Scoring Methods for Ordinal Multidimensional Forced-Choice Items

Anton L. M. de Vries
Maastricht University

L. Andries van der Ark
Tilburg University

March 31, 2008

Abstract

In most psychological tests and questionnaires, a test score is obtained by taking the sum of the item scores. In virtually all cases where the test or questionnaire contains multidimensional forced-choice items, this traditional scoring method is also applied. We argue that the summation of scores obtained with multidimensional forced-choice items produces uninterpretable test scores. Therefore, we propose three alternative scoring methods: a weak and a strict rank preserving scoring method, which both allow an ordinal interpretation of test scores; and a ratio preserving scoring method, which allows a proportional interpretation of test scores. Each proposed scoring method yields an index for each respondent indicating the degree to which the response pattern is inconsistent. Analysis of real data showed that with respect to rank preservation, the weak and strict rank preserving method resulted in lower inconsistency indices than the traditional scoring method; with respect to ratio preservation, the ratio preserving scoring method resulted in lower inconsistency indices than the traditional scoring method.

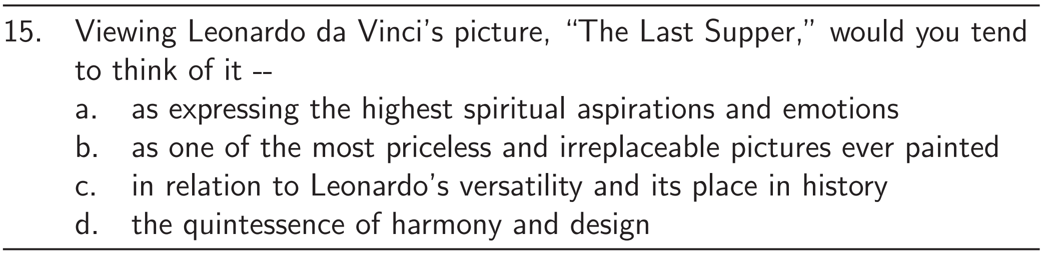
keywords: forced choice, testing method; ipsative; multidimensional forced choice response format; preference measures; scoring, testing

Correspondence: Anton L. M. de Vries
Dept. of Neurocognition, Fac. of Psychology, Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands
phone:+31 43 388 4043, fax:+31 43 388 4125,
email:a.devries@psychology.unimaas.nl

1 Introduction

A multidimensional forced-choice (MFC) item consists of $m \geq 2$ *statements*; each statement is an indicator of a different trait or dimension. For example, Figure 1 shows an MFC item from the questionnaire the Study of Values Part II (SOV; Kopelman, Rovenpor, & Guan, 2003). The SOV measures six traits: (1) theoretical value, (2) aesthetic value, (3) political value, (4) religious value, (5) economic value, and (6) social value. In the item of Figure 1, statement a is an indicator of religious value, b of economic value, c of theoretical value, and d of aesthetic value. A respondent is instructed to rank all statements according to preference by assigning score 4 to the most preferred statement down to score 1 to the least preferred statement. The *statement score* pertaining to trait q in item j is denoted Y_{jq}. Notice that for the item in Figure 1, the statement scores for political value and social value are not available.

Questionnaires have the *ordinal MFC format* if a respondent is instructed to rank all m statements in the item (as for the item in Figure 1), or to rank k out of m statements ($k < m$). Questionnaires that employ the ordinal MFC format are, for example, the Canfield Learning Styles Inventory (CLSI; Canfield, 1980), the Survey of Interpersonal Values (SIV; Gordon, 1976), the Survey of Personal Values (SPV; Gordon, 1984), the Edwards Personal Preference Schedule (EPPS; Edwards, 1954), the Occupational Preference Questionnaire (OPQ; Saville, Sik, Nyfield, Hackston, & MacIver, 1996), and the Beroepen Interessen Test [Vocational Interests Test] (BIT; Evers, Lucassen, & Wiegersma, 1999; Irle, 1955). Occasionally, MFC questionnaires may adopt alternative instructions for assigning scores to statements; for example, an item containing two statements over which three points must be distributed, allows one of four responses: (3, 0), (2, 1), (1, 2), or (0, 3) (SOV Part I; Kopelman et al., 2003). Statement scores obtained using these alternative instructions are not discussed in this paper. One important reason



15. Viewing Leonardo da Vinci's picture, "The Last Supper," would you tend to think of it --
    a. as expressing the highest spiritual aspirations and emotions
    b. as one of the most priceless and irreplaceable pictures ever painted
    c. in relation to Leonardo's versatility and its place in history
    d. the quintessence of harmony and design

Note. Item derived from Kopelman et al. (2003).

Figure 1: *An MFC Item from the Study of Values Part II.*

Table 1: *The Responses of a Single Respondent to the 15 Items of the SOV Part II and the Corresponding Statement Scores, Traditional Test Scores, and Several Alternative Test Scores (See Text).*

			Statement scores						Total
Item	Statements	Responses	T	A	P	R	E	S	
1.	SERP	3214			4	1	2	3	10
2.	TPAR	2431	2	3	4	1			10
3.	ASTE	2413	1	2			3	4	10
4.	ERPA	4321		1	2	3	4		10
5.	ERTS	3214	1			2	3	4	10
6.	PAST	3241	1	2	3			4	10
7.	TERP	1432	1		2	3	4		10
8.	ASPE	1423		1	2		3	4	10
9.	RTAS	3124	1	2		3		4	10
10.	TAPE	1234	1	2	3		4		10
11.	PTSR	2143	1		2	3		4	10
12.	RAES	2134		1		2	3	4	10
13.	SPET	4231	1		2		3	4	10
14.	PSRA	2431		1	2	3		4	10
15.	RETA	3412	1	2		3	4		10
Traditional test scores			11	17	26	24	33	39	150

Note. The six traits are indicated by: T = Theoretical value, A = Aesthetic value, P = Political value, R = Religious value, E = Economic value, and S = Social value; a score of 4 indicates 'preferred most' and a score of 1 'preferred least'.

for using the ordinal MFC response format is that it might be more resistant to the social desirability response bias (e.g., Martin, Bowen, & Hunt, 2002; Nederhof, 1985; Stanush, 1997), although this view is not universally accepted (De Vries, 2006, chap. 8; Heggestad, Morrison, Reeve, & McCloy, 2006). The SOV Part II consists of 15 items, which all have the same ordinal MFC format as the item in Figure 1, covering all possible combinations of four out of six traits. Table 1 shows the responses of one respondent to the 15 items of the SOV Part II (third column) and the corresponding statement scores (fourth to ninth column).

The traditional scoring method for ordinal MFC items is to compute the sum of the available statement scores over all items (Canfield, 1980; Edwards, 1954; Evers et al., 1999; Gordon, 1976, 1984; Irle, 1955; Saville et al., 1996). The resulting test scores are used for further data analysis. This scoring method is also common for items with other response formats such as a

Likert scale.

For MFC items, the traditional scoring method has two undesirable features:

1. The traditional test scores are *ipsative* (e.g., Cattell, 1944; Clemans, 1966; Hicks, 1970; Radcliffe, 1963). Ipsative scores add up to a constant value. For the example in Table 1 the traditional test scores add up to 150, irrespective of the responses. Ipsative scores cannot be analyzed readily using standard statistical methods based on correlations or covariances, such as regression or factor analysis (Baron, 1996; Closs, 1996; Cornwell & Dunlap, 1994; Dunlap & Cornwell, 1994; Guilford, 1952; Johnson, Wood, & Blinkhorn, 1988; also see, e.g., Aitchison, 1986/2003; Brady, 1989; Chan & Bentler, 1993, 1996, 1998; Ten Berge, 1999); and ipsative scores yield relative information rather than absolute information about the traits measured (e.g., Broverman, 1962; Closs, 1976, 1996; Johnson et al., 1988). Consequently, ipsative scores allow valid comparisons of traits within a respondent but not between respondents (Fedorak & Coles, 1979; Katz, 1962).

2. Traditional test scores do not allow valid comparisons between traits within a respondent. Even the relative interpretation within a respondent, which is the only way ipsative scores can be interpreted is hampered by the way the traditional test scores are constructed: Statement scores in item j, $Y_{j1}, \ldots, Y_{jQ}$, are a mixture of rank numbers and missing values (cf. Table 1). The traditional scoring method implies that the missing values are replaced with zeros, that is, $Y^*_{jq} = 0$ if Y_{jq} is missing and $Y^*_{jq} = Y_{jq}$ otherwise; and the test score for trait q, denoted X_q is computed as $X_q = \sum_j Y^*_{jq}$. Therefore, the traditional scoring method uses zeros as estimates of the missing rank orders. For a low-ranking trait these zeros may be reasonable estimates, but for high-ranking traits these zeros may be far off, yielding heavily biased test scores.

This paper aims at finding alternative scoring methods.

2 Requirements for alternative scoring methods

Test scores produced by alternative scoring methods must satisfy practical requirements. Consider the traits political value and religious value in Table 1. From the statement scores in Table 1 it can be derived that in four items (items 4, 7, 11, and 14) religious value was preferred over political value; in two items (items 1 and 2) political value was preferred over religious value; and in the remaining nine items preference is not clear because at least one of the statement scores is missing. These observations are used to define practical requirements for test scores.

2.1 Rank order preservation

The first requirement, called *rank order preservation*, conveys the idea that the test scores of respondent i on political value, X_{iP}, and religious value, X_{iR}, should express that respondent i preferred more statements expressing religious value than statements expressing political value. Test sores X_{iR} and X_{iP} satisfy rank order preservation if $X_{iR} > X_{iP}$. It may be verified that for the statement scores shown in Table 1, the traditional test scores do not satisfy rank order preservation, because $X_{iR} = 24 < X_{iP} = 26$.

In general, let S_q be the set of statements pertaining to trait q ($q = 1, \ldots, Q$), and let $F_i(S_q \succ S_r)$ be the number of times that respondent i preferred a statement from set S_q over a statement from set S_r in his or her responses to the MFC questionnaire. Weak rank order preservation is defined as

$$X_{iq} > X_{ir} \Leftrightarrow F_i(S_q \succ S_r) \geq F_i(S_r \succ S_q) \text{ for all } q \neq r. \tag{1}$$

Strict rank order preservation is defined as Equation 1 with a strict inequality in the right-hand side. Note that there are $Q(Q-1)/2$ pairs of test scores (X_{iq}, X_{ir}), and investigating rank order preservation requires checking the inequality constraint in Equation 1 for all pairs. If it is possible to construct Q test scores for respondent i that satisfy the $Q(Q-1)/2$ inequality constraints imposed by Equation 1, then we say that respondent i has a response pattern *consistent* with respect to rank order preservation.

2.2 Ratio preservation

The second requirement, called *ratio preservation*, conveys the idea that test scores X_{iP} and X_{iR} should express that the preference ratio political value to religious value equals $2 : 4 = .5$. It may be verified that for the statement scores shown in Table 1, the test scores obtained with the traditional scoring method do not satisfy ratio preservation, because $X_{iP} : X_{iR} = 26 : 24 = 1.08$.

Ratio preservation is defined as

$$\frac{X_{iq}}{X_{ir}} = \frac{F_i(S_q \succ S_r)}{F_i(S_r \succ S_q)} \text{ for all } q \neq r. \tag{2}$$

Table 2: *Dominance matrix* $\mathbf{D}_i$ *for the Scores in Table 1, the Row Sum, the Initial Test Scores, Weak Rank Order Preserving Test Scores, and Strict Rank Order Preserving Test Scores.*

	T	A	P	R	E	S
Theoretical value	0	+1	+1	+1	+1	+1
Aesthetic value	−1	0	+1	+1	+1	+1
Political value	−1	−1	0	+1	+1	+1
Religious value	−1	−1	−1	0	+1	+1
Economic value	−1	−1	−1	−1	0	+1
Social value	−1	−1	−1	−1	−1	0
Row sum	−5	−3	−1	1	3	5
Initial test scores	1	2	3	4	5	6
WRP test scores	1	2	3	4	5	6
SRP test scores	1	2	3	4	5	6

Note. WRP = weak rank preserving; SRP = strict rank preserving.

Investigating ratio preservation requires checking the equality constraints in Equation 2 for all $Q(Q-1)/2$ pairs of test scores. Note that ratio preservation implies rank order preservation.

If it is possible to construct Q test scores for respondent i that satisfy the $Q(Q-1)/2$ equality constraints imposed by Equation 2, then we say that respondent i has a response pattern consistent with respect to ratio preservation. Only in some very rare cases will it be possible that Q test scores exactly satisfy the $Q(Q-1)/2$ equality constraints in Equation 2, because there are more constraints than test scores. For practical situations ratio preservation will only hold approximately.

3 Two rank order preserving scoring methods

We propose two simple scoring methods which aim at producing test scores that satisfy weak and strict rank order preservation, respectively (cf. Equation 1).

For respondent i, for each pair of traits it is investigated which of the two traits is preferred most often by comparing the statement scores in the items that have a statement for both traits. The results are collected in a $Q \times Q$ dominance matrix $\mathbf{D}_i$ with all diagonal elements equal to zero, and off-diagonal elements $D_{iqr} = +1$ if trait r is preferred more often than trait q, $D_{iqr} = -1$ if trait r is preferred less often than trait q, and $D_{iqr} = 0$ otherwise ($q = 1, \ldots, Q; r = 1 \ldots, Q$). For the statement scores in Table 1, $\mathbf{D}_i$ is shown in Table 2. The *initial test scores* are the rank numbers of the column sums of $\mathbf{D}_i$ (Table 2).

Weak rank order preserving scoring method. If the initial test scores satisfy the inequality constraints in Equation 1 and thus are weak rank order preserving, then the initial test scores are also the final test scores. Whether or not the initial test scores are weak rank order preserving can be derived from $\mathbf{D}_i$ in the following way. Order the rows and columns of $\mathbf{D}_i$ by the initial test scores. If the test scores are weakly rank order preserving then all upper diagonal elements should be nonnegative and all lower diagonal elements should be nonpositive. It may be verified that this is the case for the test scores in Table 2. If the initial test scores are not weak rank order preserving (Table 3 shows an example), then the following adjustment of the test scores is proposed. Partition the Q traits in as many subsets as possible such that for traits from different subsets weak rank order preservation holds. In Table 3, these subsets are {T, A, P, R}, {E}, and {S}. Traits in the same subset receive the same test score, that is, the average initial test score (see Table 3, last row but one).

The weak rank order preserving test scores can be interpreted as follows. If the test score pertaining to trait q is greater than the test score pertaining to trait r, then the respondent has preferred trait q over trait r at least as many times as he or she has preferred trait r over trait q in the items that allow a direct comparison. The preference order between two traits is undetermined if the corresponding two test scores are equal.

Strict rank order preserving scoring method. If the initial test scores satisfy strict rank order preservation, then the initial test scores are also the final test scores. Whether or not the initial test scores are strict rank order preserving can be derived from $\mathbf{D}_i$ in a similar way as for weak rank order preservation. Order the rows and columns of $\mathbf{D}_i$ by the initial test scores. If the test scores are strict rank order preserving then all upper diagonal elements should be strictly positive and all lower diagonal elements should be strictly negative. It may be verified that this is the case for the test scores in Table 2. If the initial test scores are not strict rank order preserving (Table 3), then an adjustment of the test scores is proposed, which is similar to the adjustment of test scores that were not weak order preserving. Partition the Q traits in as many subsets as possible such that for traits from different subsets strict rank order preservation holds. In Table 3, these subsets are {T, A, P, R, E} and {S}. Traits in the same subset receive the same test score, that is, the average initial test score (see Table 3, last row).

Table 3: *Dominance matrix $\mathbf{D}_i$ for an Inconsistent Response Pattern, the Row Sum, the Initial Test Scores, Weak Rank Order Preserving Test Scores, and Strict Rank Order Preserving Test Scores*

	T	A	P	R	E	S
Theoretical value	0	+1	+1	−1	+1	+1
Aesthetic value	−1	0	+1	+1	+1	+1
Political value	−1	−1	0	+1	+1	+1
Religious value	+1	−1	−1	0	0	+1
Economic value	−1	−1	−1	0	0	+1
Social value	−1	−1	−1	−1	−1	0
Row sum	−3	−3	−1	0	2	5
Initial test scores	1.5	1.5	3	4	5	6
WRP test scores	2.5	2.5	2.5	2.5	5	6
SRP test scores	3	3	3	3	3	6

Note. WRP = weak rank preserving; SRP = strict rank preserving.

The strict rank order preserving test scores can be interpreted as follows. If the test score pertaining to trait q is greater than the test score pertaining to trait r, then the respondent has preferred trait q over trait r more often than he or she has preferred trait r over trait q in the items that allow a direct comparison. The preference order between two traits cannot be interpreted if the corresponding two test scores are equal. On the one hand, the strict rank order preserving test scores have a stronger interpretation than the weak rank order preserving test scores, and on the other hand, the probability that two test scores are equal is greater for strict rank order preserving test scores than for weak rank order preserving test scores (see, Table 3, for an example).

It may be noted that both rank order preserving test scores can be viewed as ipsative scores because they represent the order of the trait preferences within a respondent (cf. Chan, 2003, who called these scores ordinal ipsative data). However, contrary to the traditional test scores, the rank order preserving test scores have a sound ordinal interpretation within the limits of ipsative data.

Indices of inconsistency. The degree of inconsistency of test scores with respect to weak rank order preservation, denoted I_i^{weak}, is expressed by the number of pairs of test scores that do not satisfy weak rank order preservation (Equation 1). Note that for the initial test scores in Table 3 $I_i^{\text{weak}} = 1$, and for the test scores produced by the weak and strict rank order preserving scoring method $I_i^{\text{weak}} = 0$ by definition. The degree of inconsistency of test scores with respect to strict rank order preservation, denoted I_i^{strict}, is expressed by the number of pairs of test scores that do not satisfy strict rank order preservation. For the initial test scores in Table 3 $I_i^{\text{strict}} = 2$, and for test scores in Table 3 produced by the weak rank order preserving scoring method $I_i^{\text{strict}} = 1$, for test scores in Table 3 produced by the strict rank order preserving scoring method $I_i^{\text{strict}} = 0$ by definition.

4 Ratio preserving scoring method

A more elaborate method is proposed that aims at producing test scores that satisfy ratio preservation (Equation 2). Let

$$R_{iqr} = \frac{F_i(S_q \succ S_r)}{F_i(S_r \succ S_q)} \tag{3}$$

be the preference ratio of respondent i with respect to trait q and trait r $(q \neq r)$. Let ε be a positive value smaller than the smallest statement score. If $F_i(S_r \succ S_q)$ in Equation 3 equals zero, the preference ratio does not exist, and we advocate to replace R_{iqr} by a maximum ratio

$$R^*_{iqr} = \frac{F_i(S_q \succ S_r) - \varepsilon}{\varepsilon}.$$

Similarly, if $F_i(S_q \succ S_r)$ in Equation 3 equals zero, we advocate to replace R_{iqr} by a minimum ratio

$$R^*_{iqr} = \frac{\varepsilon}{F_i(S_r \succ S_q) - \varepsilon}.$$

It can be shown that this replacement strategy equals the multiplicative replacement strategy advocated by Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003). As a rule of thumb, Hornung and Reed (1990) suggested to take $\varepsilon = 1/\sqrt{2}$.

The preference ratios of respondent i are collected in a $Q \times Q$ ratio matrix $\mathbf{R}_i$. By definition, $R_{iqq} = 1$ $(q = 1, \ldots, Q)$. For the statement scores in Table 1, $\mathbf{R}_i$ is shown in Table 4. Each row of $\mathbf{R}_i$ is a vector of ratios with the row trait as the reference. Only if all rows are linearly dependent (i.e., $\mathbf{R}_i$ has rank 1), the response patterns are consistent with respect to ratio preservation. This means that any row of $\mathbf{R}_i$ can be obtained by multiplying another row with a constant value.

Consistent response patterns. The ratios in $\mathbf{R}_i$ are unchanged if they are multiplied or divided by a constant value. For constructing test scores for

Table 4: *Ratio matrix* $\mathbf{R}_i$ *for the Scores in Table 1, the Exact Geometric Mean* ($\mathbf{g}_i$)*, the Geometric Mean for* $\varepsilon = 1/\sqrt{2}$*, and the Resulting Ratio Preserving Test Scores on the Same Scale as the Traditional Test Scores (i.e,* $\mathbf{X}_i = \mathbf{g}_i \times 150/\sum_q g_{iq}$*).*

	T	A	P	R	E	S
Theoretical value	1	a	a	5	a	a
Aesthetic value	$\frac{1}{a}$	1	a	5	a	a
Political value	$\frac{1}{a}$	$\frac{1}{a}$	1	2	5	5
Religious value	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{2}$	1	a	a
Economic value	$\frac{1}{a}$	$\frac{1}{a}$	$\frac{1}{5}$	$\frac{1}{a}$	1	a
Social value	$\frac{1}{a}$	$\frac{1}{a}$	$\frac{1}{5}$	$\frac{1}{a}$	$\frac{1}{a}$	1
$\mathbf{g}_i$	$\frac{1}{\sqrt[6]{5a^4}}$	$\frac{1}{\sqrt[6]{5a^2}}$	$\sqrt[6]{\frac{a^2}{50}}$	$\sqrt[6]{\frac{50}{a^2}}$	$\sqrt[6]{5a^2}$	$\sqrt[6]{5a^4}$
$\mathbf{g}_i$ ($\varepsilon = 1/\sqrt{2}$)	0.20	0.39	1.02	0.98	2.56	5.00
RP test scores	3	6	15	14	38	74

Note. a = the maximum ratio = $\frac{6-\varepsilon}{\varepsilon}$; RP = ratio preserving.

a consistent response pattern it suffices to take an arbitrary row from $\mathbf{R}_i$, and multiply each element with a conveniently chosen constant c. Practical values of c are $c = 1/(\sum_r R_{iqr})$, so that the test scores are proportions that add up to 1; $c = 100/(\sum_r R_{iqr})$, so that the test scores are percentages; or, for the statement scores in Table 1, $c = 150/(\sum_r R_{iqr})$, so that the ratio preserving test scores add up to the same value as the traditional test scores. Because all rows are linearly dependent, each row will give the same result when the elements are multiplied by $c/(\sum_r R_{iqr})$. The obtained test scores can be interpreted at a ratio level; that is, if $X_{iq}/X_{ir} = c$, then the preference of trait q over trait r was c times the preference of trait r over trait q. It may be noted that the ratio preserving test scores can be viewed as ipsative scores because any ratio of scores represents the preference ratio of two traits within a respondent (cf. Chan, 2003, who called these scores multiplicative ipsative data). However, contrary to the traditional test scores, the ratio preserving test scores have a sound proportional interpretation within the limits of ipsative data.

Inconsistent response patterns. In case of an inconsistent response pattern, the rows of $\mathbf{R}_i$ are not linearly dependent, and some average of the rows of $\mathbf{R}_i$ should be taken as estimated test scores. For vectors whose elements can be interpreted as ratios, Aitchison (1992) and Pawlowsky-Glahn and Egozcue (2002) advocated the geometric mean as an adequate average. Let $\mathbf{r}_{iq} = (R_{iq1}, \ldots, R_{iqr}, \ldots, R_{iqQ})^T$ denote row q of $\mathbf{R}_i$ $(q = 1, \ldots, Q)$, then the geometric mean over the Q rows of $\mathbf{R}_i$ is

$$\mathbf{g}_i = \left(\sqrt[Q]{\prod_q R_{iq1}}, \ldots, \sqrt[Q]{\prod_q R_{iqr}}, \ldots, \sqrt[Q]{\prod_q R_{iqQ}} \right).$$

The rationale for using the geometric mean is its relation to the *Aitchison distance*, a measure that appreciates that the elements of a vector are ratios. For example, the Aitchison distance between two vectors is unaffected if one of the vectors is multiplied by a constant value. The Aitchison distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted $d_a(\mathbf{x}, \mathbf{y})$, and defined as

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{Q} \sum_{q<r} \left(\ln \frac{x_q}{x_r} - \ln \frac{y_q}{y_r} \right)^2}. \tag{4}$$

The vector $\tilde{\mathbf{x}}$ that minimizes the sum of squared Aitchison distances between the rows of $\mathbf{R}_i$ and $\tilde{\mathbf{x}}$, $\sum_q d_a^2(\mathbf{r}_{iq}, \tilde{\mathbf{x}})$, equals $\mathbf{g}_i$ (Pawlowsky-Glahn & Egozcue, 2002). For a detailed discussion of the Aitchison distance we refer the interested reader to Aitchison (1992) and Pawlowsky-Glahn and Egozcue (2002).

Index of inconsistency. The closer the rows of $\mathbf{R}_i$ in terms of Aitchison distances (Equation 4), the more consistent a respondent has answered. If respondent i has a consistent response pattern, then all rows of $\mathbf{R}_i$ are linearly dependent and all Aitchison distances are zero. The average of the Aitchison distances between the rows and the geometric mean can serve as an unnormed index of inconsistency I_i^{ratio}; that is,

$$I_i^{\text{ratio}} = \frac{1}{Q} \sum_{q=1}^{Q} d_a^2(\mathbf{r}_{iq}, \mathbf{g}_i). \tag{5}$$

For the response pattern in Table 1 $I_i^{\text{ratio}} = 19.31$. To decide whether this is an unacceptably large value, the distribution of I^{ratio} can be computed. For 10,000 simulated random response patterns we found that 94.18% had an inconsistency index less than 19.31, which indicates that the response pattern of respondent i is rather inconsistent; and any test scores derived from this response pattern should be interpreted with care.

Table 5: *Results from the Empirical Example Using Six Traits: Percentage of Consistent Test Scores and Summary Statistics of the Distribution of the Inconsistency Index.*

Test scores	Weak rank preservation					
	Perc.	Min.	Q_1	Median	Q_3	Max.
Traditional	57.8%	0	0	0	1	5
Weak rank preserving	100.0%	0	0	0	0	0
Strict rank preserving	100.0%	0	0	0	0	0
Ratio preserving	49.8%	0	0	1	1	5
	Strict rank preservation					
	Perc.	Min.	Q_1	Median	Q_3	Max.
Traditional	2.6%	0	2	3	4	8
Weak rank preserving	12.4%	0	1	2	3	7
Strict rank preserving	100.0%	0	0	0	0	0
Ratio preserving	1.6%	0	3	4	5	9
	Ratio preservation					
	Perc.	Min.	Q_1	Median	Q_3	Max.
Traditional	0.0%	2.24	13.80	16.66	20.15	37.79
Ratio preserving	0.0%	1.28	6.37	9.04	12.17	24.67

Note. Perc. = Percentage of consistent responses; Min. = minimum value of inconsistency index; Q_1 = First quartile of inconsistency index; Median = Median of inconsistency index; Q_3 = Third quartile of inconsistency index; and Max. = maximum value of inconsistency index.

5 Empirical example

The SOV Part II (cf. Table 1) was administered to 386 first-year Psychology students from the University of Amsterdam, resulting in 386 sets of 60 statement scores. There were no missing values. First, for these 386 respondents, we computed the test scores using the traditional scoring method, the weak rank preserving scoring method, the strict rank preserving scoring method, and the ratio preserving scoring method. Hence, every respondent had four sets of test scores. Second, we verified for each set of test scores, whether it was weakly rank preserving, strictly rank preserving, and ratio preserving, and we computed the corresponding inconsistency indices I^{weak}, I^{strict}, and I^{ratio}. Table 5 shows the percentages of consistent sets of test scores, and summary statistics (minimum, maximum, and quartile scores) of the distributions of the inconsistency indices.

Test scores obtained using the strict and weak rank preserving scoring method satisfy weak rank preservation by definition. Test scores obtained using the traditional and ratio preserving scoring method satisfied weak rank preservation for approximately half the sample. Most often one violation was encountered. Test scores obtained using the strict rank preserving scoring method satisfy strict rank preservation by definition. For all other scoring methods, the percentage of test scores satisfying strict rank order preservation was small. Test scores obtained using the weak and strict rank order preserving scoring methods are ranks and, therefore, excluded from the results for ratio preservation. No set of test scores satisfied ratio preservation, but the inconsistency indices were smaller for test scores obtained using the ratio preserving scoring method than for test scores obtained using the traditional scoring method.

Inspection of the items of the SOV Part II suggests that religious value may consist of two subtraits (ideological: items 1, 4, 9, and 11; and ecclesiastical: items 2, 5, 7, 12) and this may be a reason to exclude religious value. Table 6 shows the values of the statistics in Table 5 without the trait religious value. The number of consistent sets of test scores increased, and the values of the inconsistency indices decreased. One set of test scores obtained using the ratio preserving scoring method was completely ratio preserving.

6 Discussion

We have argued that the traditional scoring method of MFC items, the summation of the available statement scores, which is suggested in several test manuals yields test scores that cannot be interpreted. Three alternative scoring methods for MFC items were proposed. Software in R (R Development Core Team, 2006) to compute the alternative test scores is available from the second author upon request.

The weak and strict rank preserving scoring methods are useful if a rank order of the traits is required that expresses the preference of the traits for a particular respondent. Test scores with the same value cannot be readily compared. There is a tradeoff between weak and strict rank preserving test scores: Weak rank order test scores have a weaker interpretation but have a smaller probability that test scores of different traits receive the same value. Strict rank order test scores have a stronger interpretation but have a greater probability that test scores of different traits receive the same value. It will depend on the purpose of the test and the consistency of the response patterns which of the two scoring methods is preferred.

The ratio preserving scoring method is useful if a ratio interpretation of the traits within a respondent is required. The resulting test scores are

Table 6: *Results from the Empirical Example Using Five Traits: Percentage of Consistent Test Scores and Summary Statistics of the Distribution of the Inconsistency Index.*

Test scores	Weak rank preservation					
	Perc.	Min.	Q_1	Median	Q_3	Max.
Traditional	70.5%	0	0	0	1	4
Weak rank preserving	100.0%	0	0	0	0	0
Strict rank preserving	100.0%	0	0	0	0	0
Ratio preserving	67.6%	0	0	0	1	3
	Strict rank preservation					
	Perc.	Min.	Q_1	Median	Q_3	Max.
Traditional	9.1%	0	1	2	3	5
Weak rank preserving	30.6%	0	0	2	2	5
Strict rank preserving	100.0%	0	0	0	0	0
Ratio preserving	8.5%	0	1	2	3	6
	Ratio preservation					
	Perc.	Min.	Q_1	Median	Q_3	Max.
Traditional	0.0%	0.96	7.87	9.68	11.89	22.46
Ratio preserving	0.3%	0.00	3.17	4.81	7.12	17.28

Note. Perc. = Percentage of consistent responses; Min. = minimum value of inconsistency index; Q_1 = First quartile of inconsistency index; Median = Median of inconsistency index; Q_3 = Third quartile of inconsistency index; and Max. = maximum value of inconsistency index.

seldomly consistent with respect to ratio preservation because it requires that the Q test scores satisfy $Q(Q-1)/2$ equality constraints. Inconsistency index I^{ratio} can be used to evaluate the consistency of the response pattern of a respondent. Respondents with relatively large values may be possible outliers and their test scores should be interpreted with care. Very popular or very unpopular statements in an item may increase the average value of the inconsistency index.

A pitfall of the ratio preserving scoring method is the handling of zeros, which only disappears if the respondent is administered a very large number of items. The value of ε is always arbitrary and can have a large effect on the resulting test scores. The problem is well known in the related field of compositional data analysis (e.g., Fry, Fry, & McLaren, 2000; Martín-Fernández et al., 2003).

By proposing these alternative scoring methods, we do not intend to advocate the use of ordinal MFC items in future tests or questionnaires. We believe that the problems of ipsative test scores (no absolute interpretation, biased correlation structure, no norm tables possible) are serious. However, there are many existing tests and questionnaires that (1) have an ordinal MFC format and (2) are frequently used. Those tests can benefit from the alternative scoring methods.

Some authors have suggested different scoring methods for MFC items. Unfortunately, these scoring methods are not applicable to existing tests and questionnaires

1. Some authors (De Vries, 2006, chap. 6; Heggestad et al., 2006) have changed the format of the MFC items so that the statement scores do not add up to a constant value per item. They applied the traditional scoring method but the test scores are no longer ipsative. For existing questionnaires this procedure cannot be applied because the MFC item format cannot be changed anymore.

2. McCloy et al. (2005) suggested to use a multidimensional unfolding model for scoring statement scores of MFC items. This is an inventive idea but it requires that the normative P-values are known in advance. McCloy et al. (2005) used P-values obtained from a Likert scale version of their test. However, these are usually not available.

References

Aitchison, J. (1986/2003). *The statistical analysis of compositional data.* London: Chapman and Hall.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology, 24*, 365–379.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69*, 49–56.

Brady, H. E. (1989). Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika, 54*, 181–202.

Broverman, D. M. (1962). Normative and ipsative measurement in psychology. *Psychological Review, 69*, 295–305.

Canfield, A. A. (1980). *Learning Styles Inventory: Manual.* Ann Arbor, MI: Humanics Media.

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

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika, 30*, 99–121.

Chan, W., & Bentler, P. M. (1993). The covariance structure analysis of ipsative data. *Sociological Methods & Research, 22*, 214–247.

Chan, W., & Bentler, P. M. (1996). Covariance structure analysis of partially additive ipsative data using restricted maximum likelihood estimation. *Multivariate Behavioral Research, 31*, 289–312.

Chan, W., & Bentler, P. M. (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika, 63*, 360–369.

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monograph, 14*, vi–56.

Closs, S. J. (1976). Ipsative vs. normative interpretation of interest test scores or ‘What do you mean by “like”?’ *Bulletin of the British Psychological Society, 28*, 289–299.

Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational and Organizational Psychology, 69*, 41–47.

Cornwell, J. M. & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson. *Journal of Occupational and Organizational Psychology, 67*, 89–100.

De Vries, A. L. M. (2006). *The merit of ipsative measurement: Second thoughts and minute doubts*. Unpublished doctoral dissertation, Maastricht University, Maastricht, The Netherlands.

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

Edwards, A. L. (1954). *Edwards Personal Preference Schedule: Manual.* New York: The Psychological Corporation.

Evers, A., Lucassen, W., & Wiegersma, S. (1999). *Beroepen Interessen Test (BIT) versie 1997: Handleiding* [Vocational Interests Test (VIT) version 1997: Manual]. Lisse, The Netherlands: Swets & Zeitlinger.

Fedorak, S. & Coles, E. M. (1979). Ipsative vs. normative interpretation of test scores: A comment on Allen and Foreman's (1976) norms on Edwards Personal Preference Schedule for female Australian therapy students. *Perceptual and Motor Skills, 48*, 919–922.

Fry, J. M., Fry, T. R. L., & McLaren, K. R. (2000). Compositional data analysis and zeros in micro data. *Applied Economics, 32*, 953–959.

Gordon, L. V. (1976). *Survey of Interpersonal Values (revised).* Chicago: Science Research Associates.

Gordon, L. V. (1984). *Survey of Personal Values.* Chicago: Pearson Performance Solutions.

Guilford, J. P. (1952). When not to factor analyze. *Psychological Bulletin, 49*, 26–37.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9–24.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167–184.

Hornung, R. W. & Reed, L. D. (1990). Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene, 5*, 46–51.

Irle, M. (1955). *Berufs-Interessen-Test (B-I-T): Handanweisung* [Vocational Interests Test (V-I-T): Manual]. Göttingen, Germany: Hogrefe.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153–162.

Katz, M. (1962). Interpreting Kuder Preference Record scores: Ipsative or normative. *Vocational Guidance Quarterly, 10*, 96–100.

Kopelman, R. E., Rovenpor, J. L., & Guan, M. (2003). The Study of Values: Construction of the fourth edition. *Journal of Vocational Behavior, 62*, 203–220.

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*, 247–256.

Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology, 35*, 253–278.

McCloy, R. A., Heggestad, E. D. & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8*, 222–248.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*, 263–280.

Pawlowsky-Glahn, V. & Egozcue, J. J. (2002). BLU Estimators and compositional data. *Mathematical Geology, 34*, 259–274.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved, January 14, 2008, from `http://www.R-project.org`.

Radcliffe, J. A. (1963). Some properties of ipsative score matrices and their relevance for some current interest tests. *Australian Journal of Psychology, 15*, 1–11.

Saville, P., Sik, G., Nyfield, G., Hackston, J., & MacIver, R. (1996). A demonstration of the validity of the Occupational Personality Questionnaire (OPQ) in the measurement of job competencies across time and in separate organizations. *Applied Psychology: An International Review, 45*, 243–262.

Stanush, P. L. (1997). Factors that influence the susceptibility of self-report inventories to distortion: A meta-analytic investigation (Doctoral dissertation, Texas A&M University, 1997). *Dissertations Abstracts International, Section B: The Sciences and Engineering, 58*, 2167.

Ten Berge, J. M. F. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research, 34*, 89–102.