Applied Latent Class Analysis

Edited by

JACQUES A. HAGENAARS

Tilburg University

ALLAN L. McCUTCHEON

Gallup Research Center
University of Nebraska, Lincoln

CAMBRIDGE
UNIVERSITY PRESS

equation models for response-based segmentation and unobserved heterogeneity." *Marketing Science*, **16**, 39–59.
Jorgensen, M., & Hunt, L. (1996). "Mixture model clustering of data sets with categorical and continuous variables." In *Proceedings of the Conference ISIS '96, Australia, 1996*, pp. 375–84.
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.
Krzanowski, W. J. (1983). "Distance between populations using mixed continuous and categorical variables." *Biometrika*, **70**, 235–43.
Lawrence C. J., & Krzanowski, W. J. (1996). "Mixture separation for mixed-mode data." *Statistics and Computing*, **6**, 85–92.
Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mill.
McLachlan, G. J., & Basford, K. E. (1988). *Mixture Models: Inference and Application to Clustering*. New York: Marcel Dekker.
McLachlan, G. J., & Peel, D. (1996). "An algorithm for unsupervised learning via normal mixture models." In D. L. Dowe, K. B. Korb, & J. J. Oliver (eds.), *Information, Statistics and Induction in Science*. Singapore: World Scientific, pp. XXX–XX.
McLachlan, G. J., & Peel, D. (1999). *Modelling Nonlinearity by Mixtures of Factor Analysers via Extension of the EM Algorithm*. Technical Report, Australia, Center for Statistics, University of Queensland.
McLachlan, G. J., Peel, D., Basford, K. E., & Adams, P. (1999). "The EMMIX software for the fitting of mixtures of normal and *t*-components," *Journal of Statistical Software*, **4**, No. 2.
Reaven, G. M., & Miller, R. G. (1979). "An attempt to define the nature of chemical diabetes using multidimensional analysis," *Diabetologia*, **16**, 17–24.
Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD's User's Guide*. Boston: Statistical Innovations, Inc.
Wallace, C. S., & Dowe, D. L. (Forthcoming). "MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions," *Statistics and Computing*, **X**, X–X.
Wedel, M., DeSarbo, W. S., Bult, J. R., & Ramaswamy, V. (1993). "A latent class Poisson regression model for heterogeneous count data with an application to direct mail." *Journal of Applied Econometrics*, **8**, 397–411.
Wolfe, J. H. (1970). "Pattern clustering by mulltivariate cluster analysis," *Multivariate Behavioral Research*, **5**, 329–50.
Yung, Y. F. (1997). "Finite mixtures in confirmatory factor-analysis models," *Psychometrika*, **62**, 297–330.

FOUR

Some Examples of Latent Budget Analysis and Its Extensions

Peter G. M. van der Heijden, L. Andries van der Ark, and Ab Mooijaart

1. INTRODUCTION

Latent budget analysis is a tool for the analysis of two-way contingency tables. The idea was initiated by Goodman (1974). Clogg (1981) extended this idea to an asymmetrical latent class model for the analysis of social mobility tables. Clogg used the following example: Let profession of the father be variable A, with categories indexed by i $(i = 1, \ldots, I)$; let profession of the son be variable B, with categories indexed by j $(j = 1, \ldots, J)$; let the latent social class variable be X, with categories indexed by t $(t = 1, \ldots, T)$. Let π_{ij} be the joint probability of profession i of the son and profession j of the father. Let π_t^X be the probability that a son belongs to the tth latent social class; $\pi_{it}^{\bar{A}X}$ the conditional probability that a son has a father with profession i given that he belongs to latent social class t; and $\pi_{jt}^{\bar{B}X}$ the conditional probability that a son has profession j given that he belongs to latent social class t.

The latent class model with T latent classes for a two-way table with probabilities p_{ij} is

$$\pi_{ij} = \sum_{t=1}^{T} \pi_t^X \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X}, \tag{1}$$

with all parameters nonnegative and restricted by

$$\sum_{t=1}^{T} \pi_t^X = 1, \quad \sum_{l=1}^{I} \pi_{it}^{\bar{A}X} = 1, \quad \sum_{i=1}^{J} \pi_{jt}^{\bar{B}X} = 1.$$

In this example, the explanatory variable is profession of the father and the response variable is profession of the son. Clogg assumed that there was a mediating (latent) variable, which he interpreted as social class.

He assumed that this latent variable was categorical. By rescaling the parameters $\pi_{it}^{\bar{A}X}$ into parameters $\pi_{it}^{A\bar{X}}$ by

$$\pi_{it}^{A\bar{X}} = \pi_t^X \pi_{it}^{\bar{A}X} \Big/ \sum_{t=1}^{T} \pi_t^X \pi_{it}^{\bar{A}X},$$

Goodman (1974) and Clogg (1981) noticed that it is possible to rewrite Equation (1) into

$$\frac{\pi_{ij}}{\pi_{i+}} = \sum_{t=1}^{T} \pi_{it}^{A\bar{X}} \pi_{jt}^{\bar{B}X}. \tag{2}$$

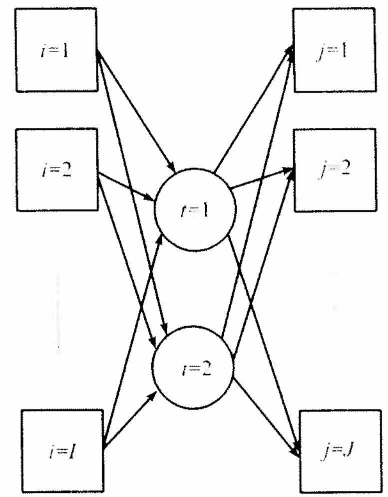
with parameter restrictions

$$\sum_{t=1}^{T} \pi_{it}^{A\bar{X}} = 1, \qquad \sum_{j=1}^{J} \pi_{jt}^{\bar{B}X} = 1. \tag{3}$$

Compared with Model 1, in Model 2 the probabilities that are decomposed are conditional probabilities rather than joint probabilities. That is, the conditional probability π_{ij}/π_{i+} is the probability that the son has profession j given that the father has profession i. The parameters are interpreted as follows: The parameters π_{it}^{AX} are the probabilities that a father with profession i belongs to the tth latent social class, and π_{jt}^{BX} are the probabilities that the son has profession j given that he belongs to the tth social latent class. It may be noted that the parameters π_{jt}^{BX} have the same interpretation in Model 1 and Model 2.

Model 2 is illustrated graphically in Figure 1. In the social sciences, the representation in this figure is known as a MIMIC model (i.e., the Multiple Indicator Multiple Cause model; Goodman, 1974). It may be noted that the squares in Figure 1 represent the levels of the professions, whereas the T circles represent the levels of the latent variable. (This should not be confused with representations of structural equations models often used in the social sciences, where both circles and squares always represent variables, and not levels of variables.)

Independently, de Leeuw and van der Heijden (1988) reinvented Model 2 in the context of an analysis of time budgets. A time budget of an individual i is the distribution of time over J mutually exclusive activities. Hence, the J elements add up to 1 and they are nonnegative, just like the conditional probabilities (π_{ij}/π_{i+}) in Model 2. The word *budget* emphasizes that if time is spent on one activity, it cannot be spent on another activity at the same time. Therefore, they termed Model 2 the *latent budget model* (LBM). The T vectors of parameters $(\pi_{1t}^{BX}, \ldots, \pi_{Jt}^{BX})$ are



MIMIC model

Figure 1. Graphic representation of a MIMIC model.

called *latent budgets*. Similarly, the I vectors of conditional probabilities $(\pi_{i1}/\pi_{i+}, \ldots, \pi_{iJ}/\pi_{i+})$ are called *expected budgets*.

In 1988, the authors were unaware of the fact that the idea of the LBM had been introduced much earlier by Goodman (1974). Van der Heijden, Mooijaart, and de Leeuw (1992) pointed out the equivalence between the LBM and Goodman's and Clogg's work. However, they emphasized a mixture-model interpretation of the LBM. The expected budgets are mixtures of T latent budgets. The mixture interpretation is illustrated graphically in Figure 2. In Figure 2 only the expected budgets i and the latent budgets t are shown. The figure shows that an expected budget i is a mixture of the T latent budgets. The T *mixing parameters* for row i are provided by the parameters $\pi_{it}^{A\bar{X}}$. These mixing parameters show for which proportion the expected budgets are built up from the latent budgets. The mixing parameters are not revealed by Figure 2.

The LBM with T latent budgets has $(I-T)(J-T)$ degrees of freedom. For $T=1$, the LBM is equivalent to the independence model because then $p_{ij}/p_{i+} = \pi_{jt}^{\bar{B}X} = p_{+j}$. For $T = \min(I, J)$, the LBM is saturated, and estimates of expected proportions are equal to observed proportions.

The LBM is usually estimated by the method of maximum likelihood under the assumption that the frequencies are generated by a product-multinomial distribution (although we have also been working on other estimation methods; see Mooijaart, van der Heijden, and van der Ark,

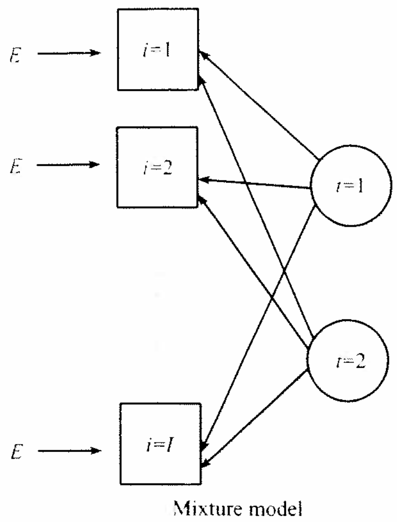


Figure 2. Graphic representation of a mixture model.

1999; van der Ark, 1999). Likelihood ratio tests are used to assess the fit of the LBM against the data and to determine the number of latent budgets (i.e., T) needed to describe the data adequately.

Clogg (1981) noted that Model 2 is not identified. De Leeuw, van der Heijden, and Verboon (1990) also discussed the identification problem of the LBM, and they worked out the situation for $T = 2$ in some detail. Writing Model 2 in matrix notation shows the identification problem: Collect the conditional proportions π_{ij}/π_{i+} in a matrix $\mathbf{\Pi}$, the mixing parameters $\pi_{it}^{A\bar{X}}$ in a matrix $\mathbf{A}$, and the latent budget parameters $\pi_{jt}^{\bar{B}X}$ in a matrix $\mathbf{B}$; then Model 2 equals

$$\mathbf{\Pi} = \mathbf{AB}'. \tag{4}$$

It is always possible to rewrite Equation 4 into $\mathbf{\Pi} = (\mathbf{AS}^{-1})(\mathbf{SB}') = \mathbf{A}^*\mathbf{A}^{*\prime}$, where $\mathbf{S}$ is a $K \times K$ matrix with each row adding up to 1. The parameters $\mathbf{A}$ and $\mathbf{B}$ yield the same expected budgets as $\mathbf{A}^*$ and $\mathbf{B}^*$. Because the elements of each row of $\mathbf{S}$ add up to 1, the parameters $\mathbf{A}^*$ and $\mathbf{B}^*$ are also subject to the equality restrictions in Equation (3). Furthermore, $\mathbf{S}$ can be chosen freely as long as all elements of $\mathbf{A}^*$ and $\mathbf{B}^*$ are nonnegative. De Leeuw et al. (1990) choose $\mathbf{S}$ such that as many parameters as possible from either $\mathbf{A}^*$ or $\mathbf{B}^*$ are zero, because this facilitated the interpretation. Van der Ark, van der Heijden, and Sikkel (1999) extended this work for $T > 2$. Their view of the identification problem for the LBM is similar to the identification problem in factor analysis, in which unidentified solutions are usually rotated to simplify interpretation. The common factor model is called *identified* because the varimax-rotated solutions are always

unique in practical situations. Similarly, Van der Ark et al. (1999) called the LBM identified for some specific choices of $\mathbf{S}$. They proposed an *inner extreme* solution, that is, choosing $\mathbf{S}$ such that the mixing parameters are as distinct as possible, which facilitates the interpretation in terms of the explanatory variable (e.g., the example of Section 2), and *outer extreme* solution, that is, choosing $\mathbf{S}$ such that the latent budgets are as distinct as possible, which facilitates the interpretation in terms of the response variable (e.g., the example in Section 3).

Van der Heijden et al. (1992) discuss various ways in which the parameters of the LBM can be constrained. They distinguish fixed-value constraints (e.g., some parameters are fixed to some constant), equality constraints (see, for some estimation problems, Mooijaart and van der Heijden, 1992), and situations in which the parameters $\pi_{it}^{A\bar{X}}$ and $\pi_{jt}^{\bar{B}X}$ are functions of external information. Sometimes these constraints can also be used as well to identify the LBM (e.g., the example in Section 4). A later development was to study how latent budget analyses of different groups could be compared; this was termed *simultaneous latent budget analysis* (see Siciliano and van der Heijden, 1994).

The LBM is closely related to correspondence analysis, and de Leeuw and van der Heijden (1991) describe under what circumstances the LBM is equivalent to correspondence analysis (see van der Ark and van der Heijden, 1997; van der Ark et al., 1999; van der Heijden, Gilula, and van der Ark, 1999). Latent budget analysis is also used in geology, where it is known as *end-member analysis* (see Renner, 1993; Weltje, 1997; van der Heijden, 1994). Many other results, in particular concerning least-squares estimates, standard errors, and testing procedures, can be found in van der Ark (1999).

In this chapter, by discussing some examples, we demonstrate some of the possibilities of the LBM and its extensions. Section 2 shows an example of latent budget analysis of a two-way table dealing with sentence endings of the books of Plato. Section 3 illustrates the possibilities of the LBM for comparing contingency tables in the context of trades started by different ethnic groups; here, the city of Amsterdam is compared with the city of Rotterdam. Section 4 shows the possibilities of the LBM for studying how the school success of pupils is related to explanatory variables such as IQ, sex, and the profession of the father.

2. THE WORKS OF PLATO

We start with a straightforward application of the LBM. The Greek philosopher Plato wrote forty-five books. The exact order in which these works were written is known approximately, except for the books *Critias*, *Philebus*, *Politicus*, *Sophist*, and *Timaeus*. The objective of this example is to show that the LBM can be used for seriation, that is, to find the chronological order in which all 45 books were written. For this purpose, we used data obtained by Kaluscha (1904), who collected all "sentence endings" in the 45 books. Each of the last five syllables of a sentence ending is scored as being "short" or "long," so that each sentence of each book belongs to one of $2^5 = 32$ categories.

The idea underlying the determination of the chronological order of the books from the distributions of sentence endings is that the style and rhythm of the texts changed through time, and that sentence endings are considered highly relevant with regard to rhythm (Boneva, 1970). For each book, we had the frequencies of sentence endings, yielding a matrix of 45 books by 32 sentence endings. The data are in Table 1, where the chronological order of the 40 "known" books is preserved. The 45 books are considered to be 45 budgets, each containing 32 categories. The frequencies in these 32 categories express the writing style of the particular book.

The LBM takes typical styles of writing as latent budgets, and the different books are then approximated by a mixture of these typical styles. The mixture-model interpretation (see Figure 2) is most appropriate in this context. The data, or an aggregated version, were studied earlier by, for example, Cox and Brandwood (1959), Atkinson (1970), and Greenacre (1984).

Latent budget analysis considers the frequencies of sentence endings of each book as a sample from a multinomial distribution. The LBM with $T = 1$ latent budget (independence of works and sentence endings) has a likelihood ratio chi square of $L^2 = 3{,}678$ (the degrees of freedom, df, is 1,364). This model implies that the writing styles in all books are identical. The LBM with $T = 2$ latent budgets has a fit of $L^2 = 2{,}022$ (df is 1,290). This model implies that there are two typical writing styles. The two estimated latent budgets show what these typical writing styles are. For each book i, the two mixing parameters $\pi_{it}^{A\bar{X}}$ $(t = 1, 2)$ show how the budget of book i is built up from these two typical writing styles. This model described $(3{,}678 - 2{,}022)/3{,}678 = 0.45$ of the departure from independence. The LBM with $T = 3$ latent budgets assumes that there are three typical writing styles. The fit was $L^2 = 1{,}661$ (df is 1,218), and this explained 0.55 of the departure from independence. For the LBM with $T = 4$ latent budgets, the fit was $L^2 = 1{,}440$ (df is 1,148), and this explained 0.61 of the departure from independence. The model with $T = 2$ described a considerable part of the departure from independence.

Not much more information was extracted from the data by consideration of more latent budgets.

For seriation, the LBM with $T = 2$ latent budgets is most appropriate. This model yields a unidimensional chronological order of the books because each book has two mixing parameters, $\pi_{i1}^{A\bar{X}}$ and $\pi_{i2}^{A\bar{X}}$, where $\pi_{i1}^{A\bar{X}} + \pi_{i2}^{A\bar{X}} = 1.0$. Therefore, it suffices to interpret only the 45 estimates $\hat{\pi}_{i1}$ for studying the differences between the books. We give a graphical representation of these 45 mixing parameter estimates in Figure 3 because this simplifies the interpretation. A graphical interpretation of the LBM with two latent budgets is that the 45 books are on a line segment. The latent budgets are the endpoints of the line segment: If the writing style of book i matches the typical writing style of latent budget 1 exactly, then $\pi_{i1}^{A\bar{X}} = 1.0$, and $\pi_{i2}^{A\bar{X}} = 0.0$. Book i is plotted on the endpoint of the line segment that coincides with latent budget 1. If the writing style of book i' is built up for 0.5 from the first typical writing style and for 0.5 from the second typical writing style, then book i' is plotted in the middle of the line segment, exactly in between the two latent budgets. If one writing style is typical for the earlier years of Plato's writings, and the second writing style is typical for the later years, then the line segment represents the chronological order of the books.

Not all the individual books could be printed into Figure 3. Therefore, two shaded areas are given. In the shaded area on the left side (close to the older writing style budget) are all the works with known chronological order up until *Republic 10*, with the exception of *Laches* and *Cratylus*. In the shaded area on the right side (close to the newer writing style budget) are the works with known chronological order from *Laws 2* onward. Figure 2 shows that there are clearly two distinct

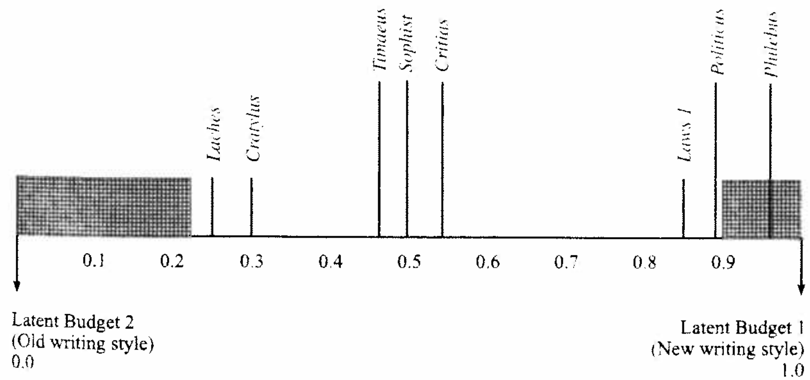


Figure 3 Graphic representation of mixing-parameter estimates for Plato data

Sentence Endings

Books	..... –....	....– –...–	...–. –..–.	...–– –..––	..–.. –.–..	..–.– –.–.–	..––. –.––.	..––– –.–––	.–... ––...	.–..– ––..–	.–.–. ––.–.	.–.–– ––.––	.––.. –––..	.––.– –––.–	.–––. ––––.	.–––– –––––
Charmides	5	6	11	11	8	20	8	15	6	17	17	9	14	19	6	20
	4	10	10	14	9	31	16	20	4	17	8	21	9	23	8	19
Laches	4	5	8	11	1	8	6	13	4	16	6	9	14	19	6	24
	7	16	12	13	8	22	7	18	8	9	22	12	10	30	8	20
Lysis	3	14	17	17	11	15	4	6	10	15	17	10	15	19	3	10
	7	6	5	16	13	19	9	5	11	11	16	8	13	14	4	4
Euthyphro	10	3	14	9	5	10	4	11	3	10	6	10	7	9	3	9
	7	10	5	12	8	9	5	12	2	9	5	10	5	11	3	9
Gorgias	16	27	22	44	19	39	27	45	24	38	29	40	36	73	24	42
	18	42	30	52	21	61	34	58	22	59	32	43	39	46	37	46
Hippas Minor	2	6	4	15	8	9	1	6	2	6	10	7	7	8	3	11
	5	5	4	11	3	11	5	7	1	2	8	5	2	9	6	9
Euthydemus	8	9	13	24	11	37	21	27	10	21	35	25	12	26	18	28
	11	12	20	20	17	51	12	21	9	32	20	26	16	36	16	28
Cratylus	29	26	23	42	17	25	13	32	17	29	28	34	38	47	9	39
	27	22	18	36	21	35	17	25	25	22	52	31	27	34	21	38
Meno	5	5	11	11	9	14	8	27	8	23	18	23	21	22	12	24
	4	25	13	24	6	26	16	17	5	19	13	15	14	18	11	20
Menexenus	1	3	3	0	1	9	6	6	6	17	5	6	5	8	3	11
	1	6	4	6	6	4	4	6	3	8	4	6	6	14	9	7
Phaedrus	9	9	10	23	18	28	10	21	17	33	23	40	18	26	9	26
	13	22	25	42	29	29	18	19	14	39	17	37	20	39	11	19
Symposium	10	11	27	26	16	33	13	22	11	28	16	39	19	35	22	27
	11	20	25	29	10	48	26	24	12	23	26	25	18	45	13	32
Phaedo	17	15	21	30	10	32	23	30	15	38	21	34	15	32	15	30
	16	41	10	43	19	44	18	34	16	28	28	33	26	47	11	32
Theaetetus	19	20	21	35	20	39	29	40	18	25	32	35	39	37	17	37
	23	28	38	43	23	36	37	45	16	39	34	37	42	55	18	49
Parmenides	22	15	19	31	9	23	7	15	17	26	21	30	15	16	9	26
	18	12	25	28	13	24	13	18	14	24	19	26	13	21	7	25
Protagoras	9	13	14	8	10	11	15	21	12	15	14	19	20	16	13	24
	11	9	21	17	11	31	17	25	8	25	18	22	12	28	13	20
Crito	2	2	4	6	1	5	4	6	1	4	6	9	5	10	3	8
	0	3	5	6	3	9	6	5	2	6	3	7	3	13	1	10
Apology	2	10	11	13	2	13	11	12	8	14	11	13	7	13	17	11
	3	10	14	17	5	22	16	15	6	16	16	11	6	14	9	15
Republic 1	12	11	10	9	10	15	13	22	10	27	15	20	28	23	10	25
	15	17	13	26	7	30	7	14	11	11	11	18	11	18	13	29
Republic 2	3	6	9	11	7	13	11	12	8	23	9	18	12	13	5	19
	9	12	6	10	7	20	5	12	16	12	13	15	15	13	10	15
Republic 3	2	2	4	12	6	19	14	13	6	10	7	13	15	19	19	15
	2	13	9	12	9	25	19	13	7	11	10	18	21	22	4	22
Republic 4	3	13	9	10	4	16	7	11	4	24	7	12	10	18	11	11
	3	17	10	18	8	22	8	11	10	11	6	18	14	6	6	17
Republic 5	5	13	9	14	13	19	10	15	13	22	9	18	19	18	10	15
	10	19	17	23	9	27	7	16	9	19	15	18	14	22	11	18
Republic 6	5	8	9	11	10	13	19	7	5	16	12	17	7	21	5	6
	7	19	13	16	10	22	6	18	4	16	13	17	8	10	6	17

(*continued*)

Table 1 ***(continued)***

Books	Sentence Endings															
	⏑⏑⏑⏑⏑	⏑⏑⏑⏑–	⏑⏑⏑–⏑	⏑⏑⏑––	⏑⏑–⏑⏑	⏑⏑–⏑–	⏑⏑––⏑	⏑⏑–––	⏑–⏑⏑⏑	⏑–⏑⏑–	⏑–⏑–⏑	⏑–⏑––	⏑––⏑⏑	⏑––⏑–	⏑–––⏑	⏑––––
	–⏑⏑⏑⏑	–⏑⏑⏑–	–⏑⏑–⏑	–⏑⏑––	–⏑–⏑⏑	–⏑–⏑–	–⏑––⏑	–⏑–––	––⏑⏑⏑	––⏑⏑–	––⏑–⏑	––⏑––	–––⏑⏑	–––⏑–	––––⏑	–––––
Republic 7	3	8	6	8	6	16	3	8	6	15	12	15	8	15	3	5
	3	24	5	14	10	29	7	16	9	18	6	12	6	21	11	6
Republic 8	2	5	9	9	5	23	5	6	5	11	9	18	5	15	7	8
	1	17	12	15	6	21	5	25	3	7	11	12	8	9	7	10
Republic 9	1	4	6	3	1	19	7	8	2	16	6	13	13	15	11	12
	5	17	12	16	8	21	5	8	3	11	16	9	6	20	4	9
Republic 10	5	4	8	6	10	14	11	5	6	9	12	15	9	23	8	15
	5	19	8	9	7	23	8	19	7	11	13	19	6	15	5	15
Laws 1	6	13	13	21	5	10	6	15	11	3	4	4	14	24	5	15
	5	25	3	3	5	8	13	13	16	3	8	12	13	27	8	10
Laws 2	6	11	10	20	7	3	3	5	4	4	3	1	10	17	1	19
	7	17	0	3	6	6	8	6	5	2	6	10	11	26	8	21
Laws 3	8	19	10	11	8	11	7	8	6	7	3	3	13	22	7	14
	16	37	2	3	4	5	14	8	6	1	5	5	15	22	7	12
Laws 4	7	11	12	17	10	9	5	3	4	2	3	5	7	22	5	8
	13	20	3	2	9	6	1	15	9	4	5	2	15	17	17	13
Laws 5	8	14	9	15	11	8	1	5	6	3	2	0	5	18	5	3
	11	19	1	1	5	9	3	10	9	1	3	6	13	15	13	11
Laws 6	9	20	12	32	6	13	9	14	4	1	1	3	15	29	5	15
	22	30	1	2	4	5	7	16	11	8	3	5	12	22	9	22
Laws 7	13	17	13	24	10	11	7	20	9	1	7	6	12	46	9	17
	13	43	3	3	12	13	6	13	11	1	5	11	21	32	12	29
Laws 8	4	10	10	10	8	8	3	11	3	1	2	3	3	21	6	17
	8	23	1	1	5	4	3	14	3	3	6	6	11	28	4	12
Laws 9	10	7	6	19	7	11	3	7	4	5	3	6	15	36	9	15
	17	31	3	1	7	9	4	14	7	2	2	8	8	28	16	13
Laws 10	4	9	9	21	14	9	4	8	8	6	3	2	11	28	3	9
	13	40	2	3	4	11	9	9	8	3	8	7	19	46	8	20
Laws 11	8	6	3	13	6	8	5	5	4	3	1	5	13	17	13	13
	7	20	1	0	2	7	9	9	9	7	2	4	7	40	5	15
Laws 12	7	8	8	12	8	12	5	9	8	5	4	6	12	29	5	11
	12	28	2	1	4	7	9	12	8	3	0	4	12	29	7	17
Critias	5	6	10	10	2	5	4	2	3	3	4	4	4	8	5	3
	3	9	2	7	3	2	3	7	5	1	3	9	4	3	5	6
Philebus	24	46	38	52	25	18	30	25	20	7	6	12	64	51	32	32
	27	62	7	7	14	27	11	32	41	4	7	23	53	86	28	47
Politicus	13	22	25	33	20	23	24	35	24	8	7	8	41	34	19	29
	19	31	3	6	18	14	26	38	25	5	21	16	53	52	22	56
Sophist	26	23	22	47	24	28	28	31	31	28	15	12	30	42	23	27
	33	21	19	28	37	19	34	32	30	21	24	38	48	43	24	31
Timaeus	18	27	26	30	14	17	23	23	46	20	17	25	26	23	17	18
	30	25	13	21	26	23	25	23	26	25	25	49	23	29	17	14

subgroups within the books with known chronological order: the earlier works (up until *Republic*), and the later works (*Laws*). Within these two subgroups, however, the chronological order was not clearly shown by the LBM. From the five undated books, *Philebus* and *Politicus* are mostly built up from the newer writing style budget. From their sentence endings, these books appear to be similar to the later works. The remaining books, *Critias*, *Sophist*, and *Timeaus*, do not belong to the later works or to the earlier works. Their writing style is a mixture of the older writing style and the newer writing style. This may suggest that these books were written in between.

In this example, we did not interpret the latent budgets because we lack the knowledge of sentence endings.

3. ETHNIC DIFFERENCES AMONG PEOPLE STARTING A TRADE: AN EXAMPLE OF SIMULTANEOUS LATENT BUDGET ANALYSIS

In The Netherlands, trades are registered with Chambers of Commerce. Kloosterman and van der Leun (1998), who investigated the way ethnic groups differ in the types of trades they start, concentrated on the so-called sheltered sector and on two large cities in The Netherlands, namely Rotterdam and Amsterdam. The data are presented in Table 2a.

Table 2a. Trades Started in Amsterdam and Rotterdam: Cross-Classification by Ethnic Group and Type of Trade

	Amsterdam						Rotterdam					
Group	**1**	**2**	**3**	**4**	**5**	**Total**	**1**	**2**	**3**	**4**	**5**	**Total**
Dutch	382	367	788	113	28	1933	323	209	459	91	153	1235
Turks	14	21	3	8	10	56	29	30	2	15	14	90
Moroccans	12	36	2	5	7	62	8	17	2	13	5	45
Antilleans	8	6	2	1	2	19	5	4	3	4	3	19
Surinamese	44	33	33	17	24	151	35	31	28	19	33	146
Cape Verd.	0	0	0	0	0	0	5	1	0	0	3	9
Ghanaians	23	4	4	2	4	37	3	1	0	0	1	5
Other	185	93	82	24	35	419	74	16	19	16	8	133
No info.	146	116	119	39	61	481	42	23	31	7	7	110
Total	814	676	1033	209	426	3158	524	332	544	165	227	1792
Proportions	0.257	0.214	0.327	0.066	0.135	1.000	0.292	0.185	0.304	0.092	0.127	1.000

Note: Types of trade are 1 = wholesale trade; 2 = retail trade; 3 = producer services; 4 = catering and restaurants; 5 = personal services.

Surinam was a former Dutch colony, and the Antilles are still closely linked administratively to The Netherlands. By their educational system (language, history), this makes it easier for members of these groups to integrate into Dutch society. The Turks and Moroccans are large ethnic groups that originally came in the 1960s and 1970s as so-called guest workers. The Cape Verdeans and the Ghanaians are relatively small ethnic minorities. The trades speak for themselves. Amsterdam and Rotterdam differ in that the port of Rotterdam generates considerable employment, specifically in the wholesale trade and catering services (compare the marginal column proportions in Table 2a), whereas Amsterdam is both a tourist and an industrial center. The two cities thus provide the ethnic groups with a different opportunity structure. One could argue that the success of the different ethnic groups with respect to the opportunities offered depends on their network in specific trades, for example, the number of clients of the same ethnic group, and on their human capital, for instance, knowledge of the Dutch language or knowing how the trade as a whole operates in The Netherlands. These different types of human capital and networks ensure that some ethnic groups are more likely to start certain specific trades rather than others.

This is where the usefulness of latent budget analysis becomes apparent: As shown in Figure 1, the LBM assumes the existence of a categorical latent variable, with T states between ethnic group i and trade j, and these latent states could very well be reflecting human capital and the networks. In terms of Figure 2, the LBM approximates the distribution of each ethnic group (observed budget) by a mixture of a number of latent distributions (latent budgets). The latent budgets may be interpreted as typical extreme distributions that deviate from the marginal distribution of trades started in Rotterdam and Amsterdam. The way in which they deviate reveals how typical sources of human capital and networks create specific opportunities to start specific trades.

It should be noted that the absolute sizes of the ethnic groups are not reflected in the parameter estimates. For completeness, absolute sizes are provided for some of the groups in Table 2b. We concentrated here on the type of trade that people from ethnic groups choose when they start a trade, that is, the information provided in Table 2a. Another study would be to look at the relative proportions of ethnic groups that start trades at all and then to compare Amsterdam and Rotterdam. The relevant data are shown in Table 2b.

For Amsterdam, the LBM with $T = 1$ latent budget (i.e., the independence model) has $L^2 = 299$ (df is 28); for $T = 2$, $L^2 = 69$ (df is 18); for

Table 2b. Trades Started in Amsterdam and Rotterdam: Absolute Sizes of the Ethnic Groups

	Amsterdam			Rotterdam		
	Trades			Trades		
Group	Sample	Prop.	No. of Inhab.	Sample	Prop.	No. of Inhab.
Dutch	1,933	0.612	419,698	1,235	0.689	358,425
Turks	56	0.018	30,992	90	0.050	35,598
Moroccans	62	0.020	47,202	45	0.025	24,550
Antilleans	19	0.006	10,501	19	0.011	11,708
Surinamese	151	0.048	69,011	146	0.081	46,679
Cape Verd.	0	0.000	not spec.	9	0.005	not spec.
Ghanaians	37	0.012	not spec.	5	0.003	not spec.
Other	419	0.133	not spec.	133	0.074	not spec.
No info.	481	0.152	not spec.	110	0.061	not spec.

$T = 3$, $L^2 = 13$ (df is 10). For Rotterdam, for $T = 1$, $L^2 = 218$ (df is 32); for $T = 2$, $L^2 = 75$ (df is 21); for $T = 3$, $L^2 = 22$ (df is 12; $0.025 < p < .05$). The fit of the LBMs with $T = 3$ therefore seems adequate. In terms of Figure 1, the latent states represent three types of human capital and networks that lead to specific patterns of trade that are started. The fit indices should be interpreted with care because many observed frequencies equal zero. We studied the parameter estimates for the solutions with $T = 3$, given in Table 3. We have identified the solution by making the latent budgets as extreme as possible, that is, by making as many latent budget parameters (π_{it}^{BX}) equal to zero as possible (see van der Ark et al., 1999).

The latent budgets are most easily interpreted by comparing parameter estimates with the marginal proportions p_{+j}. This shows that for Amsterdam, the first latent budget is characterized by wholesale trade (i.e., estimate 0.933 is greater than the marginal proportion 0.257). In terms of human capital and networks, the first latent state represents *knowledge of the supply side*. The second latent budget is characterized by retail trade ($0.635 > 0.214$), catering industry ($0.175 > 0.066$), and personal services ($0.190 > 0.135$); this latent state represents *knowledge of the demand side of economy*. The third latent budget is characterized by producer services ($0.805 > 0.327$) and personal services ($0.184 > 0.135$); this latent state represents a *good education and access to relevant Dutch networks*.

We interpreted the mixing-parameter estimates $\hat{\pi}_{it}^{A\bar{X}}$ from graphical displays similar to Figure 3. Because $T = 3$, we now use *ternary diagrams*

Table 3. Parameter Estimates for LBMs with $T = 3$ for Amsterdam and Rotterdam

	Amsterdam				Rotterdam			
Mixing Parameters	$T=1$	$T=2$	$T=3$		$T=1$	$T=2$	$T=3$	
Dutch	0.212	0.282	0.506		0.329	0.144	0.527	
Turks	0.267	0.661	0.071		0.407	0.561	0.032	
Moroccans	0.207	0.755	0.038		0.240	0.701	0.058	
Antilleans	0.449	0.420	0.131		0.341	0.446	0.213	
Surinamese	0.313	0.399	0.288		0.292	0.428	0.280	
Cape Verd.					0.582	0.418	0.000	
Ghanaians	0.661	0.179	0.160		0.661	0.339	0.000	
Other	0.474	0.292	0.235		0.707	0.095	0.198	
No information	0.326	0.367	0.308		0.488	0.110	0.402	
Latent budgets	$T=1$	$T=2$	$T=3$	p_{+j}	$T=1$	$T=2$	$T=3$	p_{+j}
Wholesale trade	0.933	0.000	0.000	0.257	0.795	0.000	0.000	0.292
Retail trade	0.045	0.635	0.000	0.214	0.096	0.431	0.146	0.185
Producer serv.	0.000	0.000	0.805	0.327	0.000	0.000	0.705	0.304
Catering & rest.	0.022	0.175	0.011	0.066	0.108	0.259	0.000	0.092
Personal serv.	0.000	0.190	0.184	0.135	0.000	0.310	0.149	0.127

(see van der Ark and van der Heijden, 1997). Figure 4(a) gives the plot of the parameter estimates for the LBM with $T = 3$ latent budgets for the ethnic groups in Amsterdam. The vertices of the triangle represent the latent budgets. The upper vertex represents the first latent budget, the right-hand vertex represents the second latent budget, and the left-hand vertex represents the third latent budget. The side opposite a vertex represents the area where the corresponding mixing parameters ($\pi_{it}^{A\bar{X}}$) are zero. The expected budgets can be depicted in the diagram, and their mixing parameters determine the position in the diagram; that is, the position of an expected budget in the diagram is $\pi_{i1}^{A\bar{X}}$ times the distance from the bottom side to the upper vertex, $\pi_{i2}^{A\bar{X}}$ times the distance from the left-hand side to the right-hand vertex, and $\pi_{i3}^{A\bar{X}}$ times the distance from the right-hand side to the left-hand vertex.

Figure 4(a) reveals that, more than average, the Dutch currently start in the third latent budget (latent state for good education and access to Dutch networks), whereas Ghanaians, Antilleans, Turks, and Moroccans are ordered between the first latent budget (latent state for supply side) and the second latent budget (latent state for the demand side). The Surinamese are intermediate between the Dutch and the other ethnic groups. This might be explained by the fact that the Surinamese form an ethnic group that are reasonably well integrated in Dutch society.

In Rotterdam [Figure 4(b)], the graphical representation is very similar to that in Amsterdam. The first latent budget is characterized by wholesale trade and to some extent by catering; the second latent budget by retail trade, catering, and personal services; and the third latent budget by producer services and some personal services. Again, the Dutch start trades predominantly in the third latent budget; the Ghanaians, Cape Verdeans, Turks, and Moroccans are ordered between the first and second latent budgets; and the Surinamese and now also the Antilleans are intermediate between the Dutch and the other ethnic groups.

Although there are differences between the solutions of Amsterdam and Rotterdam, the similarities are striking. Therefore, we investigated whether a more parsimonious solution, obtained by imposing equality restrictions to the parameter estimates, could describe the data. This is done in simultaneous latent budget analysis (Siciliano and van der Heijden, 1994). Because the Cape Verdeans did not start any trades in Amsterdam, we deleted them from the table of Rotterdam, and we analyzed a table of 2 (cities) × 8 (ethnic groups) × 5 (trades).

In a first analysis, we imposed the latent budget parameters (π_{jt}^{BX}) to be equal for Rotterdam and Amsterdam. Thus, the latent budgets for Amsterdam and Rotterdam are equivalent, but the way in which ethnic groups make use of them may differ. In terms of Figure 1, this implies that the ethnic groups in Amsterdam have different sources of human capital and networks than the ethnic groups in Rotterdam, but the way in which this human capital leads to starting trades is the same in both cities. The LBM with $T = 3$ has a fit of $L^2 = 48.3$ (df is 26).

In a second analysis, we imposed equality of the mixing parameters ($\pi_{it}^{A\bar{X}}$) for both Rotterdam and Amsterdam. Thus, the latent budgets of Amsterdam and Rotterdam are different, but the way in which they are mixed by $\pi_{it}^{A\bar{X}}$ is identical. In terms of Figure 1, this means that the ethnic groups in both cities have the same human capital and networks, but this leads to different trades in Amsterdam than in Rotterdam. Because the opportunities of the two cities differ (compare their marginal proportions), the specific latent budget estimates for Amsterdam and Rotterdam are not expected to be equal when we define the estimates of the mixing probabilities as equal. The LBM with $T = 3$ latent budgets has an adequate fit of $L^2 = 41.8$ (df is 30; $p > .05$). Given the worse fit of the solution with equality restrictions on latent budget parameters,

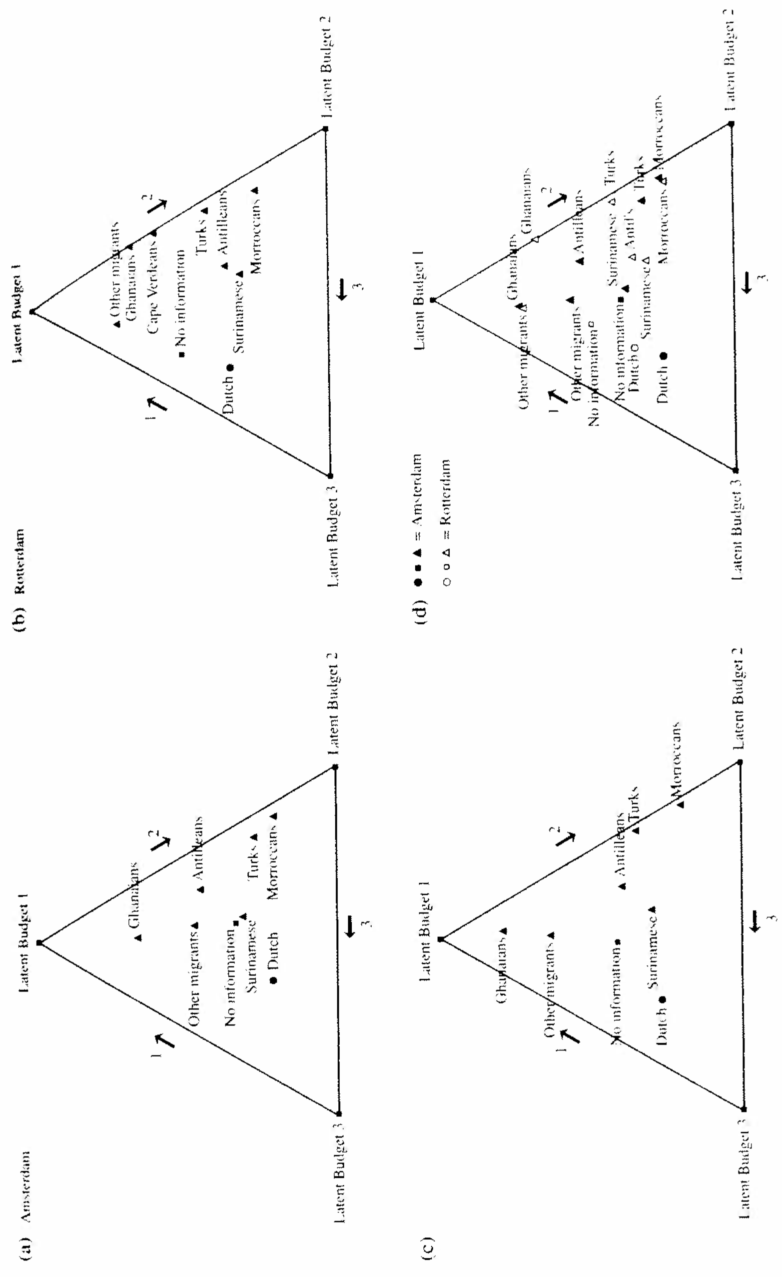


Figure 4. LBM of ethnic differences among people starting a trade in (a) Amsterdam; (b) Rotterdam; (c) Amsterdam and Rotterdam, with homogeneous mixing parameters; (d) Amsterdam and Rotterdam, with homogeneous latent budgets.

it comes as no surprise that the fit for $T = 3$ was not adequate if we imposed the restriction that both the mixing parameters and the latent budget parameters are equal in Amsterdam and Rotterdam: $L^2 = 84.1$ (df is 42).

First, we interpreted the solution with equality restrictions on the mixing parameters (π_{it}^{AX}: see Table 4a), and next the solution with equality restrictions on the latent budget parameters ($\pi_{jt}^{\bar{B}X}$: see Table 4b).

In Table 4a, the first latent budget is characterized by wholesale trade, although to a larger extent for Rotterdam than for Amsterdam. In Amsterdam, this is compensated for by larger estimates for all other trades, except for catering. In the second latent budget, retail dominates, particularly in Amsterdam (together with wholesale trade), whereas in Rotterdam catering and personal services are larger. The third latent budget is characterized by producer services, with personal services a bit larger in Amsterdam, whereas retail is a bit larger in Rotterdam. We found it difficult to interpret these small (but significant) differences between Amsterdam and Rotterdam substantively. Figure 4(c), which shows the mixing-parameter estimates, is quite similar to Figures 4(a) and 4(b). The

Table 4a. Homogeneous Mixing Parameters in Amsterdam and Rotterdam

	Amsterdam				Rotterdam			
Mixing Parameters	$T=1$	$T=2$	$T=3$		$T=1$	$T=2$	$T=3$	
Dutch	0.264	0.185	0.551		0.264	0.185	0.551	
Turks	0.349	0.627	0.024		0.349	0.627	0.024	
Moroccans	0.196	0.775	0.029		0.196	0.775	0.029	
Antilleans	0.396	0.444	0.160		0.396	0.444	0.160	
Surinamese	0.295	0.416	0.289		0.295	0.416	0.289	
Ghanaians	0.793	0.120	0.086		0.793	0.120	0.086	
Other	0.635	0.159	0.206		0.635	0.159	0.206	
No info.	0.409	0.262	0.329		0.409	0.262	0.329	
Latent budgets	$T=1$	$T=2$	$T=3$	p_{+j}	$T=1$	$T=2$	$T=3$	p_{+j}
Wholesale trade	0.682	0.102	0.000	0.257	0.855	0.000	0.062	0.292
Retail trade	0.164	0.543	0.083	0.214	0.052	0.461	0.126	0.185
Producer serv.	0.072	0.000	0.701	0.327	0.000	0.020	0.674	0.304
Catering & rest.	0.044	0.167	0.031	0.066	0.093	0.259	0.000	0.092
Personal serv.	0.038	0.188	0.185	0.135	0.000	0.259	0.139	0.127

Note: Table gives parameter estimates for simultaneous latent budget analysis with $T = 3$ for Rotterdam and Amsterdam.

Table 4b. Homogeneous Latent Budgets in Amsterdam and Rotterdam

	Amsterdam				Rotterdam			
Mixing Parameters	$T=1$	$T=2$	$T=3$		$T=1$	$T=2$	$T=3$	
Dutch	0.243	0.215	0.542		0.326	0.189	0.486	
Turks	0.307	0.622	0.072		0.398	0.572	0.030	
Moroccans	0.240	0.720	0.041		0.233	0.714	0.054	
Antilleans	0.510	0.349	0.141		0.334	0.461	0.205	
Surinamese	0.358	0.346	0.296		0.290	0.437	0.274	
Ghanaians	0.716	0.112	0.172		0.657	0.343	0.000	
Other	0.542	0.201	0.257		0.702	0.122	0.186	
No info.	0.374	0.299	0.327		0.480	0.168	0.352	
Latent budgets	$T=1$	$T=2$	$T=3$	p_{+j}	$T=1$	$T=2$	$T=3$	p_{+j}
Wholesale trade	0.811	0.000	0.000	0.257	0.811	0.000	0.000	0.292
Retail trade	0.113	0.545	0.078	0.214	0.113	0.545	0.078	0.185
Producer serv.	0.000	0.000	0.756	0.327	0.000	0.000	0.756	0.304
Catering & rest.	0.076	0.205	0.000	0.066	0.076	0.205	0.000	0.092
Personal serv.	0.000	0.250	0.167	0.135	0.000	0.250	0.167	0.127

Note: Table gives parameter estimates for simultaneous latent budget analysis with $T = 3$ for Rotterdam and Amsterdam.

Dutch predominate particularly in the third latent budget: Surinamese are situated between the Dutch and the other ethnic groups, ordered as Moroccans, Turks, Antilleans, then Ghanaians.

Table 4b shows the parameter estimates for the LBM with homogeneous latent budgets. Again, latent budget 1 is characterized by wholesale trade: latent budget 2 by retail trade, catering and restaurants, and personal services: and latent budget 3 by producer services and personal services. The mixing-parameter estimates are displayed in Figure 4(d). For each ethnic group, we found the Amsterdam label close to the Rotterdam label. For interpreting small distinctions, we concentrated on more specific characterizations by specific budgets. The Amsterdam Dutch are to a larger extent characterized by latent budget 3, the Amsterdam Turks to a larger extent by latent budget 2, and the Amsterdam Antilleans by latent budget 1, whereas the Rotterdam Antilleans are characterized more by latent budget 2, the Rotterdam Surinamese by latent budgets 2 and 3, the Amsterdam Ghanaians by latent budget 1. Rotterdam other migrants more by latent budget 1, and Rotterdam "no information" more by latent budgets 1 and 3. For more information, we refer to Kloosterman and van der Leun (1998).

4. SOCIAL MILIEU AND SECONDARY EDUCATION: AN EXAMPLE OF CONSTRAINED LATENT BUDGET ANALYSIS

At the age of 11–12 years, children in The Netherlands go from primary school to secondary school. Distinct types of secondary education can be chosen, with two main types: vocational types of education and general types of education. Choice depends on such aspects as capacities of children, interests, advice of the primary school teacher, and advice of parents. In educational research, much interest is directed to the way in which the social milieu of a child influences this choice. In this example, we investigated this question by using the LBM. The best interpretation of the LBM in this context is in terms of the MIMIC model (see Figure 2). Three explanatory variables, that is, sex, social milieu, and IQ, yield (as shown by the mixing parameters, $\pi_{it}^{A\bar{X}}$) an individual's human capital (the latent variable, having T classes), and this human capital provides opportunities to go to a specific level of education (as shown by the latent budget parameters, $\pi_{jt}^{\bar{B}X}$).

In 1977 and 1981, data were collected from more than 37,000 children about their social milieu and aspects regarding their secondary education. Distinct variables were collected: for a description, see Statistics Netherlands (1982) and Meester and de Leeuw (1983). The variables we used in our analysis are the scores on an intelligence test, social milieu (profession of father), sex, and the level of education attained in 1981, that is, after 4 years of secondary education. The intelligence test used was the (Dutch) Test for Intellectual Capacity (TIC), a figure exclusion test that consists of 33 items. The TIC scores were recoded by Meester and de Leeuw (1983) as 1 for 1–14 correct items, 2 for 15–17 correct, 3 for 18–20 correct, 4 for 21–23 correct, 5 for 24–26 correct, 6 for 27–29 correct, and 7 for 30–33 correct items. The social milieu of the family is measured by the profession of the father, in six categories: category 1 is skilled and unskilled laborers, 2 is farmers and farm laborers, 3 is shopkeepers, 4 is lower employees, 5 is middle employees, and 6 is higher employees and scientific and free professions. The last explanatory variable is the dichotomous variable of sex. The response variable is the level of education attained after 4 years, and these levels are 1, dropped out; 2, junior vocational education (LBO); 3, general education, medium level (MAVO); 4, general education, high level (HAVO); 5, general education, preparatory to university (VWO); and 6, senior vocational training ((M)BO). Meester and de Leeuw (1983) eliminated all children having no TIC score (16,433 children). According to them, this elimination is not crucial because

having no TIC score seemed to have been a random process. Furthermore, children with a value missing on the level of education attained (38) or on a type of education called *extraordinary lower education* (646) were eliminated from the sample. Children having a father who is unemployed or medically unfit for work were also eliminated (6,190). This last elimination is more crucial, and it should be kept in mind that our analysis does not discuss children having these characteristic. Following these selections there remained a sample of 16,236 children. The data are given in Table 5.

We analyzed the data with the LBM by coding the levels of the explanatory variables sex, social milieu, and TIC as $2 \times 6 \times 7 = 84$ rows and the levels of the response variable level of education attained as six columns. Let the variable sex be A, indexed by i; let social milieu be C, indexed by k; let TIC be F, indexed by p; and let the response variable level of education attained be B, indexed by j. Thus, the LBM can be rewritten by replacing the index i in Model 2 by ikp, so that the LBM becomes

$$\frac{\pi_{ikpj}}{\pi_{ikp+}} = \sum_{t=1}^{T} \pi_{ikpt}^{ACF\bar{X}} \pi_{jt}^{\bar{B}X}.$$

The LBM with $T = 1$ (independence) is equivalent to the model in which the variables sex, social milieu, and TIC are dependent, and independent from level of education attained. This model may be considered as our baseline model. It has a fit of $L^2 = 4{,}612$, with df = 415. The LBM with two or more latent budgets can be interpreted as a MIMIC model (Figure 1). The MIMIC model emphasizes that each child has T probabilities $\pi_{ikpt}^{ACF\bar{X}}$ of falling into the latent classes, which can be interpreted as the individual's human capital. These T probabilities are determined by the levels of explanatory variables A, C, and F. Once a child is in one of the T latent classes, there are J probabilities $\pi_{jt}^{\bar{B}X}$ of attaining each of the levels of education.

A sensible approach to the analysis is first to determine the number of latent classes T that is needed to give an adequate description of the data. For $T = 2$, $L^2 = 1{,}113$ (df is 328); for $T = 3$, $L^2 = 441$ (df is 243); for $T = 4$, $L^2 = 226$ (df is 160); and for $T = 5$, $L^2 = 116$ (df is 79). All the models have to be rejected at $p = .05$. To check whether this could be due to the specific form of our models, we studied the residuals of the least restricted LBM, that is, the LBM with $T = 5$. We found no intelligible patterns in the residuals or specific outlier cells, so we assumed that the misfit of the models is due to a large sample size.

Table 5. The SMVO Data

School Type		Boys						Girls					
SES	TIC	1	2	3	4	5	6	1	2	3	4	5	6
1	1	43	126	23	5	2	17	28	87	24	13	3	35
	2	41	172	58	20	9	28	29	131	57	15	0	74
	3	50	271	83	58	24	87	67	209	128	59	6	141
	4	64	268	131	93	44	111	64	200	157	95	34	194
	5	43	202	121	113	47	109	35	163	177	105	39	201
	6	11	78	60	62	43	78	20	54	106	92	48	103
	7	4	15	20	23	27	19	2	10	22	40	38	28
2	1	3	13	1	1	1	8	2	8	5	1	0	5
	2	3	18	9	0	0	10	2	14	10	4	0	12
	3	2	18	12	15	3	23	5	18	16	19	3	26
	4	8	25	15	14	9	47	0	18	23	21	8	46
	5	5	25	16	12	16	35	0	13	28	21	15	39
	6	2	4	7	20	11	22	5	6	19	37	15	30
	7	0	3	2	5	7	9	0	4	4	12	17	10
3	1	11	17	6	1	1	10	7	12	11	2	0	8
	2	9	37	11	6	2	10	6	29	11	5	1	11
	3	23	59	26	12	6	29	16	43	30	19	4	38
	4	12	72	34	23	14	38	18	39	39	36	13	49
	5	11	40	26	37	25	36	16	32	54	54	25	39
	6	7	20	26	25	30	25	11	12	28	41	20	24
	7	3	1	7	9	12	9	2	3	3	16	7	3
4	1	9	29	13	4	1	4	3	15	6	3	0	10
	2	9	38	21	5	4	13	10	24	26	7	2	29
	3	12	56	47	37	15	27	12	54	40	37	15	35
	4	11	62	52	54	26	43	15	39	64	56	27	61
	5	12	48	62	55	37	30	9	31	54	87	44	52
	6	6	15	33	40	45	24	7	11	35	49	39	39
	7	3	4	7	17	23	7	2	3	5	23	26	9
5	1	5	25	14	9	3	9	6	20	8	3	1	12
	2	8	26	30	23	7	11	9	22	24	19	4	30
	3	13	60	65	39	35	50	10	42	50	44	33	59
	4	20	79	91	94	71	70	17	58	97	82	55	79
	5	11	58	70	95	95	63	11	44	89	103	101	70
	6	9	39	44	71	107	40	5	17	46	117	104	47
	7	4	7	9	28	57	12	2	3	28	49	70	21
6	1	4	6	10	6	4	3	5	2	6	1	1	5
	2	7	14	15	11	5	12	4	3	6	18	2	11
	3	5	31	34	39	21	23	5	16	24	33	16	21
	4	10	16	45	54	52	36	9	16	44	83	46	29
	5	7	16	44	71	105	28	7	7	40	80	83	27
	6	3	12	24	40	85	19	8	7	32	66	100	15
	7	3	4	9	16	52	9	1	3	10	29	51	1

Notes: School types are 1. drop out; 2. LBO; 3. MAVO; 4. HAVO; 5. VWO; 6. (M)BO. Social milieu is 1. skilled and unskilled laborers; 2. farmers and farm laborers; 3. shopkeepers; 4. lower employees; 5. middle employees; 6. higher employees. TIC scores are number of items correct: 1, 1–14; 2, 15–17; 3, 18–20; 4, 21–23; 5, 24–26; 6, 27–29; 7, 30–33.

Given the large sample size, we were satisfied with the description that the LBM offers with three latent budgets. Although significant, the discrepancy between the $L^2 = 441$ and df $= 243$ is not enormous; the model describes 0.904 of the departure from the independence model ($T = 1$) (i.e., $0.904 = (4{,}612 - 441)/4{,}612$). Because the gain in percentage moving from the LBM with three to the LBM with four latent budgets is relatively small, we choose the LBM with three latent budgets to examine more carefully.

The latent budget parameter estimates ($\pi_{jt}^{\bar{B}X}$) are shown in Table 6. In the first latent budget children go predominantly into *lower vocational training (LBO) or drop out*, and to a lesser extent they go into medium general education (MAVO) and (M)BO. In the second latent budget, children go predominantly into *higher general education (HAVO and VWO)*, and in the third latent budget they go predominantly into *medium and higher general education (MAVO and HAVO) and higher vocational training (MBO)*, but not to general, university preparatory education (VWO).

For the study of the mixing-parameter estimates, we give plots of the estimates separately for each TIC score p and each sex i. This gives $7 \times 2 = 14$ plots, shown in Figure 5. In each plot, we have set out horizontally the six levels of social milieu k and vertically the probability of going to one of the latent budgets t. Each plot has 18 points: namely, children in each of the six levels of social milieu can go to each of the three latent budgets; points belonging to the same latent budgets are connected, so that each plot has three lines. In Figure 5, the first latent budget is indicated by the line with the circles, the second latent budget is indicated by the

Table 6. Latent Budgets Estimates for $T = 1$ (Independence) and $T = 3$ for Educational Level after 4 Years of Secondary School

		Panel 1			Panel 2		
Group	$T = 1$	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$
1. Drop out	0.063	0.160	0.014	0.011	0.177	0.025	0.005
2. LBO	0.226	0.658	0.000	0.000	0.701	0.038	0.006
3. MAVO	0.192	0.121	0.090	0.325	0.092	0.090	0.331
4. HAVO	0.188	0.000	0.367	0.232	0.000	0.337	0.228
5 VWO	0.142	0.000	0.530	0.000	0.015	0.500	0.000
6. (M)BO	0.189	0.061	0.000	0.432	0.015	0.011	0.430
π_t^X	1.000	0.343	0.267	0.389	0.304	0.274	0.422

Note: Panel 1, unconstrained estimates $T = 3$; panel 2, estimates $T = 3$ constrained by the multinomial logit model.

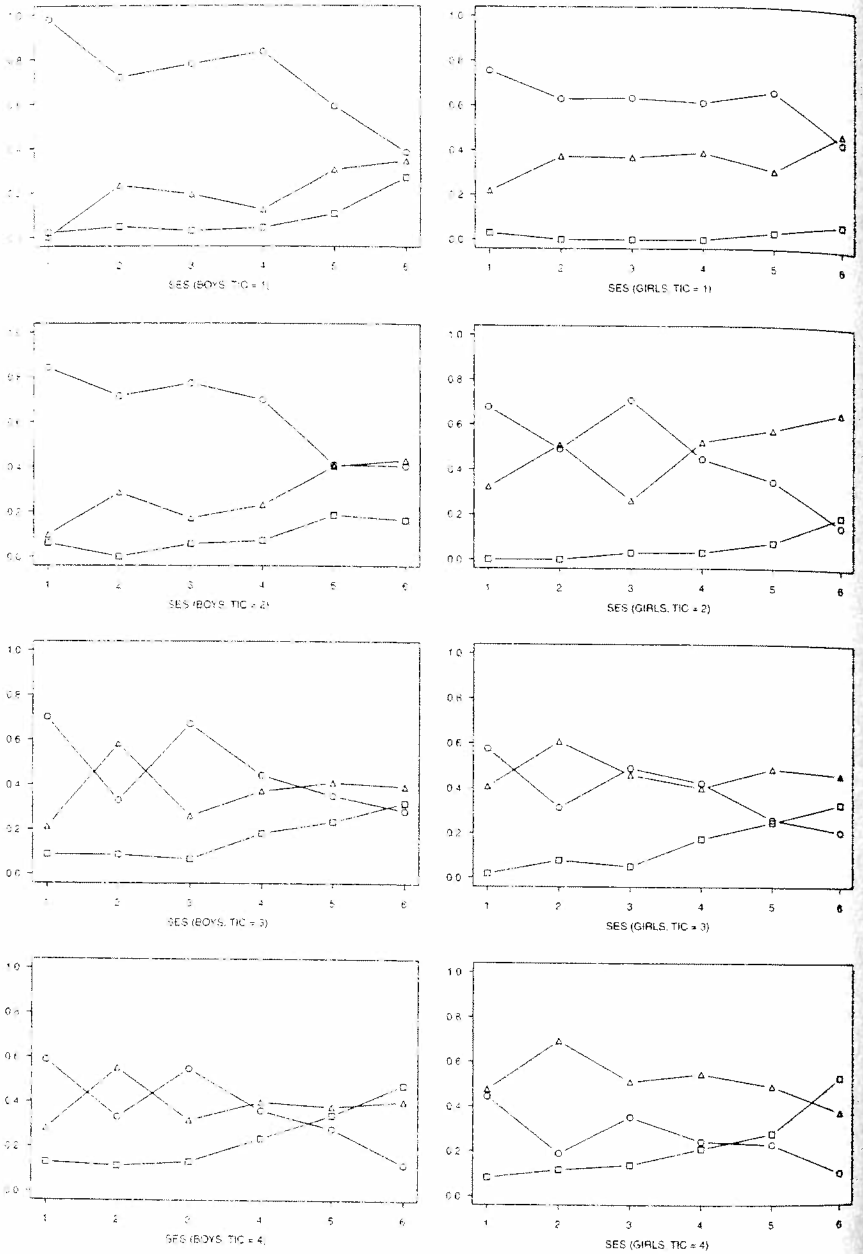


Figure 5. Unconstrained mixing-parameter estimates for each TIC score group and each sex.

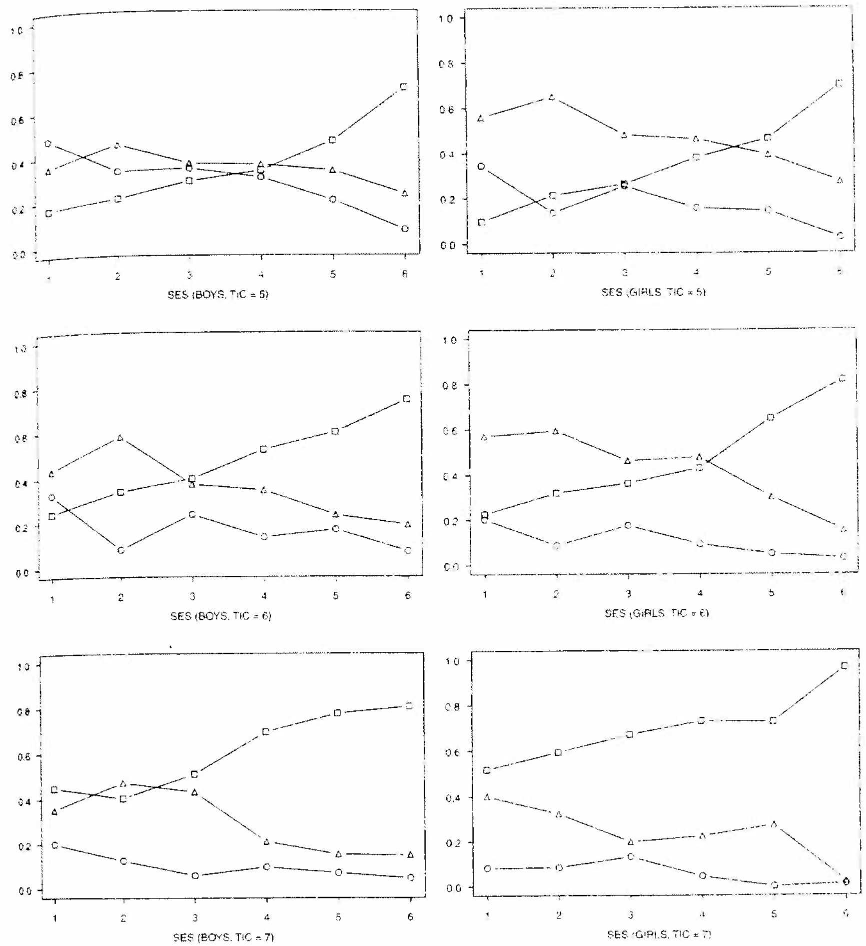


Figure 5 (*continued*)

line with the squares, and the third latent budget is indicated by the line with the triangles.

We chose to display the parameters in this way for the following reasons: First, if sex had no influence on the probability of going to latent budgets, then plots on the left (boys) would be identical to plots on the right (girls). This way of displaying the parameters clearly shows the influence of sex if we look at how each pair of plots differs. Second, if the social milieu had no influence on the probability of going to the latent budgets, then all lines would be horizontal, and departures from this would be easily displayed.

It is clear that the probability of going to a latent budget will be strongly influenced by the TIC score because the levels attained not only reveal differences between different types of education (general vs. vocational), but also between higher and lower types of education. Therefore, in going from the plots at the top (TIC score equals 1) to the bottom (TIC score equals 7), the line with circles drops generally: this is not surprising because this line shows the probability of going to latent budget 1, which is the budget in which 0.658 of the children go to LBO and 0.160 drop out: children more often drop out or go to LBO when their TIC score is lower.

There are many interesting aspects in these plots. For instance, in all levels of TIC, children with fathers who are medium or higher employees (5 and 6) have a much higher probability than average of going to latent budget 2, which is the budget for higher general education (HAVO, 0.367) and university preparatory education (VWO, 0.530). Their probability of going to budget 1 (drop out and LBO) is much lower. The reverse holds for children whose father is a skilled or unskilled laborer: Given their TIC score, their probability of going to budget 1 is in general the highest. On average, children whose fathers are farmers (2) are more likely than average, given their TIC score, to go to latent budget 3, where they have a high probability of following medium vocational training ((M)BO). It may be noted that the latent budget parameters, being probabilities, can be interpreted easily; they not only show tendencies in the data (e.g., girls go on average less to budget 1 than boys), but also show how strong the effects are.

Van der Heijden et al. (1992) showed how the factorial structure in the explanatory variables could be used to investigate the effects of each of the factors and their interactions. This is done by means of a multinomial logit model for the mixing parameters $\pi_{ikpt}^{ACF\bar{X}}$, that is,

$$\pi_{ikpt}^{AXF\bar{X}} = \exp\left(\sum_{m=1}^{M} x_{ikpm}\gamma_{mt}\right) \Bigg/ \sum_{t=1}^{T} \exp\left(\sum_{m=1}^{M} x_{ikpm}\gamma_{mt}\right). \qquad (5)$$

The design matrix **X** has $I \times K \times P$ rows and M columns, and these M columns represent dummy variables for the main effects for factors A, C, and F; for their two-way interaction effects $A \times C$, $A \times F$, and $C \times F$; and their three-way interaction $A \times C \times F$. The elements γ_{mt} are parameters for column m and latent budget t. To identify the model, $\gamma_{m1} = 0$. For more details, see van der Heijden et al. (1992), who also explain the relationship between these models and loglinear models with latent variables.

We have systematically imposed all possible constraints on the mixing parameters ($\pi^{ACF\bar{X}}$). The most restrictive LBM has only main effects

A, C, and F. This LBM turns out to fit reasonably well, with $L^2 = 627$ (df is 379). Forsaking the model with unconstrained mixing parameters ($\pi_{ikpt}^{ACF\bar{X}}$) for the model with only main effects thus gains us $379 - 243 = 136$ df, at the expense of a loss of fit of $627 - 441 = 186$.

The latent budget parameter estimates are similar to those for the unconstrained model with three latent budgets (see Table 6). We studied the estimates by deriving averages of mixing parameters $\hat{\pi}_{it}^{A\bar{X}} = \hat{\pi}_{i+++t}/\hat{\pi}_{i++++}$, $\hat{\pi}_{kt}^{C\bar{X}} = \hat{\pi}_{+k++t}/\hat{\pi}_{+k+++}$, and $\hat{\pi}_{pt}^{F\bar{X}} = \hat{\pi}_{++p+t}/\hat{\pi}_{++p++}$. Thus, we obtained parameters for sex only, for social milieu only, and for TIC only. Plots of these parameter estimates are given in Figure 6. The plot for TIC score shows that the probability of going to latent budget 1 (mainly LBO, drop out) decreased as TIC increased; the probability of going to latent budget 2 (mainly VWO, HAVO) increased as TIC increased; and the probability of going to latent budget 3 (mainly MAVO, HAVO, (M)BO) increased from TIC 1 to 4, and then decreased smoothly. In the plot for social milieu, the probability of going to budget 1 is low for children of farmers (2) and medium and higher employees (5, 6), the probability to go to budget 2 increased rapidly for children of lower to higher employees, and the probability of going to budget 3 was above average for children of farmers and below average for children of higher employees. In the plot for sex, we found that there is no difference in the probability of boys and girls going to latent budget 1. However, there was a difference in their probability of going to latent budgets 2 and 3: for boys, these probabilities were approximately equal, whereas girls went more often to latent budget 3 and less often to latent budget 2.

Latent budget analysis offered considerable insight into these data. The MIMIC-model interpretation that we used showed with which probabilities children, given a specific background, go to specific latent budgets. These latent budgets specified the probabilities of reaching specific final levels of education. In this example, the parameters were also very easy to interpret, so that it was easy to indicate the processes that operate in the relationship between explanatory variables such as TIC, sex, and social milieu on the one hand, and secondary education on the other. The constraints allowed a simplified interpretation.

5. CONCLUSIONS

The LBM closely answered specific research questions that are interesting from a substantive point of view. For Plato's data, we assumed that there were a few typical (latent) writing styles, and each book is a

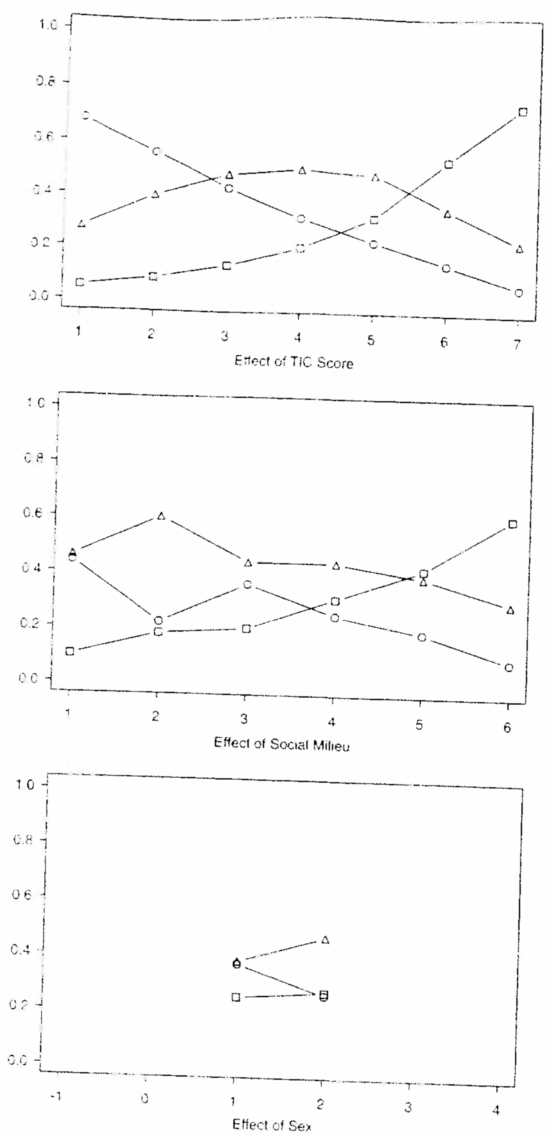


Figure 6. Constrained mixing-parameter estimates for the TIC score groups, for the social milieu groups, and the sexes.

mixture of these typical styles. This assumption is related directly to the parameterization of LBM. In this example, the LBM was interpreted as a mixture model. In the ethnic entrepreneur example and the secondary education example, we assumed the existence of a latent variable mediating between the explanatory variable and the response variable

In both examples, the interpretation of this latent variable was human capital. The LBM was interpreted as a MIMIC model.

An important asset of the LBM is the simple interpretation of the parameters, also for nonstatisticians. The merits of the LBM are most evident when the row variable can be interpreted as the explanatory variable and the column variable can be interpreted as the response variable. Otherwise, a (symmetrical) latent class model is more suitable.

REFERENCES

Atkinson, A. C. (1970). "A method for discriminating between models." *Journal of the Royal Statistical Society, Series B*, **32**, 323–53.

Boneva, L. (1970). "A new approach to a problem of chronological seriation associated with the works of Plato." In F. R. Hodson, D. G. Kendall, & P. Tautu, *Mathematics in the Archeological and Historical Sciences*. Edinburgh: Edinburgh University Press, pp. 173–85.

Clogg, C. C. (1981). "Latent structure models of mobility," *American Journal of Sociology*, **86**, 836–68.

Cox, D. R., & Brandwood, L. (1959). "On a discriminatory problem connected with the works of Plato." *Journal of the Royal Statistical Society, Series B*, **21**, 195–200.

de Leeuw, J., & van der Heijden, P. G. M. (1988). "The analysis of time-budgets with a latent time-budget model." In E. Diday et al. (eds.), *Data Analysis and Informatics 5*. Amsterdam: North-Holland, pp. 159–66.

de Leeuw, J., & van der Heijden, P. G. M. (1991). "Reduced rank models for contingency tables," *Biometrika*, **78**, 229–32.

de Leeuw, J., van der Heijden, P. G. M., & Verboon, P. (1990). "A latent time budget model," *Statistica Neerlandica*, **44**, 1–22.

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

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. New York: Academic Press.

Kaluscha, W. (1904). "Zur Chronologie der Platonischen Dialoge." *Wiener Studien*, pp. 25–7.

Kloosterman, R., & van der Leun, J. (1998). "The same musical chairs? Urban opportunity structures and business starters by immigrant entrepreneurs in Amsterdam and Rotterdam." In J. Rath & R. Kloosterman (eds.), *Immigrant Entrepreneurs in the Netherlands* (in Dutch). Amsterdam: het Spinhuis, pp. XXX–XX.

Meester, A., & de Leeuw, J. (1983). *Intelligence, Social Milieu and the School Career* (in Dutch). Leiden: Department of Data Theory.

Mooijaart, A., & van der Heijden, P. G. M. (1992). "The EM-algorithm for latent class analysis with constraints," *Psychometrika*, **57**, 261–69.

Mooijaart, A., van der Heijden, P. G. M., & van der Ark, L. A. (1999). "A least-squares algorithm for a mixture model for compositional data," *Computational Statistics and Data Analysis*, **30**, 359–79.

Renner, R. M. (1993). "The resolution of a compositional data set into mixtures of fixed source compositions." *Applied Statistics*, **42**, 615–31.

Siciliano, R., & van der Heijden, P. G. M. (1994). "Simultaneous latent budget analysis of a set of two way tables with constant row sum data." *Metron*, **53**, 155–79.

Statistics Netherlands (1982). "School career and origin of pupils in secondary education. Part 2: cohort 1977, choice of school type" (in Dutch). The Hague: Staatsuitgeverij.

van der Ark, L. A. (1999). *Contributions to Latent Budget Analysis. A Tool for the Analysis of Compositional Data*. Leiden: DSWO Press.

van der Ark, L. A., & van der Heijden, P. G. M. (1997). "Graphical display of latent budget analysis and latent class analysis, with special reference to correspondence analysis." In M. Greenacre & J. Blasius (eds.), *Visualization of Categorical Data*. San Diego: Academic Press, pp. 489–508.

van der Ark, L. A., van der Heijden, P. G. M., & Sikkel, D. (1999). "On the identifiability in the latent budget model." *Journal of Classification*, **16**, 117–37.

van der Heijden, P. G. M. (1994). "End-member analysis and latent budget analysis." *Applied Statistics*, **43**, 527–28.

van der Heijden, P. G. M., Gilula, Z., & van der Ark, L. A. (1999). "An extended study into the relationships between correspondence analysis and latent class analysis." In M. Sobel & M. Becker (eds.), *Sociological Methodology 1999*, Vol. 29. Cambridge: Basil Blackwell, pp. 147–86.

van der Heijden, P. G. M., Mooijaart, A., & de Leeuw, J. (1992). "Constrained latent budget analysis." In C. C. Clogg (ed.), *Sociological Methodology 1992*, Vol. 22. Cambridge: Basil Blackwell, pp. 279–320.

Weltje, G. J. (1997). "End member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem." *Mathematical Geology*, **29**, 503–49.

FIVE

Ordering the Classes

Marcel Croon

I. INTRODUCTION

A. The Problem of Ordering the Classes in Latent Class Analysis

Latent class analysis was conceived by Lazarsfeld (1950a, 1950b, 1954) as a method for analyzing the association among manifest qualitative variables in terms of an unobserved nominal variable whose values represent different subpopulations. The latent class model assumes that the heterogeneous population from which the respondents were sampled can be partitioned in a restricted number of homogeneous subpopulations: the latent classes. Individual respondents belong to one of these classes, and respondents belonging to the same class share with each other the same set of response probabilities for the manifest variables. Furthermore, within each latent class the responses to different manifest variables are supposed to be stochastically independent. This assumption of *local indepedence*, as it is called, implies that the classes in the latent class model can be considered to be the values of a nominal latent variable in such a way that if one conditions on the value of the latent variable, the association between the manifest variables disappears completely. The association among the manifest variables is then completely explained by their relationships with the latent variable.

In this basic formulation of the latent class model, nothing impels the classes to be ordered along some continuum because the set of latent classes only represents a partition of the original population into disjoint subpopulations. The latent variable, whose levels correspond to the classes, is clearly a nominal variable. In social research, however, the response categories of the manifest variables can often be ordered on some common attributional continuum. In many applications of the latent class