	−0.445*	0.228*	−0.011	0.834	0.800
Useful for society	−0.579*	−0.599*	−0.429*	**−0.279***	−0.497*	−0.403*	−0.657*	**−0.255***	0.020	−0.094
Meeting people	0.025	−0.237*	**0.279***	−0.158*	0.128	0.256*	**0.286***	−0.181*	0.262	−0.128
People equally treated	0.910*	1.178*	**1.218***	0.399*	0.597*	1.340*	**1.459***	0.501*	−0.268	−0.743
Family friendly	−1.144*	−0.407*	−0.817*	**−0.262***	−0.999*	−0.309*	−1.185*	**−0.076**	−0.737	−0.690
Holidays	−1.372*	**−0.205***	−0.901*	−0.284*	−1.143*	−0.233*	−1.157*	**−0.243***	−1.167	−0.910
Pleasant people	0.965*	**1.700***	1.612*	0.341*	0.435*	**1.451***	1.274*	0.723*	−0.735	−1.016
Pay	0.018	**0.775***	0.431*	0.136*	−0.117*	**0.632***	0.327*	0.240*	−0.757	−0.749
Job security	−0.725*	**0.958***	0.392*	−0.039	−0.597*	**1.046***	−0.233*	0.063	−1.683	−1.643
Good hours	−0.675*	**0.838***	0.188*	0.062	−0.535*	**0.623***	−0.213*	0.151*	−1.513	−1.158
No pressure	−2.002*	**0.399***	−1.023*	−0.423*	−1.731*	**0.136***	−1.595*	−0.111	-2.401	−1.867
Class size (proportion)	0.161	0.310	0.288	0.133	0.225	0.386	0.285	0.103		
Subsamples		(3) + (4)				(2) + (4)				

Values in bold indicate for each item the highest preference value over all latent classes in each model

* $p < 0.05$

“pleasant people to work with” is evidently linked to a job – just like any of the items listed – but it is not an inherent aspect of the job as such. That is why it is less prominent – although still relatively important – among the intrinsically oriented respondents.

Whereas the comparison across occasions within each measurement mode produces very similar results, the comparison across measurement modes reveals some specific findings for some of the extrinsic items. Most noticeable is the large effect parameter observed for “good pay” in the ranking assignment, which is smaller in the rating task. It is still consistent with the theoretical expectation that this would be part of the extrinsic qualification of work values, but the difference is pronounced. “Good pay” is clearly ranked highest in the ranking data but not in the rating data. Keep in mind that this result is after controlling for primacy in the ranking assignment and controlling for overall agreement in the rating task. Hence location of the item in the set is most likely not a prime reason. On the “why” of this finding we can only speculate. One plausible reason might be that when ranking work values is concerned it is socially acceptable to rank it among the top three items. After all, who does not work ‘for a living’? In a rating task respondents might be more reluctant to rate its importance higher than other job values since it is less socially desirable. Hence, when “good pay” is rated highly other related extrinsic work values will be rated equally high. We have no means of checking socially desirable responding in both methods (other than overall agreement and primacy that are included in the model) with the current dataset. A second difference between ratings and rankings is observed in the case of “generous holidays”. In the ranking data it is linked to extrinsic values; in the rating task both the extrinsic and the non-differentiation class assign similar importance to this issue. Regardless of these particularities, the contrast of the effect parameters of the extrinsic values on the second latent class compared to their estimated effect on the intrinsic latent class is pronounced. The reverse is true for the intrinsic items. This is highlighted in the two last columns in which we report the differences between the estimated effect parameter for the intrinsic latent class and the corresponding parameter for the extrinsic latent class. These differences indicate the increase in the logit of preferring the particular item when going from the extrinsic to the intrinsic latent class and define a contrast in preference of items. These contrast values are very similar across occasions (T1 and T2) and across measurements (ratings and rankings). Hence it is safe to conclude that these two classes have a distinct view on the intrinsic versus extrinsic work values inventory.

The remaining classes are difficult to compare across measurement methods since they are divergent. Part of this is – of course – by very nature of the measurement method itself. A non-differentiation class can only be observed in a rating assignment. In a ranking assignment the potential non-differentiators are forced to make their choices. How non-differentiators react to a ranking assignment is a key topic in the following section. It is the fourth latent class in the rating task that can be labeled as the class of non-differentiators since its effect sizes are small or even not significantly different from average. Even the “higher” positive scores observed such as for “pleasant people to work with” and “people treated equally at the workplace” are still the lowest observed scores for these two very popular items in the rating assignment. Only the items “a useful job for society” and “family friendly” score relatively higher than in other classes but the small negative values observed indicate their low overall preference in all classes. A content label can be assigned to the third latent class in the rating and the ranking assignment but the label is different. For the ranking approach the items “a useful job for society”, “meeting people”, “people treated equally at the workplace” and “family friendly” are the most preferred by respondents belonging to this third class and therefore we called this class the social work values class. In the rating approach only the items “meeting people” and especially “people treated equally at the workplace” are more preferred by respondents belonging to the third class and therefore this class receives an adjusted label in which social is restricted to other people (not society or the own family). Within each measurement method the results from first and second measurement are highly similar which was confirmed when testing measurement invariance in case of the two subsamples that had the same measurement mode across waves.

5.3 Two of a Kind: Similarities between Ranking and Rating Data in Classifications into Work Values Profiles

Using particular methods to model choice preferences in ranking and rating data revealed similar latent class profiles as far as intrinsic and extrinsic work values are concerned, irrespective of whether rating or ranking questions were used. That was the key finding reported in the previous section. Now the question is: To what extent will respondents be classified in the same latent class when alternative measurement methods are used on two different occasions?

In our design we defined four subsamples. Two of these subsamples received the same measurement method at both occasions and two subsamples received different methods. The inclusion of two subsamples that were measured twice with the same instrument is used as a kind of standard to the comparison of similarity in classification when different methods are used on both occasions. After all, even when the same measurement is used we can hardly expect perfect correspondence between two measures. Random error causes variation. Furthermore, in our dataset there is a time-lag of two months between first and second measurement.

Although it is hard to imagine why "true" work values orientations would change in a short framework of only two months we cannot exclude the possibility of a true change in orientation. Consequently, if we want to evaluate the consistency in classification across measurement methods we need to compare it with a cross-classification when the same measurement method is used.

Input for Table 4 were the saved posterior membership probabilities from the four separate analyses per subsample. As mentioned in section 3.3 we used the three-step approach to adequately estimate associations between the two waves. A table with the estimated parameters is presented in appendix A. In this section we report the estimated values that indicate the cell percentages in the T1 by T2 table. We first elaborate on the two subsamples (Tables 4.1 and 4.2) that were administered the same measurement method. Columns in this case refer to the classification from the second measurement and rows to the classification from the first measurement. The two subsamples that changed measurement method (Tables 4.3 and 4.4) differ in the order of which each method was administered. To facilitate comparison, the row variable refers to the ranking assignment and the column variable to the rating task in both tables.

Values presented in the tables are observed cell percentages and their residuals that indicate the deviation compared to the expected cell percentages with statistical independence. The reason why we present these cell percentages is that it allows us to make comparisons with both marginal distributions (column and row total percentages). This is necessary for two reasons. First and foremost because our primary interest is in the comparison of classification into the intrinsic and extrinsic work values class in each subsample. Both methods have an unequal number of latent classes and each cell percentage should be compared to its lowest observed marginal that defines the highest possible percentage within each cell. Second, relative class sizes differ between T1 and T2. In each of the four subsamples we observe that the number of respondents classified in the intrinsic class decreases whereas the number of extrinsic classified respondents increases. To evaluate whether the percentage of respondents classified into the intrinsic class is in a fixed ratio with the percentage in the extrinsic class we need to take this T1-T2 shift into account.

The cross-classification of respondents in the same latent class across repeated measures (Tables 4.1 and 4.2) is highly consistent. Row and column marginal percentages differ, which can be either due to differences in measurement between T1 and T2 or due to change in time. The maximum cell percentage possible is thus limited to the smallest corresponding row or column marginal.

Subsample 1 who received the ranking assignment twice shows a cell percentage of 31.5% that is classified in the intrinsic class on both occasions. This value is almost the same as the corresponding column marginal of 31.9% and also close to the 39.8% row value. Similarly, the cell percentage of 36.7% in the extrinsic class on both waves is highly similar to the respective column (43.6%) and row (40.8%) marginal. A similar observation is made in case of the third social latent class. The large positive residuals on the diagonal of consistent classifications confirm this interpretation.

The rating latent class model includes 4 latent classes. Here, cross-classification (subsample 4) is a little bit more diffuse than in the case of the ranking assignment. The larger positive residuals on the diagonal of consistent classifications are still observed but are less pronounced in the case of the third "people" latent class and in particular when looking at the classification of the non-differentiators. The consistent scoring on the intrinsic and extrinsic latent class, however, is again clearly observed. 16.6% of respondents are classified as intrinsic on both occasions. This value needs to be compared to the 22.1 column percentage and the 26.8 row percentage of the marginal distribution. Consistent scoring is similar in case of the extrinsic class with 24.2 cell percentage, 38.2 column percentage and 30.4 row percentage. Our overall interpretation of the results in Tables 4.1 and 4.2 is that when intrinsically oriented respondents and extrinsically oriented respondents answer to either ranking on rating questionnaires they tend to respond consistently across occasions. The intriguing next question is now: Will they score consistently when the question format changes from first to second measurement?

The answer to the latter question is boldly: "yes". Whether rankings were administered after (Table 4.3) or before (Table 4.4) the rating questionnaire, in each case we observe positive residual values of consistent scoring in the case of the intrinsic latent class (+10.5% and +9.9%) and in the case of the extrinsic latent class (+8.7% and +10.0%). Comparing the cell percentages with column and row percentages is less straightforward than in Tables 4.1 and 4.2 since row and column distributions refer to two different measurement instruments (ratings and ratings) that have a different number of latent classes. Comparison of cell percentages with the column equivalence is a logical choice since the four latent classes model indicates the maximum percentage that might return in the table of cell percentages. This comparison reveals that the intrinsic cell percentages are closer to the column percentages than the extrinsic cell percentages are compared to their column percentages. Hence, the consistency in classifying respondents in the intrinsic latent class across measurement method (rating versus ranking) is somewhat higher than in the case of the extrinsic latent class.

A final issue in the comparison of the intrinsic versus extrinsic latent class across the four subsamples needs to be addressed. At first glance a reader might find that the relative class sizes differ across occasions and across methods. This comparison is however somewhat misleading since one

Table 4
Estimated cell % and residual % (= deviance from expected cell % with statistical independence) per test-condition T1 × T2

4.1 Ranking × Ranking (subsample 1)

	Ranking T1						
	Intrinsic		Extrinsic		Social		
Ranking T2	cell %	residual %	cell %	residual %	cell %	residual %	Column total %
Intrinsic	31.5	18.8	0.1	−12.9	0.3	−5.9	31.9
Extrinsic	5.7	−11.7	36.8	19.0	1.1	−7.3	43.6
Social	2.6	−7.2	4.0	−6.0	17.9	13.2	24.5
Row total %	39.8		40.9		19.3		100.0

4.2 Rating × Rating (subsample 4)

	Rating T1								
	Intrinsic		Extrinsic		People		Non-differentiation		
Rating T2	cell %	residual %	cell %	residual %	cell %	residual %	cell %	residual %	Column total %
Intrinsic	16.6	10.6	0.6	−6.2	0.6	−5.9	4.4	1.4	22.1
Extrinsic	0.7	−9.5	24.2	12.6	11.2	−0.1	2.1	−3.0	38.2
People	9.1	1.3	2.7	−6.1	16.4	7.8	0.8	−3.0	29.0
Non-differentiation	0.4	−2.4	2.9	−0.3	1.3	−1.8	6.0	4.6	10.6
Row total %	26.8		30.4		29.6		13.3		100.0

4.3 Ranking × Rating (subsample 3)

	Rating T1								
	Intrinsic		Extrinsic		Social		Non-differentiation		
Ranking T2	cell %	residual %	cell %	residual %	cell %	residual %	cell %	residual %	Column total %
Intrinsic	18.7	10.5	0.1	−10.1	8.7	−0.1	3.8	−0.3	31.3
Extrinsic	3.4	−8.1	23.0	8.7	11.7	−0.7	5.9	0.1	43.9
Social	4.1	−2.4	9.5	1.4	7.7	0.8	3.5	0.2	24.8
Row total %	26.2		32.6		28.1		13.1		100.0

4.4 Rating × Ranking (subsample 2)

	Rating T2								
	Intrinsic		Extrinsic		People		Non-differentiation		
Ranking T1	cell %	residual %	cell %	residual %	cell %	residual %	cell %	residual %	Column total %
Intrinsic	19.1	9.9	2.2	−13.4	16.2	5.3	2.2	−1.8	39.7
Extrinsic	1.7	−7.9	26.2	10.0	8.2	−3.2	5.2	1.1	41.3
Social	2.5	−2.0	10.9	3.4	3.1	−2.2	2.6	0.7	19.1
Row total %	23.3		39.3		27.5		10.0		100.0

needs to take into account that both methods produce an unequal number of latent classes and that class size changes from T1 to T2. To account for that we suggest calculating average percentages within each class across occasions – thus neutralizing the time-shift effect – and adjusting the sum of percentages in the intrinsic and extrinsic class to a common scale of 100 – thus neutralizing differences in number of latent classes across methods. For instance, in subsample 1 the average percentage classified within the intrinsic class equals 35.85 and the average percentage within the extrinsic class equals 42.20. The percentage of intrinsic respondents in the sum of all intrinsic and extrinsic respondents is then 35.85/(35.85 + 42.20) = 45.9. In subsamples 4, 3 and 2 this "adjusted" percentage of respondents within the intrinsic class is 41.6, 42.9 and 43.8 respectively. Hence, adjusting for differences in number of latent classes and changes in class sizes from T1 to T2 more clearly shows the high resemblance between ratings and rankings in the identification of an intrinsic and extrinsic work values class.

In the previous section we already argued that the third latent class in the ranking assignment and the third and fourth latent class in the rating assignment seem to be typical to each method separately. As such we did not expect particular relationships between them when the measurement method changed from first to second measurement. This is confirmed by the results. We like to underscore the results regarding the fourth "non-differentiators" latent class. One criticism to the use of ranking data is that respondents are arbitrarily forced to make choices and hence random choices might occur in case respondents do not make difference in assigning importance to the different items (Davis, Dowley, & Silver, 1999). Our results show that respondents that were classified as non-differentiators at first measurement (Table 4.3) contribute proportional to each of the latent classes of the ranking assignment at T2 since residual cell percentages are smaller than 1. Similarly, respondents classified in one of the three latent classes in the ranking assignment at T1 are proportionally allocated to the class of non-differentiators at T2. Hence, it is safe to conclude that "forcing" non-differentiating respondents to make choices does not bias latent class identification in a ranking assignment.

6 Summary and Discussion

The key argument in this study is that there are segments within a population that respond similarly to rating and ranking questions that are used to measure work values. To that purpose we investigated whether the answers given by respondents at two measurement occasions are comparable, irrespective of whether the respondents received a rating or a ranking measurement procedure, and how consistent these results were over time. A modified form-resistant hypothesis was adopted by arguing that it is important to take into account the format-specific features of each measurement procedure, which, if not controlled, can make it hard to match the results of different measurement methods. The method-specific features controlled for in the current study are the primacy effect for the ranking data and the overall liking for the rating data.

In searching for segments that reveal similar preferences in work values we needed to adopt a research approach that deviates from what has been used in previous research. First, instead of using a factor-analytic approach we used a latent class choice modeling approach; this allowed us to distinguish between groups of respondents with similar response patterns in both the ranking and the rating method. These groups constitute homogeneous segments in the population that share a similar preference structure. Second, instead of adjusting the covariance structure of ranking data to eliminate the ipsativity of the data – a procedure suggested by Jackson and Alwin (1980) – we directly modeled the raw data in such a way that it reveals relative preferences. We also used a model that allowed to research relative preferences with rating data. This model implied the use of a random intercept to control for overall agreement. The measurement part of the model then also identifies relative preference structures similar to the model used with ranking data. The principal finding of this research is that respondents classified into either the intrinsic or extrinsic work values classes are consistently classified across occasions, even if the measurement method – ranking versus rating – changes in time. Other latent classes were method-specific: A social work values class for the ranking assignment and a people oriented work values class and non-differentiating class for the rating assignment. These method-specific classes were found consistently over the two measurement occasions when the same measurement method (rating or ranking) was used.

The within-subjects design thus enabled us to investigate how consistent the classifications were across measurement methods on two measurement occasions. We found it to be surprisingly high. Our modified form-resistant hypothesis stated that specific segments could be expected to emerge from either ranking or rating. We were particularly interested in finding out how non-differentiators in the rating assignment would respond to a ranking assignment in which they are forced to make a priority ranking of work items. The cross-classifications showed that non-differentiating respondents contributed proportionally to each of the latent classes in the ranking approach, irrespective of whether they first rated and then ranked or vice versa. Thus, forcing non-differentiating respondents to choose does not lead to biases in the ranking results.

There are some important messages that our research signals to applied researchers. First, our research indicates that the assumption that ranking and rating questions trigger different latent traits within individuals is not justified. The majority of respondents even tend to answer ranking and rating versions of the same questionnaire from the same underlying latent trait, showing either an intrinsic or extrinsic work orientation. Second, the method that was evoked provides researchers with a tool to detect values preferences even when rating data is used. This is particular relevant to secondary data analysts that have no choice on the measurement mode used in the survey. Third, an attractive feature of the approach used in this research was its semi-exploratory nature. It is not completely exploratory since the research starts with a preconceived measurement model. In the ranking assignment, for instance, we included an effect of primacy and checked whether it improved measurement fit. This is typical to what is called confirmatory measurement modeling. Our models are exploratory at the same time, in the sense that specific response patterns are revealed when adding latent classes to previous models. We regard this as a strength of our approach. Non-differentiating, for instance, was a response pattern that emerged from the rating data. We did not explicitly model it.

An inevitable limitation of this study was that we compared ratings and rankings in one particular context, namely work values. The question remains to what extent these findings can be generalized to other types of concepts for which both the ranking and rating approach can be used. We also used a long item list which we thought would be most challenging in finding similarity in results. Whether similarity in results depends on the length of the items list remains to be researched. The methods used in this research, however, are also applicable with shorter lists of items.

We believe that the current study has shown the usefulness of the latent class segmentation approach for the comparison of rating and ranking data and for checking the consistency of measurements over time. Using the approach in which we transformed ratings into relative preferences to compare this data to the ranking data, we were able to show that rankings and ratings do produce results that are actually more similar than was previously assumed.

Acknowledements

This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO), grant number 400-09-373

References

Allison, P. D. & Christakis, N. A. (1994). Logit models for sets of ranked items. *Sociological Methodology*, *24*, 199–228.

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.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43(1)*, 272–311.

Becker, S. L. (1954). Why an order effect. *Public Opinion Quarterly*, *18(3)*, 271–278.

Böckenholt, U. (2002). Comparison and choice: analyzing discrete preference data by latent class scaling models. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 163–182). Cambridge: Cambridge University Press.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: one-step versus three-step estimators. *Political Analysis*, *12(1)*, 3–27.

Braithwaite, V. A. & Law, H. G. (1985). Structure of human values: testing the adequacy of the Rokeach Value Survey. *Journal of Personality and Social Psychology*, *49(1)*, 250–263.

Campbell, D. T. & J., M. P. (1950). The effect of ordinal position upon responses to items in a check list. *Journal of Applied Psychology*, *34(1)*, 62–67.

Cattell, R. B. (1944). Psychological measurement: normative, ipsative, interactive. *Psychological Review*, *51(5)*, 292–303.

Cheung, M. W. L. (2006). Recovering preipsative information from additive ipsatized data: a factor score approach. *Educational and Psychological Measurement*, *66(4)*, 565–588.

Cheung, M. W. L. & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *9(1)*, 55–77.

Croon, M. A. (1989). Latent class models for the analysis of rankings. In G. de Soete, H. Feger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 99–121). Amsterdam: Elsevier Science Publishers.

Cunningham, W. H., Cunningham, I. C. M., & Green, R. T. (1977). The ipsative process to reduce response set bias. *Public Opinion Quarterly*, *41(3)*, 379–384.

Davis, D. W., Dowley, K. M., & Silver, B. D. (1999). Post-materialism in world societies: is it really a value dimension. *American Journal of Political Science*, *43(3)*, 935–962.

de Chiusole, D. & Stefanutti, L. (2011). Rating, ranking, or both? A joint application of two probabilistic models for the measurement of values. *Testing, Psychometrics, Methodology in Applied Psychology*, *18(1)*, 49–60.

Dunlap, W. P. & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research*, *29(1)*, 115–126.

Elizur, D. (1984). Facets of work values: a structural analysis of work outcomes. *Journal of Applied Psychology*, *69(3)*, 379–389.

Elizur, D., Borg, I., Hunt, R., & Beck, I. M. (1991). The structure of work values: a cross cultural comparison. *Journal of Organizational Behavior*, *12(1)*, 21–38.

Fuchs, M. (2005). Children and adolescents as respondents: experiments on question order, response order, scale effects and the effect of numeric values associated with response options. *Journal of Official Statistics*, *21(4)*, 701–725.

Furnham, A., Petrides, K. V., Tsaousis, I., Pappas, K., & Garrod, D. (2005). A cross-cultural investigation into the relationships between personality traits and work values. *Journal of Psychology*, *139(1)*, 5–32.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I – a modified latent structure approach. *The American Journal of Sociology*, *79(5)*, 1179–1259.

Harzing, A. W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., … Zander, L. (2009). Rating versus ranking: what is the best way to reduce response and language bias in cross-national research? *International Business Review*, *18(4)*, 417–432.

Jackson, D. J. & Alwin, D. F. (1980). The factor analysis of ipsative measures. *Sociological Methods & Research*, *9*, 218–238.

Jacoby, W. G. (2011). Measuring value choices: are rank orders valid indicators? Paper presented at the 2011 Annual Meetings of the Midwest Political Science Association in Chicago, IL.

Kalleberg, A. L. (1977). Work values and job rewards: a theory of job satisfaction. *American Sociological Review*, *42(1)*, 124–143.

Kamakura, W. A., Wedel, M., & Agrawal, J. (1994). Concomitant variable latent class models for the external analysis of choice data. *International Journal of Research in Marketing*, *11(5)*, 451–464.

Klein, M., Dülmer, H., Ohr, D., Quandt, M., & Rosar, U. (2004). Response sets in the measurement of values: a comparison of rating and ranking procedures. *International Journal of Public Opinion Research*, *16(4)*, 474–483.

Knoop, R. (1994). Work values and job satisfaction. *Journal of Psychology*, *128(6)*, 683–690.

Krosnick, J. A. (1992). The impact of cognitive sophistication and attitude importance on response-order and question-order effects. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 203–218). New York: Springer-Verlag.

Krosnick, J. A. (2000). The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Newsletter*, *20(1)*, 4–7.

Krosnick, J. A. & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51(2)*, 201–219.

Krosnick, J. A. & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52(4)*, 526–538.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. Stouffer (Ed.), *Measurement and prediction* (pp. 362–412). Princeton, N.J.: Princeton University Press.

Magidson, J., Eagle, T., & Vermunt, J. K. (2003). New developments in latent class choice modeling. Proceedings Sawtooth Software Conference 2003.

Magidson, J. & Vermunt, J. K. (2006). Use of latent class regression models with a random intercept to remove the effects of the overall response rating level. In A. Rizzi & M. Vichi (Eds.), *COMPSTAT 2006: proceedings in computational statistics* (pp. 351–360). Heidelberg: Springer.

Maio, G. R., Roese, N. J., Seligman, C., & Katz, A. (1996). Rankings, ratings, and the measurement of values: evidence for the superior validity of ratings. *Basic and Applied Social Psychology*, *18(2)*, 171–181.

McCarty, J. A. & Shrum, L. J. (1997). Measuring the importance of positive constructs: a test of alternative rating procedures. *Marketing Letters*, *8(2)*, 239–250.

McCarty, J. A. & Shrum, L. J. (2000). The measurement of personal values in survey research: a test of alternative rating procedures. *Public Opinion Quarterly*, *64(3)*, 271–298.

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly*, *67(1)*, 205–211.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research*, *20(1)*, 60–103.

McFadden, D. (1986). The choice theory approach to marketing research. *Marketing Science*, *5(4)*, 275–297.

McFadden, D. & Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, *15(5)*, 447–470.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58(4)*, 525–543.

Miethe, T. D. (1985). The validity and reliability of value measurements. *The Journal of Psychology*, *119(5)*, 441–453.

Moore, M. (1975). Rating versus ranking in the Rokeach Value Survey: an israeli comparison. *European Journal of Social Psychology*, *5(3)*, 405–408.

Moors, G. (2010). Ranking the ratings: a latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research*, *22(1)*, 93–119.

Moors, G. & Vermunt, J. K. (2007). Heterogeneity in post-materialist value priorities. evidence from a latent class discrete choice approach. *European Sociological Review*, *23(5)*, 631–648.

Munson, J. M. & McIntyre, S. H. (1979). Developing practical procedures for the measurement of personal values in cross-cultural marketing. *Journal of Marketing Research*, *16(1)*, 48–52.

Ovadia, S. (2004). Ratings and rankings: reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology*, *7(5)*, 403–414.

Parsons, T. & Shils, E. A. (1962). *Toward a general theory of action*. New York: Harper.

Rankin, W. L. & Grube, J. W. (1980). A comparison of ranking and rating procedures for values system measurement. *European Journal of Social Psychology*, *10*, 233–246.

Reynolds, T. J. & Jolly, J. P. (1980). Measuring personal values: an evaluation of alternative methods. *Journal of Marketing Research*, *17(4)*, 531–536.

Rokeach, M. (1973). *The nature of human values*. New York: The Free Press.

Ros, M., Schwartz, S. H., & Surkiss, S. (1999). Basic individual values, work values, and the meaning of work. *Applied Psychology: An International Review*, *48(1)*, 49–71.

Schuman, H. & Presser, S. (1996). *Questions and answers in attitude surveys: experiments on question form, wording, and context*. Thousand Oaks, California: Sage Publications.

Stern, M. J., Dillman, D. A., & Smyth, J. D. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey. *Survey Research Methods*, *1(3)*, 121–138.

Super, D. E. (1962). The structure of work values in relation to status, achievement, interests, and adjustment. *Journal of Applied Psychology*, *46(4)*, 231–239.

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(8)*, 1096–1105.

Vermunt, J. K. (2010). Latent class modeling with covariates: two improved three-step approaches. *Political Analysis*, *18(4)*, 450–469.

Vermunt, J. K. & Magidson, J. (2005). *Latent GOLD Choice 4.0 user's guide*. Belmont: Statistical Innovations, Inc.

Vermunt, J. K. & Magidson, J. (2013). *Latent Gold 5.0 upgrade manual*. Belmont, MA: Statistical Innovations, Inc.

Vriens, I. (2015). *Two of a kind? Comparing ratings and rankings for measuring work values using latent class modeling*. s'-Hertogenbosch: Box-Press BV.

Vriens, I., Moors, G., Gelissen, J., & Vermunt, J. K. (2015). Controlling for response order effects in ranking items using latent choice factor modeling. *Sociological Methods & Research, first published online*, 1–24. doi:10.1177/0049124115588997

Appendix
Tables

(*Appendix tables follow on next page*)

Table A1

Estimated effect parameters regressing T2- on T1-probabilities-Results from the step-3 proportional ML approach

1 Ranking × Ranking (subsample 1)

	Ranking T1					
	Intrinsic		Extrinsic		Social	
Ranking T2	beta	s.e.	beta	s.e.	beta	s.e.
Intrinsic	2.550	0.955	−1.860	1.470	−0.695	1.450
Extrinsic	−0.915	0.664	2.000	0.804	−1.080	0.986
Social	−1.640	0.670	−0.142	0.799	1.780	0.796

2 Rating × Rating (subsample 4)

	Rating T1							
	Intrinsic		Extrinsic		People		Non-differentiation	
Rating T2	beta	s.e.	beta	s.e.	beta	s.e.	beta	s.e.
Intrinsic	2.136	0.410	−1.477	0.690	−1.445	0.620	0.786	0.306
Extrinsic	−1.668	0.630	1.586	0.314	0.725	0.301	−0.643	0.303
People	0.889	0.331	−0.538	0.331	1.157	0.256	−1.508	0.332
Non-differentiation	−1.356	0.728	0.429	0.373	−0.437	0.391	1.365	0.284

3 Ranking × Rating (subsample 3)

	Rating T1							
Ranking T2	Intrinsic		Extrinsic		People		Non-differentiation	
	beta	s.e.	beta	s.e.	beta	s.e.	beta	s.e.
Intrinsic	1.704	0.610	−2.782	1.716	0.568	0.598	0.510	0.615
Extrinsic	−1.147	0.432	1.627	0.868	−0.281	0.338	−0.199	0.356
Social	−0.558	0.417	1.155	0.876	−0.287	0.349	−0.311	0.378

4 Rating × Ranking (subsample 2)

	Rating T2							
	Intrinsic		Extrinsic		People		Non-differentiation	
Ranking T1	beta	s.e.	beta	s.e.	beta	s.e.	beta	s.e.
Intrinsic	1.360	0.386	−1.510	0.550	0.639	0.296	−0.485	0.428
Extrinsic	−1.150	0.527	0.926	0.317	−0.100	0.265	0.322	0.309
Social	−0.211	0.415	0.588	0.337	−0.540	0.323	0.163	0.339

Note: numbering of sub-tables reflects correspondence with results presented in Table 3.