Journal of Classification 26:329-360 (2009)
DOI: 10.1007/s00357-009-9039-7

Unfolding Incomplete Data: Guidelines for Unfolding Row-Conditional Rank Order Data with Random Missings

Frank M.T.A. Busing

Department of Psychology, Leiden University

Mark de Rooij

Department of Psychology, Leiden University

Abstract: Unfolding creates configurations from preference information. In this paper, it is argued that not all preference information needs to be collected and that good solutions are still obtained, even when more than half of the data is missing. Simulation studies are conducted to compare missing data treatments, sources of missing data, and designs for the specification of missing data. Guidelines are provided and used in actual practice.

Keywords: Unfolding; Incomplete data; Missing data; BIBD; PREFSCAL.

This research was conducted while the second author was sponsored by the Netherlands Organization for Scientific Research (NWO), Innovational Grant, no. 452-06-002. The authors would like to thank the DIOS for their invaluable computer support and Willem Heiser and Graham Cleaver and the three anonymous referees for their helpful comments and suggestions to improve the quality of this work.

Corresponding Author's Address: Frank M.T.A. Busing, Department of Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, the Netherlands, e-mail: busing@fsw.leidenuniv.nl

Published online 5 January 2010

1. Introduction

Multidimensional unfolding methods create perceptual spaces well-suited for consumers research (DeSarbo, Kim, Choi, and Spaulding 2002) and marketing research (Balabanis and Diamantopoulos 2004; DeSarbo, Young, and Rangaswamy 1997). Unfolding represents both consumers and products as points in Euclidean space. The distance relation between consumers and products provides information about the preference structure of the consumers in such a way that consumers are closer to the products they prefer. The geometrical properties of the Euclidean space allow for simple and comprehensible interpretation of the relationships.

In consumer or marketing research, it would be more than convenient if consumer respondents only evaluate a subset of products. Respondents may be unable or unwilling to comply and fail to complete the evaluation of the full set of products. For example, in memory-based evaluations, respondents must have knowledge or at least be aware of the products under consideration to provide a useful evaluation. Without providing the respondents with additional information, which may be undesirable for several reasons, the evaluation set for each respondent might differ: certain familiar products are evaluated more often than other products and some respondents evaluate more products than other respondents. Free-choice profiling (Arnold and Williams 1986; Dijksterhuis and Gower 1991) and the repertory grid method (Kelly 1955; Rowe, Lambert, Bowling, Ebrahim, Wakeling, and Thomson 2005) also provide unequally distributed incomplete data, as respondents exploit different vocabularies. On the other hand, in studies involving tasting food products, sensory fatigue is a real issue which means that one wants to restrict the number of products tasted. All respondents evaluate an equally sized subset of all products under evaluation. The expensive alternative is asking respondents to return on another occasion to complete the entire evaluation.

From a technical point of view, it is not at all necessary that the respondents evaluate all products. Most unfolding procedures allow for missing data without falling back on complete case analysis or listwise deletion. Consumer respondents may judge a subset of products and the unfolding procedures only use the valid, non-missing data, either by skipping the missing observations during computations (pairwise deletion) or by inserting "valid" data at the missing observation, i.e., by imputation of missing data. Obtaining a good solution with incomplete data can help researchers get the most out of limited resources.

The focus of this paper is (1) to investigate missing data designs for unfolding, (2) to determine key success factors for unfolding incomplete data, and (3) to provide guidelines for the proportion of non-missing data, required for a good correspondence with the results of a complete data analysis. In the following, we will first present unfolding, then briefly discuss the degeneracy problem that haunted this technique for decades, and discuss how this problem is currently solved. Then, incomplete data designs are presented that will be used in the Monte Carlo simulation study that aims to provide guidelines for researchers and data collectors, as to the amount of data that is still sufficient for proper solutions. An example with empirical data will be shown and we conclude with some general remarks.

2. Unfolding

Multidimensional unfolding is a technique that finds low-dimensional configurations for two sets of objects, the consumers and the products. The distances in the configuration between consumers and products should correspond as closely as possible with the preference ratings of the consumers for these products, in such a way that consumers are closest to the products they prefer the most. Unfolding in general consists of several different models. In this paper, we use the model initiated by Coombs (1950) and generalized to the multidimensional case by Bennett and Hayes (1960). In this model, n consumers and m products are represented as points in multidimensional space. The coordinate $\mathbf{x}_i$ of a consumer is generally referred to as its ideal point; hence, this model is called the ideal-point model. The closer a product is to a consumers ideal point, the more this product is preferred by this consumer. Specific models have been suggested (external unfolding, weighted unfolding, Carroll 1972), but the most influential model was the nonmetric model, which joined nonmetric data and metric distances. Typical "unfolding data" consist of rankings of products. These ranking data only contain ordinal information (i.e., no metric information) and the data are thus called nonmetric. Shepard (1962) showed that transformations, specifically ordinal transformations, can be used to shape this nonmetric relation in multidimensional scaling. Keeping the order relations of the original data intact, ordinal or nonmetric data are transformed into intermediate ratio data, which in turn are used to construct a metric Euclidean space (Kruskal 1964). Kruskal proposed to use the standardized residual sum of squares, abbreviated "Stress", with Stress-I (Kruskal and Carroll 1969) given as

$$\sigma_I(\boldsymbol{\Delta}, \mathbf{X}, \mathbf{Y}) = \|\boldsymbol{\gamma} - \mathbf{d}\|^2 / \|\mathbf{d}\|^2,$$

where $\boldsymbol{\Delta}$ is an $n \times m$ matrix with preferences and $\mathbf{X}$ and $\mathbf{Y}$ are the $n \times p$ and $m \times p$ coordinate matrices for consumers and products, respectively. In the unfolding case, $\|\boldsymbol{\gamma} - \mathbf{d}\|^2$ is the squared Euclidean norm $\|\cdot\|^2$ of the differences between some monotonic transformation $f(\cdot)$ of the consumer's preferences $\boldsymbol{\Delta}$, with $\boldsymbol{\gamma} = f(\boldsymbol{\Delta})$, and the distances $\mathbf{d} = d(\mathbf{X}, \mathbf{Y})$, where

$\boldsymbol{\gamma} = \text{vec}\,(\boldsymbol{\Gamma})$ and $\mathbf{d} = \text{vec}\,(\mathbf{D})$. The vec operator stacks the columns of its matrix argument. Standardization is regulated by $\|\mathbf{d}\|^2$, the sum-of-squares of the distances.

Nonmetric multidimensional scaling was one of the biggest breakthroughs in psychological research methods, but it ultimately caused unfolding's existential crisis: the freedom of the coordinates in space *and* the almost unrestricted transformations ensured that the thus weakly constrained unfolding model (Lingoes 1977) was no longer identifiable (Busing 2006). As a result, analyses tend to produce perfect (in terms of loss function) but meaningless (in terms of interpretation) configurations of points (Kruskal and Carroll 1969; Roskam 1968). Attempts to resolve the degeneracy problem often ended up in relatively unknown procedures or procedures with still uncertain results (see Borg and Groenen 2005; Busing, Groenen, and Heiser 2005, for an overview). Currently, there is a revival of attempts to set afloat unfolding with more prominent results (Kim, Rangaswamy, and DeSarbo 1999); Busing, Groenen, and Heiser 2005; Busing 2006; van Deun, Groenen, and Delbeke 2005; van Deun, Heiser and Delbeke 2007). All these attempts somehow restrict the model, either by restricting the transformations or by restricting the coordinates. An overview of the history of unfolding, the degeneracy problem, and currently available (computer) procedures can be found in van Deun (2005).

To avoid the degeneracy problem, researchers often restrict themselves to a metric unfolding analysis. Although this paper focusses on nonmetric unfolding analyses, a comparison is made between both types of analysis, but the results are relegated to Appendix A. For the nonmetric unfolding analyses, we use the penalty approach of Busing et al. (2005) implemented in PREFSCAL (available in SPSS), which avoids degeneracy by penalizing on the coefficient of variation $v(\cdot)$ (Pearson 1896). Solutions with no or low variation in the transformed preferences and/or distances, characteristics of degenerate solutions, are avoided by dividing normalized raw Stress $\sigma_n(\boldsymbol{\Delta}, \mathbf{X}, \mathbf{Y})$ (see Appendix B) with a function of the variation coefficient $v(\boldsymbol{\gamma})$, i.e., the standard deviation of $\boldsymbol{\gamma}$ divided by its mean. The division causes the loss function to attain minimum values only in combination with a definite non-zero variation coefficient, that is, with sufficient variation in transformed preferences. Penalized Stress is defined as

$$\sigma_p(\boldsymbol{\Delta}, \mathbf{X}, \mathbf{Y}, \omega, \lambda) = \sigma_n(\boldsymbol{\Delta}, \mathbf{X}, \mathbf{Y})/\mu(\boldsymbol{\Delta}, \omega, \lambda)$$

where $\mu(\boldsymbol{\Delta}, \omega, \lambda) = 1 + \omega v(\boldsymbol{\gamma})^{2/\lambda}$ with penalty parameters $\omega \geq 0$ and $0 < \lambda \leq 1$. Strong penalty parameters, with high values for ω and values for λ closer to zero, tend to produce linear transformations, whereas weak penalty parameters, with ω close to zero and λ closer to one, are prone to result in degenerate solutions. Details can be found in Busing et al. (2005), although the function currently implemented in SPSS PREFSCAL deviates slightly from the function presented in Busing et al. (2005): Normalized raw Stress (normalization done with the sum-of-squares of the transformed preferences) is used instead of raw Stress (no normalization) and an additional constant ($\upsilon(\boldsymbol{\delta})^2$) is used in combination with ω. The default value for ω changed from 0.5 to 1.0 under the influence of this last addition.

In subsequent sections, the default settings of PREFSCAL are used: classical scaling start with data imputation based on the triangle inequality (Heiser and de Leeuw 1979), row-conditional, ordinal (ties are kept tied) transformations, and default values for penalty parameters and convergence criteria, except for the maximum number of iterations, which was doubled to prevent imprecision due to premature termination of the iterative algorithm. Important in the present context is that PREFSCAL allows for a preference weight matrix with fixed nonnegative weights. When this matrix is specified as an incidence matrix (a matrix with solely zeros and ones) it allows for the specification of missing data.

3. Missing Data

Missing data can be initiated by the researcher when only a subset of the products is presented to a respondent for evaluation. Proper factorial designs can be used to define subsets which in turn can be randomly assigned to respondents. On the other hand, missing data may be a consequence of the knowledge set of the respondent. In this case, missing data might be irregularly distributed over both respondents and products. Whatever the source of the missing data, unfolding needs to cope with the fact that some data is absent.

3.1 Handling Missing Data

In general, there are two common approaches for dealing with missing data: deletion and imputation. The first approach simply excludes cases containing missing data, either for all computations (listwise deletion), or only for those computations where a missing for that case is involved (pairwise deletion). Either deletion scheme, listwise or pairwise, ignores possible systematic differences between complete and incomplete samples and produces unbiased estimates only if deleted records are a random sub-sample of the original sample. Data imputation, on the other hand, replaces missing data with "valid" data through either single (deterministic) or multiple (stochastic) imputation and could lead to the minimization of bias. However, no imputation model is free of assumptions and the imputation results should hence be thoroughly checked for their statistical properties, such as

Table 1. Incomplete Block Design for $v = 15$, $k = 5$, $r = 14$, and $b = 42$, taken from Design Computing* Specifying the Column Numbers with Either Missing or Non-Missing Data for a 42×15 Incidence Matrix. The Entries Indicate the 5 Column Numbers, Displayed in 6 Blocks of 7 Rows.

12	3	9	13	14	9	13	1	4	12	8	5	1	12	6	14	10	3	6	12	12	14	2	10	7	15	4	1	2	13
6	11	15	4	3	1	10	9	2	15	11	12	13	14	15	13	7	5	9	15	3	8	13	10	11	3	1	4	10	5
8	1	3	9	15	12	9	2	5	6	15	4	5	14	8	13	3	12	5	2	14	7	2	5	1	3	6	13	9	7
11	2	5	7	3	4	9	7	8	12	14	7	1	9	11	5	10	6	9	15	1	6	13	14	8	11	1	13	6	5
1	4	12	3	10	7	15	11	8	12	4	7	13	2	10	13	15	8	10	2	7	13	8	10	6	11	6	9	4	2
6	7	11	1	10	9	4	10	14	11	6	12	15	7	4	1	7	15	14	3	3	11	8	9	2	5	10	12	15	11
2	6	3	14	15	2	4	8	14	6	7	4	8	5	3	10	14	8	9	5	11	12	8	1	2	11	13	4	5	14

* see http://www.designcomputing.net/gendex/bib/b4.html

distributional characteristics as well as heuristically for their meaningfulness (e.g., whether, for example, negative imputed values are possible). See Little and Rubin (1987) for well-documented drawbacks of either approach. We will now compare the two methods in a small simulation study.

3.2 Simulation Study: Imputation versus Deletion

The breakfast data (Green and Rao 1972; DeSarbo et al. 1997; Borg and Groenen 2005; Busing et al. 2005; Van Deun 2005) for which 21 MBA students and their wives rank ordered 15 breakfast items, are used to compare the recovery of the complete data solution using three methods: deletion (no imputation), respondent average imputation, and product average imputation. To create incomplete data, 5 out of 15 items per respondent were set missing by specifying an incidence matrix using a known balanced incomplete block design (see Table 1). Each method was replicated 1000 times with permutated rows and columns of the incidence matrix on each instance.

The quality of the equivalence between the distances of the unfolding analyses with and without missing data, i.e., the recovery of the unfolding solution, is quantified using Tucker's congruence coefficient ϕ_{xy} comparing both sets (respondents and products) and ϕ_y comparing the product sets only (Burt 1948; Tucker 1951). By using ϕ, a scale-independent similarity measure for ratio data, a Procrustes analysis to match configurations becomes superfluous, since the distances are independent of rotation and translation and the congruence coefficient is independent of dilation (see Appendix B). At the individual level, the influence of missing data is determined with Kendall's rank order correlation τ_b (Kendall 1948), comparing the rank ordered distances of the complete and incomplete data solutions for identical respondents, averaged over respondents. The comparison measures ϕ_{xy}, ϕ_y,

Table 2. Descriptive Statistics (Upper Part, with Means and Standard Deviations in Parentheses) and MANOVA Tests of Between-Subjects Effects (Lower Part, with F-Statistics, Significance in Parenthesis, and Effect Sizes on the Second Line) Comparing Recovery of Unfolding Solutions using Deletion (No Imputation), Respondent Average Imputation, and Product Average Imputation Methods.

Method	ϕ_{xy}	ϕ_y	τ_b
Deletion	.957 (.015)	.967 (.019)	.661 (.054)
Respondent Average Imputation	.964 (.008)	.957 (.025)	.645 (.055)
Product Average Imputation	.965 (.008)	.962 (.023)	.658 (.047)
Between-Subjects Effects	ϕ_{xy}	ϕ_y	τ_b
F (p)	157.536 (.000)	48.601 (.000)	27.169 (.000)
η_p^2	.096	.032	.018

and τ_b, take all distances into account, also the distances associated with missing data.

A multivariate analysis of variance indicates a significant overall difference between the recovery capabilities of the three imputation methods (using Wilks' Lambda: $F(6, 5944) = 227.990; p < .001; \eta_p^2 = .187$). Table 2 provides descriptive statistics and the tests of the between-subject effects, including effect sizes, expressed as partial eta squared (η_p^2). For the simulation studies, emphasis is on the effect sizes as the number of replications can always be increased to obtain significant results. Here, all differences are significant, but the descriptive statistics and the effect sizes indicate that the differences in recovery are not very serious. According to Cohen (1988), a partial eta squared of .010 indicates a small effect, .059 a medium effect, and .138 a large effect. The deletion method is slightly better than the imputation methods for ϕ_y and τ_b, but worse for ϕ_{xy}.

The actual solutions from the incomplete data are superior for the deletion method. Table 3 shows Stress-I, indicated by σ_I^- based on the valid data only, and the rank order correlations (τ_b^-). The overall difference is significant (using Wilk's Lambda: $F(4, 5946) = 6145.930; p < .001; \eta_p^2 = .805$) and the tests of the between-subject effects show large effects for all measures in favor of the deletion method. Where the deletion method improves considerably on Stress-I and rank order correlations as compared to the complete data solution (with $\sigma_I = .241$ and $\tau_b = .701$, respectively), the imputation methods worsen (see descriptives from Table 3). The introduction of additional error by imputation, causes higher Stress-I values for the imputation methods. The deletion method uses its freedom to find a better solution, mainly in the transformation part of the loss function, but without deviating from the imputation methods concerning the recovery of the unfolding solution. In conclusion, due to the inconclusive recovery results,

Table 3. Descriptive Statistics (Upper Part, with Means and Standard Deviations in Parentheses) and MANOVA Tests of Between-Subjects Effects (Lower Part, with F-Statistics, Significance in Parenthesis, and Effect Sizes on the Second Line) Comparing Fit of Unfolding Solutions using Deletion (No Imputation), Respondent Average Imputation, and Product Average Imputation Methods.

Method	σ_I^-	τ_b^-
Deletion	.164 (.025)	.770 (.022)
Respondent Average Imputation	.298 (.013)	.545 (.023)
Product Average Imputation	.273 (.015)	.618 (.023)
Between-Subjects Effects	σ_I^-	τ_b^-
F (p)	15160.740 (.000)	25362.813 (.000)
η_p^2	.911	.945

the better actual solutions for the deletion method, and the absence of assumptions concerning the missing data, the deletion method is preferred for further analysis.

3.3 Missing Data by Researcher

Researchers may only want to provide a subset of products to a respondent for evaluation. These planned missings both reduce the burden on respondents, improving the quality of their evaluations, and save time and money. With the missing data under the control of the researcher, the missing completely at random (MCAR) assumptions apply, if the missings are properly randomized (Little and Rubin 1987). To determine which subset of products is presented to a respondent, simple missing data designs can be considered, but since the relations between objects of different sets are in order, rather than just means, more complicated fractional block designs might be necessary. A balanced incomplete block design (BIBD) (Cochran and Cox 1957) is such a sophisticated fractional block design. A BIBD is usually defined as an arrangement of v distinct objects in b blocks, such that each block contains k distinct objects, each object occurs in exactly r different blocks, and every two objects occur together in exactly λ blocks (definition by Prestwich 2001). Convenient for the current topic, with the same set of parameters, a BIBD can be defined in terms of an incidence matrix $\mathbf{I}$, which is then a binary matrix with v rows and b columns where each row sums to r and each column sums to k (see, for an example, Figure 1, left panel). Any pair of distinct rows has scalar product $\lambda = \mathbf{I}_i'\mathbf{I}_j$ for all $i \neq j$. Since the parameters are not independent ($vr = bk$ and $\lambda(v-1) = r(k-1)$), a BIBD is commonly expressed in terms of v, k, and λ. The term "incomplete" stems from the fact that $k < v$.

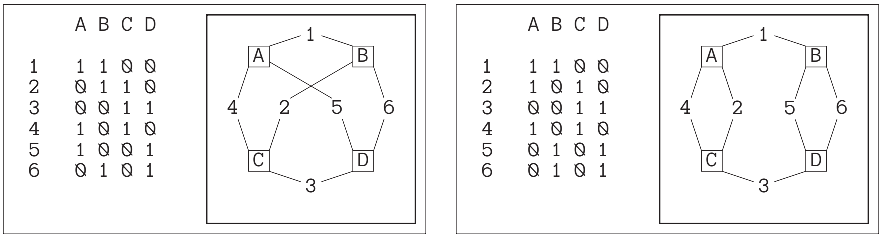


Figure 1. Example of a Balanced Incomplete Block Design (BIBD) (left panel) and a Row-Balanced Incomplete Block Design (row-BIBD) (right panel), where valid data is represented by a connection (line) between respondents (numbers) and products (letters).

Although the description of a BIBD is relatively simple, the generation of a BIBD is a complex problem. BIBD's are only available for a limited series of particular parameter values (see for example Clatworthy's (1973) catalog) and solvable for small parameter values within an acceptable period of time (see Nguyen 1994). These features pose a serious problem for a Monte Carlo study, which depends on fast problem solving, handling thousands of computational problems within a limited time period. To overcome this practical problem, a design is explored where each row sums to r, but where the requirement of sum k per column and scalar product λ between rows is relaxed. Still, $vr = bk$, but now k is a random variable with mean vr/b and some variation. This design is referred to as a row-balanced incomplete block design (row-BIBD), indicating that every respondent evaluates the same number (r) of products, but products might not be evaluated the same number (k) of times (an example is given in Figure 1, right panel). Different k's for different columns seems to be a minor problem, since the number of consumers is typically large compared to the number of products.

3.4 Simulation Study: BIBD versus Row-BIBD

The breakfast data are used to evaluate the difference between a BIBD and a row-BIBD concerning the recovery of unfolding solutions. A known BIBD, with $v = 15$, $k = 5$, and $\lambda = 8$ (Nguyen 1993, 1994), is used that matched the breakfast data. The resulting incidence matrix is transposed and zero's and ones are interchanged to get a BIBD with $v = 42$, $k = 14$, and $\lambda = 18$ (see Table 1). The incomplete unfolding using the BIBD is replicated 1000 times, each time randomly interchanging rows and columns of the incidence matrix. For the analyses using the row-BIBD, 1000 runs are conducted creating another incidence structure on each instance. Both designs exclude 5 out of 15 products per respondent.

A multivariate analysis of variance indicates a significant overall difference between the recovery capabilities of both designs (using Wilks'

Table 4. Descriptive Statistics (Upper Part, with Means and Standard Deviations in Parentheses) and MANOVA Tests of Between-Subjects Effects (Lower Part, with F-Statistics, Significance in Parenthesis, and Effect Sizes on the Second Line) Comparing Recovery of Unfolding Solutions using Balanced Incomplete Block Designs (BIBD) and Row-Balanced Incomplete Block Designs (row-BIBD).

Design	ϕ_{xy}	ϕ_y	τ_b
BIBD	.957 (.014)	.967 (.018)	.659 (.055)
row-BIBD	.957 (.016)	.968 (.017)	.662 (.054)
Between-Subjects Effects	ϕ_{xy}	ϕ_y	τ_b
F (p)	.155 (.694)	2.777 (.096)	2.583 (.108)
η_p^2	.000	.001	.001

Lambda: $F(3, 1996) = 2.840; p = .037$). The effect size, however, is very small ($\eta_p^2 = .004$), which is reflected in the tests of the between-subject effects provided in Table 4: none of the three statistics shows a significant result. The descriptive statistics also show that the differences are very small, which leads to the conclusion that both designs perform alike. Subsequently, the more flexible and faster row-BIBD is used to specify the missing data by researcher.

3.5 Missing Data by Respondent

In memory-based evaluations, only products that are known to the respondents are available for evaluation. If a researcher still offers all products to the respondents, the results for the unknown products will mostly be neutral, random, invalid, or missing. Shocker, Ben-Akiva, Boccara, and Nedungadi (1991) discuss a hierarchical chain of sets modeling decision-making. In their view, consumers use an universal set, which contains an awareness or knowledge set, which in turn contains a consideration set, which contains a choice set, which finally contains the product of choice. Each set is smaller in number of products than or equal to the previous set. For the present research this means that the researcher offers a universal set for evaluation to all respondents, but respondents only evaluate the products they know, i.e., products from their knowledge set. For an even higher quality of their evaluations, the researcher might persuade respondents to use their consideration set or even their choice set. As a result of the use of knowledge sets, different respondents may evaluate a different number of products, simply because some respondents know more products than others. A simulation study is used to determine the consequences of the variation in the number of products per respondent for the recovery of the unfolding solutions. The products in the knowledge sets might be uniformly distributed over the entire

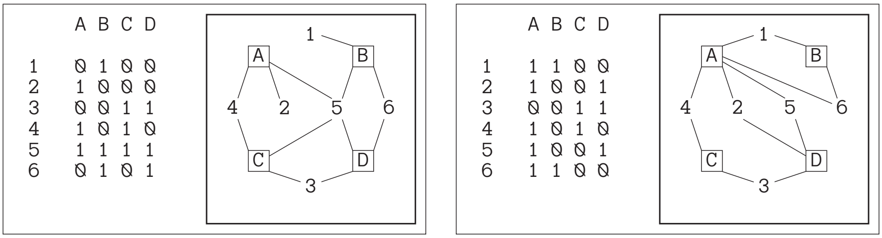


Figure 2. Example of a Knowledge Set Design (left panel) and a Product Familiarity Design (right panel), where valid data is represented by a connection (line) between respondents (numbers) and products (letters).

product range, but this is not expected in practice. Some products are more familiar than other products, for example due to more advertising, longer existence, or wider availability. As a result, the knowledge sets are expected to be unequally distributed over the products. The impact of the unequal number of respondents per product on the recovery of unfolding solutions is investigated in another simulation study.

The missing data in the present case is related to the respondents (knowledge sets) or to the products (familiarity), hence the missing data is not completely at random (MCAR), but only missing at random (MAR). Since the MCAR assumptions no longer apply, it is imperative that additional analyses are performed to get a thorough insight in the distribution of the missing data over both respondents and products.

3.6 Simulation Study: Knowledge Sets

The breakfast data are used to determine the influence of knowledge sets on the recovery of unfolding solutions using incomplete data. To simulate knowledge sets, the number of evaluations per respondent is varied. The variation is set by drawing the number of non-missings from a normal distribution with mean 10 (as in the previous simulation studies) and standard deviation a. The minimum number of non-missings is set to 2. An example of a *knowledge set design* is given in Figure 2 (left panel), where different respondents know a different number of products, represented by a line between respondents (numbers) and products (letters). The levels of factor a are set from 0 to 10 with steps of 1, where $a = 0$ specifies no variation, thus 10 evaluations per respondent exactly. This study uses 1000 replications of incomplete data per level of factor a.

A multivariate analysis of variance indicates a significant difference (using Wilks's Lambda: $F(30, 32247) = 304.186; p < .001; \eta_p^2 = .216$) between the variation levels a. The between-subject effects (Table 5, lower

Table 5. Descriptive Statistics (Upper Part, with Means and Standard Deviations in Parentheses) and MANOVA Tests of Between-Subjects Effects (Lower Part, with F-Statistics, Significance in Parenthesis, and Effect Sizes on the Second Line) Comparing Recovery of Unfolding Solutions using Missing Data Designs with Different Levels of Variation in Number of Products per Respondent (a).

Variation Level a	ϕ_{xy}	ϕ_y	τ_b
0	.957 (.015)	.966 (.019)	.657 (.055)
1	.952 (.015)	.964 (.019)	.649 (.056)
2	.941 (.019)	.960 (.020)	.633 (.054)
3	.923 (.026)	.950 (.022)	.604 (.056)
4	.909 (.028)	.941 (.023)	.579 (.054)
5	.892 (.037)	.931 (.025)	.559 (.056)
6	.883 (.040)	.928 (.026)	.545 (.058)
7	.875 (.042)	.923 (.027)	.537 (.058)
8	.869 (.046)	.921 (.029)	.529 (.060)
9	.866 (.047)	.920 (.029)	.527 (.061)
10	.863 (.049)	.918 (.027)	.521 (.060)
Between-Subjects Effects	ϕ_{xy}	ϕ_y	τ_b
F (p)	1006.808 (.000)	587.935 (.000)	806.964 (.000)
η_p^2	.478	.349	.423

part) show significant differences with large effects and the descriptive statistics of all recovery measures give the same result: as the variation level a increases, i.e., as the differences in number of evaluations per respondent increase, the recovery of the unfolding solutions worsens. Since the total number of missings is equal for all levels of a, the variation in number of missings definitely influences recovery. Especially the respondent points suffer from the variation, as can be concluded from the effect sizes and the differences in decrease of ϕ_{xy}, ϕ_y, and τ_b, for increasing a.

3.7 Simulation Study: Product Familiarity

The breakfast data are used to determine the influence of product familiarity on the recovery of unfolding solutions using incomplete data. Product familiarity is reflected by increasing the chance that a product is chosen for evaluation, which differs from the approach taken by Chatterjee and DeSarbo (1992), where familiarity is linked with reliability and preferences require additional uncertainty information. For the current simulation study, the chance to choose the first 3 products (20%) for evaluation is b times greater than the chance to choose the remaining products, thus defining high familiar and low familiar products. Corresponding comparison measures ϕ_{xy}^{high} and ϕ_{xy}^{low} only use the distances between the respondents and the products under consideration, that is, the first 20% or the last 80% of the prod-

Table 6. Descriptive Statistics (Upper Part, with Means and Standard Deviations in Parentheses) and MANOVA Tests of Between-Subjects Effects (Lower Part, with F-Statistics, Significance in Parenthesis, and Effect Sizes on the Second Line) Comparing Recovery of Unfolding Solution using Missing Data Designs with Different Familiarity Level of the Products (b).

Familiarity Level b	ϕ_{xy}	ϕ_y	τ_b
1	.957 (.014)	.967 (.018)	.659 (.056)
2	.957 (.015)	.969 (.017)	.662 (.052)
3	.957 (.016)	.968 (.017)	.665 (.051)
4	.958 (.014)	.970 (.016)	.668 (.048)
5	.957 (.017)	.969 (.016)	.665 (.050)
6	.957 (.017)	.968 (.016)	.666 (.049)
7	.957 (.017)	.969 (.016)	.668 (.050)
8	.957 (.017)	.969 (.017)	.667 (.051)
9	.957 (.016)	.968 (.017)	.667 (.049)
10	.958 (.014)	.969 (.016)	.668 (.048)
Between-Subjects Effects	ϕ_{xy}	ϕ_y	τ_b
F (p)	.846 (.574)	2.613 (.005)	3.061 (.001)
η_p^2	.001	.002	.003

ucts. The levels of factor b are set from 1 to 10 with steps of 1, with equal chances for $b = 1$. An example of the missing data design is given in Figure 2 (right panel). This study uses 1000 replications of incomplete data for each level of factor b.

A multivariate analysis of variance indicates significant differences (using Wilks's Lambda: $F(27, 29165) = 2.186; p < .001$) between the familiarity levels b, but with an effect size close to zero ($\eta_p^2 = .002$). The between-subject effects indicate that the differences are due to ϕ_y and τ_b, but also with an effect sizes close to zero (see lower part of Table 6). Comparing the familiarity levels shows that only $b = 1$ is responsible for the differences, and not even with all other levels of b. It is nevertheless save to conclude that the familiarity level has no influence on the recovery of the unfolding solutions.

However, some of the products are more familiar than others and it is the difference between these two sets, high familiar and low familiar products, we are interested in. Table 7 (lower part) gives the results of a two-way analysis of variance with familiarity level and high-low familiarity as fixed factors. As for the multivariate analysis of variance, the familiarity level has a significant effect on recovery, but an effect size close to zero. The difference, however, between high familiar products and low familiar products is significant, together with a large effect size ($\eta_p^2 = .245$). It is therefor important to make a distinction between high familiar and low familiar products, whereas the familiarity level is a matter of secondary significance.

Table 7. Average Congruence Coefficients for High Familiar and Low Familiar Products (Upper Part, with Standard Deviations in Parentheses), and the Univariate Two-Way Analysis of Variance (Lower Part) Comparing Recovery of High Familiar and Low Familiar Products in Unfolding Solutions using Missing Data Designs with Different Familiarity Levels of the Products (b).

Descriptive Statistics

Familiarity Level b	ϕ_{xy}^{high}	ϕ_{xy}^{low}
1	.965 (.014)	.953 (.015)
2	.968 (.013)	.953 (.016)
3	.970 (.014)	.953 (.017)
4	.971 (.012)	.953 (.016)
5	.970 (.015)	.952 (.018)
6	.971 (.015)	.952 (.018)
7	.971 (.015)	.952 (.019)
8	.971 (.014)	.952 (.018)
9	.971 (.014)	.952 (.017)
10	.972 (.012)	.953 (.015)

Univariate Two-Way Analysis of Variance

Source	SS	df	MS	F	p	η_p^2
Familiarity Level b	.016	9	.002	7.468	.000	.003
High-Low	1.539	1	1.539	6485.668	.000	.245
Interaction	.027	9	.003	12.470	.000	.006

3.8 Impossible Missing Data

Unfolding is unable to compute a solution from an unconnected block design and it is therefore required that the incidence graph of any block design previously discussed is connected (i.e., that there exists a path joining any two of its vertices). In Figure 3, an example is shown of an unconnected design, as one small block, with respondent `4` and product `C`, is not connected with the large block of respondents (`1`, `2`, `3`, `5`, and `6`) and products (`A`, `B`, and `D`). Determining the positions of both blocks with respect to one another is impossible. Thus, we will ensure in the following that each design is connected.

4. Monte Carlo Simulation Study

A comparison is made between unfolding on a complete and an incomplete set of data, for which an incidence matrix is used to specify the incomplete set of data. The current Monte Carlo simulation study attempts to determine key success factors for unfolding with incomplete data and aims at providing guidelines for researchers and data collectors.

Data is generated according to the model of Wagenaar and Padmos (1971), that is, $\delta_{ij} = \|\mathbf{x}_i - \mathbf{y}_j\| \times \exp^{N(0,e)}$. After generating $i = 1, \ldots, n$

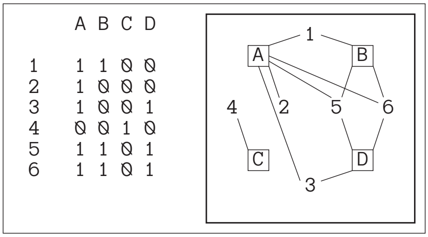


Figure 3. Example of an Unconnected Design, where valid data is represented by a connection (line) between respondents (numbers) and products (letters).

points for the respondents and $j = 1, \ldots, m$ points for the products in a p-dimensional space from a uniform distribution, 5% outliers are created in each set. Using the distances from the centroid of the configuration, points are shifted 1.5–3.0 times the interquartile range of the distances outside the maximum distance from the centroid to become an outlier. This choice is similar to the outlier definition in boxplots (see, for example, SPSS 2006), when applied to the distances of points to the origin. Next, the distances between the sets are computed and perturbed by multiplying them with a log normal distribution. The levels of error are roughly equivalent to Kruskal's Stress-I values corresponding with a perfect to a very poor fit (Kruskal 1964), with slightly higher Stress-I values for the three dimensional case. For each respondent, the (error-perturbed) distances are replaced with their rank number. The variation in the rank numbers, expressed in values of Kendall's rank order correlation τ_b, average 0.87 and 0.70 for the error levels 0.10 and 0.25, respectively. The levels for the independent factors in the simulation study are summarized in Table 8.

For each generated data set, a complete unfolding as well as $m - p$ incomplete unfolding solutions are computed, as the number of inclusions (i, i.e., the number of non-missing products), starts at p (the dimensionality) and ends at $m - 1$ (the total number of products minus one, i.e., with one missing per respondent). The factors from Table 8 are studied in a fully crossed factorial design with 1000 replications for each cell. Cases for which the incidence matrix is not connected or with insufficient free parameters are excluded from further analyses.

Based on the results of the simulation studies from the previous section, two types of incidence matrices are used to specify the incomplete data. The first type specifies missing data by researcher with a row-BIBD, where each respondent evaluates the same number of products and products are evaluated about the same number of times. The second type of incidence

Table 8. Summary of Independent Factors and Accompanying Levels for the Simulation Study.

Factor	Description	#	Levels
n	number of respondents	5	10, 20, 40, 80, 160
m	number of products	4	5, 10, 20, 40
p	number of dimensions	2	2, 3
e	error level	3	0.00, 0.10, 0.25

matrices specifies missing data by respondent, where the number of evaluations per respondent varies depending on the number of products ($a = m/4$) and 20% of the products (high familiar products) are evaluated $b = 10$ times more often than others (low familiar products).

4.1 Results and Guidelines for Handling Missing Data by Design

The influence on recovery for the factors from Table 8 are determined with a multivariate analysis of covariance (main effects and 2-way interactions only), where the continuous variable inclusion proportion ($\text{prop}(i) = i/m$) is specified as a covariate. All multivariate tests are significant ($p < .001$), but with varying effect sizes. As indicated by the effect sizes of the multivariate effects (Table 9, second column), there is better recovery for data with fewer missings, more products (m), and more observations ($n \times m$). It is also beneficial to have data with a low level of error (e), while increasing the number of respondents (n) or changing the dimensionality (p) only has a marginal effect on recovery. The tests of the between-subject effects are also significant ($p < .001$) for all factors and for all recovery measures. Table 9 shows the effect sizes in the last three columns. These results lead to the same key success factors. Additionally to the multivariate effects, the number of respondents (n) does influence the recovery of the product configuration (ϕ_y) and the rank order recovery per respondent (τ_b), as $\eta_p^2 = .067$ and $\eta_p^2 = .068$, respectively. The number of observations ($n \times m$) has a large effect on the rank order correlation with $\eta_p^2 = .173$.

Figure 4 provides guidelines for applied research when the researcher is in control of the missing data. The panels show I-beams and markers for all factors of the Monte Carlo simulation study, except for dimensionality, which has an insufficient effect on recovery to be included. We first explain the elements of such I-beam plots and then indicate how they should be read. The I-beam and markers, i.e., the high, low, and close in high-low graphs, indicate high, low and medium recovery. For the congruence coefficients ϕ, these indicators correspond with the values .99, .95, and .98, respectively. Although Tucker (1951) employs .80 and Cureton and D'Agostino (1983)

Table 9. Effect Sizes for the Main Effects (Wilks' Lambda) and Effect Sizes for the Tests of the Between-Subject Effects of the Multivariate Covariance Analysis Comparing the Recovery of Unfolding Solutions for Different Number of Respondents (n), Number of Products (m), Number of Dimensions (p), and Error Levels (e), with Inclusion Proportion ($\text{prop}(i)$) as Covariate.

Source	Wilks' λ	ϕ_{xy}	ϕ_y	τ_b
$\text{prop}(i)$	$.551^{***}$	$.214^{***}$	$.172^{***}$	$.538^{***}$
n	$.041^{*}$	.004	$.067^{**}$	$.068^{**}$
m	$.071^{**}$	$.105^{**}$	$.024^{*}$	$.134^{**}$
p	$.039^{*}$	$.018^{*}$	.008	.002
e	$.061^{**}$	$.066^{**}$	$.027^{*}$	$.096^{**}$
$n \times m$	$.072^{**}$	$.083^{**}$	$.106^{**}$	$.173^{***}$
$n \times p$	$.016^{*}$	.005	$.016^{*}$	.005
$m \times p$	.005	.005	.002	.008
$n \times e$	$.016^{*}$	$.010^{*}$	$.037^{*}$	$.016^{*}$
$m \times e$	$.018^{*}$	$.032^{*}$	.004	.007
$p \times e$	.007	.000	.001	.006
R^2		.407	.455	.672

*, **, and *** indicate small, medium, and large effect sizes.

and Mulaik (1972) advocate .90 to identify congruent factors or component loadings, the relation between ϕ and σ_I as discussed in Appendix B combined with the rules-of-thumb by Kruskal (1964) (although not specified for unfolding) called for much stricter values for ϕ. For the rank order correlation τ_b, values of .90, .70, and .80 are considered sufficiently high in actual practice, also considering the variation in rank order correlations for the different error levels. The actual values for the three recovery measures are reached with 95% accuracy, providing a common 5% type-I error.

Figure 4 can be read as follows. Suppose we have about 10 products, 20 respondents, and we expect almost errorfree data. Suppose we are interested in the rank order correlations for which we are satisfied with only $\tau_b = .70$ recovery. In this case, we use the upper left panel for 10 products and 20 respondents and the left-hand side cluster for error level 0.0. The rank order correlation is on the right-hand side of the cluster, indicated with a square marker. The lower part of the I-beam provides the minimal $\tau_b = .70$ rank order correlation, which in this case allows for an inclusion proportion of .70. Thus, with a 95% chance that the rank orders corresponds at least $\tau_b = .70$ with the complete unfolding solution 3 products can be set missing per respondent.

The three different foundations (ϕ_{xy}, ϕ_y, and τ_b) for the inclusion proportions in Figure 4 and the multivariate covariance analysis provide similar results: more products (m), more observations ($n \times m$), and less error (e)

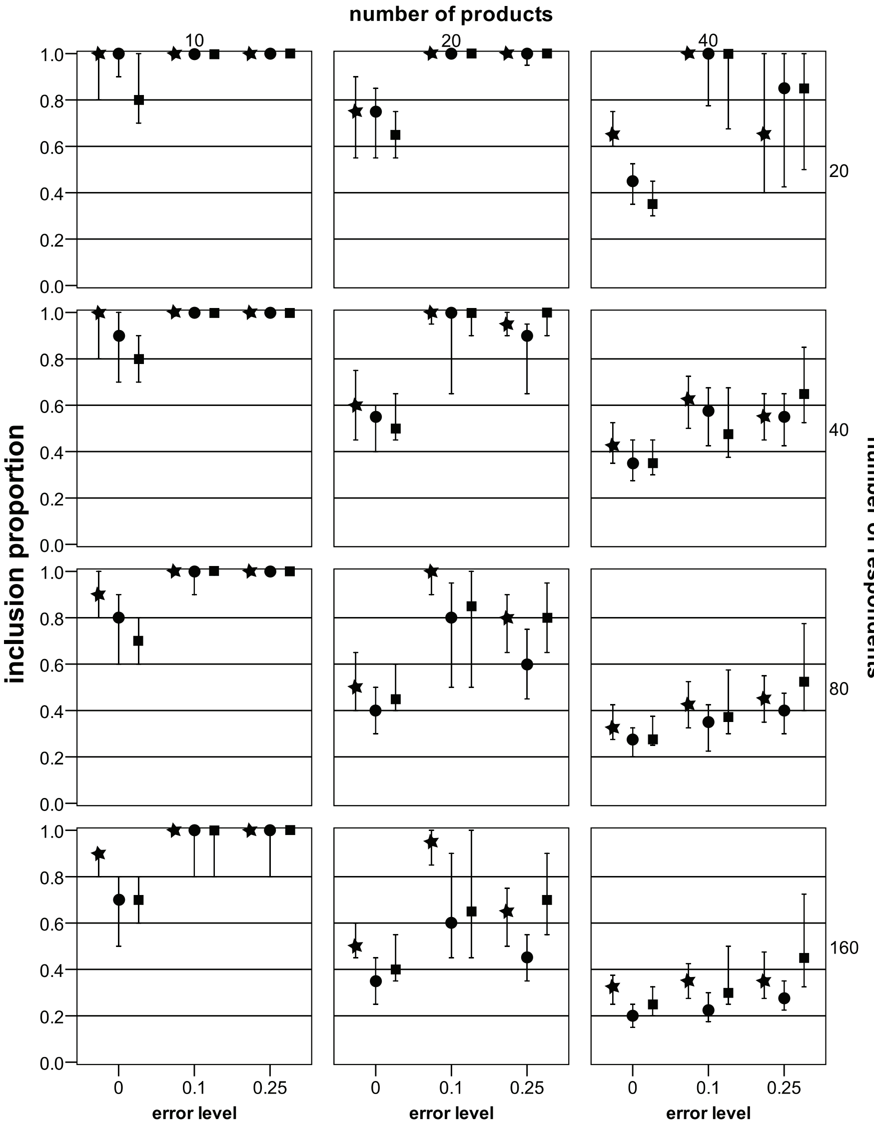


Figure 4. High-Low Graphs for Inclusion Proportions when Data are Missing by Researcher, where the I-Beams (Low-Close-High) indicate 95% Chances on Minimal Values for ϕ_{xy} (.95–.98–.99), ϕ_y (.95–.98–.99), and τ_b (.70–.80–.90), indicated by Stars, Dots and Squares, respectively.

Table 10. Main Effects (Wilks' Lambda) and Effect Sizes for the Tests of the Between-Subject Effects of Two Multivariate Covariance Analyses, One for Missing Data by Researcher and One for Missing Data by Respondent, Comparing the Recovery of Unfolding Solutions for Number of Respondents (n), Number of Products (m), Number of Dimensions (p), and Error Level (e), with Inclusion Proportion ($\text{prop}(i)$) as Covariate.

Source	Wilks' λ	ϕ_{xy}^{high}	ϕ_{xy}^{low}	ϕ_y	τ_b
$\text{prop}(i)$	.679***	.424***	.434***	.261***	.662***
n	.036*	.032*	.028*	.027*	.048*
m	.069**	.086**	.088**	.004	.099**
p	.037*	.019*	.021*	.004	.001
e	.057*	.003	.003	.009	.095**
$n \times m$	.063**	.028*	.036*	.071**	.204***
$n \times p$	.010*	.013*	.014*	.003	.018*
$m \times p$	.006	.001	.001	.003	.005
$n \times e$	.020*	.050*	.058*	.031*	.018*
$m \times e$	.015*	.034*	.037*	.001	.004
$p \times e$	.007	.004	.004	.000	.009
R^2		.539	.549	.402	.737

*, **, and *** indicate small, medium, and large effect sizes.

allow for lower inclusion proportions. Figure 4 lacks small sample sizes with $n = 10$ and $m = 5$, because for these cases the inclusion proportion is always equal to 1.0. The recovery of the product configuration, quantified with ϕ_y, and situated in the middle of the clusters of three with the dot marker, allows for the lowest inclusion proportions. This is plausible considering the number of parameters to be estimated and the amount of data available. Notable is the fact that the high error levels often allow for a lower inclusion proportion as compared with the medium error levels, as can be seen in Figure 4 for $n = 20$ and $m = 40$ and for $n = 80$ and $m = 20$.

4.2 Results and Guidelines for Handling Missing Data by Respondent

The influence on recovery when the data are missing by respondent was determined with a multivariate covariance analysis. Recovery of the entire configuration (ϕ_{xy}) is split up into the recovery of a high familiar set of products (ϕ_{xy}^{high}) and the recovery of a low familiar set of products (ϕ_{xy}^{low}).

All tests (multivariate and between-subjects) are significant ($p < .001$) and Table 10 shows the effect sizes only. The conclusions are similar to the conclusions from the missing data by researcher design, although less pronounced: unfolding solutions are better recovered for data with fewer missings ($\text{prop}(i)$), more products (m), less error(e), and more observations ($n \times m$). There are small effects for the number of respondents (n), dimensionality (p), and some of the interactions. Considering the between-subject effects (last four columns of Table 10), the rank order correlation benefits exceptionally well from additional observations, but recovery of the correlation is also sensitive to error. Finally, it should be noted that although high familiar products are better recovered than low familiar products (significantly with very small effect sizes (not shown here)), the independent factors have similar effects on both sets of products.

Guidelines for applied research when the data is missing due to respondents are given in Figure 5. In general, the inclusion proportions are seriously higher than for the missing data by researcher design (Figure 4). Only for a large number of observations, and then even with a large number of products, the inclusion proportions approach 50%. Compare, for example, $n = 160$ and $m = 10$ with $n = 40$ and $m = 40$: both samples have the same number of observations, but the latter, with more products, allows for more missing data.

5. Example

The results of the Monte Carlo simulation study are used to determine the inclusion proportion for the breakfast data. The breakfast data consists of 42 respondents and 15 products (breakfast items) and the inclusion proportion is determined by taking the average between inclusion proportions of $m = 10$ and $m = 20$ for $n = 40$ and $e = .25$. In this case, the error level is known from the complete set of data, which is something to be guessed at in other circumstances. The number of missing preferences per respondent can be chosen, depending on the quality of recovery (low, medium, or high), on the primary interest of the researcher (the product configuration, the respondents rank orders, or the entire configuration), and on the missing data design (by researcher or by respondent). For the current illustration, we are interested in the product configuration and thus focuss on ϕ_y. The inclusion proportions for low, medium, and high recoverability are .825, .95, and .975, for the missing data by researcher design and .90, .975, and .975 for the missing data by respondent design. With 15 products, this leads to 0–3 missing preferences per respondent. Since the complete set of data is available, multiple incomplete data analyses are possible and 1000 replications are used to create the configurations and boxplots.

Figure 6 shows the unit standard deviation confidence ellipses (Meulman and Heiser 1983) or confidence regions for the incomplete data solutions after 1000 replications. The incomplete data solutions are optimally rotated, translated, and dilated by orthogonal Procrustes analysis (Cliff 1966) to match the complete unfolding solution. It is obvious, even by sight, that the solutions with fewer missing preferences per respondent and the solutions from the missing data by researcher design contain smaller regions.

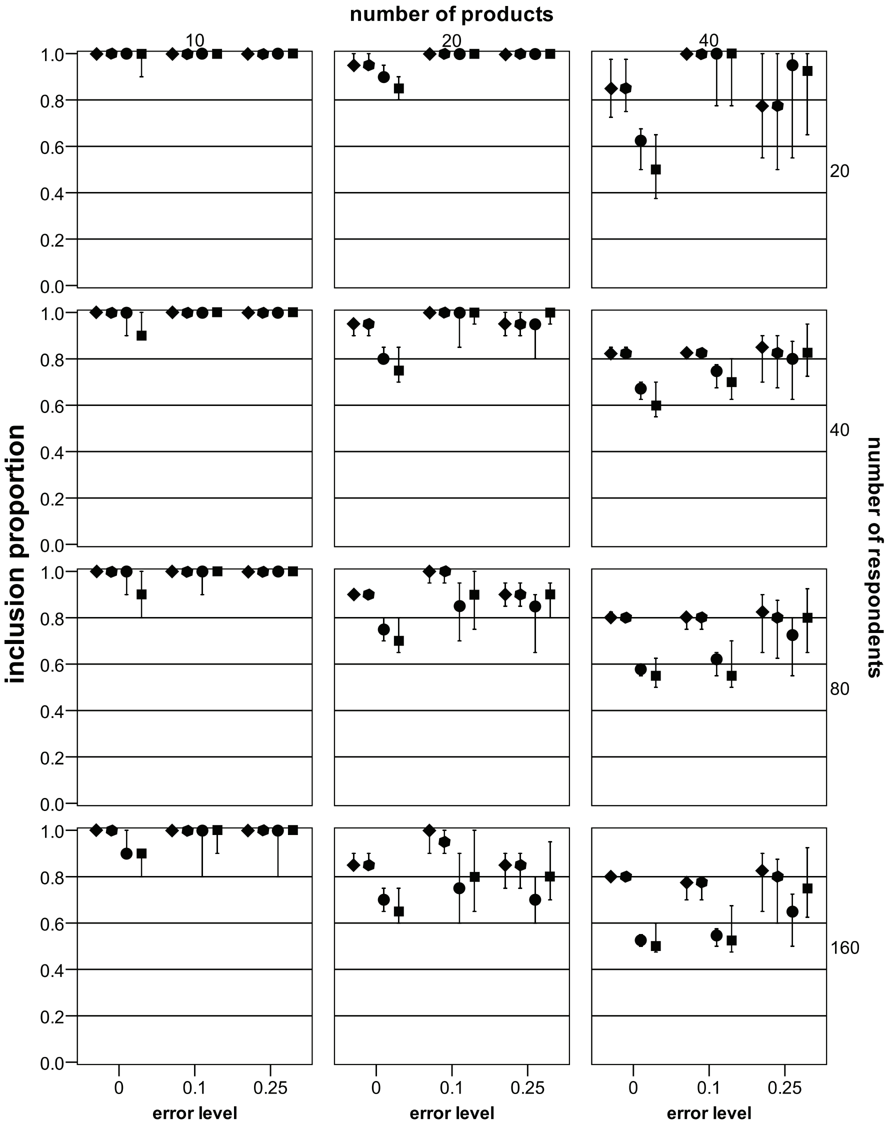
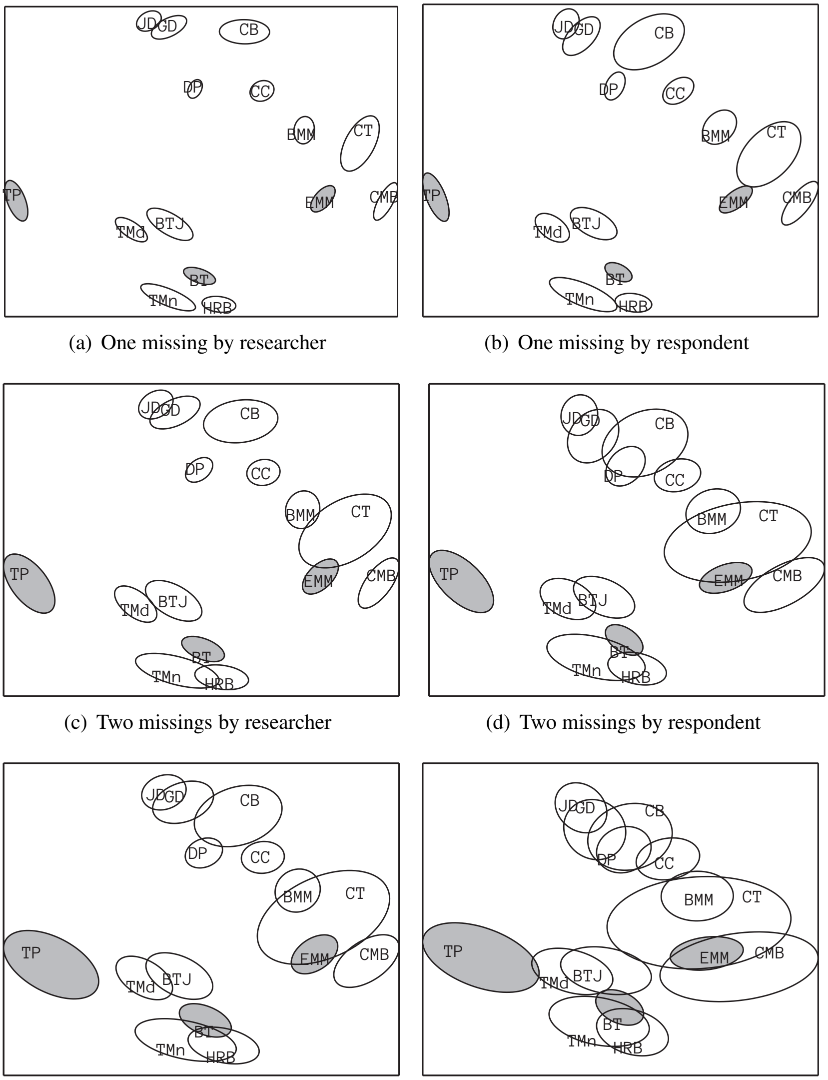


Figure 5. High-Low Graphs for Inclusion Proportions when Data are Missing by Researcher, where the I-Beams (Low-Close-High) indicate 95% Chances on Minimal Values for ϕ_{xy}^{high} (.95–.98–.99), ϕ_{xy}^{low} (.95–.98–.99), ϕ_y (.95–.98–.99), and τ_b (.70–.80–.90), indicated by Diamonds, Polygons, Dots and Squares, respectively.



(a) One missing by researcher

(b) One missing by respondent

(c) Two missings by researcher

(d) Two missings by respondent

(e) Three missings by researcher

(f) Three missings by respondent

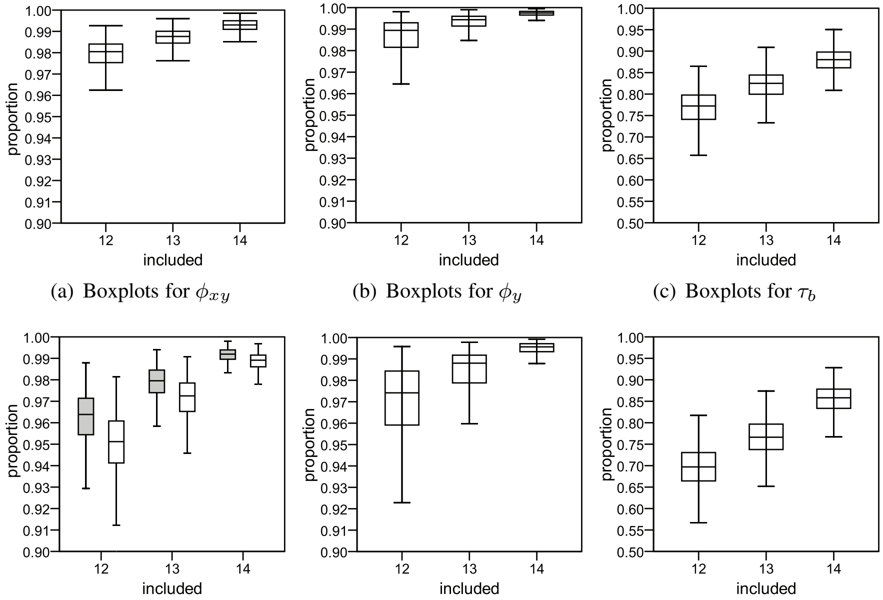
Figure 6. Configurations with unit standard deviation confidence ellipses for the incomplete breakfast data using two different designs for specifying missing data: missing data by researcher design (left panels) and missing data by respondent design (right panels). The breakfast items (and plotting codes) are: toast pop-up (TP), buttered toast (BT), English muffin and margarine (EMM), jelly donut (JD), cinnamon toast (CT), blueberry muffin and margarine (BMM), hard rolls and butter (HRB), toast and marmalade (TMd), buttered toast and jelly (BTJ), toast and margarine (TMn), cinnamon bun (CB), Danish pastry (DP), glazed donut (GD), coffee cake (CC), and corn muffin and butter (CMB).

These solutions are more alike and provide better recovery of the complete data solution. Nevertheless, the three high familiar products in the missing data by respondent design, toast pop-up (TP), buttered toast (BT), and English muffin and margarine (EMM), indicated in the configurations by filled confidence regions, deviate from this general observation by maintaining their small regions, such that these products are comparable with the missing data by researcher design. Compare, for example, the confidence regions of CT and EMM, where the region of the latter remains small, while the region of the former increases considerably with each additional missing preference per respondent. In all cases, the true product points (indicated by the plotting codes) lie within the boundaries of their confidence region. This indicates that the incomplete data configurations are indeed very similar to the complete data configuration, although the variation of the coordinates from the incomplete data solutions increases for additional missing data.

The boxplots in Figure 7 display the distributions of the recovery measures. For the missing data by researcher design, nearly all congruence coefficients are greater than .98 (panel *a* and *b*), and even greater than .99 considering only the product configuration (panel *b*). It seems that the guidelines from Table 4 are somewhat conservative, since $\phi_y \geq .99$ was expected for data without missings and $\phi_y \geq .98$ for data with only one missing per respondent. For the missing data by respondent design, the recovery is acceptable for one or two missing preferences per respondent, but recovery quickly worsens for additional missing data. High familiar products are better recovered than low familiar products (panel *d*), but extra missing data results in inferior configurations for the high familiar products too. However, returning to where we started from, the product configuration is recovered quite well, also for two and even three missing preferences, which is better than predicted from the Monte Carlo simulation study results.

6. Conclusion

An extensive study was performed that investigated the effects of incomplete data on the results of a multidimensional unfolding analysis. We focused on two research designs that are often utilized in consumer and marketing research. In the first, the missing data pattern is imposed by the researcher, while in the second design the respondent "controls" the missing data pattern. The goal of the study was to propose guidelines to researchers about the amount of missing data that unfolding can handle without corrupting the results of the analysis. Therefore, we compared all incomplete data solutions with solutions obtained on complete data using two resemblance measures: Tucker's congruence coefficient (ϕ) and Kendall's rank order correlation (τ_b).



(a) Boxplots for ϕ_{xy} (b) Boxplots for ϕ_y (c) Boxplots for τ_b

(d) Boxplots for ϕ_{xy}^{high} (gray) and ϕ_{xy}^{low} (e) Boxplots for ϕ_y (f) Boxplots for τ_b

Figure 7. Boxplots of the recovery measures for the incomplete breakfast data using two different designs for specifying missing data: missing data by researcher design (upper panels) and missing data by respondent design (lower panels).

Unfolding analysis has the possibility to include a weight matrix. When this weight matrix is coded as a zero-one matrix, it can be used to handle missing data. This option is equal to the pairwise deletion scheme, as for a zero weight both the (missing) data and the corresponding distance are ignored in computations. Often, researchers choose to impute data for the missings. We compared the pairwise deletion scheme with two simple imputation methods, and it can be concluded that pairwise deletion works better (Tables 2 and 3). Of course, more elaborate imputation schemes could be thought off, but this is left for future research.

The first design, where the researcher controls the missing data, conforms to a situation where the data are missing completely at random (MCAR). In this case, often a balanced incomplete block design is utilized in order to make the study as efficient as possible. For our simulation study, however, such a BIBD poses insurmountable problems and we proposed a row-balanced incomplete block design (row-BIBD). We investigated efficiency loss using a row-BIBD compared to a BIBD and it can be concluded that this loss is negligible (Table 4). All our further studies used the row-BIBD. In practical settings, where only a single analysis has to be performed, we advice to use a real BIBD (if available) to have a somewhat higher efficiency though.

The second design, where the respondent controls the missing data, conforms to a situation where the data are missing at random (MAR). Two factors determine the missing data pattern: the knowledge set of the respondent and product familiarity. The knowledge set corresponds to the set of products the respondent knows and so is able to judge. This knowledge set may differ over subjects in size and content. Product familiarity corresponds to the fact that some products are very well known (and thus in the knowledge set of every respondent) and others are less well known. We distinguished between high and low familiarity. We found that variance in knowledge set has a large influence on the recoverability (Table 5), while product familiarity on average has only a minor influence (Table 6). However, high familiar products are recovered better than low familiar products (Table 7).

Knowing this, we investigated recovery of the complete data unfolding solution using the two designs, where in the second design knowledge set variance and familiarity were used as additional factors. We varied the proportion of included data, the number of respondents, the number of products, the number of dimensions, and the error level. Key success factors in the recovery of the unfolding solution using incomplete data are (in order of importance): the proportion of non-missing data, the number of observations, the number of products, and the error in the data.

Figures 4 and 5 can be used as guidelines for researchers to choose on the amount of information to be collected. The first figure is for the case where the researcher can determine the missing data pattern. This case is the least sensitive to missing data and it was concluded that up to 80% of the data could be missing without real deterioration. In the second design, where respondents control the missing data pattern, unfolding is more sensitive. Researchers should be careful to include at least 50% of the data to have a good recovery. In both situations, when there are less respondents, or less products, the more the percentage of valid data should increase. In all cases, the researcher is advised to be careful in the research design to keep the error at the lowest possible level. The guidelines presented are in a sense conservative: even for lower inclusion proportions than presented, and thus with lower recovery measures, the incomplete data solutions remain similar such that substantive conclusions would not change.

The guidelines were illustrated with an example on empirical data. In Figures 6 and 7, the effects can be seen of various amounts of missing data and the influence of both designs. It was concluded that the solutions are all very similar. Substantive conclusions of these different solutions will be equal.

The two set-ups studied correspond to situations where the data are missing completely at random or missing at random. A third case exists when data are not missing at random (NMAR). Such a situation would occur, for example, if the respondent is asked to indicate the top r ranked products (see, for example, DeSarbo et al. 1997). The possibilities to construct data in this manner are enormous and a thorough investigation would require (computer) time, (journal) space, and a serious (human) effort. It is beyond the scope of this paper and left for future research.

The problem of local minima was not addressed in this paper, although predecessors of the Monte Carlo simulation study used both random and close starts. Using random starts introduced unwanted variation in the somewhat more conservative results, while the close starts, using the results of the complete data solution, provided unrealistically good results. Results from these studies (not reported here) indicate that it is always better to use a good start, using available information about locations of products whenever possible. Random starts can not match the guideline results presented here, although PREFSCAL performed fairly well under both circumstances.

The results obtained are very promising. Although really small samples have low recoverability when data are missing for each respondent, for studies involving more products, less than half of the data has to be included without danger of changing the conclusions. This is of major importance for all consumer research: a lot of time and money can be saved.

Appendix A
Simulation Study:
Metric versus Nonmetric Unfolding of Incomplete Data

In many cases, and certainly due to the huge problem of degenerated solutions in unfolding, nonmetric data are often analyzed metrically. In a simulation study using the breakfast data (Green and Rao 1972), metric unfolding, only estimating a dilation parameter for each respondent, is compared with nonmetric unfolding, where the preferences of each respondent are transformed monotonically. The study was replicated 1000 times with a shuffled BIBD (see Table 1) on each instance. The metric and nonmetric unfolding solutions are both compared on their own merits and concerning their recovery capabilities.

The original metric unfolding solution for the complete data set is not a particular good solution with $\sigma_I^+ = .299$ and $\tau_b^+ = .608$, while the variation of the distances is quite good. The incomplete data solutions, with 33% less data to fit, only improves a little over the complete data solution

Table A1. Descriptive Statistics (Upper Part, with Means and Standard Deviations in Parentheses) and MANOVA Tests of the Between-Subjects Effects (Lower Part, with F-Statistics, Significances in Parenthesis, and Effect Sizes on the Second Line) Comparing the Fit of Metric and Nonmetric Unfolding Solutions with Incomplete Data.

Analysis	σ_I^-	τ_b^-
Metric Unfolding	.272 (.009)	.647 (.019)
Nonmetric Unfolding	.162 (.024)	.772 (.020)
Between-Subjects Effects	σ_I^-	τ_b^-
F (p)	18206.414 (.000)	19862.169 (.000)
η_p^2	.901	.909

with $\sigma_I^- = .272$ and $\tau_b^- = .647$ (see Table A1). The nonmetric unfolding solution for the complete data is definitely better than its metric counterpart with $\sigma_I^+ = .241$ and $\tau_b^+ = .701$, while the variation of both transformed preferences and distances are comparable. The nonmetric unfolding solution is certainly not degenerated. However, in contrast with the incomplete data solutions of the metric unfolding, the nonmetric unfolding solutions improve considerably with one third of the data missing, especially on Stress, with $\sigma_I^- = .162$ and $\tau_b^- = .772$. A multivariate analysis of variance indicates a significant difference (using Wilks' Lambda: $F = 11868.735; p = .000; \eta_p^2 = .922$) between the metric and the nonmetric unfolding solutions with incomplete data. The between-subject effects for both σ_I^- and τ_b^- are significant with large effect sizes (see Table A1).

For the discussion of the recovery of both unfolding methods, we have to keep in mind that different solutions need to be recovered. A multivariate analysis of variance indicates a significant difference (using Wilks' Lambda: $F = 630.381; p = .000; \eta_p^2 = .487$) between the recovery of the metric and the nonmetric unfolding solution. Although the differences in recovery are only minor for the whole configuration (ϕ_{xy}) and the configuration of the products (ϕ_y), there is a large effect for the differences in rank order recovery (τ_b): the metric unfolding recovers the rank orders better than the nonmetric unfolding does (see Table A2, last column). This is noticeable, but appearances may be deceiving. The complete data metric unfolding has a worse fit for the rank orders ($\tau_b^+ = .608$) as mentioned before, but it recovers these less fitting rank orders better ($\tau_b = .714$) than the nonmetric unfolding does ($\tau_b = .661$) for its better fitting rank orders ($\tau_b^+ = .701$).

In conclusion, it is safe to state that the metric unfolding on incomplete data recovers the inferior complete data solution about equally well as the nonmetric unfolding does for the superior complete data solution.

Table A2. Descriptive Statistics (Upper Part, with Means and Standard Deviations in Parentheses) and the Tests of the Between-Subjects Effects (Lower Part, with F-Statistics, Significances in Parenthesis, and Effect Sizes on the Second Line) Comparing the Recovery of Metric and Nonmetric Unfolding Solutions.

Analysis	ϕ_{xy}	ϕ_y	τ_b
Metric Unfolding	.953 (.021)	.971 (.019)	.714 (.054)
Nonmetric Unfolding	.957 (.015)	.968 (.018)	.661 (.055)
Between-Subjects Effects	ϕ_{xy}	ϕ_y	τ_b
F (p)	27.180 (.000)	13.270 (.000)	478.766 (.000)
η_p^2	.013	.007	.193

Appendix B
Tucker's Congruence Coefficient and Other Least-squares Measures

The least-squares loss function in multidimensional scaling, commonly known as (raw) *Stress* (Kruskal 1964), is defined as $\sigma_r = \|\boldsymbol{\delta} - \mathbf{d}\|^2$, where $\| \cdot \|^2$ represents the squared Euclidean norm and $\boldsymbol{\delta}$ and $\mathbf{d}$ are vector representations of the preferences and the distances, respectively. For an objective comparison between different configurations, note that the optimal scaling factor α for the distances in σ_r is found by minimizing $\sigma_r(\alpha)$ with respect to α (taking the derivative and setting it equal to zero):

$$\text{function}: \quad \sigma_r(\alpha) = \|\boldsymbol{\delta} - \alpha\mathbf{d}\|^2 = \boldsymbol{\delta}'\boldsymbol{\delta} + \alpha^2\mathbf{d}'\mathbf{d} - 2\alpha\boldsymbol{\delta}'\mathbf{d} \tag{1}$$

$$\text{derivative}: \quad \frac{\partial\sigma_r(\alpha)}{\partial\alpha} = 2\alpha\mathbf{d}'\mathbf{d} - 2\boldsymbol{\delta}'\mathbf{d}$$

$$\text{minimum}: \quad \alpha = \frac{\boldsymbol{\delta}'\mathbf{d}}{\mathbf{d}'\mathbf{d}} \tag{2}$$

de Leeuw and Heiser (1977) prove that for a local minimum of Stress, it holds that $\mathbf{d}'\mathbf{d} = \boldsymbol{\delta}'\mathbf{d}$ and thus (2) equals one. Substituting (2) in (1) and dividing both sides by $\boldsymbol{\delta}'\boldsymbol{\delta}$ gives

$$\frac{\sigma_r(\alpha)}{\boldsymbol{\delta}'\boldsymbol{\delta}} + \frac{(\boldsymbol{\delta}'\mathbf{d})^2}{(\mathbf{d}'\mathbf{d})(\boldsymbol{\delta}'\boldsymbol{\delta})} = 1, \tag{3}$$

which links *normalized* raw Stress σ_n (the first term on the left-hand side) with the sum-of-squares accounted for (SSAF) (the second term on the left-hand side) (after Commandeur and Heiser 1993, p. 70). Normalized raw Stress is the least-squares loss function minimized by SPSS PROXSCAL (Meulman, Heiser, and SPSS Inc. 1999), while the sum-of-squares accounted for is used as one of the convergence criteria in GENFOLD (DeSarbo and Carroll 1985; DeSarbo and Rao 1984). In multidimensional scaling literature the sum-of-squares accounted for is also referred to as dispersion accounted for (DAF) (Heiser and Groenen 1997). The square root of SSAF (or

DAF) is equal to Tucker's congruence coefficient ϕ. A link with Kruskal's Stress-I (Kruskal and Carroll 1969) is established by minimizing its square σ_I^2 with respect to the optimal scaling factor α:

$$\text{function :} \qquad \sigma_I^2(\alpha) = \frac{\sigma_r(\alpha)}{\alpha^2 \mathbf{d}'\mathbf{d}} = \frac{\boldsymbol{\delta}'\boldsymbol{\delta} + \alpha^2 \mathbf{d}'\mathbf{d} - 2\alpha\boldsymbol{\delta}'\mathbf{d}}{\alpha^2 \mathbf{d}'\mathbf{d}} \tag{4}$$

$$\text{derivative :} \qquad \frac{\partial \sigma_I^2(\alpha)}{\partial \alpha} = \frac{2\alpha^2(\mathbf{d}'\mathbf{d})(\boldsymbol{\delta}'\mathbf{d}) - 2\alpha(\boldsymbol{\delta}'\boldsymbol{\delta})(\mathbf{d}'\mathbf{d})}{\alpha^4(\mathbf{d}'\mathbf{d})^2}$$

$$\text{minimum :} \qquad \alpha = \frac{\boldsymbol{\delta}'\boldsymbol{\delta}}{\boldsymbol{\delta}'\mathbf{d}} \tag{5}$$

Substituting (5) in (4) and using (3) shows that the square of Kruskal's Stress-I is equal to normalized raw Stress as

$$\begin{aligned}
\sigma_I^2(\alpha) &= \frac{\boldsymbol{\delta}'\boldsymbol{\delta} + (\boldsymbol{\delta}'\mathbf{d})^{-2}(\boldsymbol{\delta}'\boldsymbol{\delta})^2\mathbf{d}'\mathbf{d} - 2(\boldsymbol{\delta}'\mathbf{d})^{-1}(\boldsymbol{\delta}'\boldsymbol{\delta})\boldsymbol{\delta}'\mathbf{d}}{(\boldsymbol{\delta}'\mathbf{d})^{-2}(\boldsymbol{\delta}'\boldsymbol{\delta})^2\mathbf{d}'\mathbf{d}} \\
&= 1 - \frac{(\boldsymbol{\delta}'\mathbf{d})^2}{(\mathbf{d}'\mathbf{d})(\boldsymbol{\delta}'\boldsymbol{\delta})} \\
&= \frac{\sigma_r(\alpha)}{\boldsymbol{\delta}'\boldsymbol{\delta}} \\
&= \sigma_n.
\end{aligned}$$

Due to the normalization factors (the denominators), the function values of σ_n, SSAF (or DAF), ϕ, and σ_I^2 are insensitive to differences in scale and/or sample size and thus well-suited to objectively compare distances of different multidimensional scaling configurations. Similar derivations can be found in Borg and Groenen (2005).

References

ARNOLD, G.M., and WILLIAMS, A.A. (1986), "The Use of Generalized Procrustes Analysis in Sensory Analysis", in *Statistical Procedures in Food Research*, ed. R. Piggott, London: Elsevier, pp. 233–253.

BALABANIS, G., and DIAMANTOPOULOS, A. (2004),"Domestic Country Bias, Country-of-Origin Effects, and Consumer Ethnocentrism: A Multidimensional Unfolding Approach", *Journal of the Academy of Marketing Science, 32(1)*, 80–95.

BENNETT, J., and HAYES, W.L. (1960), "Multidimensional Unfolding: Determining the Dimensionality of Ranked Preference Data", *Psychometrika, 25,* 27–43.

BORG, I., and GROENEN, P.F. (2005), *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.), New York: Springer-Verlag.

BURT, C. (1948),"The Factorial Study of Temperamental Traits", *British Journal of Psychology (Statistical Section), 1,* 178–203.

BUSING, F.M.T.A. (2006), "Avoiding Degeneracy In Metric Unfolding By Penalizing the Intercept", *British Journal of Mathematical and Statistical Psychology, 59*, 419–427.

BUSING, F.M.T.A., GROENEN, P.J.F., and HEISER, W.J. (2005), "Avoiding Degeneracy in Multidimensional Unfolding by Penalizing on the Coefficient of Variation", *Psychometrika, 70(1)*, 71–98.

CARROLL, J.D. (1972), "Individual Differences and Multidimensional Scaling", in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences* (Vol. 1), eds. R.N. Shepard, A.K. Romney, and S. B. Nerlove, New York: Seminar Press, pgs. 105–155.

CHATTERJEE, R. and DESARBO, W.S. (1992), "Accommodating the Effects of Brand Unfamiliarity in the Multidimensional Scaling of Preference Data", *Marketing Letters, 3(1)*, 85–99.

CLATWORTHY, W.H. (1973), "Tables of Two-Associate-Class Partially Balanced Designs", *Applied Mathematics Series 63*.

CLIFF, N. (1966), "Orthogonal Rotation to Congruence", *Psychometrika, 31*, 33–42.

COCHRAN, W.G., and COX, G.M. (1957), *Experimental Designs* (2nd ed.), New York: Wiley.

COHEN, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New York: Academic Press.

COMMANDEUR, J.J.F., and HEISER, W.J. (1993), "Mathematical Derivations in the Proximity Scaling (PROXSCAL) of Symmetric Data Matrices", Technical Report RR-93-04, Leiden, The Netherlands: Leiden University, Department of Data Theory.

COOMBS, C.H. (1950), "Psychological Scaling Without a Unit of Measurement", *Psychological Review, 57*, 148–158.

CURETON, E.E., and D'AGOSTINO, R.B. (1983), *Factor Analysis: An Applied Approach*, Hillsdale, NJ: Lawrence Erlbaum.

DE LEEUW, J., and HEISER, W.J. (1977), "Convergence of Correction-matrix Algorithms for Multidimensional Scaling", in *Geometric Representations of Relational Data*, eds. J.C. Lingoes, E.E.C.I. Roskam, and I. Borg, Ann Arbor, MI: Mathesis Press, pp. 735–752.

DESARBO, W.S., and CARROLL, J.D. (1985), "Three-Way Metric Unfolding via Alternating Weighted Least Squares", *Psychometrika, 50 3)*, 275–300.

DESARBO, W.S., KIM, J., CHOI, S.C., and SPAULDING, M. (2002), "A Gravity-Based Multidimensional Scaling Model for Deriving Spatial Structures Underlying Consumer Preference/Choice Judgments", *Journal of Consumer Research, 29*, 91–100.

DESARBO, W.S., and RAO, V.R. (1984), "GENFOLD2: A Set of Models and Algorithms for the GENeral unFOLDing Analysis Of Preference/Dominance Data", *Journal of Classification, 1*, 147–186.

DESARBO, W.S., YOUNG, M.R., and RANGASWAMY, A. (1997), "A Parametric Multidimensional Unfolding Procedure for Incomplete Nonmetric Preference/Choice Set Data in Marketing Research", *Journal of Marketing Research, 34(4)*, 499–516.

DIJKSTERHUIS, G.B., GOWER, J.C. (1991), "The Interpretation of Generalized Procrustes Analysis and Allied Methods", *Food, Quality and Preference, 3*, 67–87.

GREEN, P.E., RAO, V. (1972), *Applied Multidimensional Scaling*, Hinsdale, IL: Dryden Press.

HEISER, W.J., and DE LEEUW, J. (1979), *How to use SMACOF-III: A Program for Metric Multidimensional Unfolding*, Leiden, The Netherlands: Leiden University, Department of Data Theory.

HEISER, W.J., and GROENEN, P.J.F. (1997), “Cluster Differences Scaling with a Within-clusters Loss Component and a Fuzzy Successive Approximation Strategy to Avoid Local Minima”, *Psychometrika, 62*, 63–83.

KELLY, G A. (1955), *The Psychology of Personal Constructs: Theory of Personality*, New York: Norton.

KENDALL, M.G. (1948), *Rank Correlation Methods*, London: Charles Griffin and Company Limited.

KIM, C., RANGASWAMY, A., and DESARBO, W.S.(1999), “A Quasi-metric Approach to Multidimensional Unfolding for Reducing the Occurrence of Degenerate Solutions”, *Multivariate Behavioral Research, 34*, 143–180.

KRUSKAL, J.B. (1964), “Multidimensional Scaling by Optimizing Goodness-of-fit to a Nonmetric Hypothesis”, *Psychometrika, 29*, 1–27.

KRUSKAL, J.B., and CARROLL, J.D. (1969), “Geometrical Models and Badness-of-fit Functions, in *Multivariate Analysis*, ed. P. R. Krishnaiah, New York: Academic Press, pp. 639–671.

LINGOES, J.C. (1977), “A General Nonparametric Model for Representing Objects and Attributes in a Joint Metric Space”, in *Geometric Representations of Relational Data*, ed. J. C. Lingoes, Ann Arbor, Michigan: Mathesis Press, pp. 475–496.

LITTLE, R.J.A., and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, New York: J. Wiley & Sons.

MEULMAN, J.J., and HEISER, W.J. (1983), *The Display of Bootstrap Solutions in Multidimensional Scaling*, Technical Report, Leiden, The Netherlands: Leiden University, Department of Data Theory.

MEULMAN, J.J., HEISER, W.J., and SPSS Inc. (1999), *SPSS Categories 10.0*, Chicago, IL, USA: SPSS Inc.

MULAIK, S.A. (1972), *The Foundations of Factor Analysis*, New York: McGraw-Hill.

NGUYEN, N. (1993), “An Algorithm for Constructing Optimal Resolvable Incomplete Block Designs”, *Communications in Statistics, Simulation & Computation, 22*, 911–923.

NGUYEN, N. (1994), “Construction of Optimal Block Design by Computer”, *Technometrics, 36*, 300–307.

PEARSON, K. (1896), “ Regression, Heredity, and Panmixia”, *Philosophical Transactions of the Royal Society of London, Series A, 187*, 253–318.

PRESTWICH, S.D. (2001), “Balanced Incomplete Block Design as Satisfiability”, in *Proceedings of the 12th Irish Conference on AI and Cognitive Science (AICS 2001)*, ed. D. O’Donoghue, NUI Maynooth, Ireland.

ROSKAM, E.E. CH I. (1968), *Metric Analysis of Ordinal Data*, Voorschoten: VAM.

ROWE, G., LAMBERT, N., BOWLING, A., EBRAHIM, S., WAKELING, I., and THOMSON, R. (2005), “Assessing Patients Preferences for Treatments for Angina Using a Modified Repertory Grid Method”, *Social Science & Medicine*, 2585–2595.

SHEPARD, R.N. (1962), “The Analysis of Proximities: Multidimensional Scaling with an nknown distance fun ction. I.”, *Psychometrika, 27*, 125–140.

SHOCKER, A.D., BEN-AKIVA, M., BOCCARA, B., and NEDUNGADI, P. (1991), “Consideration Set Influences on Consumer Decision-Making and Choice: Issues, Models, and Suggestions”, *Marketing Letters, 2:3*, 181–197.

SPSS (2006), *SPSS for Windows, Release 15.0*, [Computer Software].

TUCKER, L. R. (1951), *A Method for Synthesis of Factor Analytic ”S” Studies*, Washington, D.C.: Department of the Army.

VAN DEUN, K. (2005), *Degeneracies in Multidimensional Unfolding*, Unpublished doctoral dissertation, Catholik University Leuven.

VAN DEUN, K., GROENEN, P.J.F., and DELBEKE, L. (2005), " VIPSCAL: A Combined Vector Ideal Point Model for Preference Data", Econometric Institute Report EI 2005-03, Rotterdam: Erasmus University Rotterdam.

VAN DEUN, K., HEISER, W.J., and DELBEKE, L. (2007), "Multidimensional Unfolding by Nonmetric Multidimensional Scaling of Spearman distances in the Extended Permutation Polytope", *Multivariate Behavioral Research, 42(1)*, 103–132.

WAGENAAR, W.A., and PADMOS, P. (1971), "Quantitative Interpretation of Stress in Kruskal's Method Multidimensional Scaling Technique", *British Journal of Mathematical and Statistical Psychology, 24*, 101–110.