Geometric Data Analysis of Individual Differences

by

Brigitte Le ROUX[1]
MAP5 (CNRS)

and

Henry ROUANET[2]
CRIP5

Department of Mathematics and Computer Science
Université René Descartes, PARIS, France

[1]e-mail: lerb@math-info.univ-paris5.fr
http://www.math-info.univ-paris5.fr/~lerb/
[2]e-mail: rouanet@math-info.univ-paris5.fr
http://www.math-info.univ-paris5.fr/~rouanet/

Contents

List of Figures

List of Tables

Chapter 1

Geometric Data Analysis of Individual Differences

1.1 Introduction

The objective of this chapter is to study the multidimensional structure of the individual differences of an EPGY file (courses of mathematics, grade 3), by means of Multiple Correspondence Analysis (MCA). MCA is applicable to an Individuals×Variables table where variables are categorized, that is, have a finite number of categories. The procedure of MCA is part of *Geometric Data Analysis* (GDA) for which multivariate data sets are conceptualized as clouds of points, and the interpretation of data is essentially based on these clouds[1]: See Benzécri & al (1973), Lebart & al. (1984), Greenacre (1984), Benzécri (1992), Gower & Hand (1995). Benzécri and his colleagues have developed GDA mostly around Correspondence Analysis (CA), but GDA also covers Principal Component Analysis (PCA) recast in geometric terms. Multiple Correspondence Analysis (MCA), which grew up as a special case of CA, is the analog of PCA for categorized variables. The domain of application of MCA is extremely wide. As far as Educational Research is concerned, MCA is routinely used by the Evaluation Deparment of the French Ministry of Education, as reflected in the journal *Education et Formations*. For an

[1] Geometric Data Analysis is known in France as "Analyse des données"; the name "Geometric Data Analysis", which denotes the specificity of the approach, has been suggested by P. Suppes. *Cloud of points* ("nuage de points") to designate the set of points in a Euclidean space, is now a well accepted phrase in English. As we proceed in the chapter, we will briefly recall the meaning of the words belonging to the specific vocabulary of MCA.

example of MCA applied to the teaching of Mathematics see Murtagh (1981).

The MCA procedure leads to constructing two clouds of points: a cloud of categories (or of modalities[2]), and a cloud of individuals. In the analyses of the present chapter, we emphasize the study of the cloud of individuals and Structured Data Analysis, as has been done in Rouanet & Le Roux (1993), Le Roux & Rouanet (1998), Bourdieu (1999), Chiche & al. (2000), and Le Roux & Rouanet (forthcoming).

The chapter is organized as follows. After the present introduction (§1.1), we describe the data set and its coding (§1.2), then we proceed to MCA (§1.3), we interpret axes (§1.4); we study the cloud of individuals (§1.5). As is usual in geometric data analyses, we complement the study by a Euclidean classification (§1.6). We close the chapter with a discussion and conclusions (§1.7).

1.2 Data and Coding

The data studied are those of 533 EPGY students in the third grade; they concern the following 5 strands: Integers, Fractions, Geometry, Logic and Measurement.

1.2.1 Variables Retained for Analysis

Active variables. The variables that serve to define the distance between individuals, that is, to construct the cloud of individuals, are called active variables. For each strand, we have taken three types of variables:

1. Error rates.

2. Latencies for correct answers.

3. Number of exercises to master the concepts of strand[3].

Crossing the three types of variables with the five strands, we get fifteen active variables in all.

[2] *Response modality* ("modalité de réponse") is often used as a synonym of category, in the context of questionnaires.

[3] `See Suppes & Tod for a description of how the number of exercises to mastery is determined.`

Structuring factors of individuals. In addition to the fifteen active variables, we will study the number of hours spent on the computer, Gender and Age, as structuring factors of individuals.

1.2.2 Univariate Analyses and Coding of Variables

The distributions of *Error rates* differ among the strands (see Figures 1.1-a to 1.1-e). The Integers, Fractions and Measurement distributions are $\mathcal{I}$–shaped whereas Geometry and Logic ones are more bell–shaped.

Remark. The numbers of students who make *no error* are 9 in Integers, 70 in Fractions, 4 in Geometry, 0 in Logic, 49 in Measurement.

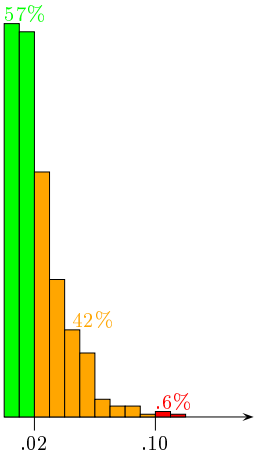

Figure 1.1-a. Integers

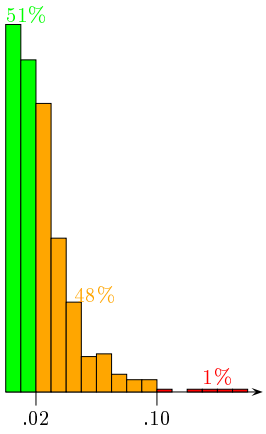

Figure 1.1-b. Fractions

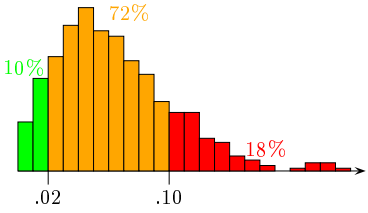

Figure 1.1-c. Geometry

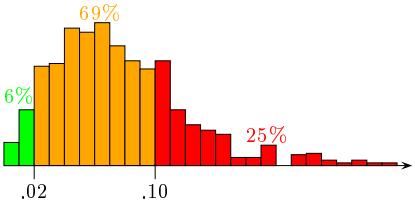

Figure 1.1-d. Logic

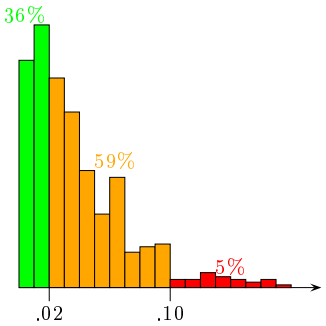

Figure 1.1-e. Measurement

Figure 1.1: Distributions of *Error rates* and percentages for the 3 grouped categories.

The five *Latency* distributions are more or less bell–shaped (see Figures 1.2-a to 1.2-e).

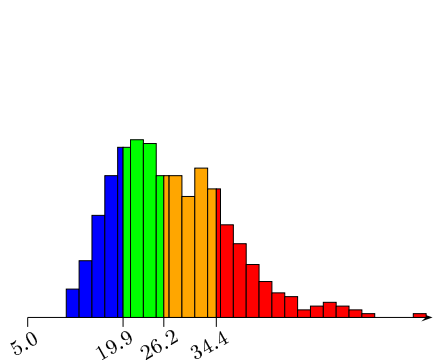

Figure 1.2-a. Integers

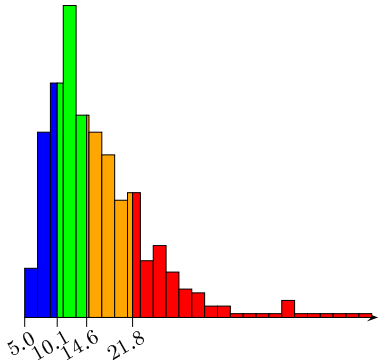

Figure 1.2-b. Fractions

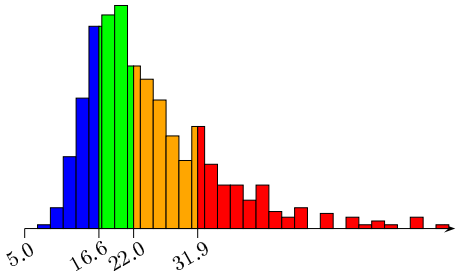

Figure 1.2-c. Geometry

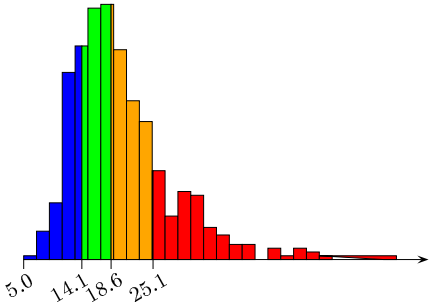

Figure 1.2-d. Logic

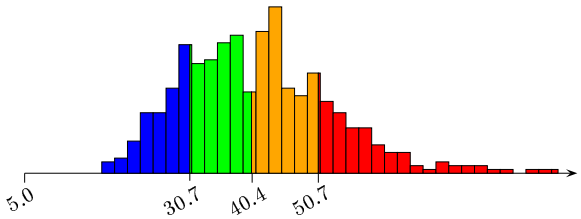

Figure 1.2-e. Measurement

Figure 1.2: Distributions of Latencies

The *Number of Exercises* is a discrete variable (see Table 1.1, p.5).

Given the heterogeneity of variables and of their distributions, the most appropriate Geometric Data Analysis is Multiple Correspondence Analysis (MCA). We therefore proceed to the coding of variables into a number of categories (2, 3 or 4) that we describe below. The phase of coding in GDA is

Integers

4	*5*
480	53

Fractions

4	*5*	*6*	*7*
432	49	1	1

Geometry

6	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
1	1	19	61	118	203	100	20	7	3

Logic

4	*5*	*6*	*7*	*9*	*10*
165	301	55	9	2	1

Measurement

4	*5*	*6*
387	125	21

Table 1.1: For each strand, absolute frequencies of the *Number of exercises* (in italics) to mastery.

always crucial for efficient analyses and must be carefully performed according to the specificities of variables in order to attain as much homogeneity as possible, which is required to define a distance between individuals.

Error rates. We have taken a common coding defined by two cuts at 2% and 10%, generating 3 categories, namely less than 2 errors per 100 exercises, between 2 and 10 errors, more than 10 errors.

Category	Integers	Fractions	Geometry	Logic	Measurement
1 $\cdot < .02$	305	274	55	31	192
2 $.02 \leq \cdot < .10$	225	254	382	367	316
3 $\cdot \geq .10$	3	5	96	135	25

Table 1.2: Absolute frequencies of error rates

Latencies. *Latencies* widely differ among strands[4]. Consequently we have, for each strand, taken a 4–category coding defined by the inferior quintile (20%), median (50%), and superior quintile (80%); see Table 1.3 (p.6) and Figure 1.2 (p.4). Absolute frequencies are 106 for category 1, 160 for categories 2 and 3, 107 for category 4: For instance, 106 students have a latency less than 19.84 in Integers, less than 10.06 in Fractions etc. Quintiles have been taken in order to give more importance to extreme individuals in MCA (cf. distance definition p. 7). Hence we obtain $4 \times 5 = 20$ categories.

Number of exercises to mastery. For Integers, Fractions, and Measurement, we code two categories: four exercises (1) and more than four

[4]This discrepancy may be attributable to the differing organizations of exercises among the strands.

Category	1	2	3	4
Integers	$\cdot < 19.84$	$19.84 \leq \cdot < 26.18$	$26.18 \leq \cdot < 34.35$	$\cdot > 34.35$
Fractions	$\cdot < 10.06$	$10.06 \leq \cdot < 14.64$	$16.64 \leq \cdot < 21.79$	$\cdot > 21.79$
Geometry	$\cdot < 16.55$	$16.55 \leq \cdot < 21.95$	$21.95 \leq \cdot < 31.92$	$\cdot > 31.92$
Logic	$\cdot < 14.05$	$14.05 \leq \cdot < 18.60$	$18.60 \leq \cdot < 25.13$	$\cdot > 25.13$
Measurement	$\cdot < 30.66$	$30.66 \leq \cdot < 40.35$	$40.35 \leq \cdot < 50.71$	$\cdot > 50.71$

Table 1.3: Categories and Cut values (in seconds) for *Latencies* coding.

exercises to mastery (2). For Geometry and Logic, we code three categories each. In Geometry, the categories are less than eleven exercises (1), eleven or twelve exercises (2), and more than twelve exercises (3)[5]. In Logic, the categories are four exercises (1), five exercises (2), and more than five exercises (3).

Category	Integers	Fractions	Geometry	Logic	Measurement
1	480	482	82	165	387
2	53	51	321	301	146
3			130	67	

Table 1.4: Absolute frequencies of number of exercises

One thus obtains 47 categories (or modalities).

1.2.3 Response Patterns

With each individual, there is associated a response pattern defined by the individual's responses to the 15 coded variables.

Example. For the first individual in the file, table below gives the raw responses and the corresponding pattern:

	Integers	Fractions	Geometry	Logic	Measurement
Raw $\rightarrow$ Coded Error rates	$.014 \rightarrow 1$	$.015 \rightarrow 1$	$.034 \rightarrow 2$	$.054 \rightarrow 2$	$.069 \rightarrow 2$
Raw $\rightarrow$ Coded Latencies	$27.2 \rightarrow 3$	$13.7 \rightarrow 2$	$26.5 \rightarrow 3$	$17.5 \rightarrow 2$	$37.5 \rightarrow 2$
Raw $\rightarrow$ Coded Exercises	$4 \rightarrow 1$	$4 \rightarrow 1$	$12 \rightarrow 2$	$5 \rightarrow 2$	$6 \rightarrow 2$

The corresponding response pattern is thus 11222 32322 11222.

[5]The Number of Exercises in geometry is different from the one in the other strands because some concept classes in Geometry have no criterion to mastery. This is due to the fact that in those concept classes the student is given a geometrical construction consisting of a number of steps, none of which can be omitted.

The number of possible patterns is equal to

$$3^5 \times 4^5 \times (2 \times 2 \times 3 \times 3 \times 2) = 243 \times 1024 \times 72 = 17\,915\,904.$$

The number of observed patterns is 520, quite close to 533 (total number of individuals), which expresses almost the maximum of individual differences at the level of coded variables[6].

1.3 Multiple Correspondence Analysis

1.3.1 Theoretical Sketch of MCA

For a general presentation of MCA, it is convenient to adopt the language of questionnaire in a standard form. On one hand, there is a set Q of questions, question $q \in Q$ consisting in a set K_q of response categories; we let K be the set of all response categories: $K = \cup_{q \in Q} K_q$. On the other hand, there is a set I of n individuals, and each individual $i \in I$ chooses one and only one response category of each question. Hence the basic Individuals×Questions data table. We present hereafter MCA as a full–fledged geometric method for categorized variables[7].

Distance between individuals. In any Geometric Data Analysis, the first step always consists in defining the distance between individuals. The MCA distance between two individuals is determined by the questions to which they choose different response categories (i.e. for which they disagree). Let n_k be the number of individuals who choose response category k; let $f_k = n_k/n$ be the relative frequency of k. If for question q, individuals i and i' disagree, i choosing k and i' choosing $k' \neq k$, the quantity $1/f_k + 1/f_{k'}$ represents the amount of the square of the distance accounted for by question q, and the overall distance $d(i, i')$ between individuals i and i' is given by the formula:

$$d^2(i, i') = \frac{1}{Q} \sum \left(\frac{1}{f_k} + \frac{1}{f_{k'}} \right)$$

[6] As active variables, we could have taken the averages over strands of the three types of variables; however in doing so, the distance between individuals would only depend on the three types of variables, and we would lose an important source of variation between individuals. As a matter of comparison, if we had taken only the three types of variables as active variables, we would have much fewer possible patterns (say about $3 \times 4 \times 3 = 36$) than individuals.

[7] MCA is often presented as a special case of CA applied to the Individuals×Categories table after the disjunctive encoding of the Individuals×Questions table (see p.8).

where the sum is taken over the questions for which the two individuals disagree (denoting the cardinality of Q by the same symbol Q). Observe that an unfrequent category creates more distance than a frequent one, and that the distances blow up when the frequencies f_k are too small. For this reason, it is mandatory in MCA to group "unfrequent categories" — less than 5% as a rule of thumb — with other ones. This is what we have done above Error rates (cf. hereafter p.10).

Cloud of individuals. The $n(n-1)/2$ distances between individuals define a set of n points in a multidimensional space, called a cloud of individuals. The Q responses of an individual can be replaced by a K–ple of $\{1,0\}$: 1 if category is chosen and 0 if not (disjunctive encoding), hence the individual points lie in a K–dimensional space. Since there is only one category chosen per question, there are Q linear constraints, therefore the cloud of individuals actually lies in a $K-Q$ dimensional subspace.

Cloud of categories. The MCA distance $d(k,k')$ between two categories k and k' is given by the formula:

$$d^2(k,k') = \frac{n_k + n_{k'} - 2n_{kk'}}{n_k\, n_{k'}/n}$$

where $n_{kk'}$ is the number of individuals who have chosen both k and k'. These distances define a cloud of K category points, called a cloud of categories. The categories of question q lie in a $K_q - 1$ dimensional subspace (K_q denoting the cardinality of the set K_q), therefore the cloud of K category points lies in a $K-Q$ dimensional subspace (like the cloud of individuals).

Variance of cloud. The variance of a cloud is the weighted mean of the square of the distances of the points of the cloud from the mean point of the cloud. In MCA, the cloud of individuals and the cloud of categories have the same variance equal to $(K-Q)/Q$.

Principal axes. The principal axes of a cloud are defined by successively fitting lines to the cloud by the method of orthogonal least squares. The orientations of axes are arbitrary. The projections of the points of a cloud onto the first principal axis provide the best one–dimensional adjustment of the cloud (in the least square sense), or equivalently provide the maximum variance one–dimensional approximation of the cloud. The projections onto the first two principal axes, that is, onto the first principal plane, provide the best two–dimensional adjustment, and so forth. The usual geometric representations of the cloud are the orthogonal projections of the cloud onto the principal planes 1-2, 1-3, 2-3, etc.

Eigenvalues (denoted λ_1, λ_2, etc.). The eigenvalue associated with an axis is the variance of the projected cloud onto this axis; one also says variance of the axis or inertia of the axis[8]. There are $K-Q$ eigenvalues. The sum of the eigenvalues is equal to the variance of the cloud, i.e. $(K-Q)/Q$, hence the *variance rates* (or inertia rates) defined for each axis as the ratio of the eigenvalue by their sum.

Principal Coordinates. The coordinates of the points of a cloud along principal axes with unit–norm direction vectors are called principal coordinates. In GDA, geometric representations are done from principal coordinates. The coordinate of individual i along a principal axis is denoted y^i, and the set $(y^i)_{i\in I}$ defines the principal variable on I, whose variance is equal to λ (variance of axis). Similarly, one defines the principal coordinate y^k of response category k, and the principal variable $(y^k)_{k\in K}$ on K, whose variance is also equal to λ.

Transition Formulas in MCA. Let $K_i \subset K$ be the subset of categories chosen by individual i; then the coordinate y^i of individual point i on a principal axis is the mean of the Q coordinates y^k of categories $k \in K_i$ divided by $\sqrt{\lambda}$:

$$y^i = (1/\sqrt{\lambda})\Big(\sum_{k\in K_i} y^k/Q\Big)$$

Similarly, let $I_k \subset I$ be the subset of n_k individuals having chosen category k; then the coordinate y^k of category k on a principal axis is the mean of the coordinates y^i of individuals $i \in I_k$ divided by $\sqrt{\lambda}$:

$$y^k = (1/\sqrt{\lambda})\Big(\sum_{i\in I_k} y^i/n_k\Big)$$

Subclouds, Category mean–points. Given a category k, the subset of individuals that have chosen category k defines a subcloud of individuals with which there is associated its mean point, called a category mean–point. For a given axis, the principal coordinate of the category mean point is equal to $\sqrt{\lambda}\,y^k$ (y^k being the coordinate of category point k).

[8] Searching the axis of maximum variance amounts to determining the eigenvalues and eigenvectors of a symmetric endomorphism, hence the name of eigenvalues for the variances of axes.

1.3.2 Application to the EPGY Data Set

For the EPGY data set, we have $Q = 5 \times 3 = 15$ active variables. For Integers and Fractions, Category 3 of Error rates has a frequency less than 1% (cf. Table 1.2, p.5), therefore we have grouped this category with Category 2. As a result, the total number of categories is $K = 47 - 2 = 45$.

Distances. The distance between two individuals is the MCA distance defined earlier (p.7).

For example, for the two response patterns 11111 11111 11111 (low error rates, short latencies, and small number of exercises), and 22333 11111 22332 (high error rates, short latencies, and large numbers of exercises), the square of the distance is equal to (cf. Tables 1.2 p.5 and 1.4 p.1.4):
$\frac{533}{15}\Big(\big((\frac{1}{305}+\frac{1}{228})+(\frac{1}{274}+\frac{1}{259})+(\frac{1}{55}+\frac{1}{96})+(\frac{1}{31}+\frac{1}{135})+(\frac{1}{192}+\frac{1}{25})\big)+(0+0+0+0+0)+\big((\frac{1}{480}+\frac{1}{53})+(\frac{1}{482}+\frac{1}{51})+(\frac{1}{82}+\frac{1}{130})+(\frac{1}{165}+\frac{1}{67})+(\frac{1}{387}+\frac{1}{146})\big)\Big)$

Basic results of MCA. The basic results of MCA are the following: (i) the eigenvalues; (ii) the principal coordinates of the 45 categories (see Appendix Table 1.1, p.35) and of the 533 individuals (or of the 520 response patterns); (iii) the contributions of categories to axes (see Appendix Table 1.2 p.36 and §1.4.1 p.11); (iv) the geometric representations of the two clouds (categories and individuals).

Eigenvalues. There are $K - Q = 45 - 15 = 30$ eigenvalues, and the sum of eigenvalues $(K - Q)/Q$ is equal to 2.

	Axis 1	Axis 2	Axis 3	Axis 4
Eigenvalues (λ)	.3061	.2184	.1460	.1199
Raw rates of inertia	15.3%	10.9%	7.3%	6.0%
Modified rates	63.1%	25.4%	6.9%	3.1%
Cumulated modified rates	63.1%	88.5%	95.4%	98.6%

Table 1.5: Eigenvalues; raw and modified rates

How many axes to interpret? To assess the importance of axes, the modified rates of inertia (cf. Benzécri, 1992, p.412) give a better assessment of the importance of axes than the raw rates. Let $\lambda_m = 1/Q$ (mean eigenvalue, here .067), and $\lambda' = (\lambda - \lambda_m)^2$. Then the modified rate is equal to

$\lambda'/\sum\lambda'$, the sum being taken over the eigenvalues greater than λ_m[9]. The modified rates indicate how the cloud deviates from a *spherical cloud* (with all eigenvalues equal to the mean eigenvalue).

Looking at modified rates, it is clear that one axis is not sufficient (63.1%); whereas taking two axes brings the rate up to 88.5%. The following two axes have eigenvalues near to each other; taking four axes brings the rate to 99%. We will in any case interpret the first two axes, and attempt to interpret the next two ones.

1.4 Interpretation of Axes

Benzécri (1992, p.405) gives the following guideline: "Interpreting an axis amounts to finding out what is similar, on the one hand, between all the elements figuring on the right of the origin and, on the other hand between all that is written on the left; and expressing with conciseness and precision, the contrast (or opposition) between the two extremes". The Method of Contribution of Points and Deviations that we have devised (Le Roux & Rouanet, 1998) offers a guide along this line.

1.4.1 Aids to Interpretation

Contributions of points to the variance of a principal axis (Ctr). The proportion of the variance of an axis due to a point, denoted Ctr, is called the contribution of point to axis. This contribution is equal to the product of weight p by the square of coordinate y, divided by the variance λ of axis: $\text{Ctr} = p\,y^2/\lambda$. The weights of categories are proportional to their frequencies, that is, the weight p of category k is equal to f_k/Q[10].

Contribution of the deviation between two points. Let p and p' denote the weights of two points, y and y' their coordinates along a principal axis. The contribution of the deviation is given by the following formula:

$$\frac{pp'}{p+p'}\,(y-y')^2/\lambda$$

[9]For $Q = 2$, the modified rates are equal to the rates of the CA of the associated contingency table.

[10]The definition of contribution also applies to individuals, but is less interesting in MCA, because all individuals have the same weight $1/n$.

Contribution of a subset of points of the cloud. The contribution notions readily extend to a subset of points (subcloud). With a subcloud there are associated its weight (sum of the weights of its points), its mean point (with weight equal to the weight of the subcloud) and its variance. Three types of contribution are accordingly defined: (i) the global contribution of subcloud, which is the sum of the contributions of points; (ii) the contribution of the mean point of the subcloud; (iii) the within–contribution, which the product of its weight by its variance divided by λ.

Active vs supplementary elements. The individuals and the categories that have participated in the determination of axes are called active elements. One can also study supplementary individuals or categories which have not participated in the determination of principal axes, but whose corresponding points are projected on them, and whose coordinates are given by the transition formula.

1.4.2 First Interpretation of Axes of EPGY Data Set

Contributions of the 3 types of variables. Table 1.6 shows the relative contributions of the 3 types of variables. For Axis 1, Error rates account for 43% of the variance, the Number of Exercises for 30%, then the Latencies for 27%. For axis 2, latencies account for 68% of the variance; error rates and Number of Exercises account for only 18% and 14% respectively. Therefore the first axis is mainly the axis of error rates, and the second axis is mainly the axis of latencies. From Table 1.6 (axes 3 and 4), one notices that latencies account for 92% of the variance of axis 3, and error rates for 67% of the variance of axis 4. The third axis, which is specific of latencies, is a refinement of the second axis, and the fourth axis a refinement of the first one (cf. p.13).

	Axis 1	Axis 2	Axis 3	Axis 4
Error rate	.433	.184	.029	.671
Latency	.271	.677	.916	.053
Exercises	.296	.139	.055	.276
Total	1.	1.	1.	1.

Table 1.6: Contributions of the 3 types of variables

Contributions of 5 strands. Table 1.7 shows the contributions of the 5 strands. For the first 3 axes, the 5 strand contributions range from 12% to 26%, whereas for Axis 4, the contribution of Logic predominates (49%)[11].

	Axis 1	Axis 2	Axis 3	Axis 4
Integers	.196	.225	.191	.045
Fractions	.122	.219	.209	.020
Geometry	.186	.182	.254	.287
Logic	.259	.228	.208	.493
Measurement	.237	.146	.138	.155
Total	1.	1.	1.	1.

Table 1.7: Contributions of the 5 strands

Contributions of 15 variables. If one looks at the contributions of the 15 variables (Table 1.8, p.14) in detail, one sees that the Number of Exercises in Fractions and in Geometry hardly contribute (3% and 2%) to axis 1, that latencies of the 5 strands contribute almost equally to axis 2. If one now examines Figures 1.3-a, 1.3-b, 1.3c (p.15), corresponding to the 3 types of variables in plane 1-2, one notes that for each type variable there is a coherence between strands except for the Number of Exercises in Geometry (Figure 1.3-c).

If one examines plane 2-3 (Figure 1.4-b, p.16), one observes a Guttman effect for latencies, that is, the horizontal axis (axis 2) orders categories from right to left according to decreasing latencies; and axis 3 opposes medium categories (2 and 3) to extreme ones (1 and 4)[12]. Similarly, one observes a Guttman effect of error rates in plane 1-4 (Figure 1.4-a). These Guttman effects observed in planes 2-3 (latencies) and 1-4 (error rates) show that the data are essentially articulated around 2 scales: one of error rates (axis 1) and one of latencies (axis 2). We will now interpret axes 1 and 2 in detail.

1.4.3 Interpretation of Axis 1 ($\lambda_1 = .3061$)

The number of categories whose contributions to axis 1 are greater than average ($1/Q = 1/45 = .022 = 2.2\%$) is equal to 20 (cf. Appendix Table

[11] `The specificity of Logic appears also on the data about tests (see Andrew).`

[12] Strictly speaking, there is a Guttman effect if for axes ℓ and ℓ' one has the following relation between the principal variables $\forall i \in I : y^i_{\ell'} = a\big((y^i_\ell)^2 - \lambda_\ell\big)$ (Benzécri, 1992, p. 94). Finding a Guttman effect reveals that there is an ordinal scale underlying the data.

		Axis 1	Axis 2	Axis 3	Axis 4
Error Rate	Integers	.083	.043	.002	.004
	Fraction	.051	.034	.013	.001
	Geometry	.104	.035	.003	.273
	Logic	.096	.053	.005	.251
	Measuremt	.098	.018	.006	.142
Latency	Integers	.048	.159	.189	.004
	Fraction	.043	.157	.196	.013
	Geometry	.059	.123	.199	.008
	Logic	.066	.131	.200	.015
	Measuremt	.054	.106	.132	.012
Exercises	Integers	.065	.022	.000	.037
	Fraction	.028	.028	.000	.005
	Geometry	.023	.023	.051	.006
	Logic	.097	.044	.003	.227
	Measuremt	.084	.022	.000	.000
Total		1.	1.	1.	1.

Table 1.8: Contributions of the 15 variables

1.2, p.36); to which we will add the low error rate category for Logic, hence 21 categories, depicted on Figure 1.5 (p.17). On one side of axis (left on Figure 1.5), one finds the 5 low error rate categories (circles), and the small number of exercises categories in Logic and Measurement (squares). On the other side (right), one finds the 5 high error rate categories, the large Number of Exercises categories (except in Geometry), and the 5 short latencies categories (diamonds).

The interpretation of axis 1 will be based on these 21 categories, which account for 81% of the variance of axis. The opposition between high and low error rates is very important, and accounts for 35% of the variance of axis 1 (out of the 43% accounted for by all error rate categories). The contributions of short latency categories for the 5 strands are greater than average contribution (see Appendix, Table 1.2, p.36). These 5 categories are located on the right of origin (cf. Figure 1.5 and Table 1 in Appendix); which shows a link between high error rates and short latencies. The opposition between low error rates and short latencies accounts for 28% of the variance of axis 1, and the one between small and large numbers of exercises for 24%. The opposition between the 7 categories on the left of origin and the 14 on the right of origin accounts for 67% of the variance of axis 1.

The first axis is the axis of error rates and numbers of exercises.
It opposes on one side low error rates and small numbers of

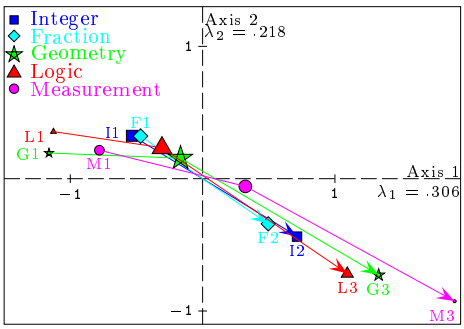

Figure 1.3-a. *Error rates* (plane 1-2).

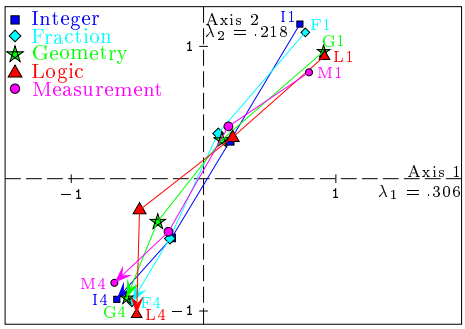

Figure 1.3-b. *Latencies* (plane 1-2).

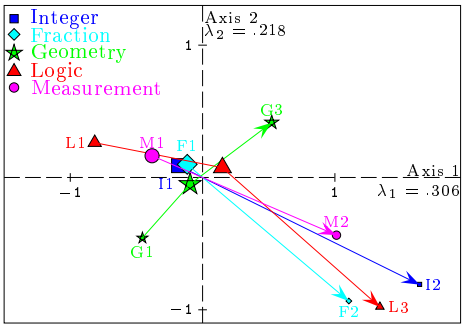

Figure 1.3-c. *Number of exercises* to mastery (plane 1-2).

Figure 1.3: Clouds of modalities (3 types of variables) in plane 1- 2

exercises and on the other side high error rates and large numbers of exercises, the latter being associated with short latencies.

1.4.4 Interpretation of Axis 2 ($(\lambda_2 = .2184)$

On Figure 1.6 (p.17), the 15 categories whose contributions to axis 2 are greater than average are depicted. At the top of the figure, one finds the 5 short latency categories. At the bottom of the figure, one finds the 5 long latency categories, the 3 high error rate categories in Integers, Logic and Geometry and the two large Number of Exercises categories in Fractions and Logic. These 15 categories account for 72% of the variance of axis 2. The interpretation of axis 2 will be based on these 15 categories. The opposition between short and long latency categories account for 55% of the variance of axis 2 (out of the 68% accounted for by all latency categories). The opposition between the 5 short latency categories and the 3 high error rate

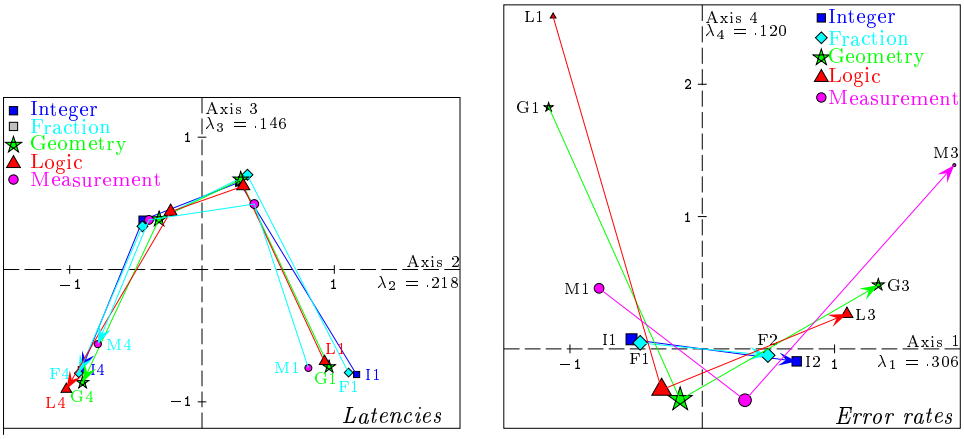

Figure 1.4: Guttman effects in cloud of categories in planes 2-3 (left) and 1-4 (right).

categories in Integers, Geometry and Logic accounts for 35% of axis 2; the opposition between the 5 short latency categories and the 2 large Number of Exercises in Fractions and Logic accounts for 21% of the variance of axis 2. More generally, the opposition between the 5 short latency categories and the 10 aforementioned categories accounts for 65% of the variance of axis 2.

> *The second axis is the axis of latencies.* It opposes short latencies and long latencies, the latter being associated with high error rates and large numbers of exercises.

1.4.5 Typical Response Patterns Emerging from Analysis

From the interpretations of axes and the distribution of categories in plane 1-2, the following response patterns emerge as typical patterns:
pattern 11111 11111 11111 (point A) (low error rates, short latencies, small number of exercises); pattern 11111 44444 11111 (point B) (Low error rates, long latencies, small number of exercises); pattern 22332 11111 22332 (point D) (high error rates, short latencies, large number of exercises); and Pattern 22332 44444 22332 (point C) (high error rates, long latencies, large number of exercises). Notice that none of the 533 individuals matches any one of these typical patterns.

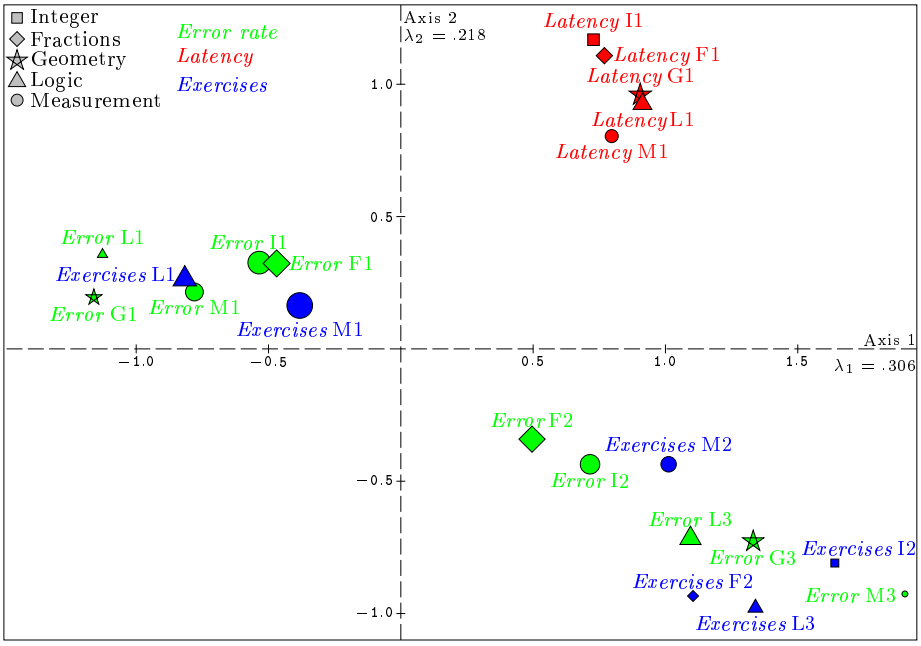

Figure 1.5: Interpretation of Axis 1.

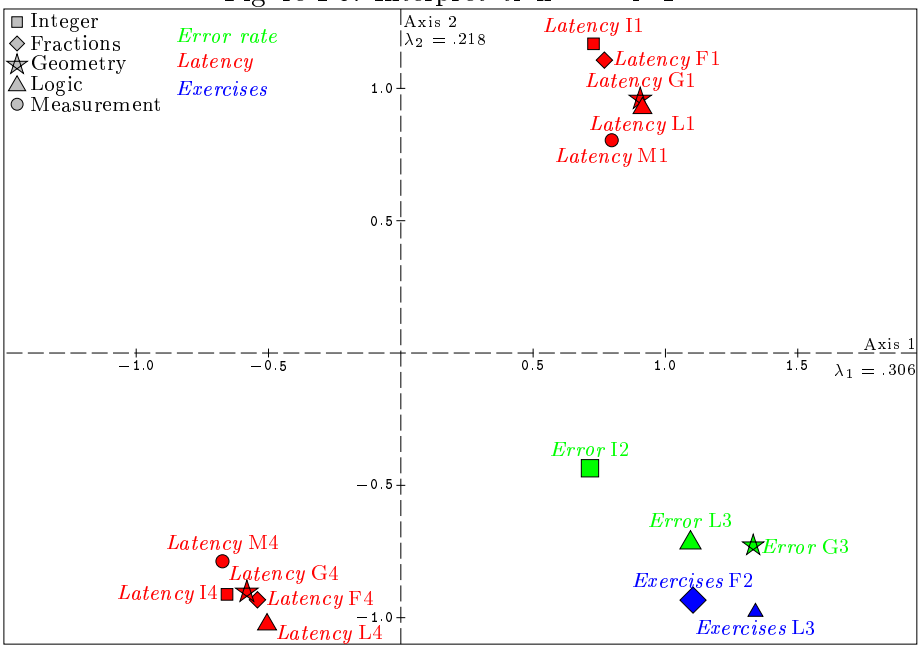

Figure 1.6: Interpretation of Axis 2.

1.5 Cloud of Individuals

1.5.1 Description of Cloud

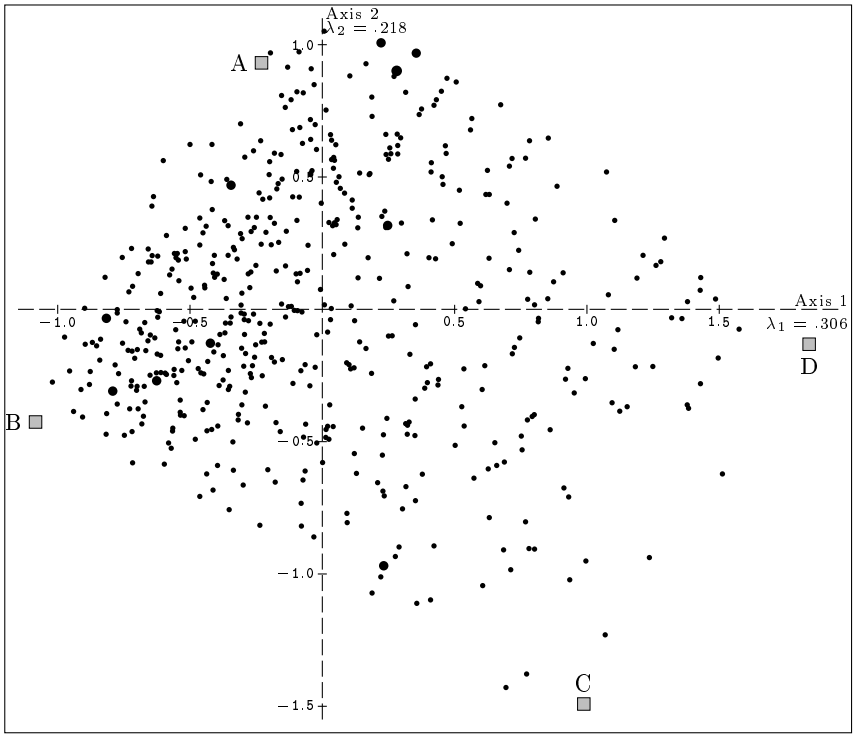

Figure 1.7: Cloud of individuals (patterns) with typical patterns.

Cloud in plane 1-2 (Figure 1.7). The cloud of individuals (533 students) is represented on Figure 1.7; it consists in 520 observed response patterns, to which we add the 4 typical response patterns. The individuals are roughly scattered inside the quadrilateral ABCD defined by the 4 typical patterns, with a high density of points along the side AB and a low density along the opposed side. This shows there are many students who make few errors whatever their latencies. On the other hand, students with high error rates are less numerous and very dispersed. The Table 1.9 (p.19) gives the variances of subclouds of individuals with low and high error rates in plane 1-2.

Error rates	Integers	Fractions	Geometry	Logic	Measurement
low	0.307	0.362	0.211	0.204	0.257
high	0.469	0.500	0.387	0.399	0.360

Table 1.9: Variances of subgroups of individuals with low and high error rates in plane 1-2

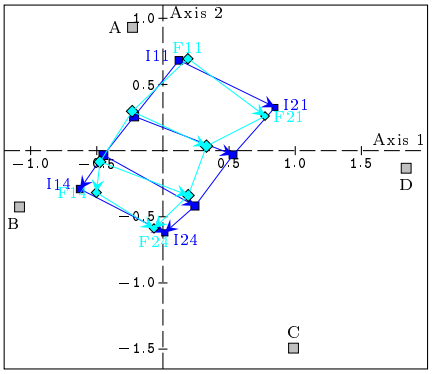

Figure 1.8-a: Error rates×Latencies for Integers and Fractions.

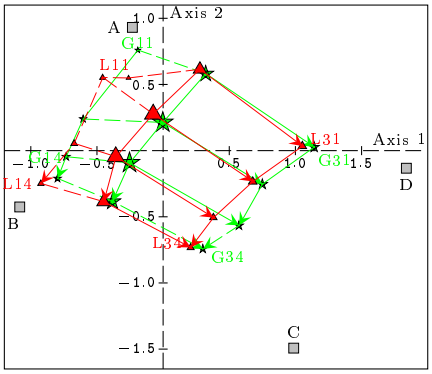

Figure 1.8-b: Error rates×Latencies for *Geometry* and Logic.

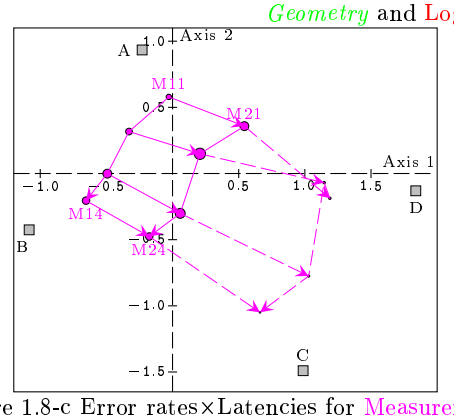

Figure 1.8-c Error rates×Latencies for Measurement

Figure 1.8: Cloud of individuals. Error rates×Latencies in plane 1-2.

If for each strand, we cross error rate and latency categories, with each composite category there is associated a subcloud of individuals with its mean point. Figure 1.8-a shows the 2×4 mean points for Integers and also the 2×4 mean points for Fractions; the marker sizes are proportional to the frequencies of subgroups. One notices, on Figure 1.8-a, that the 8 mean points for Integers are very close to the 8 mean points for Fractions, and that

these points are closer to side AB than to side CD of the quadrilateral; this means that error rates are globally low. Similarly, Figure 1.8-b shows the 3×4 mean points for Geometry and also for Logic. One notices the proximity of homologous points, with some mean points of small groups quite close to side AB, that is, corresponding to students with globally low error rates. Similarly, Figure 1.8-c shows the 3×4 mean points for Measurement. There are very few individuals with high error rates whose mean points are close to side CD, that is, corresponding to students with globally high error rates.

These three figures show well that quadrilateral ABCD is a frame that brings forth the following geometric model. When one goes down along the AB direction, latencies increase, while error rates remain constant; when one goes down along the AD direction, error rates increase, while latencies remain constant.

More on Guttman Effects. Guttman effects can be investigated in the cloud of individuals. Figure 1.9-a shows for *Error rates* in plane 1-4 the 45 observed category mean points (among the possible $2 \times 2 \times 3 \times 3 \times 3 = 108$ category mean points). The error rate scale appears very distinctly, showing a strong homogeneity across strands. Similarly Figure 1.9-b shows, for *Latencies* in plane 2-3, the 214 (among $4^5 = 1024$) observed category mean points (labels are written for patterns with frequency ≥ 9). The scale of latencies is not so sharply distinct, as some subjects have both short and long latencies across strands (for example 43441 and 34124).

Comparing Integers and Geometry. Is there any difference in students' behavior in Integers vs Geometry?

Among the 55 students who have a low *error rate* in Geometry, 52 have also a low error rate in Integers. More than half of these 52 students have long latencies (categories 3 and 4) both in Integers and in Geometry.

We have depicted in Figure 1.10 (plane 1-2) the category mean points associated with the observed combinations of *latencies* in Integers and Geometry (15 observed out of $4\times4= 16$ possible). If, for each latency category in Integers, one joins category points 1, 2, 3, 4 in Geometry (lines 11 through 14, 21 through 24, 31 through 34, 42 through 44 on graph on the left), the segments are roughly parallel to axis 1, that is, error rates decrease and latencies weakly increase; whereas if we do a similar construction for each error rate category (graph on the right), the segments are roughly parallel to axis 2, that is, latencies increase and error rates remain about steady.

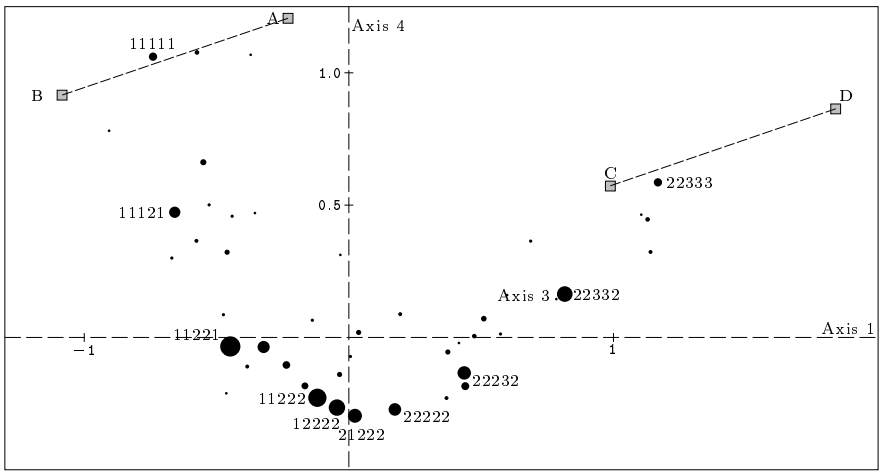

Figure 1.9-a: 44 Error rate patterns in plane 1-4

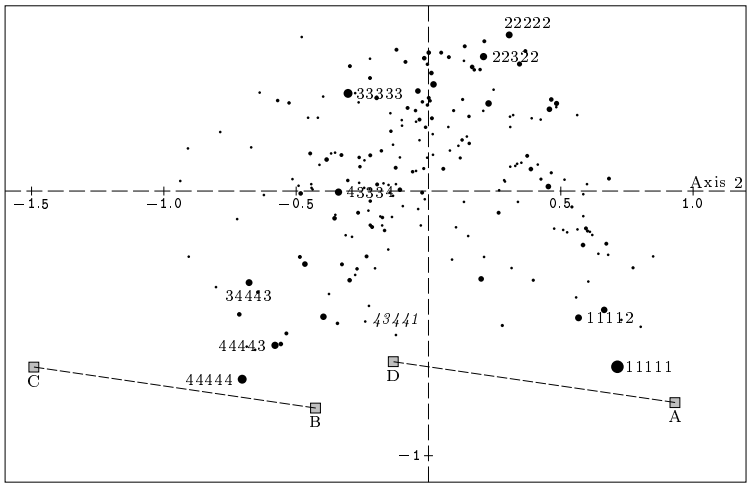

Figure 1.9-b: 214 latency patterns in plane 2-3

Figure 1.9: Guttman effects in cloud of individuals.

Such findings lead one to conclude that in order to improve performance in Geometry, some increase in latency is needed.

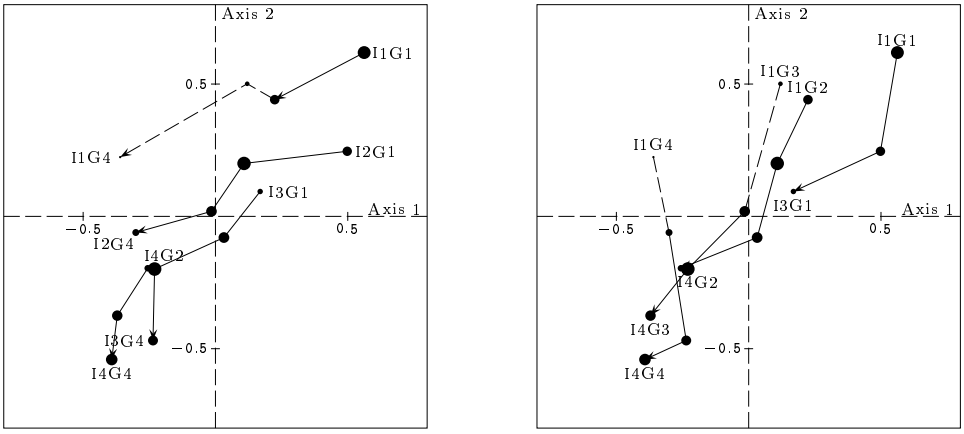

Figure 1.10: Comparing Integer and Geometry latencies.

1.5.2 Structured Data Analysis

Structuring factors. In an Individuals×Variables table, there are usually variables, such as individual identification variables (age, gender etc.), that are not used to construct the geometric space, but have a status similar to factors in a designed experiment. We call such variables Structuring Factors, and by Structured Data Analysis, we mean the integration, in Geometric Data Analysis, of the major procedures of analysis of variance (sources of variation, between–within decomposition, main effects, interaction effects, etc.), while allowing for the features specific to observational data, in the first place the fact that structuring factors are non–orthogonal, as a rule.

Subclouds, Category mean–point. The class of individuals that are in category k of a structuring factor defines a subcloud of individuals, with its mean point (category mean–point), its variance, its principal axes etc. For a given axis, the principal coordinate of the category mean–point is equal to $\sqrt{\lambda}\, y^k$, where y^k is the coordinate of category k in the cloud of categories.

Concentration Ellipses. The concept of concentration ellipse stems from classical statistics (Cramér, 1946)[13]. The concentration ellipse of a subcloud in a principal plane provides a geometric summary of the subcloud; its center

[13] For a bivariate normal distribution, the concentration ellipse contains 86% of the distribution.

is the mean point of the subcloud; the axes of the ellipse are the principal axes of the subcloud, each half–axis has for length 2 standard deviations of the subcloud in this direction.

Between and Within Variances. Given a partition of a cloud of individuals, the mean points of the classes of the partition defines a cloud whose variance is called between–variance of the partition. The weighted average of the variances of subclouds is called within–variance of the partition. The overall variance of cloud decomposes itself additively in between–variance plus within–variance.

Double decomposition of Variance. Given a set of sources of variation and the set of principal axes, the double decomposition consists in calculating the part of variance of each source of variation on each principal axis (for example, between and within).

1.5.3 Structured Analysis of EPGY Data Set

We now study the following structuring factors: Time spent on computer, Age and Gender, allowing for missing data (133, 57, 46 respectively).

Number of hours spent on computer. The number of hours spent on computer ranges from 6h30 through 248h, the median being equal to 21h20. For about 20% of students, time is less than 15h, and for about 20% it is greater than 34h. Eight students spent between 70h and 122h30, one student spent 248h.

The number of hours is correlated with the second axis ($-.731$, Spearman); globally, the shorter the latencies, the smaller the number of hours, as would be expected. If we code the number of hours into 4 classes (cuts at inferior quintile, median and superior quintile), one sees, on the table of double decomposition of the variance, that for the first axis, the within variance is much larger than the between one, and that for the second axis, the between variance is almost equal to the within variance.

variances	Axis 1	Axis 2
between	.0176	.1009
within	.2748	.1087

Table 1.10: Between and within Variances for Number of Hours.

Age and Gender. The ages of students, at the end of the course, range from 5 years to $11\frac{1}{2}$ years with a mode between 8 and 9 years. If one codes the age in 4 classes from the quartiles (7.76, 8.415 and 9.05), one observes an age effect with the 4 age categories ordered on axes 1 and 2. The deviation between the extreme age classes (85% of the variance of the age on the axis 1 and 78% on the axis 2) is equal to 0.89 SD of axis 1 (an important deviation) but only to 0.46 SD of axis 2 (a medium deviation). One can therefore say that, when age increases, error rates definitely increase and latencies slightly decrease (cf. Figure 1.11).

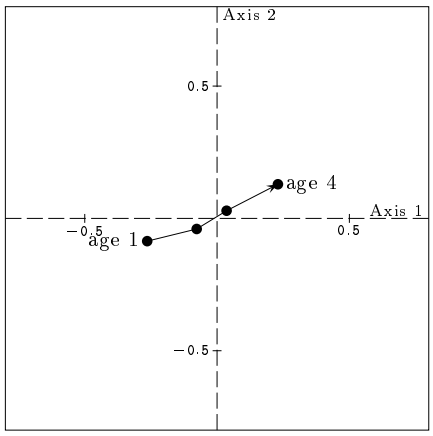

Figure 1.11: Mean points of the 4 age categories.

Nevertheless, within each age class, students are very scattered as shown in Table 1.11 of double decomposition of variance.

variances	Axis 1	Axis 2
between	.0317	.0065
within	.2684	.2102

Table 1.11: Between and within Variances for Age.

The students are composed of 283 boys and 204 girls. There is virtually no difference between boys and girls as shown in Table 1.12.

If we cross Age and Gender, generating $4 \times 2 = 8$ classes, one notices the large dispersion within the 8 classes in plane 1-2 — the within variance is equal to .4738, with a between–variance only equal to .0368. The interaction between Age and Gender is very low (see Table 1.13).

variances	Axis 1	Axis 2
between	.0001	.0012
within	.2927	.2112

Table 1.12: Between and within Variances for Gender.

	Axis 1	Axis 2
Age×Gender	.0306	.0099
Age	.0301	.0067
Gender	.0000	.0012
Interaction	.0006	.0016
Within (Age×Gender)	.2683	.2055
Total variance ($n = 468$)	.2989	.2154

Table 1.13: Double decomposition of variances for the crossing Age×Gender.

For each axis, the sum of the variances of Age and Gender factors and of their interaction is almost equal to the one of their crossing, which reflects that the crossing is nearly orthogonal (see Table 1.14).

	Boys	Girls	
$\cdot < 7.76$	68	49	117
$7.76 < \cdot < 8.415$	74	46	120
$8.415 < \cdot < 9.05$	66	51	117
$\cdot > 9.05$	60	54	114
Total	268	200	468

Table 1.14: Absolute frequencies for the crossing Age×Gender (468 students).

1.6 Euclidean Classification

1.6.1 Theoretical Sketch of Euclidean Classification

Purpose of Classification. The methods of classification — also called *cluster analysis* — consist in constructing classes (or clusters) of a set of objects, so that the objects within a same class are as close together as possible (compactness) whereas those belonging to different classes are as remote from one another as possible (separability). In GDA, the objects to be classified are the points of a cloud, and the classes of a classification are subclouds of the cloud. Cf. Benzécri (1992).

Hierarchical classification. It is a system of *nested classes*, represented by a *hierarchical tree* (after the pattern of natural science).

Agglomerative (or ascending) Hierarchical Classification (AHC). One starts with one–element classes, from which one proceeds to successive aggregations, until all objects are grouped within a single class. At each step of the construction, one starts with a partition and groups two classes of this partition.

Variance index. Given two classes c and c', the variance index is defined by $\delta(c, c') = \big(f_c f_{c'}/(f_c + f_{c'})\big) d^2(c, c')$, where $d^2(c, c')$ denotes the Euclidean distance between the centers of classes c and c'. When two classes c and c' of a partition are grouped together, the between–variance decreases from an amount equal to $\delta(c, c')$.

Euclidean classification. It is the AHC method performed on a Euclidean cloud and taking the variance index as aggregation index.

Level index. In a *Euclidean classification*, at each step ℓ, one starts with a partition, and the aggregated classes are those for which the variance index is minimum. The value of this minimum is called *level index* δ_ℓ. Thus at each step, the construction process leads to the minimal decrease of the between–variance of the partition (or equivalently to the minimal increase of the within– variance). As one ascends the construction, the level indices form an *increasing sequence*; the higher in the hierarchy, the higher the heterogeneity level where aggregation is made. The sum of the successive level indices is equal to the total variance. One thus gets a *decomposition of variance according to the hierarchy of classes.*

1.6.2 Classification of the EPGY Data Set

We have made a Euclidean classification of individuals. Figure 1.12 shows the superior tree resulting in the partition in 6 classes (ce1, ... ce6) as well as the sequence of level indices. Clearly, two partitions emerge: a three–class partition generated by 2 successive dichotomies, and a six–class partition generated by 5 dichotomies.

We will comment on the successive dichotomies leading to these two partitions; then, we amend the partition in 6 classes of the AHC to get a

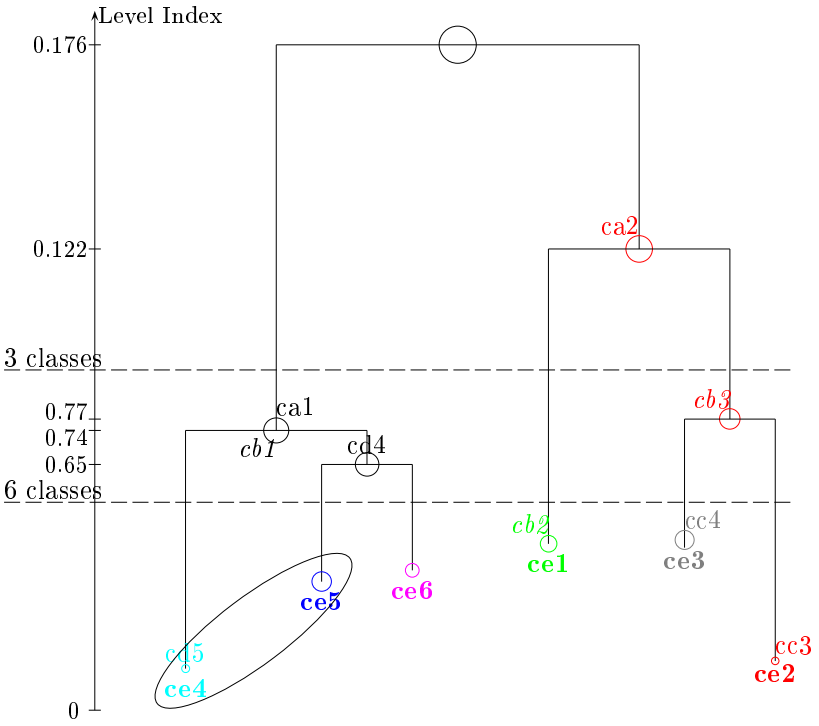

Figure 1.12: Superior hierarchical tree resulting in six-class partition.

partition in 5 classes that we will retain as a final summary. The 5 successive partitions of AHC will be designated by CA (2 classes ca1 and ca2), CB (3 classes cb1, cb2 and cb3), CC (4 classes cc1, ... cc4), CD (5 classes) and CE (6 classes), the final partition in 5 classes will be designated by C.

Partition in 3 classes

The *first dichotomy* (level index .176, partition CA) separates the students in 2 classes ca1 of 255 students and ca2 of 278 students (cf. Figure 1.13- a), and is characterized as follows:

Error rates are twice as small in class ca1 as in class ca2. The students that have low error rates are a large majority in class ca1: they are 196/305 for Integers, 169/274 for Fractions, 48/55 for Geometry, 28/31 for Logic and 160/192 for Measurement.

The *latencies* are almost all superior to the first quintile for class ca1 and inferior to the last quintile the class ca2. The students that have long

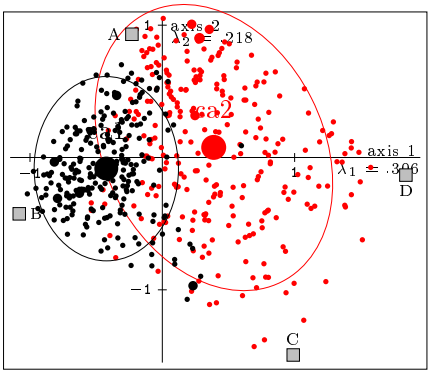

Figure 1.13-a: two–class partition

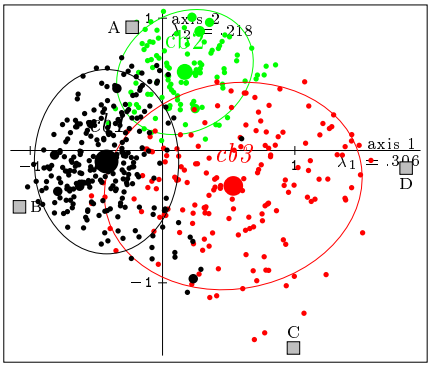

Figure 1.13-b: three–class partition (CB).

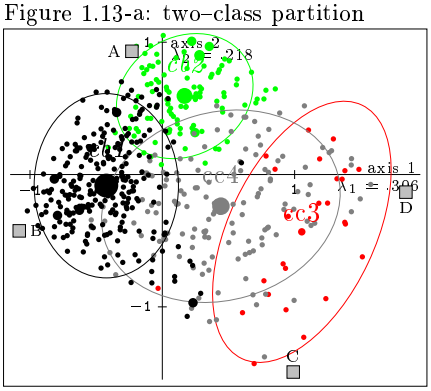

Figure 1.13-c: four–class partition (CC).

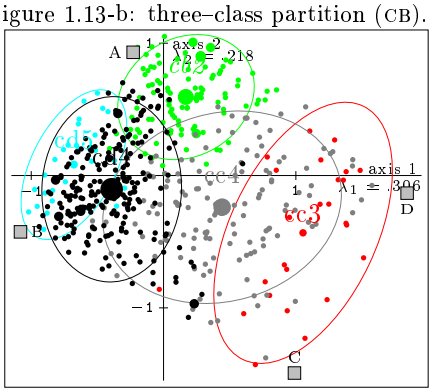

Figure 1.13-d: five–class partition (CD).

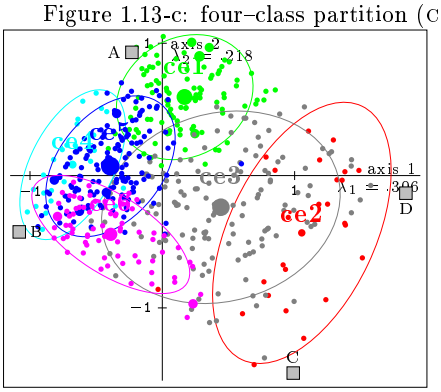

Figure 1.13-e: six–class partition (CE).

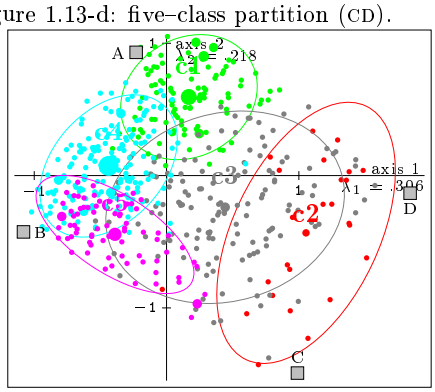

Figure 1.13-f: Final five–class partition (C).

Figure 1.13: Ellipses of classes of successive partitions.

latencies (superior to the fourth Quintile) are about 80% in class ca1. More than 70% of the students in class ca1 have latencies above median in all strands.

The average *number of exercises* for class ca1 is always inferior to the one for class ca2. Among the student that need more than 4 exercises, in class ca1, there are 2/53 for Integer, 2/51 for Fractions, 18/146 in Logic; and 38/130 need more than 12 exercises in Geometry and 6/61 do more than 5 exercises in Logic. Class ca1 is characterized by students with low error rate; it includes almost all students with long latencies and who need few exercises to mastery. An extreme pattern of this class is 11111 44444 11111 (point B).

The *second dichotomy* (level index .122) generates the three–class partition CB. It comprises class cb1 (alias ca1) and splits class ca2 into two classes cb2 of 111 students and cb3 of 167 students (Figure 1.13-b, p.28). It separates out class cb2. Class cb2 is rather compact; its first characteristic is that almost all students of this class (except 3 in Fractions and in Measurement, 9 in Geometry and 15 in Logic) have latencies below median. The average error rates are situated around the median; 2/3 of students have error rates inferior to .02 in Integers and Fractions, and are between .02 and .10 in the 3 other strands for almost all (respectively for 104, 100 and 85). Except in Geometry, the students of this group need few exercises for mastery: none does more than 4 in Integers and Fractions; 22 more than 4 in Measurement; in Logic 86 do 5 and 24 do less of 5; in Geometry, almost all (except 7) do more than 11 exercises. This class is therefore characterized by short latencies, medium error rates, and small numbers of exercises except in Geometry. For class cb3, the distribution of latencies is for each strand near the overall distribution, with under-representation of long latencies. The between and within variances of partition CB on the first 2 axes are given in Table 1.15.

	Axis 1	Axis 2
Between: CB	.1813	.0993
Within: I(CB)	.1248	.1191

Table 1.15: Between and within variances for the three–class partition CB.

The between–variance is very superior to the within–variance for axis 1 and slightly inferior for axis 2.

To sum up (cf. Table 1.16): The partition in 3 classes contains

a class (cb1) with low error rates (and rather long latencies), a class (cb2) with short latencies and small numbers of exercises to mastery and a class (cb3) with high error rates.

	Frequencies	Error rates	latencies	Exercises
cb1	255	low	rather long	rather small
cb2	111		short	small except in Geometry
cb3	167	high	medium	rather large

Table 1.16: Synopsis of three–class partition CB.

The subsequent dichotomies lead to refining the partition in 3 classes by subdividing the classes of low and high error rates and with large dispersion of latencies.

Partition in 6 classes

The *third dichotomy* (level index .077 $\ll$ .122) generates the four–class partition CC. It splits up class cb3 into two classes cc3 with 25 students and cc4 with 142 students; both classes are very dispersed (cf. Figure 1.13-c, p.28). Class cc3 is characterized by students who have almost all high error rates and who do large numbers of exercises; latencies are scattered on all categories with a weak majority with latencies inferior to the median.

Class cc4 is characterized by high error rates but on the average inferior to those of class cc3, the proportions of students with extreme latencies are comprised between 10% and 18% ($< 20\%$), except for short latencies in Logic and Measurement (23%). The students who need many exercises to master a notion are mostly in this class (34/53 in Integers, 43/51 in Fractions, 48/67 in Logic and 81/146 in Measurement).

The *fourth dichotomy* (level index .074) generates the five–class partition CD. It splits class cc1 (alias ca1, cb1) in class cd4 (227 students) and class cd5 (28 students) (cf. Figure 1.13-d, p.28). The class cd5 is composed of students who have, in large majority, low error rates (they are 21 in fractions, 28 in Logic, 17 in Geometry, and for no strand the error rate is greater than .10); with medium latencies (between the first quintile and the fourth quintile), and not very large numbers of exercises (27 or 28 do only 4 exercises, except in Geometry where 22 students do 11 or 12 exercises and only 5 students do less than11).

The *fifth dichotomy* (level index .065) generates the six–class partition CE. It divides class cd4 in two classes ce5 (150 students) and ce6

(77 students) (cf. Figure 1.13-e, p.28). For class ce5, one has low error rates, medium latencies (proportions of short latencies vary between 5% and 11% $\ll$ 20% and the long ones between 15% and 19%). This class is very close to class ce4, with nevertheless error rates slightly greater than average. Class ce6 is the one of the long latencies (above median for all the students), and with proportions of small numbers of exercises superior to the proportion over the 533 students except in Logic: cf. Table 1.17 (p.31).

Strands		Class ce6 ($n = 77$)	All (n=533)
Integer	< 4 exercises	97%	90%
Fractions	< 4 exercises	99%	90%
Geometry	< 11 exercises	34%	14%
Logic	< 5 exercises	29%	31%
Measurement	< 4 exercises	83%	73%

Table 1.17: Class ce6 of the six–class partition.

Amendment: Final partition in 5 classes

The concentration ellipses of classes ce4 and ce5 in the plane 1-2 appear to be quite near to each other, which invites grouping these two classes together (Figure 1.13-f, p.28). The partition in 5 classes (c1, c2, c3, c4, c5) thus obtained has a between-variance on the 4 first axes equal to 0.3863, which is greater than the variance of the partition in 5 classes of the hierarchical classification (= 0.3676). The between–variance and the within–variance of this partition in 5 classes on the two first axes are given in the table 1.18.

	Axis 1	Axis2
Between–Variance	.1964	.1277
Within–Variance	.1097	.0907

Table 1.18: Between and within variances for the final five–class partition.

The between–variance is greater than the within–variance. We will summarize the data with this partition in 5 classes: see Table 1.19.

There are two compact classes of well–performing students; one class is close to point A (class $c1$) with short latencies and medium error rates, and the other one is close to point B (class c4) with rather low error rates and medium to long latencies. Class c4 includes students with low error rates, especially in Geometry (46/55) and in Logic (28/31).

	frequencies	Error rates	Latencies	Exercises
c1	111		short	small except in Geometry
c2	25	high		rather large
c3	142	high		
c4	178	low		rather small
c5	77		long	rather small

Table 1.19: Synopsis of final five–class partition.

1.7 Conclusions

Starting from the three types of variables (Error Rates, Latencies, Number of Exercises) and from the five strands (Integers, Fractions, Geometry, Logic, Measurement), we have used MCA to construct a geometric space of individual differences for the gifted students of grade 3. The geometric analysis shows a good homogeneity of strands for each type of variable. It also shows that the individual differences are articulated around two scales: one of error rates and number of exercises and one of latencies. The error rate scale is clear–cut showing strong homogeneity; the one of latencies is not so sharp as some subjects have both short and long latencies across strands.

MCA provides a geometric summary of data. The individual points are scattered within a quadrilateral ABCD: When going down along the AB direction, latencies increase, while error rates remain constant; when going down along the AD direction, error rates increase, while latencies remain constant. The scattering of points within the quadrilateral is not uniform, showing a low density along side CD and a high density along AB.

A Euclidean classification of individuals has been performed leading to a five–class partition. There are two compact classes of well–performing students; one class is close to point A (class $c1$) with short latencies and medium error rates, and the other one is close to point B (class c4) with rather low error rates and medium to long latencies (a profile little encouraged by current standards of educational testing).

Once the geometric space is constructed, one can study in detail the cloud of individuals by means of structuring factors. In the present study, we have investigated Age and Gender. There is an effect of Age; when age increases, error rates increase and latencies slightly decrease. There is virtually no difference between boys and girls. One could enrich these analyses by introducing further structuring factors, such as the scores on final tests.

References

Benzécri J-P. & al (1973). *L'analyse des données. 1. Taxinomie 2. Analyse des correspondances.* Paris, Dunod.

Benzécri J-P. (1992). *Correspondence Analysis Handbook.* New York, Dekker.

Bourdieu P. (1999). Une révolution conservatrice dans l'édition. *Actes de la recherche en Sciences Sociales* p. 3-28.

Chiche J., Le Roux B. , Perrineau P. & Rouanet H. (2000). L'espace politique des électeurs français à la fin des années 1990. *Revue française de science politique*, vol. 50, n° 3, p. 463-487.

Cramér H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

Gower J-C. & Hand D. (1996). *Biplots.* London, Chapman & Hall.

Greenacre M. (1984). *Theory and Applications of Correspondence Analysis.* London, Academic Press.

Lebart L., Morineau A. & Warwick K-M. (1984). *Multivariate Descriptive Statistical Analysis : Correspondence Analysis and related Techniques for large matrices.* London, Wiley.

Le Roux B. & Rouanet H. (1984). L'analyse multidimensionnelle des données structurées. *Mathématiques et Sciences Humaines*, 85, 5-18.

Le Roux B.& Rouanet H. (1998). Interpreting axes in Multiple Correspondence Analysis : Method of the contributions of points and deviations, in Blasius & Greenacre (Eds). *Visualization of Categorical Data.* San Diego, Academic Press.

Le Roux B.& Rouanet H. (forthcoming). *Geometric Data Analysis: from Correspondence Analysis to Structured Data.* Dordrecht, Kluwer.

Murtagh F. (1981). Recherche d'un scalogramme sur les réponses de 1300 élèves à une batterie d'épreuves de mathématiques, *Les Cahiers d'Analyse des Données*, Vol. VI, n° 3, 297-318.

Rouanet H. & Le Roux B. (1993). *Analyse des Données Multidimensionnelles.* Paris, Dunod.

Software

To analyze this data set we have used: SPSS for coding and elementary statistics, ADDAD for MCA, and EyeLID for structured data analyses. An extensive (though limited in data table size) version of ADDAD (Association pour le Développement et la Diffusion de l'Analyse des Données) and a DOS–version of EyeLID (a program for graphical inspection of multivariate data) are available on the following ftp:

`ftp.math-info.univ-paris5.fr/pub/MathPsy/AGD`

ADDAD, EyeLID and `ellipse` programs can be download from the Brigitte Le Roux's homepage:

`http://www.math-info.univ-paris5.fr/~lerb`

(under the "Logiciels" heading).

APPENDIX

			$Axis1$	$Axis2$	$Axis3$	$Axis4$
Error	Integers	$I1$	−.535	.326	−.054	.071
		$I2$	.715	−.436	.073	−.095
	Fractions	$F1$	−.469	.322	−.165	.045
		$F2$	.496	−.341	.174	−.048
	Geometry	$G1$	−1.160	.194	−.191	1.826
		$G2$	−.167	.155	.049	−.384
		$G3$	1.331	−.727	−.084	.482
	Logic	$L1$	−1.128	.356	.332	2.513
		$L2$	−.307	.234	−.063	−.309
		$L3$	1.094	−.717	.095	.264
	Measure−	$M1$	−.780	.215	−.093	.458
	ment	$M2$	.323	−.058	.084	−.388
		$M3$	1.905	−.926	−.349	1.389
Latency	Integers	$I1$	.728	1.168	−.794	.043
		$I2$	.199	.283	.660	.105
		$I3$	−.242	−.448	.374	−.105
		$I4$	−.657	−.912	−.759	−.042
	Fractions	$F1$	.769	1.107	−.779	.163
		$F2$	.109	.342	.715	−.033
		$F3$	−.256	−.451	.325	.108
		$F4$	−.542	−.933	−.784	−.274
	Geometry	$G1$	.905	.960	−.737	.211
		$G2$	.137	.292	.679	−.059
		$G3$	−.348	−.325	.380	.015
		$G4$	−.582	−.902	−.854	−.143
	Logic	$L1$	.913	.925	−.695	.180
		$L2$	.218	.311	.630	−.047
		$L3$	−.484	−.237	.437	.118
		$L4$	−.506	−1.026	−.906	−.284
	Measure−	$M1$	.797	.804	−.744	.218
	ment	$M2$	.188	.395	.495	−.201
		$M3$	−.265	−.401	.375	.023
		$M4$	−.674	−.787	−.564	.050
Exercises	Integers	$I1$	−.181	.089	−.002	−.086
		$I2$	1.640	−.810	.022	.779
	Fractions	$F1$	−.117	.099	−.009	−.032
		$F2$	1.104	−.934	.084	.304
	Geometry	$G1$	−.456	−.457	−.723	−.068
		$G2$	−.095	−.051	.223	.078
		$G3$	.522	.414	−.095	−.151
	Logic	$L1$	−.816	.263	.071	.736
		$L2$	.149	.074	−.071	−.561
		$L3$	1.340	−.979	.146	.710
	Measure−	$M1$	−.382	.164	.004	−.013
	ment	$M2$	1.012	−.436	−.011	.035

Table 1.1: coordinates of the 45 categories.

			$Axis1$	$Axis2$	$Axis3$	$Axis4$
Error	Integers	$I1$	.036	.019	.001	.002
		$I2$	.048	.025	.001	.002
	Fractions	$F1$	.025	.016	.006	.001
		$F2$	.026	.017	.007	.001
	Geometry	$G1$	.030	.001	.002	.191
		$G2$	.004	.005	.001	.059
		$G3$	.069	.029	.001	.023
	Logic	$L1$	.016	.002	.003	.204
		$L2$	.014	.011	.001	.037
		$L3$	.066	.040	.001	.010
	Measure–	$M1$	.048	.005	.001	.042
	ment	$M2$	.013	.001	.002	.050
		$M3$	.037	.012	.003	.050
Latency	Integers	$I1$	.023	.083	.057	.000
		$I2$	.003	.007	.060	.002
		$I3$	.004	.018	.019	.002
		$I4$	.019	.051	.053	.000
	Fractions	$F1$	.026	.074	.055	.003
		$F2$	.001	.011	.070	.000
		$F3$	.004	.019	.015	.002
		$F4$	.013	.053	.056	.008
	Geometry	$G1$	.036	.056	.049	.005
		$G2$	.001	.008	.063	.001
		$G3$	.008	.010	.020	.000
		$G4$	.015	.050	.067	.002
	Logic	$L1$	.036	.052	.044	.004
		$L2$	.003	.009	.054	.000
		$L3$	.015	.005	.026	.002
		$L4$	.011	.065	.075	.009
	Measure–	$M1$	.028	.039	.050	.005
	ment	$M2$	.002	.014	.034	.007
		$M3$	.005	.015	.019	.000
		$M4$	.020	.038	.029	.000
Exercises	Integers	$I1$	.006	.002	.000	.004
		$I2$	.058	.020	.000	.034
	Fractions	$F1$	.003	.003	.000	.001
		$F2$	.025	.025	.000	.005
	Geometry	$G1$	.007	.010	.037	.000
		$G2$	.001	.000	.014	.002
		$G3$	.014	.013	.001	.003
	Logic	$L1$	.045	.007	.001	.093
		$L2$	.003	.001	.001	.099
		$L3$	.049	.037	.001	.035
	Measure–	$M1$	.023	.006	.000	.000
	ment	$M2$	.061	.016	.000	.000

Table 1.2: Contributions of the 45 categories.