Journal of Educational and Behavioral Statistics
Winter 1997, Vol. 22, No. 4, pp. 447–477

Multiple Correspondence Analysis as a Tool for Quantification or Classification of Career Data

Peter G. M. van der Heijden, Joop Teunissen, and **Charles van Orlé**
Utrecht University

Keywords: *career data, classification, event history data, longitudinal data, multiple correspondence analysis*

Career data of an individual are defined by counts that describe the amount of time an individual has spent in each of a number of mutually exclusive states. It is argued that correspondence analysis is particularly suited to such data, and this is illustrated by an example from educational careers. Furthermore it is shown that correspondence analysis can be used for a quantification of individual careers into one or more scores, or a classification, that can then be used in further analyses.

In the social sciences, interest often is directed toward people's careers. Conceptually, careers can play the role of either explanatory variable or of response variable. One can think of educational careers, of criminal careers, of work careers, and of clinical careers, to name but a few examples. Usually these careers are quite complex, and this can complicate the analysis. Measures that summarize individual careers are therefore called for. Taris (1994) provides a recent review of tools that are useful for a classification or quantification of careers. This article concentrates on one such tool useful for this purpose, namely, correspondence analysis.

Much attention has recently been given to statistical event history analysis (see, e.g., Blossfeld, Hamerle, and Mayer, 1989, and Yamaguchi, 1991, for recent reviews). Roughly, what is of interest here is the prediction of a transition in a career by a number of explanatory variables. For example, the time before a first job is obtained after leaving school can be predicted by age, sex, and network variables. However, useful as it may be for obtaining answers to many questions, statistical event history analysis seems to focus on only small parts of careers of individuals.

Another approach is to concentrate on finding a description of the whole career. As far as we know, all tools which do this are exploratory. Recently, some attention has been given to optimal matching methods stemming from biology

The authors gratefully acknowledge the computational efforts performed by Rafaele Huntjens. Request for reprints should be addressed to the first author.

(see Abbott and Hrycak, 1990). These methods are used to compare strings of information. Sequences of states are compared, and a count is made to measure the dissimilarity between two sequences. We come back to this method in a later section.

Another tool that was earlier used for the analysis of career data is correspondence analysis (CA). For this tool, the states of a career of each individual are represented by counts that describe the amount of time this individual has spent in each of a number of mutually exclusive states. For example, if there are eight mutually exclusive states, each individual receives eight counts describing the amount of time he or she has spent in each of the states. For a sample of individuals, CA then operates on these sets of frequencies. The analysis of event history data by CA was proposed by Deville and Saporta (1980, 1983; cf. Deville, 1982; Saporta, 1981, 1985) and later developed by de Leeuw, van der Heijden, and Kreft (1985), van der Heijden (1987), and van der Heijden and de Leeuw (1989). Recent applications can be found in van Buuren and de Leeuw (1992), van der Heijden and van den Brakel (1993), Martens (1994), and Taris (1994).

This article aims to stimulate interest in CA, since, in our view, it has not received the attention it deserves. In the next section we give an exposition of CA of a two-way contingency table. We then show why CA is a useful tool for the analysis of career data. The way that time dependence of the use of specific states can be studied by CA is demonstrated, and, as a special case, multiple CA is introduced. Finally, we discuss extensively an example of educational career data and illustrate the way that the results from CA and multiple CA can be used for finding one or more quantifications or a classification of these careers.

Correspondence Analysis

We will first describe CA and then, in the next section, discuss its application to career data, event history data, and the like. For introductions to CA we refer to Benzécri (1973), Nishisato (1980), Greenacre (1984), and Gifi (1990). We concentrate on the properties of the CA solution and omit details regarding the computation of CA.

CA will be presented as a tool for making graphical representations of contingency tables. Consider a contingency table having I rows ($i = 1, \ldots, I$) and J columns ($j = 1, \ldots, J$) having frequencies n_{ij}. Marginal frequencies are denoted by $n_{i+} = \Sigma_j n_{ij}$ and $n_{+j} = \Sigma_i n_{ij}$. The frequencies can be transformed into proportions p_{ij} by $p_{ij} = n_{ij}/n_{++}$.

In CA, interest is on so-called row profiles and column profiles. We will first explain how the row profiles are studied. The row profile for row i is defined as the vector of J conditional proportions p_{ij}/p_{i+}, adding up to 1. These values specify the J conditional proportions of observations in row i that fall into column j. There are I row profiles, and each row profile can be represented as a point in a J-dimensional space, where for row i the vector elements p_{ij}/p_{i+} are used as coordinates. In order to prevent distances between profiles being dominated by the dimensions that correspond with columns having large values p_{+j}, in this space weights are attached to the J dimensions, such that dimensions with smaller marginal proportions p_{+j} play a relatively larger role in the definition of distances between the I points. For this purpose the weight $1/p_{+j}$ is attached to dimension j. As a result, the distance $\delta(i, i')$ between row i and row i' is defined by

$$\delta(i, i')^2 = \sum_{j=1}^{J} \left(\frac{1}{p_{+j}}\right)\left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}}\right)^2. \qquad (1)$$

This is a weighted Euclidean distance between profiles i and i', with weights $(1/p_{+j})$.

The chi-square distance (1) allows the following interpretation. In the J-dimensional space, row i and row i' will be close together when for each j profile element p_{ij}/p_{i+} is close to $p_{i'j}/p_{i'+}$. Similarly, they will be far apart when there are large weighted differences $p_{ij}/p_{i+} - p_{i'j}/p_{i'+}$ between rows i and i', where a difference for a column with small p_{+j} has a relatively larger influence than a difference for a column with large p_{+j}.

In the centroid $\mathbf{O}$ of the cloud of row points lies the profile of marginal column proportions having elements p_{+j}. It is in the weighted average of the row profiles, when the weights p_{i+} are used as weights, which is shown by $\Sigma_i p_{i+}(p_{ij}/p_{i+}) = p_{+j}$. A chi-square distance of a row profile i to the origin $\mathbf{O}$ is small when the row profile elements p_{ij}/p_{i+} are all close to p_{+j}, and the distance to the origin is large when there are elements p_{ij}/p_{i+} that depart markedly from p_{+j}. Notice that when the row variable is statistically independent from the column variable, so that $p_{ij} = p_{i+}p_{+j}$, then for all i the profile elements $p_{ij}/p_{i+} = p_{+j}$, or, in other words, all profiles are equal and equal to the profile of marginal column proportions. The result is that all points will then fall in the centroid. It follows that a study of the relation between the row and the column variable is useful only when the proportions depart from independence.

The aim is to study differences between the I row profiles by studying the cloud of I points in J-dimensional space. This is a difficult task, and for this reason the I points in J-dimensional space are projected into a lower-dimensional space that simplifies this study. This projection is carried out in such a way that as much information as is available in the full J-dimensional space is projected onto the first few dimensions. Let the new coordinates be defined at r_{ia} for the coordinate of row profile i on dimension a. A projection is then carried out such that, for Dimension 1, the weighted variance of the distances to the origin $\mathbf{O}$, $\lambda_1^2 = \Sigma_i p_{i+} r_{i1}^2$, is maximized. For Dimension 2, the weighted variance of the distances to the origin, $\lambda_2^2 = \Sigma_i p_{i+} r_{i2}^2$, is maximized under the restriction that the row coordinates of the second dimension are orthogonal to those of the first dimension: $\Sigma_i p_{i+} r_{i1} r_{i2} = 0$. And so on for further dimensions.

A similar presentation can be given for the columns of the contingency table. Column profile j has elements p_{ij}/p_{+j}, and these elements can be used to represent profile j as a point in an I-dimensional space. This space consists of J column profiles. Weights $1/p_{i+}$ are attached to each of the I dimensions. As a result, the chi-square distance between column j and column j' is

$$\delta(j, j')^2 = \sum_{i=1}^{I} \left(\frac{1}{p_{i+}}\right)\left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}}\right)^2, \tag{2}$$

which has an interpretation similar to that given above for the rows.

The J points in I-dimensional space are projected into a lower-dimensional space that simplifies its study. Let the new coordinates be defined as c_{ja} for the coordinate of column profile j on dimension a. The result of a projection is that, for Dimension 1, the weighted variance of the distances to the origin $\mathbf{O}$, $\lambda_1^2 = \Sigma_j p_{+j} c_{j1}^2$, is maximized. For Dimension 2, the weighted variance of the distances to the origin, $\lambda_2^2 = \Sigma_j p_{+j} c_{j2}^2$, is maximized under the restriction that the row coordinates of the second dimension are orthogonal to those of the first dimension: $\Sigma_j p_{+j} c_{j1} c_{j2} = 0$. And so on for further dimensions.

Thus a CA yields a solution for the row profiles and a solution for the column profiles. It is a symmetrical technique for the analysis of a contingency table in the sense that the analysis of a table yields the same results as the analysis of the transpose of this table. A nice feature of CA is that the solution for the row profiles is closely related to the solution for the column profiles. The row scores r_{ia} can be derived from the column scores c_{ja}, and vice versa, in the following ways:

$$r_{ia} = \lambda_a^{-1} \sum_{j=1}^{J} \frac{p_{ij}}{p_{i+}} c_{ja} \tag{3}$$

and

$$c_{ja} = \lambda_a^{-1} \sum_{i=1}^{I} \frac{p_{ij}}{p_{+j}} r_{ia}. \tag{4}$$

Equation 3 shows that, apart from constant λ_a^{-1}, row profile i is in the weighted average of the column points, where the profile elements of row i are used as weights. Similarly, (4) shows that, apart from constant λ_a^{-1}, column profile j is in the weighted average of the row points, where the profile elements of column j are used as weights. In other words, profile elements determine where row and column points are placed in their respective representations. In fact, when we compare the representations of the row profiles and column profiles, a row i is pulled into the direction of those columns for which $p_{ij}/p_{i+} > p_{+j}$, and a column profile j is pulled into the direction of those rows for which $p_{ij}/p_{+j} > p_{i+}$.

Equations 3 and 4 also show that the dimensionality of the representation of the rows and that of the columns are equal, and this dimensionality turns out to be equal to the minimum of $I - 1$ and $J - 1$.

The total distance to the origin is called the inertia and is equal to $\Sigma_a \lambda_a^2$. It can be proven that this measure is equal to Pearson's chi-square X^2 divided by the sample size n_{++}:

$$\sum_{a=1}^{J} \lambda_a^2 = \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = \frac{X^2}{n_{++}}.$$

The inertia can be used to assess how much of the total distance is displayed on each of the dimensions by calculating $\lambda_a^2 / \Sigma_a \lambda_a^2$, which is the proportion of distance displayed on dimension a.

Correspondence Analysis of Career Data

In the previous section, we gave a technical description of CA of contingency tables. Here we discuss its relevance for career data.

One might wonder in what kind of situation CA is an appropriate tool for the analysis of data. One answer to this question is: whenever a data matrix can be constructed for which it is useful to study differences between row profiles, column profiles, or both. Notice that this is much more general than the description in the previous section, which started from a two-way contingency table. Thus, other types of data can be coded into matrices that can meaningfully be analyzed by CA. Examples are incidence data, pick-any-out-of-m-data, and quantitative data, details of which can be found in the references given in the previous section.

The relevance of CA for career data follows from the fact that each individual career can be coded into count data, where counts describe the amount of time an individual has spent in each of a number of mutually exclusive states. For example, assume that we are interested in the amount of time people have spent in each of three states, namely, *school*, *job*, and *jobless*. Assume that we have a cohort of people for which this is known from ages 12 to 24. We could then count the number of years each of the individuals was in school, had a job, and was jobless. This could result in data like this:

Individual	School	Job	Jobless	Total
1	6	6	0	12
2	5	2	5	12
3	8	0	4	12
		and so on		

Individual 1 spent 6 years in school, had a job for 6 years, and was never jobless. Individual 2 was 5 years in school, had a job for 2 years, and was jobless for 5 years. Individual 3 was 8 years in school, never had a job, and was jobless for 4

years. The profile for Individual 1 is 6/12, 6/12, 0. For Individual 2 it is 5/12, 2/12, 5/12. For Individual 3 it is 8/12, 0, 4/12. CA would study differences between these profiles by making a two-dimensional representation of the individuals and the states they are in. It would show which individuals have similar count data and which have different count data, and it would also show which states are similar or dissimilar in the sense that when individuals have spent time in one state they are likely or unlikely to have spent time in another state.

Notice that we could also measure the status of individuals more precisely by counting months, weeks, or even days. Although this would lead to large increases in the counts, it would lead to only relatively small modifications in the profile elements of the individuals. Notice also that the order in which individuals have spent time in each of the states is lost. This problem can be solved by making two matrices like the one above for two 6-year periods and concatenating them. This would lead, for example, to

	First 6 years			Second 6 years		
Individual	School	Job	Jobless	School	Job	Jobless
1	6	0	0	0	6	0
2	5	0	1	0	5	1
3	6	0	0	2	0	4
			and so on			

Now the number of columns is six, and therefore each profile has six elements. CA does not use the information that the first three states are measured earlier than the second three states, but in the interpretation this information can be used by labeling these states differently (see the examples discussed later). The CA of the matrix with only 3 states can be considered as a restricted version of the CA of the matrix with 6 states, the restriction being that the column scores are equal over the time points (see van der Heijden, 1987; van Buuren & de Leeuw, 1992). This also shows that it is useful to split up the 12-year period into two 6-year periods when the profiles of corresponding categories differ considerably over the two 6-year periods.

It is also possible to code the data into 12 periods, one for each year. Then the data consist only of ones and zeros, indicating whether an individual is in a state in a particular year (leading to a one) or not (zero). For the small example above this would lead to

Individual	1	2	3	4	5	6	7	8	9	10	11	12
1	100	100	100	100	100	100	010	010	010	010	010	010
2	100	100	100	100	100	001	001	010	010	010	010	010
3	100	100	100	100	100	100	100	100	001	001	001	001

For each year one pattern chosen from 100, 010, and 001 indicates the state an individual is in: for the 100 the individual is in school, for 010 he/she has a job,

and for 001 he/she is jobless. The matrix that is formed for each time point is called an indicator matrix, and the concatenation of such matrices is called a super indicator matrix. A CA of this matrix would yield a solution with 3 × 12 state-year points. The CA of the matrix with only three states discussed above can be considered again as a restricted version of the CA of the super indicator matrix, the restriction being that the quantifications of corresponding categories over the 12 time points are equal.

Missing Data

CA of a super indicator matrix is also known as multiple CA. This form of CA will be discussed in a separate section, but before we go to this section we would like to indicate that although the interpretability of the results is simplest when the states are mutually exclusive, they do not necessarily have to be exhaustive. That is, if for the first table the counts for the first individual were 6, 4, 0 (because for some reason we had lost track of this individual's state during the last two years), it could still be useful to calculate his profile, which would be .6, .4, 0, and compare it with the other profiles. This is one way to solve the missing data problem in CA. Another way is to define a new category, namely, "lost track"; the profiles would then be 6/12, 4/12, 0, 2/12 for Individual 1, 5/12, 2/12, 5/12, 0 for Individual 2, and 8/12, 0, 4/12, 0 for Individual 3.

A choice for one of the two missing data options should be made depending on how the missing data are generated. If the missing data are generated in a random way, then it makes sense to compare the first individual with profile .6, .4, and 0 to the profiles of the other individuals. Thus missingness for some individual is assumed to be unrelated to the information that is available for this individual, and therefore the available information is assumed to be representative for the whole career. If, on the other hand, missingness is related to the other scores an individual has, then the second option is a wiser choice. For example, if missingness occurs more often in individuals that spend more time being jobless, then using an extra category for missingness would be useful, because CA would then reveal this relation between this extra category for being missing and the category for being jobless.

In a later section we will illustrate these two possibilities and some possible consequences in an example. In this example missingness of part of pupils' careers is not a random process but is related to the available data, and so the second option, creating a separate category for missingness, is the better choice; however, we also illustrate the first option to show the impact that different choices for handling missing data can have. For a more detailed discussion of this area, and for other alternatives for dealing with missing data, we refer to Meulman (1982), van der Heijden and Escofier (1988), Gifi (1990), and van Buuren and van Rijckevorsel (1992).

Some Basic Choices in Constructing the Data Matrix

Before a CA is undertaken, some thought has to be given to the way in which

the data matrix is constructed. The main question is, How do we want to compare the careers of people, when chi-square distances are calculated? Some typical questions will be framed for two careers only:

- Is the time scale of two careers matched by their chronological year, so that, for example, the states in 1992 are compared in calculating the chi-square distance?
- Or do we use the age of respondents, so that the states of someone being 12 in 1993 are compared with the states of someone being 12 in 1990?
- Or do we use a sort of cohort principle to compare two careers—in which, for example, the life event upon which two careers are matched is the year when both started university, so that someone starting in 1992 is compared with someone starting in 1990?

The appropriate choice will depend on the particular research question, of course. Missing data patterns can emerge at the start or end of career data, as a result of particular choices. Again, the two choices discussed in the paragraph above are most appropriate for solving these missing data patterns. Again, if missingness is random, then the first missing data option is most appropriate, whereas if missingness is related to other characteristics of a career, it is better to investigate this relation by creating a separate category for being missing.

It might also be that, given the substantive research question, the way in which career data have to be matched depends on the actual data. For example, assume one career has as a sequence of states a-b-c-d, and another has the sequence b-c-d-a. Then the sequence b-c-d can be matched exactly. This would then lead to

a	b	c	d	mis
mis	b	c	d	a

For this purpose, optimal matching techniques could be used (see Abbott & Hrycak, 1990). After matching the sequences, these authors calculate a distance measure by counting the number of differences between the careers. These distances can then be studied using some multidimensional scaling or cluster analysis technique. An alternative would be to apply CA to provide a graphical representation of the careers. We know of no experience with this particular proposal.

Multiple Correspondence Analysis of Career Data

Multiple CA of career data can be interpreted in terms of chi-square distances, but this has certain drawbacks, to be discussed later. The equation for chi-square distances (1) indicates that careers will be close together when individuals have used their time in similar ways, and farther apart when they have used their time in different ways.

The origin **O** is formed by the profile of elements p_{+j} which correspond in this context to the relative frequency with which the states occur on each of the time

points. Individuals are farther from the origin when their profiles are less similar to the average profile in the origin, and multiple CA can be used to locate groups of individuals that depart in similar ways from the origin. The corresponding elements n_{+j} can be used to form a matrix itself of time points (here, 12) by states (here, 3). This matrix then shows which states occur particularly at which time points. It is important to have insight into this aspect of the data, because multiple CA focuses only on the departure from this average, and groups are formed in terms of their common departure from this average.

This led van der Heijden and de Leeuw (1989) to recommend a triple analysis in the case of career data (they speak of *event history data*, which is a bit more general), namely, CAs of (a) the matrix with margins n_{+j}, showing which states occur at which time points; (b) the super indicator matrix, showing how individuals depart from the average studied in (a); and (c) the matrix in which the career is not split up into separate concatenated indicator matrices, that can be considered as a restricted version of (b). If the configuration of career profiles is very similar in (b) and (c), then not much is gained in terms of information by studying (b), but considerable stability is gained by choosing (c) instead of (b). We will show an example of such a triple analysis later.

It is generally agreed that it is not a good idea to interpret a multiple CA in terms of chi-square distances. This is mainly due to the fact that there are restrictions in the complete pattern of 36 zeros and ones, namely, that at each time point only one value of one can occur. This leads to dimensions which are considered artificial. Another result is that there are artificial elements in the chi-square distances in full-dimensional space. Consider a much simpler situation: two time points. Then one option would be to analyze the contingency table of 3×3. This would provide a two-dimensional solution. Another option would be to analyze the super indicator matrix of order $n \times (3 + 3)$. This would provide a four-dimensional solution. However, it can be shown that the two-dimensional solution of the matrix of 3×3 can be derived from the first two dimensions of the four-dimensional solution of the matrix of $n \times (3 + 3)$ (see Gifi, 1990, pp. 272–273). Therefore, the last two dimensions can be considered to be artificial for this example, and in the four-dimensional space there are artificial elements in the chi-square distances. It is beyond the scope of this article to discuss this in more detail (but see Israels, 1987, and Greenacre, 1987). It is clear that other motivations for multiple CA should help to enhance its respectability. These are available, and one of them will be discussed, namely, its relation to principal component analysis of nominal data (see also de Leeuw & van Rijckevorsel, 1980).

Multiple CA as a Principal Component Analysis of Nominal Data

One way to define principal component analysis (PCA) is as follows. Let there be a matrix **X** with n rows and m columns having quantitative measures. Then the first principal component $\mathbf{z}_1$ consists of those scores that maximize $\phi_1 = \Sigma_m(\text{cor}(\mathbf{z}_1, \mathbf{x}_m))^2$, where ϕ_1 is the first eigenvalue of the correlation matrix

computed from **X**. Thus, the first principal component $\mathbf{z}_1$ summarizes what the m variables have in common. The second principal component $\mathbf{z}_2$ consists of those scores that maximize $\phi_2 = \Sigma_m(\text{cor}(\mathbf{z}_2, \mathbf{x}_m))^2$, under the restriction that $\text{cor}(\mathbf{z}_1, \mathbf{z}_2) = 0$. And so on for principal components $\mathbf{z}_3$ to $\mathbf{z}_m$.

Consider now the situation in which the information in the matrix **X** is categorical. For example, for the career data above there are 12 variables, each having three states: *school*, *job*, and *jobless*. Assume that we would like to do a PCA, but that we are in need of quantitative measures to replace the states. If we impute the column scores c_{j1} obtained from the first multiple CA dimension in **X**, then the row scores for the first dimension of multiple CA, collected in $\mathbf{r}_1$, are the scores that maximize the first eigenvalue $\phi_1 = \Sigma_m(\text{cor}(\mathbf{r}_1, \mathbf{x}_m))^2$. Thus the first dimension of multiple CA can be interpreted in terms of PCA of nominal data. For higher dimensions there is also a relation between PCA and multiple CA, but for this we refer to Gifi (1990, chap. 3).

Multiple CA and the Burt Matrix

Closely related to the interpretation of multiple CA as a tool for PCA of nominal variables is the relation between the super indicator matrix and the so-called Burt matrix. Assume that we denote the super indicator matrix by **G**. It can be proved that one way to find the multiple CA solution is by performing an eigenvalue-eigenvector decomposition of a function of the matrix $\mathbf{B} = \mathbf{G}'\mathbf{G}$, a square symmetric matrix which is called the Burt matrix. The Burt matrix **B** has the form shown in Figure 1. Each square in this matrix represents a submatrix.

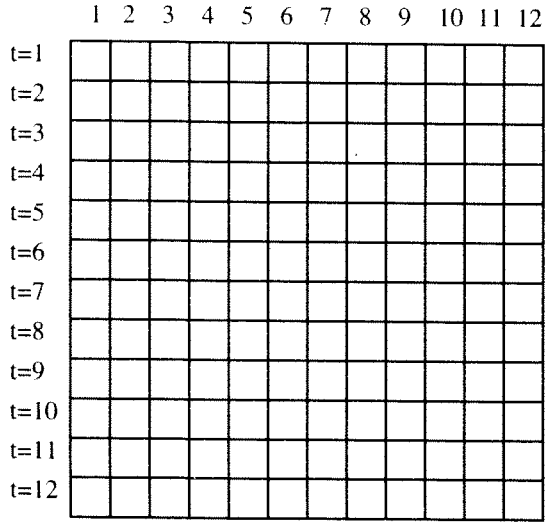

FIGURE 1. *The form of the Burt matrix*

On the diagonal we find diagonal matrices with the marginal frequencies of the three states for each of the time points. Off the diagonal we find contingency tables of time point t in the rows and t' in the columns. These contingency tables are the transition matrices showing how many individuals in state j at time point t move to state j' at time point t'.

The eigenvalue-eigenvector decomposition is computed from the matrix $\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}$, where **D** is a diagonal matrix with the marginal frequencies of **B**. The diagonal submatrices of $\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}$ are identity matrices. The section above on multiple CA and PCA for nominal data illustrates incidentally that by using the first quantification of the states for each of the time points and the transition matrices of the Burt matrix, each transition matrix can be summarized by a correlation coefficient. This strengthens the analogy to PCA.

Since the Burt matrix has all the necessary information to derive the column configuration of a multiple CA, it becomes evident that only sequences of two time points are used in an analysis, and information in transitions using three or more time points is neglected (de Leeuw, 1984). We refer to van der Heijden and de Leeuw (1989) for remarks on a relation between multiple CA and Markov-chain models.

Example: School and Work Careers of a Cohort of Pupils in a Dutch Educational Priority Area

The aim of this example is threefold. First, it is to illustrate the usefulness of CA and multiple CA for the analysis of career data. Second, it is to show how and under what circumstances it is possible to obtain quantifications or classifications of careers useful for further analyses. Third, we aim to illustrate the results of different choices for handling missing data.

Some Background on the Dutch School System

In the Netherlands, after eight years of primary education, pupils have compulsory secondary education until the age of 16. For an outline of the Dutch educational system, see the scheme in Figure 2.

There are three main streams of secondary education. However, before entering these streams nowadays, most pupils first go for one or two years to so-called bridge classes (brugklassen: BK1 or BK2), which is a first form of secondary education and is meant to stream the pupils. The main streams of secondary education are:

(1) preparatory vocational education (VBO). This form of education lasts 4 years and consists of lower preparatory vocational education (LBO) and individual preparatory vocational education (IBO).
(2) general preparatory education (AVO). AVO consists of two forms, namely, middle (MAVO) and higher (HAVO) general preparatory education, which last 4 and 5 years, respectively.
(3) preparatory scientific education (VWO). VWO prepares for scientific education. There are two forms of VWO, namely, secondary modern

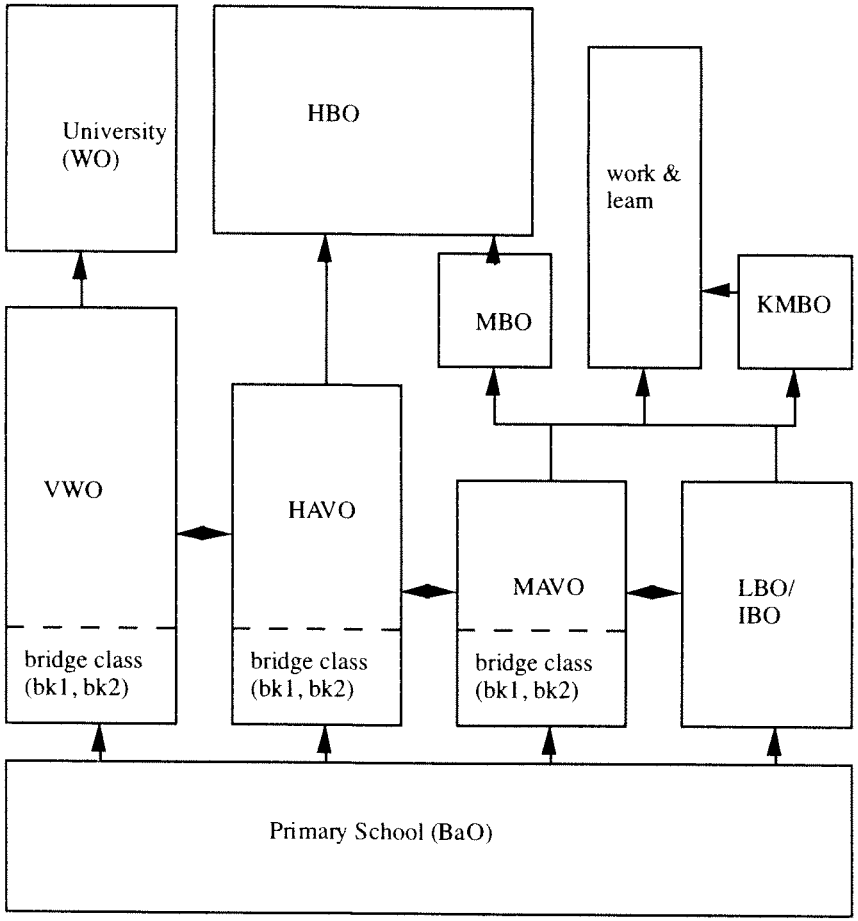

FIGURE 2. *Rough outline of the Dutch educational system*
Note. Arrows indicate changes between education types that are not uncommon.

school (atheneum) and secondary grammar school (gymnasium), each 6 years of schooling.

After secondary education, most pupils (if their school careers were successful, of course) are about 16 to 18 years old. They can continue their educational careers, as most pupils do, in all kinds of tertiary education. A small group of youngsters, however, choose jobs (which is allowed, because by then education is no longer compulsory).

Most pupils of VBO and MAVO continue their school careers in a form of further vocational education, MBO, which lasts 4 years, or in a shorter version of this kind of education, KMBO, which last 2 years. These pupils can also choose forms of combined "work and learn" education (in trainee and apprentice systems). Pupils of HAVO can follow further education in forms of higher vocational education (HBO), which lasts 2 to 5 years. VWO leads primarily to university, of course, but a substantial number of these pupils also go to forms of higher vocational education.

In Figure 2 you can also see that an accumulation of school forms is possible in the Dutch educational system and that there are all kinds of cross-stream possibilities too. In principle all kinds of transfers between streams of secondary education are possible. In practice, however, they are not always very likely. Moving up from preparatory vocational education (VBO) to general preparatory education (MAVO) is not very usual. Moving down is more likely, especially for those pupils who realize in MAVO that they prefer to work with their hands rather than with their head. There is a lot of moving up and down within general preparatory education, from middle (MAVO) to higher (HAVO) streams and vice versa. Qualified MAVO pupils can continue their school careers in HAVO4. There is also some up- and downward moving of pupils from general preparatory education (HAVO) to preparatory scientific education (VWO). However, this happens less than between MAVO and HAVO. The transfer of qualified pupils from HAVO to VWO5 is generally considered to be a difficult one.

Some Background on the Sample

Originally 688 pupils from Class 8 of 31 primary schools participated in this research project. These schools are in The Hague, Rotterdam, and Utrecht, three big cities in the western and middle part of the Netherlands. The schools have, in broad outline, the same educational and pedagogical principles.

All schools are situated in so-called *educational priority areas* (onderwijsvoorrangsgebieden). These areas have been selected by the Dutch government because on average the school populations in these areas have substantial educational deficits. Schools in these areas receive additional funding in order to organize extra activities, to appoint teachers, and to lower the rate of pupils per classroom.

School populations in educational priority areas are characterized by a large proportion of migrant pupils and indigent pupils, mainly from working-class and lower-middle-class families. For instance, half of all the schools in these areas have populations more than 33% of which are migrant children. This ethnic diversity can also be found in the research population: In total, 43% of the pupils are indigent and 57% are migrant pupils. The largest migrant subgroups are Turkish (18%), Surinamese (16%), and Moroccan (14%) pupils. This corresponds with proportions in these areas at large.

After selecting the 31 primary schools, a cohort was formed of all the pupils of Class 8. The school careers of these pupils were assessed yearly as they continued through secondary and (eventually) tertiary education (see Figure 2). In the beginning, the secondary schools concerned were asked to give information about these pupils. If pupils got "lost," however, they also were approached personally. In difficult cases local school authorities or the Municipal Office for

Compulsory Education (Gemeentelijke Afdeling Leerplichtzaken) assisted in tracing these pupils. But, of course, pupils got out of the original sample—for instance, because they got a job, went back to Morocco or Turkey, married and stayed at home, or had to serve in the army.

We have coded the school careers as follows. Starting with the school year 1985–1986, each pupil is classified in one of 38 levels, which are mostly school levels (see the rows of Table 1). Most levels are self-explanatory in Figure 2. BaO is the last year of primary education: It turned out that 13 children did not pass from the eighth level of primary education to the first level of secondary education. BK1 and BK2 indicate the bridge classes. There are four levels for IBO, four for LBO, four for MAVO, five for HAVO (though HAVO1 does not occur in the data; these pupils are then in a bridge class), and six for VWO. Then there is one level for other primary education, such as the so-called "Middenschool." We then go to levels of education that follow VWO, HAVO, MAVO, and LBO/IBO. We find KMBO, four levels for MBO, and three levels for HBO/WO. The remaining levels are then work, reeducation (i.e., choosing a different type of education in order to increase the chance of finding a job, "work and learn" (i.e., following training at work). This also holds for the number of children for whom the information is missing. Of the individuals in the category *else*, about 75% have gone back to Morocco or Turkey, about 3% are listed for the army, about 20% are absent from school for a long time (although they are supposed to go there), and about 3% are deceased. A peculiarity of the data is that the number of children in the category *else* grows steadily. The growing number of missing pupils has to do with the fact that pupils leave school and fail to provide further information on their whereabouts to the data bank.

Since we are dealing with a cohort of pupils who were in the last class of primary education in the school year 1984–1985, many of the frequencies in 1985–1986 are zero (denoted by a blank) in Table 1. The scheme in Figure 2 shows that after primary education all children are supposed to start in VWO1, HAVO1, MAVO1, LBO1, IBO1, or BK1. For example, 101 pupils start in 1985–1986 in MAVO1. In 1986–1987 there are 13 pupils in MAVO1 (probably pupils that failed to pass to MAVO2), and in 1987–1988 there is only 1 pupil left at this type of school. In later years this school level is no longer used. For the school type MAVO, the standard school career for the first 4 years is either MAVO1-MAVO2-MAVO3-MAVO4 or BK1- MAVO2-MAVO3-MAVO4, which explains the high number of pupils in MAVO2 in 1986–1987, in MAVO3 in 1987–1988, and in MAVO4 in 1988–1989. But, although these careers are standard, it is also clear that many pupils do not follow standard careers because they fail to pass or because they switch to different types of education.

The Analyses

The careers that we will study here are a special case of event history data. For this type of data van der Heijden and de Leeuw (1989; see also van der

TABLE 1
Education types by years

Level	85/6	86/7	87/8	88/9	89/0	90/1	91/2	92/3	93/4
BAO	13	1							
BK1	362	61	5						
BK2		156	52	6					
IBO1	40	7	1						
IBO2		46	8	2			1		
IBO3			46	14	3				
IBO4				35	8	3			
LBO1	118	8							
LBO2		142	43	5					
LBO3		1	167	87	10	1			
LBO4				126	49	12	1		
MAVO1	101	13	1						
MAVO2	1	149	58	5					
MAVO3			141	86	8				
MAVO4				117	55	6	2		
HAVO2		18	10						
HAVO3			31	23					
HAVO4				28	37	14	3		
HAVO5					9	25	14	1	
VWO1	2								
VWO2		27	1						
VWO3			39	2					
VWO4				29	3	1			
VWO5					16	4	7	1	
VWO6						14	4	7	2
O.PRI	24	22	19	18	4	1	1		
KMBO			2		21	20	13	4	
MBO1					20	23	7	4	
MBO2						10	13	5	
MBO3							6	9	4
MBO4								4	
HBO/WO1							1	3	
HBO/WO2								1	2
HBO/WO > 5							1		
Work				2	13	29	37	42	42
Reeducation					2	7	2	3	
Work and learn				2	15	19	13	7	3
Else	27	37	61	84	111	183	223	247	261
Missing	0	0	3	17	304	316	339	350	374

Note. Each frequency indicates the number of people at a given level in a given year. A blank indicates that no people were at a given level in a given year.

Level abbreviations are explained in the text. 85/6 = school year 1985–1986, 86/7 = school year 1986–1987, and so on.

Heijden, 1987) suggest focusing on three types of analysis, which we discuss here.

A first analysis proposed in van der Heijden and de Leeuw (1989) is the correspondence analysis (CA) of the contingency table of categories by time shown in Table 1. In the context of this example, this contingency table has the 38 school levels as rows and the nine (school) years as columns. The entries of this table are frequencies, denoting how many pupils are falling into a particular school level in a specific year. We could analyze this table with CA, but this does not seem to provide much more information than the information evident in the table on visual inspection. Therefore we refrain from showing this analysis.

Multiple correspondence analysis, missing excluded. Table 1 provides information about the use of school levels over the years. It does not provide information on individual careers. A study of such careers makes clear what sorts of switches occur between school types such as IBO, LBO, MAVO, HAVO, VWO, MBO, and HBO/WO. We study these careers by performing a multiple CA. For this purpose a table is created with 688 pupils in the rows, and the number of school levels times the number of years in the columns. Although 38 school/work levels are involved over the nine years, the number of columns is much less than 9×38, since in the first years some types of school levels are not yet used, whereas in later years specific columns are not used anymore. In fact, only those combinations of years and school/work levels in Table 1 that have frequencies larger than zero have a column, so for the year 1985–1986 there are 9 columns, for 1986–1987 there are 14 columns, and so on. The resulting table has ones and zeros entered to indicate whether a pupil falls into a specific school level in a specific year (score equals 1) or not (score equals 0). If a pupil's information is missing in a particular year, he or she receives a zero on all 38 school levels. This table is then analyzed with multiple CA. The input for such an analysis in SPSS is provided in the Appendix.

The first four eigenvalues of the multiple CA solution are .72, .65, .62, and .58. A plot of the quantifications of the pupils obtained for the first two dimensions is shown in Figure 3. It turns out that the cloud of points for the pupils form a horseshoe in the first two dimensions. Theoretical studies on the occurrence of horseshoes indicate that when such a horseshoe occurs, the first dimension reveals the main features of the data, and it is of little use to study higher dimensions (see Schriever, 1983, 1986; van Rijckevorsel, 1986). The reason is that higher dimensions are polynomial functions of the first dimension: In Figure 2 we find that, roughly, the object scores for the second dimension are a quadratic transformation of the object scores for the first dimensions. Therefore we focus for school levels on their quantification in Dimension 1. These quantifications are provided in Table 2. A graphical representation is given in Figure 4. In this graphical representation those school level–year combinations having only one or two pupils are neglected, in order to simplify interpretation. We have connected the school level–year combinations of commonly followed tracks, such as the VWO track, going from VWO2 in 1986–1987 to VWO6 in

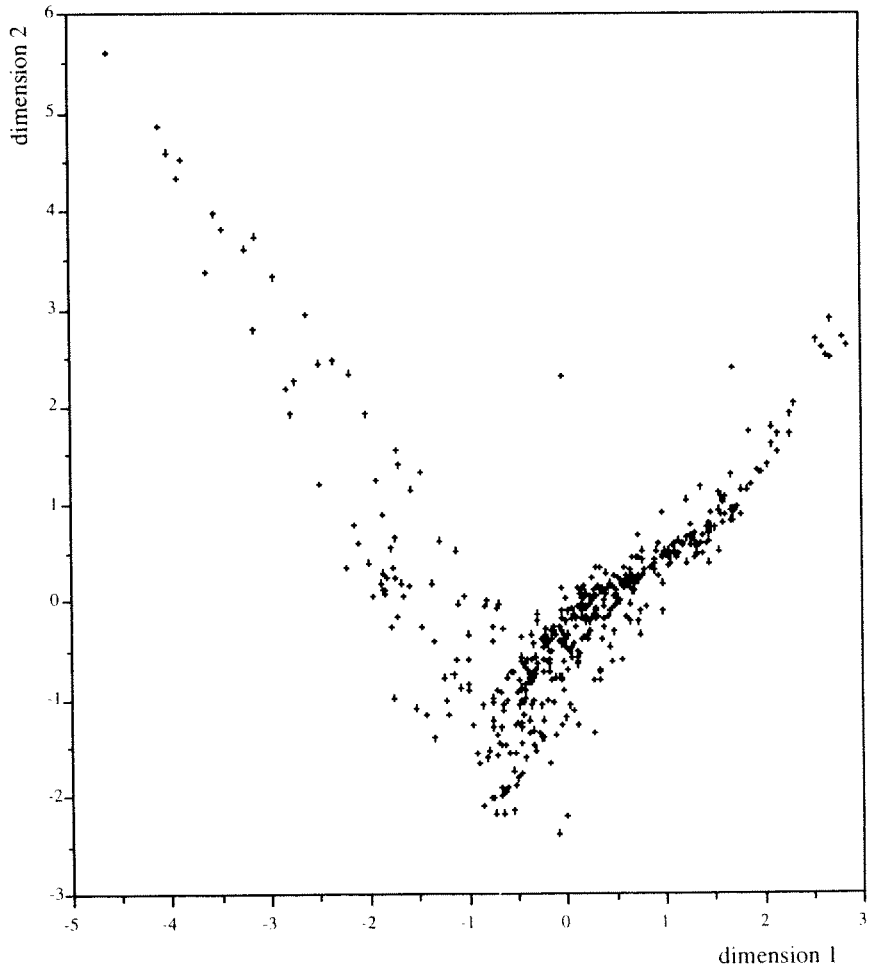

FIGURE 3. *Plot of 688 careers, using the first two dimensions of the multiple CA solution using 38 categories.*

1990–1991. In such a track pupils always pass to the next grade. The other commonly followed tracks are those of BK, HAVO, MAVO, MBO, IBO, LBO, other primary, work and learn, work, and else. The frequencies in Table 1 show that these tracks cover the school careers of a majority of the pupils. For example, for the LBO it covers the careers of 118 pupils in 1985–1986, 142 in 1986–1987, 167 in 1987–1988, and 126 in 1988–1989.

For the interpretation of Figures 3 and 4, it is useful to remember the transition formulas provided in (3) and (4) and the formula for chi-square distances provided in (1) and (2). These formulas provide the following guidelines for interpretation (compare Gifi, 1990, pp. 118–120):

TABLE 2
Education types by years

Level	85/6	86/7	87/8	88/9	89/0	90/1	91/2	92/3	93/4
BAO	.35	−.09							
BK1	−.57	−.20	.02						
BK2		−.45	−.38	−.24					
IBO1	.98	.88	.75						
IBO2		.88	.85	1.10			−1.66		
IBO3			.88	.84	1.17				
IBO4				1.01	.66	.77			
LBO1	1.40	1.10							
LBO2		1.24	.79	.26					
LBO3		.42	1.00	.54	.03	1.28			
LBO4				1.01	.59	.18	−.16		
MAVO1	−.44	−.06	−.01						
MAVO2	−.67	−.60	−.37	−.32					
MAVO3			−.68	−.44	−.14				
MAVO4				−.73	−.58	−.36	.10		
HAVO1									
HAVO2		−1.49	−1.63						
HAVO3			−1.34	−1.62					
HAVO4				−1.54	−1.65	−1.05	−.57		
HAVO5					−1.65	−1.93	−1.57	−.79	
VWO1	−3.87								
VWO2		−3.26	−2.74						
VWO3			−2.88	−2.29					
VWO4				−3.22	−2.83	−2.74			
VWO5					−3.38	−3.32	−1.87	−2.09	
VWO6						−3.33	−3.88	−2.44	−1.93
O.PRI	1.66	1.99	2.08	2.01	2.27	.97	.97		
KMBO			.86		1.40	1.23	1.25	.41	
MBO1					.09	−.12	.04	−.34	
MBO2						−.04	−.17	.25	
MBO3							−.16	−.25	−.13
MBO4								−.10	
HBO/WO1							−3.62	−1.82	
HBO/WO2								−3.62	−1.85
HBO/WO > 5							−2.21		
Work				1.67	1.50	1.24	1.18	1.12	1.12
Reeducation					−.13	.10	.19	−.28	
Work and learn				−.10	1.97	1.94	2.21	2.27	2.50
Else	.40	.27	.24	.34	.19	.08	−.02	−.08	−.10

Note. Each figure is the category quantification of the first dimension of the multiple CA. The frequency of occurrence of each combination of year and level is given in Table 1.
Level abbreviations are explained in the text. 85/6 = school year 1985–1986, 86/7 = school year 1986–1987, and so on.

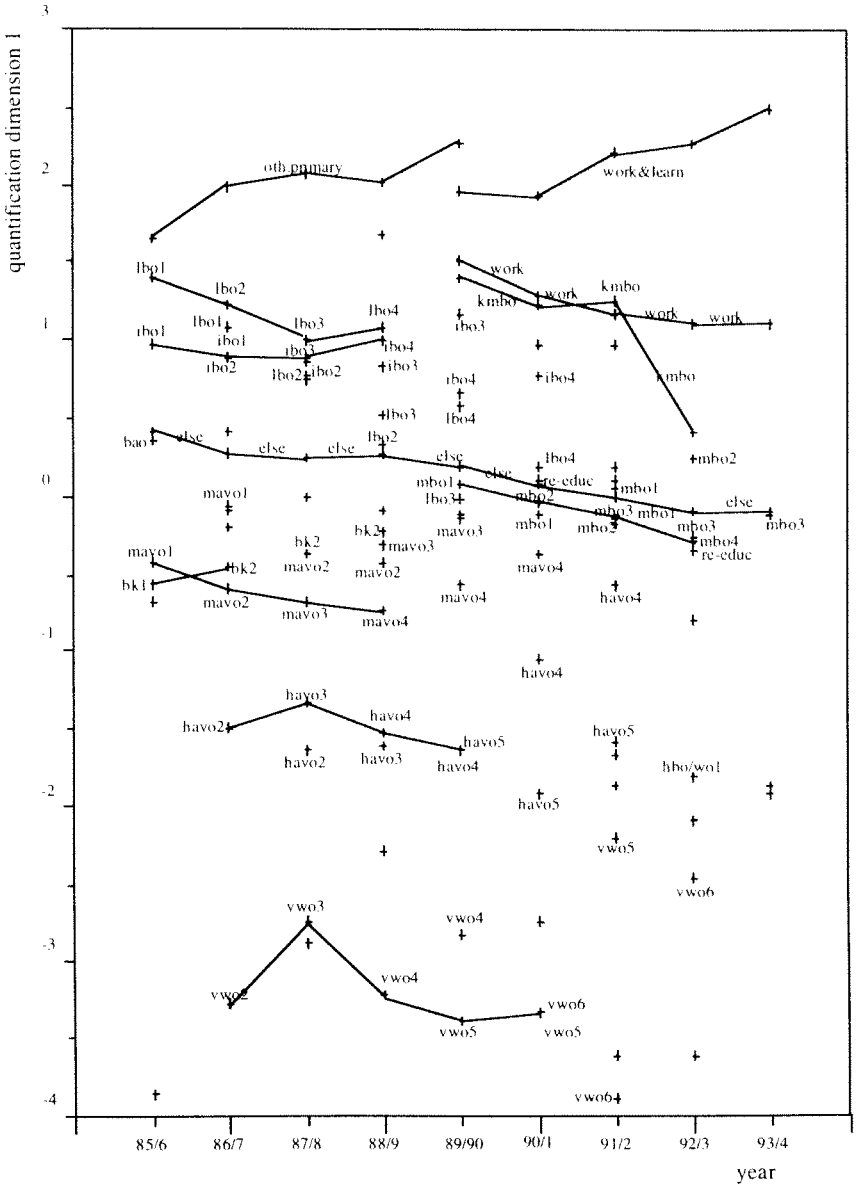

FIGURE 4. *Plot of school levels, using the first dimension of the multiple CA solution using 38 categories*
Note. Lines indicate commonly followed tracks.

(1) Pupil points in Figure 3 that are close together represent school careers that are very similar, in the sense that over many of the years these pupils are in the same school type. Pupil points that are far apart represent school careers that have hardly any elements in common. This follows from (1) for the chi-square distance between row profiles.
(2) School career–year combinations in Figure 4 whose quantifications are close together are used simultaneously by many pupils. For example, since HAVO2 in 1986–1987 and HAVO3 in 1987–1988 are close together, it follows that pupils who were in HAVO2 in 1986–1987 have a probability larger than average of being in HAVO3 in 1987–1988. In other words, the conditional probability of falling in HAVO3 in 1987–1988 given HAVO2 in 1986–1987 is larger than the unconditional probability of falling in HAVO3 in 1987–1988. This follows from the equation for the chi-square distance between column profiles.
(3) School careers in Figure 3 are close to the school type–year combinations in Figure 4 they are composed of, and vice versa. This follows from the transition formulas.

This brings us then to the following interpretation of Figure 4, in relation to Figure 3:

(1) The school careers on the extreme left in Figure 3 are made up for a large part of VWO years. Moving from the extreme left to the extreme right, the school careers are made up predominantly from VWO, HAVO, MAVO, BK1 and BK2, MBO, LBO and IBO, KMBO, work, other primary, and work and learn. This is immediately clear from the commonly followed tracks indicated by lines in Figure 4. The ordering of these common tracks on Dimension 1 is identical to what is generally considered a progression from less towards more academic school careers.
(2) The second guideline given in the previous paragraph allows the interpretation that if careers contain one or more VWO years and also some other years, these other years are predominantly HAVO years. For school careers which comprise predominantly HAVO years, possible other years are VWO years and MAVO years. Careers with MBO in the later years have predominantly MAVO and LBO years in the first years. Careers that end with work and learn started predominantly with other primary education, LBO and IBO.
(3) The previously mentioned interpretation that from bottom to top in Figure 4 (left to right in Figure 3) the school careers proceed from more academic to less academic leads to some unexpected findings in Figure 4. For example, we find that the quantification for VWO4 in 1988–1989 is more in the direction of academic careers than VWO4 in 1989–1990. This is in accordance with the more/less academic interpretation, since school careers with VWO4 in 1989–1990 have VWO4 in their fifth year, indicating that these pupils did not pass in any of the previous years. However, the reverse seems to happen for many of the HAVO years. For example, school careers that have HAVO3 in their fourth year are more academic than school careers that have HAVO3 in their third year! For MAVO and BK the order of points goes as we would expect, but for LBO the orders go in the wrong directions. We have checked what is happening in these careers for some school levels for which a reasonable amount of data is available in both years.

For HAVO3 we compared the 31 careers that had HAVO3 in 1987–1988 with the 23 that had HAVO3 in 1988–1989 (see Table 1). We calculated the number of years spent by the two groups in VWO and in VWO-HAVO. It turns out that the HAVO3-in-1988–1989 careers (careers where pupils failed to pass in any of the previous years) spent 12 of their 162 nonmissing years in VWO (proportion is .074) and 94 of their 162 nonmissing years in VWO or HAVO (proportion is .580). For the HAVO3-in- 1987–1988 careers (passing careers) this is 3 out of 179 in VWO (.017) and 94 out of 179 in VWO-HAVO (.525). This explains why the HAVO3-in-1987–1988 careers (the passers) tend more towards the less academic careers and the HAVO3-in-1988–1989 careers (the failers) tend more towards the more academic careers.

For MAVO3 we compared the 141 careers that had MAVO3 in 1987–1988 with the 86 that had MAVO3 in 1988–1989 (see Table 1). We calculated the number of years spent by the two groups in HAVO-VWO-HBO/WO and in BK-MAVO-MBO-HAVO-VWO-HBO/WO. It turns out that the MAVO3-in-1988–1989 careers (careers where pupils failed to pass in any of the previous years) spent 11 out of their 588 nonmissing years in HAVO-VWO-HBO/WO (proportion is .0187) and 400 out of their 588 nonmissing years in BK-MAVO-MBO-HAVO-VWO-HBO/WO (proportion is .680). For the MAVO3-in-1987–1988 careers (passing careers) this is 30 out of 842 in HAVO-VWO-HBO/WO (.0356) and 94 out of 179 in BK-MAVO-MBO-HAVO-VWO-HBO/WO (.722). This explains why the MAVO3-in-1987–1988 careers (the passers) are found more towards the more academic careers and the MAVO3-in-1988–1989 careers (the failers) predominantly towards the less academic careers.

For LBO2 we compared the 142 careers that had LBO2 in 1986–1987 with the 43 that had LBO2 in 1987–1988 (see Table 1). We calculated the number of years spent by the two groups in MBO and in MBO-MAVO-BK. It turns out that the LBO2-in-1987–1988 careers (careers where pupils failed to pass in any of the foregoing years) spent 7 out of their 281 nonmissing years in MBO (proportion is .0249) and 48 out of their 281 nonmissing years in MBO-MAVO-BK (proportion is .171). For the LBO2-in-1986–1987 careers (passing careers) this is 17 out of 860 in MBO (.020) and 57 out of 860 in MBO-MAVO-BK (.066). This explains

why the LBO2-in-1986–1987 careers (the passers) tend to be found more towards the less academic careers and the LBO2-in-1987–1988 careers (the failers) towards the more academic careers.

The above interpretation shows that we can use the quantifications on the first dimension as a score summarizing the academic level of the school careers of these pupils. A next step in the analysis is then to relate the quantifications of the multiple CA to some background variables of the pupils. For quantitative background variables this can be done by calculating correlations with other variables, and for categorical background variables this can be done by calculating mean scores for each of the categories.

A first variable of interest is the so-called Dutch CITO-score, which is a school achievement score measured before pupils enter secondary school. This score, available for 382 of the pupils, plays a role in the choice of a specific type of secondary school, such as vocational or general education. For the 382 pupils the correlation between this CITO-score and the first dimension of multiple CA is .58. Furthermore, for some categorical background variables that seemed relevant for school/work careers we have calculated means of the quantifications of the school careers provided by the first dimension of multiple CA. The results are shown in Figure 5. The differences between the means for sex were not

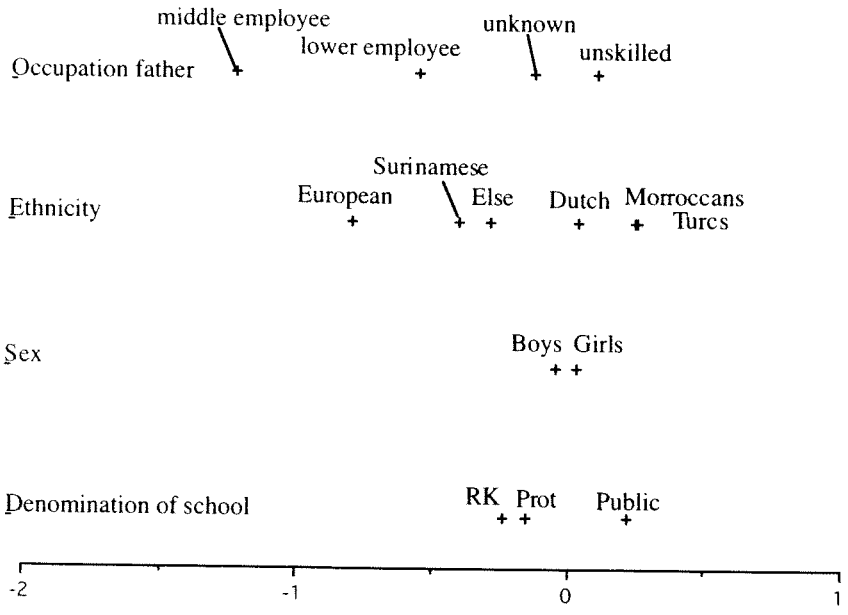

FIGURE 5. *Average object scores on first dimension of multiple CA solution for four background variables: occupation of father, ethnicity, sex, and denomination of school*
Note. RK = Roman Catholic, Prot = Protestant.

significant. More academic school careers are associated with Roman Catholic (RK) and Protestant primary schools, European boys and boys with a Surinamese ethnicity, and pupils whose fathers are middle or lower employees. Less academic school careers are associated more with public primary schools, Moroccans and Turks, and pupils whose fathers are unskilled workers.

Correspondence analysis: A multiple correspondence analysis with equality restrictions. In the former analysis we saw that the rises and falls of lines in Figure 4 correspond with interesting aspects of the data. For example, pupils who fail in MAVO end up with a less academic school career score, whereas pupils who fail in HAVO and LBO end up with a more academic school career score. However, it is not unlikely that some of the fluctuations over time are a bit unstable due to small sample size.

A more stable result is then obtained by another analysis proposed by van der Heijden and de Leeuw (1989), the CA of the matrix of 688 pupils by 38 school/work levels. Such an analysis is equivalent to a multiple CA with the additional restriction that the quantifications of identical levels are equal over time (cf. van Buuren & de Leeuw, 1992; van der Heijden & de Leeuw, 1989). The lines in Figure 4 are then restricted to being horizontal. The plot of the 688 careers in two dimensions is shown in Figure 6. We again find a horseshoe, and it turns out that the correlation between the quantifications for the school careers yielded by this CA and the quantifications for the school careers yielded by the multiple CA is .9927, which shows that for all practical purposes the quantifications of these school careers by these two analyses can be considered the same. Figure 7 shows the level quantifications for this analysis. It is clear that the quantifications in Figure 7 are very close to averages of the corresponding lines in Figure 4.

Multiple correspondence analysis, missing included. One of the main decisions to be made in applications of multiple CA concerns the way in which missing data are dealt with. We discussed this in some detail in an earlier section. Thus far we have considered two types of missing data, and we treated the two differently:

(1) The information is known but in principle irrelevant to the school career. This has been given a separate category, *else*, and mainly concerns pupils who go into the army, who go back to Morocco or Turkey, or who die. This category has been treated actively in the analysis thus far, and for each year it was placed near the origin of the two-dimensional space showing that there was no clear relation with having a more or less academic school career. Notice in Table 1 that the number of boys falling into this category increases steadily from 27 in the first year to 261 in the last year (out of 688 pupils).
(2) The information is not yet known. In Table 1 this is denoted as *missing*, and the table shows that this holds for many pupils from their fifth year. At this time they will have finished careers like IBO, LBO, and MAVO, and when they leave education they are much harder to track down. This

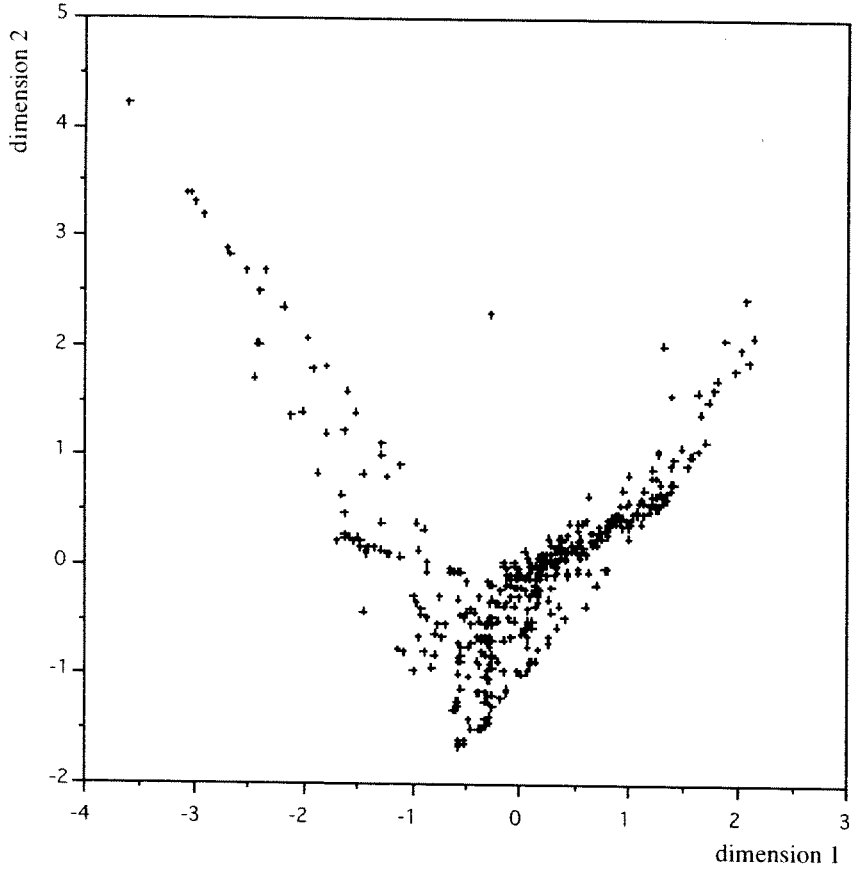

FIGURE 6. *Plot of 688 careers, using the first two dimensions of the ANACOR solution using 38 categories*

missing option is treated in the most popular way of treating missing information in multiple CA: in terms of the super indicator matrix. If a pupil is missing in a year, he or she receives zeros for all of the categories.

Earlier we pointed out that if it is assumed that missingness of part of the career is related to the nonmissing part of the career, then using a separate category for the missing part of the career is the most adequate way to deal with this problem. For this example missing data of the second type is probably related to having finished careers like IBO, LBO, and MAVO. The missingness is therefore not random, and using a separate category is the better choice to obtain an adequate description of the career data.

The multiple CA solution has as first four eigenvalues .5703, .5687, .5041, and .4514. This shows that the first and second dimension are not well identi-

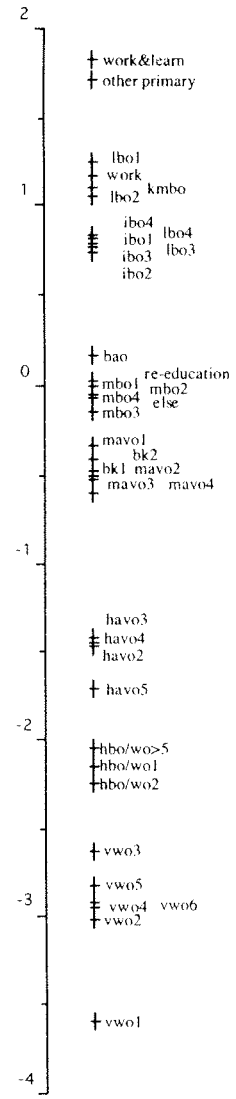

FIGURE 7. *Plot of school types, using the first dimension of the ANACOR solution using 38 categories*

fied. This means that, keeping in mind that eigenvalues can be interpreted as variances, a rotation of the cloud of points in the first two dimensions would lead to approximately the same variances.

Figure 8 shows a plot of the 688 careers, and Figure 9 shows a plot of the school levels in two dimensions. Figure 8 shows that the school careers do not form a horseshoe, so that it is not possible to summarize the careers by a single

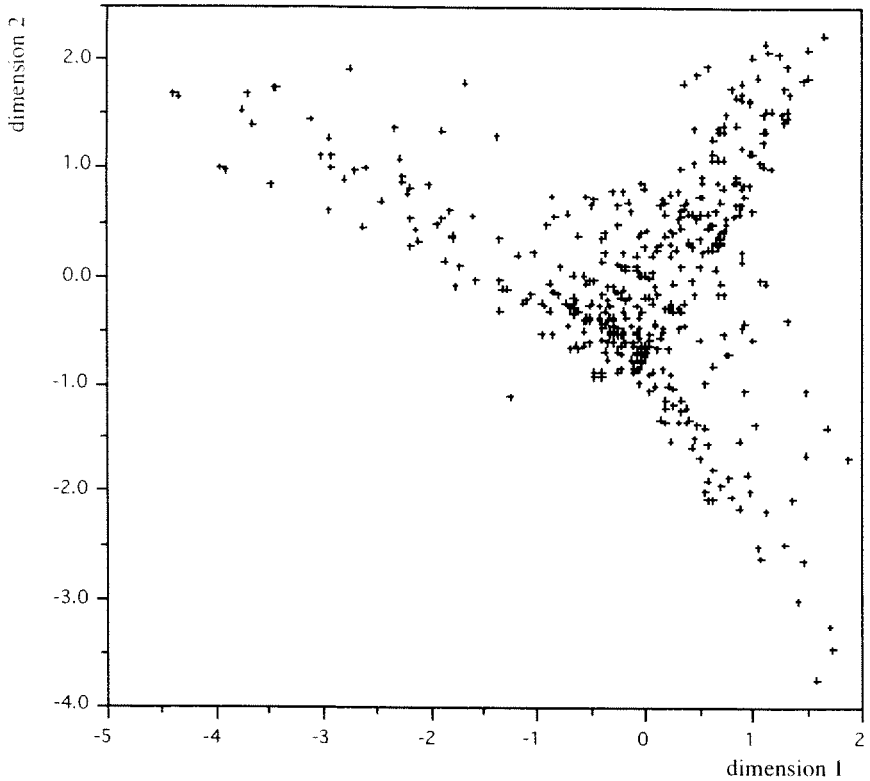

FIGURE 8. *Plot of 688 careers, using the first two dimensions of the multiple CA solution using 39 categories*

quantification. In Figure 9 we have not labeled all school levels separately, because we only want to show what happens when we include a second missing category and use this solution to show how the school careers can be related to background information. From top left to bottom right the categories are ordered from less to more academic, in much the same way as the order found in Figure 7. A special feature of Figure 8 is the cloud of points at top right from the origin. Figure 9 shows that these are school careers that use the missing category *else*, together with IBO and BK. Thus *else* happens more often than average with these school types. On the other hand, this category *else* is on the opposite side of the origin from the other missing category, labeled here as *missing*. This category is associated more often than average with MBO. It is not very surprising that *else* and *missing* are placed on opposite sides of the origin, since the marginal frequencies suggest that (a) once *else* enters a school career, this will remain the category for the remaining time points, and (b) once *missing* is in a school career, this will also remain the category for the remaining time points. In this sense both *else* and *missing* are ending states: Once you are in,

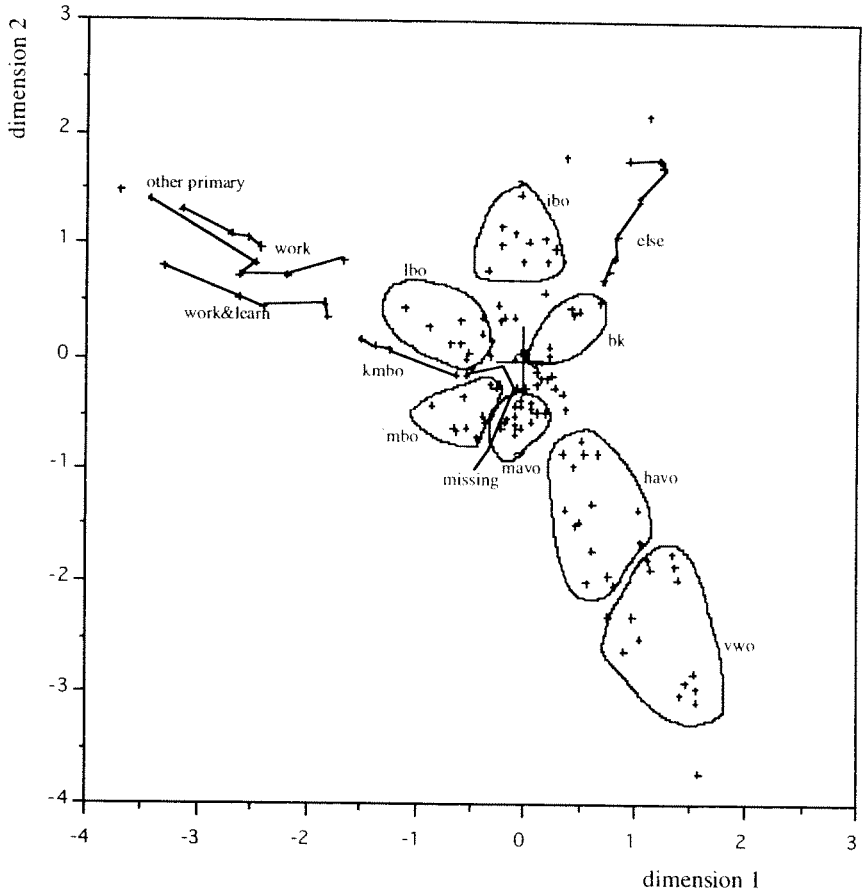

FIGURE 9. *Plot of school types, using the first two dimensions of the multiple CA solution using 39 categories*

you won't get out. At the same time, careers with both *else* and *missing* hardly exist, and therefore these careers have to be placed far away from each other.

One of the main objectives of multiple CA of school careers is to summarize the school careers by measures that can be used easily in further analyses. For the multiple CA without the category *missing* it was easy to relate the quantification obtained for the first dimension to background variables by calculating correlations or means. For the careers in Figure 8 this is not so easy. Two scores are needed to summarize the careers in this figure, and then it is not evident how to choose these two scores, since the cloud of points is barely identified. In this case we think that it is much easier to classify the careers into a number of groups. For this purpose, formal methods can be used such as cluster analysis on the coordinates in Figure 9, but it could also be done by hand. In the latter case it

makes sense to construct at least three groups, one for the careers on the left, one for bottom right, one for top right, and one for the center. In a second step this classification can then be related to other background variables by making contingency tables using the classification as one of the variables, or by calculating means for each group in the classification. Space limitations prevent us here from illustrating this in the case of our example.

Conclusion

School and other career data can be coded in such a form that (multiple) CA can be meaningfully applied, yielding a quantification for each career or a classification for each career. A quantification seems to be the most useful summary if it is possible to extract only one interpretable dimension from (multiple) CA, or if the interpretation for the first dimension can be separated from the interpretation for a second dimension. On the other hand, a classification seems to be most useful when it is difficult to attach distinct interpretations to separate sets of quantifications. Classifications then make it possible to summarize careers into a categorical variable.

CA can be applied meaningfully even when the number of states at each of the time points is large (in our example it was 38 and 39). This is not a special aspect of these data. Other illustrations are found in van der Heijden and de Leeuw (1989), where an example with 25 states is used. Of course, when the number of categories increases, the stability of the solution can decrease in the sense that small perturbation of the data results in relatively large changes in the solution. One way to investigate the stability of the solution is to apply the bootstrap (see Gifi, 1990, for examples). For the performance of the bootstrap in this context we refer to Markus (1994). The stability of the solution can be increased by decreasing the number of periods used in the data matrix.

CA provides a visual ordering of alternatives for secondary education with respect to academic difficulty. Such an ordering can be used in further analyses as quantifications or classifications of school careers. Alternative analyses are possible and useful. For example, it will probably also be useful to study transition matrices derived from two subsequent years. CA is based on these transition matrices, but the detail that such matrices provide is not given by the CA solution. On the other hand, an inspection of transition matrices alone does not provide a quantification of school careers, and this is the virtue of CA.

APPENDIX

SPSS code

The computations are performed with the SPSS-PC module Categories. The system file "file1.sys" consists of the original variables Y56 (year 1985–1986) to Y34 (year 1993–1994), each having categories numbered up to 38.

Multiple CA is performed with the SPSS-PC module HOMALS. Four dimensions are calculated, and the so-called object scores (i.e., quantifications of careers) are saved by names bim00001 to bim00004. These quantifications can be used in further analyses.


```
GET /FILE 'a:\file1.sys'.
HOMALS
/VARIABLES Y56 Y67 Y78 Y89 Y90 Y01 Y12 Y23 Y34 (38)
/DIMENSION 4 /MAXITER 1000 /CONVERGENCE .0000001
/PLOT NDIM (ALL, MAX) /SAVE bim (4).
```

The CA solution is performed with the SPSS-PC module ANACOR. Using count commands, 38 count variables are constructed. These are written to the data file a:anacor.dat. The input matrix is then read of order 688 × 38, and this matrix is analyzed directly.

```
count c1 =y56 to y34 (1).
 |
 |
count c38=y56 to y34 (38).
set results 'a:anacor.dat'/more off.
write /variables c1 to c38.
data list file 'a:anacor.dat' fixed/ c1 1-8 c2 10-17 c3 19-26 c4 28-
35 c5 37-44 c6 46-53 c7 55-62 c8 64-71/ c9 1-8 c10 10-17 c11 19-26
c12 28-35 c13 37-44 c14 46-53 c15 55-62 c16 64-71/ c17 1-8 c18 10-17
c19 19-26 c20 28-35 c21 37-44 c22 46-53 c23 55-62 c24 64-71/ c25 1-8
c26 10-17 c27 19-26 c28 28-35 c29 37-44 c30 46-53 c31 55-62 c32 64-
71/ c33 1-8 c34 10-17 c35 19-26 c36 28-35 c37 37-44 c38 46-53.
set /listing 'a:anacor.out'/more off/wkspace 500.
anacor /table all (688,38).
```

Using a text editor, the 688 rows with object scores are extracted from the output and saved as the file a:\anacor2.dat. These are then defined and linked to the original system file, which now also contains the variable bim00001, which consists of the career quantifications of the first dimension of multiple CA. Then a correlation is calculated between bim00001 and d1, the scores from the ANACOR solution.

```
DATA LIST FILE 'a:\anacor2.dat' FIXED
/ person 1-7 m 15-21 d1 23-30 d2 32-40.
DE.
join match /file * /file 'a:\file1.sys'.
set /listing 'h:\mtw21\rafaele\corr2.out'.
correlations /variables bim00001 d1.
```

References

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology, 96,* 144–185.

Benzécri, J. P. (Ed.). (1973). *L'analyse des données* (Vols. 1–2). Paris: Dunod.

Blossfeld, H. P., Hamerle, A., & Mayer, K. U. (1989). *Event history analysis: Statistical theory and applications in the social sciences.* Hillsdale, NJ: Erlbaum.

de Leeuw, J. (1984). *Canonical analysis of categorical data.* Leiden, The Netherlands: D.S.W.O.-Press.

de Leeuw, J., van der Heijden, P. G. M., & Kreft, I. (1985). Homogeneity analysis of event history data. *Methods of Operations Research, 50,* 299–316.

de Leeuw, J., & van Rijckevorsel, J. L. A. (1980). HOMALS and PRINCALS: Some generalizations of principal components analysis. In E. Diday (Ed.), *Data analysis and informatics* (pp. 231–242). Amsterdam: North-Holland.

Deville, J.-C. (1982). Analyse de données chronologiques qualitatives: Comment analyser des calendrier? *Annales de l'INSEE, 45*, 45–104.

Deville, J.-C., & Saporta, G. (1980). Analyse harmonique qualitative. In E. Diday (Ed.), *Data analysis and informatics* (pp. 375–389). Amsterdam: North-Holland.

Deville, J.-C., & Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series. *Journal of Econometrics, 22*, 169–189.

Gifi, A. (1990). *Non-linear multivariate analysis*. New York: Academic Press.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. New York: Academic Press.

Greenacre, M. J. (1987). The Carroll-Green-Schaffer scaling in correspondence analysis: A theoretical and empirical appraisal. *Journal of Marketing, 26*, 358–365.

Israels, A. (1987). *Eigenvalue techniques for qualitative data*. Leiden, The Netherlands: D.S.W.O.-Press.

Markus, M. T. (1994). *Bootstrap confidence regions in nonlinear multivariate analysis*. Leiden: D.S.W.O.-Press.

Martens, B. (1994). Analyzing event history data by cluster analysis and multiple correspondence analysis: An example using data about work and occupations of scientists and engineers. In M. Greenacre & J. Blasius (Eds.), *Correspondence analysis in the social sciences* (pp. 233–251). London: Academic Press.

Meulman, J. (1982). *Homogeneity analysis of incomplete data*. Leiden, The Netherlands: D.S.W.O.- Press.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, Ontario, Canada: University of Toronto Press.

Saporta, G. (1981). *Méthodes exploratoires d'analyse de données temorelles*. Unpublished doctoral dissertation, l'Universite P. et M. Curie, Paris.

Saporta, G. (1985). Data analysis for numerical and categorical individual time-series. *Applied Stochastic Models and Data Analysis, 1*, 109–119.

Schriever, B. F. (1983). Scaling of order dependent categorical variables with correspondence analysis. *International Statistical Review, 51*, 225–238.

Schriever, B. F. (1986). *Order dependence* (MC Tract 20). Amsterdam: Mathematics Centrum.

Taris, T. W. (1994). *Analysis of career data from a life- course perspective*. Unpublished doctoral dissertation, Free University, Amsterdam.

van Buuren, S., & de Leeuw, J. (1992). Equality constraints in multiple correspondence analysis. *Multivariate Behavioral Research, 27*, 567–583.

van Buuren, S., & van Rijckevorsel, J. L. A. (1992). Equality constraints in multiple correspondence analysis. *Multivariate Behavioral Research, 27*, 567–583.

van der Heijden, P. G. M. (1987). *Correspondence analysis of longitudinal categorical data*. Leiden, The Netherlands: D.S.W.O.-Press.

van der Heijden, P. G. M., & de Leeuw, J. (1989). Correspondence analysis, with special attention to the analysis of panel data and event history data. In C. C. Clogg (Ed.), *Sociological methodology 1989*. Oxford, England: Basil Blackwell.

van der Heijden, P. G. M., & Escofier, B. (1988). *Multiple correspondence analysis with missing data* (Publication Interne No. 423). Rennes, France: I.R.I.S.A.

van der Heijden, P. G. M., & van den Brakel, J. (1993). Three data reduction methods for the analysis of time budgets. In Istat, *Time use methodology: Toward consensus* (pp. 151–160). Rome: Istat.

van Rijckevorsel, J. L. A. (1986). About horseshoes in multiple correspondence analysis. In W. Gaul & M. Schrader (Eds.), *Classification as a tool of research* (pp. 377–388). Amsterdam: North-Holland.

Yamaguchi, K. (1991). *Event history analysis*. Newbury Park, CA: Sage.

Authors

PETER G. M. VAN DER HEIJDEN is Professor, Department of Methodology and Statistics, Utrecht University, Postbus 80.140, 3508 TC Utrecht, The Netherlands; p.vanderheijden@fsw.ruu.nl. He specializes in psychology and psychometrics.

JOOP TEUNISSEN is Assistant Professor, Department of Methodology and Statistics, Utrecht University, Postbus 80.140, 3508 TC Utrecht, The Netherlands; j.teunissen@fsw.ruu.nl. He specializes in field research, qualitative research, and education and (ethnic) minorities.

CHARLES VAN ORLÉ is Researcher at the Department of Educational Priority Areas, Rotterdam; cvorle@worldaccess.nl. He specializes in school evaluation and education and (ethnic) minorities.

Received January 3, 1996
Revision received September 9, 1996
Accepted September 9, 1996