Dottorato di ricerca in
Metodologia Statistica per la Ricerca Scientifica
XIX ciclo

Università di Bologna

Item Response Theory models for the competence evaluation: towards a multidimensional approach in the University guidance

Mariagiulia Matteucci

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2007

Università di Bologna

Dottorato di ricerca in
Metodologia Statistica per la Ricerca Scientifica
XIX ciclo

Item Response Theory models for the competence evaluation: towards a multidimensional approach in the University guidance

Mariagiulia Matteucci

Tutor
Chiar.mo Prof. Stefania Mignani

Coordinatore
Chiar.mo Prof. Daniela Cocchi

Settore Disciplinare SECS-S/01

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2007

Preface

The concept of student guidance, intended as an instrument able to give an effective support to decision in the educational training, is assuming a relevant position in the University context. This work deals with the entry guidance, that is from secondary school to University, presenting the results of a new guidance test realized in collaboration with the Guidance Service of the University of Bologna. The test is composed of general culture and faculty-specific multiple-choice items and can be classified as a competence test.

The main purpose of this work is to deal with a competence evaluation problem applied to a guidance context by using Item Response Theory (IRT), a modern methodology for item analysis. The IRT is a measurement theory which goes back to the thirties, has been developed in the sixties but has been intensively applied only recently, especially in the educational field. The basic idea of IRT is that the responses to a set of items (test) can be explained by the existence of one or more latent traits, denominated *abilities*. Typically, IRT models express the probability of a specific response alternative as a function of abilities and item parameters.

The IRT will be used in two direction: the item calibration and the multidimensional investigation.

The item calibration will be conducted under the unidimensionality assumption, that is only one latent ability is underlying the response process. The two sections of the tests (the general part and the specific one) will be considered separately to preserve unidimensionality. Two IRT models will be simultaneously implemented for the item calibration: the Multiple-choice model (MCM) and the Three parameter logistic (3PL) model. The first one works on multiple-choice data in order to understand the behavior of all the response alternatives while the second one is applied for the estimation of item parameters in terms of discrimination, difficulty and guessing. The 3PL model restricts the data to the binary case, that is "correct" and "wrong" responses. Future developments of the project will include the characterization of the abilities for examinees who produce a particular response pattern, taking the item parameters as known.

Then, we will explore the possibility of estimating the item parameters for the complete test, that is for the two sections simultaneously. The assumption of unidimensionality will be relaxed with the introduction of two abilities underlying the response process. In fact, the multidimensional approach seems to be more efficient when calibrating several unidimensional tests. Two different estimation methods will be considered: the Full-information factor analysis (FIFA) and the Markov chain Monte Carlo procedure. The first one implements the maximum likelihood estimation while the second one employs a Bayesian approach for the simulation of correlated chains. In particular, advantages of using the Gibbs sampler in the MCMC framework respect to the ML estimation will be underlined.

Finally, the problem of *incomplete design* will be considered. In fact, the general culture section of the test has been submitted with a block design. Considering an item bank of 30 items, only a randomly selected block of 10 items is presented to each individual. Therefore, we have two different types of missing data: the omitted and the not presented items. The not presented items create an incomplete design for the guidance test. A possible solution is to recode the omitted items as incorrect responses and to eliminate the not presented items from the estimation process. This procedure is already implemented in the IRT standard software for FIFA. On the other hand, the MCMC methods for IRT have been recently investigated. The incomplete design will be implemented both for unidimensional and multidimensional models in the MCMC framework.

The main contributions of this work are the extension of the IRT calibration to a guidance context, the investigation of multidimensionality and the implementation of the incomplete design in MCMC by using the Gibbs sampler algorithm. All the methods and models will be discussed together with real applications on the guidance data.

Structure of the thesis The first chapter will introduce the guidance context referred to the University guidance in the entrance phase. The first section will give some information about the situation in the University of Bologna while the second one will present the new guidance project. The design of the test and the data features will be discussed.

The second chapter will present the use of Item Response Theory (IRT) for item calibration. First of all, a concise review of IRT will be given. Then the test development and calibration will be discussed adopting a IRT perspective. Two fundamental unidimensional IRT models for the item calibration (the Multiple Choice model and the Three-parameter logistic model) will be introduced in terms of specification, interpretation and estimation. This section will also consider the problem of incomplete design. Furthermore, the calibration process applied to the guidance test will be discussed together with the results for the Psychology faculty. Finally, general conclusions on the test calibration for all the faculties involved in the project will be illustrated.

In the third chapter the IRT multidimensional approach will be introduced: main features and advantages will be explained. Then, the Full-information factor analysis (FIFA) will be described as a multidimensional method for the item parameter estimation. An application on the Psychology faculty will be considered.

The fourth chapter will describe an alternative multidimensional approach by using the Markov chain Monte Carlo (MCMC) methods. Bayesian estimation will be introduced together with a review on the MCMC methods. Then, the implementation of the Gibbs sampler to the unidimensional and multidimensional IRT models will be discussed. The simulation process will be described in detail and the incomplete design will be included in the Gibbs sampler algorithm.

In the fifth chapter conclusions and further research on applicative and methodological aspects will be discussed.

Acknowledgements

This thesis is the result of my three-year period of research at the Statistics Department of the University of Bologna.

First of all, I would like to thank Prof. Stefania Mignani for her supervision, accurate suggestions and constant guidance. She involved me in this interesting project and gave me the chance to learn and experience really much, trusting my capabilities. She always encouraged me to do my very best.

I am particularly grateful to Prof. dr. Bernard P. Veldkamp, for making possible my period of study in the Department of Research Methodology, Measurement and Data Analysis at the University of Twente (The Netherlands). I especially thank him for his advices on how to proceed, for the way he solved my doubts about IRT and guided me into the different stages of the project.

I especially thank Silvia Cagnone and Roberto Ricci for their accurate corrections, constructive criticism and their constant and encouraging support.

I wish to thank Prof. Paola Monari and the Guidance Service of Bologna University, for involving me in the project and entrusting me with the data analysis.

My period of study at Twente was crucial for my research and for the development of this thesis, so I would like to thank all the people who made it possible. A special thank to Ya-Hui Su for being a kind officemate and still a sincere friend. I also thank Rinke Klein Entink and Anke Weekers for our useful discussions about different computer programs.

I also would like to thank Yetunde Kolade and Steffi Pohl who experienced the first stages of my PhD, while I was in Germany, and provided me with encouragement and stimulus to go ahead.

I am really indebted to my friends, as well as my PhD colleagues, Laura Sardonini and Valentina Adorno, who helped me in this crucial and intensive period. Many thanks for listening to my complaints and frustrations, for supporting me and especially for tolerating my moody behaviors. I have never felt alone for all the three years and this is mainly their merit. I also would like to thank Marta Disegna, Caterina Liberati and Giula Roli for the precious relief they provided during all lunches and talks.

A special thank to Luisa Stracqualursi for understanding me, especially in the last stages of the project.

Finally, I would like to thank my family, for their patience and unconditional support, and all my friends, who were able to believe in me without understanding. The last thank goes to Stefano, for being close to me in this short but very intense period.

Contents

List of Tables

List of Figures

Chapter 1

Student guidance: the development of a new test in the University of Bologna

1.1 Student guidance situation in Bologna

Guidance can be defined as an individual process able to develop instruments for decision, education and vocational training. The definition of guidance can also be extended to the set of acts and interventions on individuals with the main aim of both supporting transitions in the educational and professional life and developing new competencies in the subjects.

In this work the University guidance will be taken into account. Mainly, there are three important phases in the student's guidance:

- Entry guidance;
- *In itinere* guidance;
- Exit guidance.

The first one refers to the comparative choice of the faculty, that is guidance from secondary school to University. The second one represents the guidance during the University period while the last one is from University to the working activity. The three steps should be separated and committed to qualified personnel.

The University guidance should not be simply information about the different faculties and courses but should provide an effective instrument to facilitate the student's decision and awareness.

The main interest of this work is on the entry guidance. To this end, the University of Bologna has developed the following instruments:

- Guidance days;
- Open day;
- Guidance Service;
- Achievement tests.

The guidance days represent the point of connection between high school students and the University of Bologna. The meeting is organized in the fair of Bologna and lasts three days. All the faculties have the possibility to present their features to students and high school teachers can be informed as well. The event is usually very successful, because students can get much information about all the faculties. If they are completely undecided, they have the opportunity to learn about different academic programmes while, if they already have a favourite faculty, they can get precise information from University enrolled students, Phd students and professors. Guidance is also oriented to high school teachers, to make them aware of all different University paths.

The Open day is a faculty-specific instrument; in fact, the single faculty may invite students that are interested in that specific subject and provide useful information.

The Guidance Service of Bologna is composed of psychologists, able to support students in their University path. The guidance is conducted through individual interviews and psychological and aptitude questionnaires.

Finally, some faculties have prepared achievement tests that are useful either for the admission tests, where scheduled, or to improve the knowledge of students about topics they will deal with. Information represents a very important starting point for a correct guidance process.

1.2 The guidance project

The student guidance in the University of Bologna is a quite complex issue. This is a direct consequence of the wide structure of the University, in terms of academic programmes and geographical configuration. The data referred to the 2006/2007 academic year show that the University of Bologna is made up of 23 different faculties containing 268 degree courses. The number of enrolled students is around 85.000. Furthermore, Bologna has a multi-campus University, with a central structure and four separated campus in Cesena, Forlì, Ravenna and Rimini.

The guidance process is very important especially to prevent students' dropping out and to introduce them into a formative path able to provide a satisfactory professional career. Many initiatives have been developed to provide a suitable help to high school students and recently, the problem of student guidance has been raised by the media. For all these reasons, in 2005 the Guidance Service of Bologna University decided to develop a new project.

The main aim of the project is to provide secondary school students with an effective instrument able to steer them into the University choice: a set of items with a faculty-specific content. From here the idea of a new type of guidance test came. The test results should be able to give students an indication on their knowledge about topics dealing with the favourite faculty(ies). Therefore the distinguishing feature is that the set of items is able to evaluate competencies instead of aptitudes. The test is not an entrance test.

Originally, only six faculties joined the project: Arts and Humanities, Economics (Forlì campus), Pharmacy, Political Science, Psychology and Statistics. In the second phase of the project also Agriculture, Foreign Languages and Literature and Education Science decided to be involved in it. Future developments include the extension of the project to other faculties of the University of Bologna.

The following sections will describe the project and the test in detail and the data collected according to the project development.

1.2.1 The project and the test development

The guidance project has been developed around the idea of building a simple and effective instrument able to investigate students' interests and capabilities.

The competence evaluation has been highlighted as a more powerful instrument in predicting the student's performance respect to the usual aptitude evaluation. This is valid both for education and for working activities (Evers et al., 2005). The efficacy of this method motivated the decision to proceed in the direction of a competence test. Nevertheless, we should note that the psychological aspect associated to the single individual is very important to provide a complete evaluation of competencies and aptitudes. In the future we hope the test will be completed in this sense. The predictive capabilities may be increased when measures of personality and ability are jointly considered.

In the guidance project, the basic idea is the creation of different tests, one for each faculty, with the following characteristics:

- Evaluating competencies;
- Faculty-specific;
- Made up of multiple-choice items.

The first specification depends directly on the aim of the project, that is to evaluate students' knowledge. The term 'competence' has a wide and complex meaning: in this context we will restrict it to the simple meaning of 'knowledge'. The second characteristic is quite innovative in the guidance context. In fact, in the classical guidance process, the student answers to a set of items and his aptitudes are evaluated. Afterwards, he is guided to different faculties he can be interested in, according to the test results. In our case, the student chooses the faculty or the faculties he is interested in before the test; after answering the faculty-specific items he gets an idea about his preparation on the subjects. Finally, multiple-choice items are used because of their flexibility.

The faculties involved in the project were asked to produce a set of items coherent with the specifications, taking into account that the test has been conceived for students in their fourth and fifth year of secondary school.

Initial test development After developing the idea and the objectives of the guidance project, each faculty was asked to produce a set of items without a prespecified size. In the first phase of the project, six different tests were developed by the single faculties, containing a set of items representing the faculty topics.

The items were divided into several groups for different subjects. The items were all multiple-choice with 5 alternatives, except the Pharmacy faculty with 4 alternatives. Furthermore, several faculties decided to include some general culture items to complete the test specification. Table 1.1 shows the initial number of items for each faculty; as we can see, the number of items is sensibly different from one faculty to another.

Faculty	Initial number of items
Arts and Humanities	89
Economics	101
Pharmacy	213
Political Science	70
Psychology	98
Statistics	49

Table 1.1. Original number of items for each faculty

Consequently, the tests were made available online and a first draft of the items was submitted to secondary school students of the region Emilia Romagna (Italy). Each student was asked to answer to 40 randomly chosen items from the item bank of the selected faculty.

A preliminary item analysis was conducted to delete all the items that were completely not coherent with the objectives of the project. Clearly, this first testing phase was not without problems. First of all, only 40 items were randomly selected from the item bank and some faculties prepared too many items to be tested. Furthermore, the number of respondents was very low (from 100 to 130 observations for all the faculties except Statistics with only 43 respondents). Consequently, the sample size resulted inadequate to perform a reliable item analysis. The number of correct responses was used to detect and delete extremely difficult or easy items and not well-formulated items.

For the second testing, we decided to create two tests for each faculty with a pre-specified length of 40 items each. The items were selected by using the results of the first analysis and some of them were included in both tests. The only exception was the Statistics faculty: with 49 starting items we were able to create only a single test. In this phase the faculties of Agriculture, Foreign Languages and Literature and Education Science decided to join the project and produced a single test with 40 items. To increase the sample size, the tests were submitted in a paper and pencil format to University freshmen in their first days of lectures (September-October 2005). The fundamental hypothesis is that freshmen have the same level of knowledge of high school students, before they start the University path. The optical character recognition was used to create the database with students' responses.

The sample size was increased because we had around 200-250 respondents for most of the faculties, except than Agriculture, Education Science and Statistics with only 48, 95, 133 involved freshmen, respectively. An item analysis was conducted to highlight problems and behaviors of the items. Afterwards, the results have been used to perform the current version of the guidance tests.

The current guidance test The current version of the guidance test is quite different from the initial one; in fact, the results of the item analysis were used to modify the original formulation of the test. We sustained the faculty-specific structure but we decided to include general culture items for all the faculties.

The design of the test is represented in Figure 1.1, where the highlighted sections reflect the items presented to the examinees.

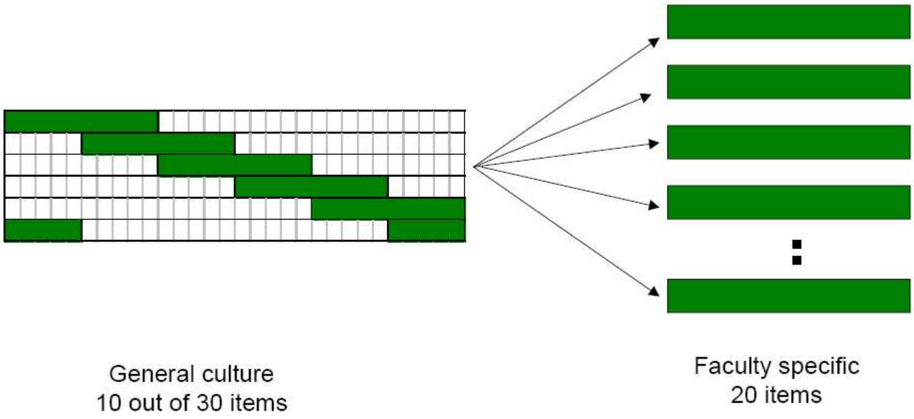
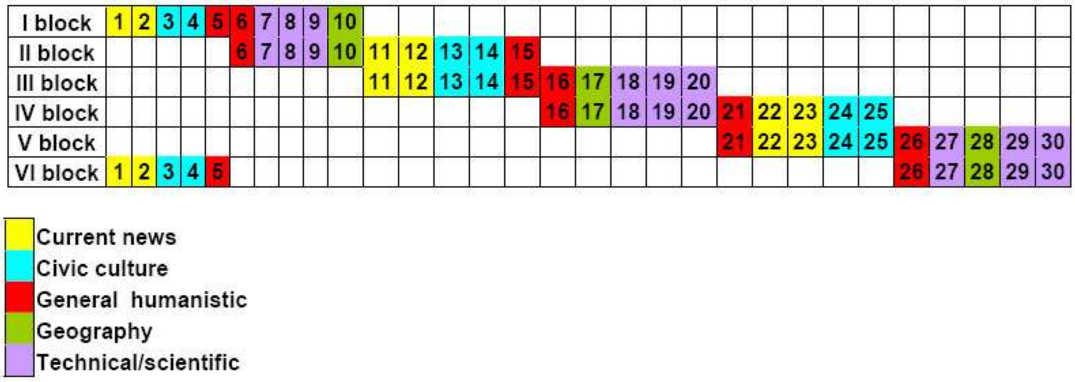


Figure 1.1. Design of the study

Each test consists of two sections: the general culture and the specific part. The first one is common to all the faculties while the second one is faculty-specific. In particular, the general culture items employed by several faculties were used to create a common item bank of thirty items. By using a block design, only 10 items are extracted for each examinees from the overall set. The general culture items are represented on the left side of the figure: each row of the rectangle symbolizes the sequence of the 30 items and the highlighted section is the set of items effectively submitted to the individual. Each row is randomly selected. On the right side of the figure the specific items are symbolized: each faculty has its own sequence of items. They are fixed and they are 20. Therefore, the test length is 30: 10 general culture and 20 specific items.

In particular, the block structure of the general culture items is presented in Figure 1.2.



I block	1	2	3	4	5	6	7	8	9	10																				
II block						6	7	8	9	10	11	12	13	14	15															
III block											11	12	13	14	15	16	17	18	19	20										
IV block																16	17	18	19	20	21	22	23	24	25					
V block																					21	22	23	24	25	26	27	28	29	30
VI block	1	2	3	4	5																					26	27	28	29	30

🟨 Current news
🟦 Civic culture
🟥 General humanistic
🟩 Geography
🟪 Technical/scientific

Figure 1.2. Block design for general culture items

The items are divided by topic, in fact we have considered five main subjects: actuality, civic culture, general humanistic, geography and technical-scientific. In particular in the humanistic area we find Italian language, literature and history while the technical-scientific one is composed of natural science, computer science and economics. The general culture items are divided into six homogeneous blocks of ten items each; each block contains two items about actuality, two about civic culture, two humanistic, two geographic and three technical-scientific. Each item is repeated into two blocks. When the student starts the test, one of the six blocks is randomly extracted, that is each item has 1/3 percent of probability to be presented. The items are multiple-choice with 5 alternatives, with only one correct answer. The only exception is the Pharmacy faculty, with 4 response alternatives for the specific items. The sequence of the general culture items and the nine faculty-specific questionnaires can be found in Appendix E.

The test is online in the web site of the Guidance Service of the Bologna University (www.orientamento.unibo.it). Before starting the test, students should get an identification code providing some personal data: year of birth, sex, attended high school and district of residence. With the identification code, they can login to the test in the following days and try different faculties or the selected one again. The online placement of the multiple-choice test allows the increment of the sample size and gives the students a simple instrument for a self-evaluation. In fact, in the end of the questionnaire the examinee knows the number of correct responses for the general culture and for the specific items and can look for mistakes reading the items again. A specific evaluation about the student's performance has not been implemented yet. This step is very important and should be conducted carefully. Before the evaluation process we should ensure that the items are coherent with the purposes of the test and we should check their properties. This phase is called *item calibration* and it is mainly what this work is about. After the item properties have been investigated, students can be evaluated on their responses. Nevertheless, we should consider the fact that the guidance test is not an examination. Even if the test is evaluating knowledge and competence, the fundamental purpose of the project is to provide the students with a qualitative comment about their level of knowledge rather than giving a strict judgment with a score. Through a faculty-specific test, individuals can also be able to self-evaluate their level of interest on different subjects. We really hope that all the faculties will join the project in the future, because this would be a fundamental condition to give students a complete support.

1.2.2 The data

The collected data are referred to the period from May to September 2006. The number of respondents for each faculty is presented in Table 1.2. Observations were selected by using the variable *year of birth*.

The students in their last year of secondary school and future freshmen are born in 1987. Nevertheless, the range 1985-1990 of year of birth was considered in the analysis to include students in their fourth year of high school, failed and one year ahead individuals. As we can note, the most selected faculties are Political Science, Arts and Humanities, Psychology and Economics. These faculties (except Psychology) have the highest number of enrolled students in Bologna too, respect to the other faculties involved in the project. On the other hand,

Faculty	Num. of respondents	Percentages
Agriculture	223	2.38
Arts and Humanities	2006	21.41
Economics	1060	11.31
Education Science	238	2.54
Foreign Languages and Literature	935	9.98
Pharmacy	939	10.02
Political Science	2403	25.65
Psychology	1385	14.78
Statistics	181	1.93
Total	9370	100.00

Table 1.2. Number and percentage of respondents for each faculty (May-September 2006)

Agriculture and Statistics have the lowest number of respondents: unfortunately both faculties are reporting a serious decrement in the number of freshmen. Also the Education Science faculty has a low number of respondents but the test was online after the other faculties. Totally in the period May-September 2006 we had 9370 respondents.

Table 1.3 shows the number of freshmen and enrolled students in the University of Bologna for the faculties joining the project. The data are referred to the 2006/2007 academic year [1].

The composition of our sample seems to reflect quite well the real composition of freshmen, in terms of chosen faculty. Nevertheless, differences may be imputed to the gap between students who are interested in the faculty and students who are also able to succeed in the admission test. In fact, some courses of Arts and Humanities, Economics, Education Science, Foreign Languages and Literature, Political Science and Psychology only accept a limited number of students.

To give an idea of the composition of the sample, the respondents have been classified by sex, attended school and district of residence. Table 1.4 shows the percentage distribution by sex for all the faculties and the total sample.

We can clearly note that the pretty humanistic faculties as Arts and Humanities, Education Science, Foreign Languages and Literature and Psychology have a larger proportion of females than males. This happens also for the Pharmacy

[1] The data have been extracted from the University database on 16th January 2007 and are only provisional for the academic year.

Faculty	Freshmen	%	Enrolled	%
Agriculture	307	3.41	1369	3.04
Arts and Humanities	2504	27.83	13591	30.15
Economics	1768	19.65	9289	20.61
Education Science	991	11.01	5173	11.48
Foreign Languages and Literature	754	8.38	3071	6.81
Pharmacy	743	8.26	2759	6.12
Political Science	1529	16.99	7530	16.7
Psychology	253	2.81	1644	3.65
Statistics	148	1.64	653	1.45
Total	8997	100.00	45079	100.00

Table 1.3. Number and percentage of freshmen and enrolled students (academic year 2006/2007)

	F	M	Missing
Agriculture	46.64	53.36	0.00
Arts and Humanities	70.74	29.16	0.10
Economics	57.92	41.98	0.09
Education Science	82.77	17.23	0.00
Foreign Languages and Literature	85.24	14.76	0.00
Pharmacy	70.39	29.50	0.11
Political Science	62.05	37.87	0.08
Psychology	74.37	25.49	0.14
Statistics	49.17	50.83	0.00
All respondents	68.32	31.59	0.09

Table 1.4. Percentage distributions of respondents by sex (F=female; M=male)

and Political Science faculties. In the remaining technical-scientific faculties the proportions of females and males are almost the same. The dataset is composed of 68% females and 32% males approximately. Missing data are negligible. The data referred to the 2006/2007 academic year show that the enrolled students are composed of 63% females and 37% males.

Table 1.5 represents the percentage distribution by attended school for the single faculties and for the entire dataset. Most of the respondents are attending high school (*Italian Liceo*) for all the faculties while only few of them are attending vocational school.

This is not surprising, however we can note that for the scientific faculties there is a substantial decreasing of the proportion of high school students in favor of

	High school	Polytechnic school	Vocational school	Missing
Agriculture	55.61	35.87	8.52	0.00
Arts and Humanities	81.66	14.16	4.04	0.15
Economics	58.87	34.91	6.04	0.19
Education Science	71.85	18.49	9.66	0.00
Foreign Languages and Literature	75.19	20.64	4.06	0.11
Pharmacy	75.72	19.70	4.47	0.11
Political Science	76.61	18.85	4.45	0.08
Psychology	72.71	20.07	7.00	0.22
Statistics	56.35	38.12	5.52	0.00
All respondents	73.86	20.88	5.13	0.13

Table 1.5. Percentage distributions of respondents by attended school

polytechnic schools. Considering the complete dataset, the 74% of respondents comes from high school while only the 21% from polytechnic and the 5% from vocational school.

Table 1.6 shows the percentage distribution by district of residence for all the faculties and the total sample.

	Emilia Romagna	Rest of Italy	Missing
Agriculture	40.81	59.19	0.00
Arts and Humanities	35.64	64.16	0.20
Economics	41.89	57.64	0.47
Education Science	42.86	57.14	0.00
Foreign Languages and Literature	36.68	63.10	0.21
Pharmacy	39.72	59.85	0.43
Political Science	36.33	63.42	0.25
Psychology	39.86	59.86	0.29
Statistics	38.12	61.88	0.00
All respondents	38.01	61.72	0.27

Table 1.6. Percentage distributions of respondents by district of residence

There are not large differences in the single results; totally the 38% of respondents comes from the region Emilia Romagna (the region where the University of Bologna is situated) while the 62% comes from the rest of Italy.

The data from the online test were collected keeping different codes for the two type of missing data: the not presented and the omitted items. Table 1.7

shows the percentage of correct, incorrect and omitted items for the 30 general culture items, calculated on the complete dataset of 9370 respondents.

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	42.57	56.10	1.33	16	69.63	28.80	1.58
2	80.19	18.51	1.30	17	63.19	34.72	2.09
3	93.56	5.43	1.02	18	41.25	57.40	1.35
4	95.78	3.30	0.92	19	73.49	25.19	1.32
5	83.46	15.56	0.98	20	44.72	51.22	4.05
6	25.73	72.95	1.31	21	93.95	5.20	0.85
7	72.45	26.33	1.22	22	76.62	22.18	1.20
8	89.26	8.90	1.84	23	56.07	42.34	1.59
9	45.97	52.97	1.06	24	57.53	40.72	1.76
10	92.44	6.50	1.06	25	92.00	7.25	0.75
11	49.03	48.04	2.93	26	34.02	62.25	3.73
12	81.97	16.63	1.40	27	83.89	14.80	1.31
13	64.99	33.96	1.05	28	32.71	64.74	2.55
14	87.54	11.16	1.30	29	97.97	1.24	0.78
15	53.58	44.04	2.38	30	63.86	34.84	1.31

Table 1.7. Percentage of correct, incorrect and omitted responses for the general culture items on the complete dataset (9370 respondents)

No missing data imputation has been applied to the data. The percentages in Table 1.7 have been calculated only on the presented items.

Appendix A contains the results on the percentages of correct, incorrect and omitted items for each faculty.

Chapter 2

Item Response Theory for item calibration

2.1 The choice of Item Response Theory

Item Response Theory (IRT) is a measurement theory that has been intensively applied only recently, especially in the educational field. The IRT roots can be traced back in the thirties and forties but the theory was first formalized in the sixties with the fundamental work of Lord and Novick (1968). The development of IRT should be imputed to the concrete necessity of overcoming the lacks of the Classical Test Theory (CTT). Psychometricians and measurement specialists needed an effective instrument for the construction of tests and the interpretation of test scores but CTT could satisfy these requirements only partially. In fact, CTT is founded on assumptions that make the framework very sensitive to the sample conditions. Furthermore, the CTT analysis are based on the test score, in terms of number of correct responses. The item characteristics are not included in the evaluation of the test properties.

On the other hand, IRT seemed to be a promising method to substitute CTT in several theoretical and application fields toward a more complete and effective framework. In order to understand the basic concept of IRT, we will start from a practical example. Table 2.1 shows the cross classification of responses to two binary items[1].

		Item 9		
		Correct	Incorrect	Total
Item 7	Correct	977 (79%)	267 (21%)	1244
	Incorrect	78 (55%)	63 (45%)	141
	Total	1055	330	1385

Table 2.1. Cross classification of responses to item 7 and item 9 for the Psychology test (row percentages in brackets)

The 79% of respondents who answer correctly to the item 7, also give a correct response to the item 9. On the other hand, the 55% of respondents who answer incorrectly to the item 7 are able to give a correct response to the item 9 while the 45% fail again. What do the data mean? How are the responses to the different items related? Suppose the existence of a latent trait, denominated *ability*, representing an unobservable characteristic of the individuals. The higher the ability level is, the more the probability to answer correctly increases. Imagine that the examinees with an high level of ability give a correct response for item 7. Do they need the same ability also to answer to item 9? If yes, then the relationship between the two items can be explained through the existence of the trait. The basic idea is that responses are related to each other: when the ability is introduced to explain the data, the responses become independent. Therefore, the subsequent question is: are the items good indicators of the latent ability? If the relationship between the items and the trait is strong, the indicators will be a good instrument to predict the ability level of the examinees. These conclusions can be extended to both the case of a set of k items and the presence of two or more abilities. The information provided by the response patterns (2^k for binary items and m^k for multiple-choice items with m alternatives) can be summarized into a contingency table reporting the counts of respondents for each pattern.

Item Response Theory is included in the latent variable modelling framework (see Bartholomew and Knott (1999), Skrondal and Rabe-Hesketh (2004)); as a measurement theory the main focus is on the relationship between the examinees'

[1] The items refer to the specific questionnaire for the Psychology faculty developed according to the guidance project. The items are multiple-choice but they have been dichotomized to make the exposure simpler. The Psychology test can be found in Appendix E.

performance in a test and the latent ability(ies). The IRT features are perfectly coherent with the focus of the guidance project.

2.1.1 IRT features and assumptions

The concept of model in Item Response Theory is expressed as a mathematical function used to describe the trace line(s) or conditional probability of a response given the latent variable, for an item with categorical responses (Thissen and Steinberg, 1986). The parametric model describes the relationship between the "observable" (the examinee's performance in the test) and the "unobservable" (the latent ability).

The choice of the IRT model depends mainly on the structure of the data. Items can have only two response categories (correct and incorrect) or more than two (multiple-choice items). In the first case we will use models for binary data while in the second one for polytomous data. Furthermore, the responses can be nominal or ordinal: IRT provides specific models for both cases. Then, the specification of dimensionality is fundamental in the model choice. When only one latent trait is underlying the performance on the test we are under the assumption of unidimensionality. On the other hand, when two or more latent abilities are needed to explain the correlation among the responses we are in the multidimensional context. The choice of the correct dimensionality is a complex issue in IRT and a simple and effective way to detect the number of dimensions still does not exist. The third chapter will deal with this argument. Finally, the model depends on both the function expressing the relation between the performance and the ability and on the number of item parameters.

Consider a set of k binary items and the existence of a single latent ability θ underlying the responses to the items. The unidimensional IRT model for binary data expresses the probability π_j of a correct response to the item j, with $j = 1, ..., k$, as a function of the ability and a set of item parameters. The two most common probability models used in IRT are the normal and the logistic distribution functions. When the distribution is normal, we obtain a probit model, i.e.

$$\pi_j = \Phi(\eta_j) \quad \Longrightarrow \quad \Phi^{-1}(\pi_j) = \eta_j, \tag{2.1}$$

where Φ is the standard normal cumulative distribution function and the predictor η_j is a function of the ability θ. As an example, consider the set of item parameters $\{\alpha_j, \beta_j\}$ describing the item characteristics. The predictor becomes $\eta_j = \alpha_j(\theta - \beta_j)$ and the model is called the two-parameter normal ogive model (Lord, 1952). On the other hand, the use of a logistic distributions leads to the well known logit model, i.e.

$$\pi_j = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)} \quad \Longrightarrow \quad \text{logit}(\pi_j) = \eta_j. \tag{2.2}$$

Model (2.2) is called two-parameter logistic model (Birnbaum, 1968). As indicated by Birnbaum (1968), the logistic curve coincides with the normal ogive one very nearly. In particular, Haley (1952) proved that the two models are equivalent in terms of predicting the same probability, through the introduction of a scaling constant $\boldsymbol{D}$=1.702 into the logistic model, as follows

$$\left| \Phi(\eta_j) - \frac{\exp(\mathbf{D}\eta_j)}{1 + \exp(\mathbf{D}\eta_j)} \right| < 0.01. \tag{2.3}$$

The derivation procedure is based on choosing $\boldsymbol{D}$ so that the maximum difference between $\Phi(\eta_j)$ and $\frac{\exp(\mathbf{D}\eta_j)}{1+\exp(\mathbf{D}\eta_j)}$ is as small as possible.

In both models the probability of a correct response is expressed as a monotonically increasing function of the trait. The curve is called *item characteristic curve* (ICC) and allows a straightforward interpretation: the higher the ability level is, the higher the probability of a correct answer is as well.

When the item response model fits, that is there is a close correspondence between the chosen model and the responses to a set of items, some useful properties are achieved in the IRT framework. First of all, item and ability estimates are said to be invariant. Item parameter estimates are independent of the group of examinees used from the population of examinees for whom the test was designed. Examinee ability estimates are not dependent on the particular choice of test items used from the population of items which were calibrated. Furthermore, estimates of standard errors for individual ability estimates are possible instead of a single estimate of error for all the examinees, as is the case in CTT.

Assumptions The main assumptions of the IRT models are the unidimensionality and the local independence. The first one refers to the presence of a single trait influencing the test performance. This condition is difficult to obtain in practice, because the examinees usually employ different abilities to answer a set of items. Nevertheless, what is required for the unidimensionality assumption is the existence of the single dominant component characterizing the responses. In the third chapter we will see how this assumption can be relaxed by using multidimensional IRT models. The contemporary presence of more that one latent trait increases the complexity of the model but allows a deeply investigation of the data structure. A particular attention should be paid in case of time-limited tests. The required speed for the answer implicity creates a further ability necessary to solve the test. Therefore, unidimensionality does not hold in case of speeded condition.

The second assumption implies that, when the latent space has been completely specified, the examinees' responses to a set of items are statistically independent. Under the unidimensionality assumption, the θ-conditional probability of a response pattern can be expressed as the product of the single conditional probabilities for all the items in the test. In the multidimensional case, the local independence holds conditioned to a vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_Q)$ of Q latent traits.

2.1.2 Estimation methods

In item response models, the probability of a correct response is a function of examinees' ability and item parameters. These two characteristics are both unknown. The only available data are the responses to a set of items given by a sample of individuals. In the estimation process, two important features should be taken into account: the nonlinearity of the response model and the impossibility of observing the latent variable θ. The estimation is analogous to performing a non-linear regression with unknown predictor values. The main focus is on the determination of the θ values for each examinee and the item parameters from the item responses.

The simultaneous estimation of the ability and the item parameters can be performed according to either maximum likelihood (ML) methods or in a Bayesian framework.

As a general rule, the estimation depends on how the probability of the observed response patterns is conceptualized. In the *stochastic subject* interpretation of probability, the observed persons are regarded as fixed. The probability represents the unpredictability of specific events, i.e. the encounter of a person with a particular item. Within this approach, the latent variables are constructed as unknown fixed parameters. In the *random sampling* interpretation of probability, the observed persons are regarded as a random sample from a population. Therefore, a specific distribution of the latent trait must be assumed to interpret the probability and the latent variables are treated as random. Three ML estimation methods are available:

- Joint maximum likelihood (JML);
- Conditional maximum likelihood (CML);
- Marginal maximum likelihood (MML).

The first two methods imply the concept of fixed latent variable while in the MML estimation the latent variables are treated as random.

The JML implements the maximum likelihood through an iterative procedure to estimate the item parameters and the abilities simultaneously. Simply, we look for the values of the parameters that jointly maximize the log-likelihood function. After the specification of the starting values, the item parameter estimates and the ability estimates are alternatively updated. This method is very simple to implement but the complexity increases as the number of observations increases. The standard limit theorems do not apply and the resulting parameter estimators are not consistent.

The CML is based on the availability of a sufficient statistic for the ability so that the likelihood function can be simplified conditioning to it. This method can be applied only in case of the one-parameter logistic (1PL) model (Rasch, 1960) considering $\eta_j = \theta - \beta_j$. In fact, a sufficient statistic for the ability is represented by the total test score in terms of sum of correct responses: $T(i) = \sum_{j=1}^{k} y_{ij}$, where y_{ij} is the binary response taking the value 1 for correct and 0 for incorrect, $j = 1, ..., k$ items and $i = 1, ..., n$ examinees. The applicability of JML and CML methods is rather limited.

The most widely used method is MML estimation, which is based on the marginal probability of observing a response pattern, obtained integrating out over the distribution of ability. The MML employs the EM iterative procedure. The method will be described in Section 2.3 and in Chapter 3 for the unidimensional and multidimensional models, respectively. A single ability value may be associated to each examinee by using maximum a posteriori (MAP) or expected a posteriori (EAP) techniques.

All the ML estimation methods refer to fixed item parameters. On the other hand, the Bayesian approach regards both the latent variable and the item parameters as random. The implementation of a Markov chain Monte Carlo (MCMC) procedure will be described in Chapter 4, both for unidimensional and multidimensional models.

2.2 Test development and calibration

The role of testing is crucial in psychological and educational measurement and Item Response Theory has been successfully applied to this context in the last decades.

The most widely used instrument is a set of items, collected in a test, submitted to respondents to provide information about the variables of interest. The development and analysis of tests are considered good ways to understand psychological features, aptitudes and competencies. Furthermore, item analysis has been widely used in job selection to evaluate knowledge and aptitude of the examinees and in customer satisfaction to find out latent behaviors and preferences of consumers. Therefore the test development represents a complex and long process.

According to Hambleton and Swaminathan (1985), the construction of a test in the IRT framework can be summarized in the following steps

1. Preparation of test specifications;

2. Preparation of the item pool;

3. Testing the items;

4. Selection of test items;

5. Compilation of norms;

6. Specification of cutoff scores;

7. Reliability studies;

8. Validity studies;

9. Final test production.

First of all, test developers should define the characteristics of both the test and the involved items, that is mainly the specification of objectives and purposes. In our guidance test the principal intention is to understand knowledge and competence of students before they start the university career, in order to give them suggestions about their possible choices.

With no doubts, one of the most difficult and time-consuming phases in a test development is the creation of the item pool: a set of items consistent with the specifications given in the previous step. To this end, multiple-choice items have been widely used because of their flexibility. Usually they consist of a stem (a question or a statement) and a list of response alternatives. We advise against this type of items only when the examinee is asked to give rich and complex argumentations supporting the answer or to suggest his or her personal and original ideas. The use of multiple-choice items can accentuate the guessing, that is the probability of randomly choosing the correct alternative. Usually this risk is overestimated, because it can be kept under control increasing the number of response options and the length of the test. The last ensures that the examinee cannot get a high score in the test just guessing.

Test developers should follow some basic and general rules in order to create the items. Firstly, a good item should have synthetic and clear stem and alternatives, with a univocal interpretation. The length of the item is dependent on the type of question and on the topic, but should be simplified as much as possible. Negative sentences in the stem should be avoided as well as questions requiring to select the wrong alternative or the worst in the set. Secondly, the response options should be homogeneous in shape, length and content. In fact, a typical mistake committed by test developers is to insert a correct alternative that is much longer or precise than the others. Consequently, the attention of the examinee is focused on that response. Furthermore, items with only one response correct, that is not a combination of other responses, are suggested. The order position of the correct alternative should vary in the test. Finally, we require the items to be coherent with the aim of the test. For example, the guidance test should not contain items with too specific and technical content. The items are faculty-specific but they have been created to give suggestions to students and not to evaluate them respect to university advanced topics.

The following step in the creation of the test is the item testing, that mainly represents the item calibration. The items are submitted to a sample of individuals and are analyzed according to an IRT model. The calibration allows to understand the role of the different distractors in the multiple-choice items and to infer the item properties, like discrimination power and difficulty level. After calibrating a set of test items, the item parameters are taken as known and used to characterize the latent ability for examinees who produce a particular response pattern.

In the guidance test, two IRT models were used in item calibration. The first one is the multiple-choice model (MCM) developed by Thissen and Steinberg (1984) especially to infer the behavior of multiple-choice items and their specific response alternatives; the second one is the three-parameter logistic (3PL) model, introduced by Birnbaum (1968) on binary data to characterize three main item properties: discrimination, difficulty and guessing. The two models will be simultaneously used in the calibration phase to point out problems and features of the items and their response alternatives and will be described in detail in the following sections.

After the item testing, the development of a test proceeds with the selection of the items. Principally, the items are chosen from the item bank respect to both the properties inferred in the previous phase and the information they provide for the examinees' evaluation. Furthermore, items can be selected for a test with pre-specified length or information.

Then the compilation of norms and the specification of cutoff scores follow. Finally, reliability and validity studies should be conducted in order to validate the test and submit it to evaluate the respondents.

Referring to Hambleton and Swaminathan (1985) scheme, we can place the guidance test in the third phase. The tests are online to allow a bigger sample size, required for correctly calibrating the items, but still we are in the "testing the items" step. Nevertheless, the complete project includes the development of different tools to investigate the properties of the test. In the future, we don't exclude the possibility to modify the test formulation, coherently with the results.

2.3 The multiple-choice model

The MCM was first formalized by Thissen and Steinberg in 1984 as an extension of Samejima's (1979) model for multiple choice items, given a set of k items with m response alternatives. These models are principally used in the item calibration for categorical response alternatives. The models are for observed item response data and the latent trait θ involved is considered a *random* variable. The original formulation allows for a different number of alternatives for each item but, coherently with our applicative context, we have decided to fix the number of answer options to m.

The starting point is the work by Bock (1972), whose main idea was to specify a response function $\eta_h = \alpha_h \theta + \delta_h$ for each category $h = 1, .., m$ as a linear function of the latent ability θ. The relationship between the responses to item j, with $j = 1, .., k$ items, and the θ's is modeled through a logistic transformation as follows

$$P(y_j = h|\theta) = \frac{\exp(\alpha_h \theta + \delta_h)}{\sum_{l=1}^{m} \exp(\alpha_l \theta + \delta_l)}. \tag{2.4}$$

Each response function depends on a slope parameter α_h and on an intercept term δ_h. Each parameter is referred to a specific item j but here we use a reduced formulation of the model instead of the complete one, to keep the specification simple. Bock (1972) justified the model with a plausible psychological interpretation: each alternative of item j is assumed to give rise to a quantitative "response tendency" in a given subject. In the population of subjects of ability θ, these tendencies are assumed to be normally and independently distributed. If each subject chooses the alternative for which is tendency is maximal, the proportion of subjects in the population who choose alternative h, is closely approximated by (2.4).

Samejima (1979) conceptually modified the (2.4) adding a completely latent response category to take into account the so called "totally undecided individuals", that is the examinees who don't know the correct answer and guess. This category is labelled with "zero" or "don't know" (DK) and its response probability is

$$P(y_j = 0|\theta) = \frac{\exp(\alpha_0\theta + \delta_0)}{\sum_{l=0}^{m} \exp(\alpha_l\theta + \delta_l)}. \tag{2.5}$$

Consequently a parameter γ_h is introduced in the model to represent the unknown proportion of individuals who randomly choose each option h, as follows

$$P(y_j = h|\theta) = \frac{\exp(\alpha_h\theta + \delta_h) + \gamma_h \exp(\alpha_0\theta + \delta_0)}{\sum_{l=0}^{m} \exp(\alpha_l\theta + \delta_l)}. \tag{2.6}$$

The γ_h is fixed to $1/m$ by Samejima (1979), under the assumption that each response category has the same probability to be randomly chosen by individuals who don't know the response. Thissen and Steinberg (1984) found this assumption unlikely and allowed each γ_h to be a function of estimated parameters.

2.3.1 Identification problems

The model in (2.6) has $3m-1$ free parameters: m slopes, m intercepts and $(m-1)$ guessing proportions (because $\sum_{h=1}^{m} \gamma_h = 1$ for each item). The parameters α_h and δ_h are not identified respect to location and Thissen and Steinberg (1984) suggested to solve the indeterminacies imposing the constraints:

$$\sum_{h=0}^{m} \alpha_h = \sum_{h=0}^{m} \delta_h = 0,$$

through the parameterizations $\boldsymbol{\alpha} = \boldsymbol{T}\boldsymbol{\alpha}^*$ and $\boldsymbol{\delta} = \boldsymbol{T}\boldsymbol{\delta}^*$. The operation consists of pre-multiplying the vectors $m \times 1$ of free parameters $\boldsymbol{\alpha}^*$ and $\boldsymbol{\delta}^*$ by a transformation matrix $\boldsymbol{T}$ of dimension $(m+1) \times m$. The $\boldsymbol{T}$ matrix can consist of deviation, polynomial or triangle contrasts. We use the deviation $\boldsymbol{T}$ matrix; as an example, for multiple-choice items with 5 alternatives it is

$$\boldsymbol{T} = \begin{pmatrix} -0.17 & -0.17 & -0.17 & -0.17 & -0.17 \\ 0.83 & -0.17 & -0.17 & -0.17 & -0.17 \\ -0.17 & 0.83 & -0.17 & -0.17 & -0.17 \\ -0.17 & -0.17 & 0.83 & -0.17 & -0.17 \\ -0.17 & -0.17 & -0.17 & 0.83 & -0.17 \\ -0.17 & -0.17 & -0.17 & -0.17 & 0.83 \end{pmatrix}.$$

Furthermore, the γ_h should be constrained to sum to unity and to lie on the interval $[0, 1]$. The γ_h are thought to be generated from a set of pseudo-parameters γ_h^*, which can take any real value, as follows

$$\gamma_h = \frac{\exp(\gamma_h^*)}{\sum_{h=1}^{m} \exp(\gamma_h^*)}. \tag{2.7}$$

The γ_h^* must sum to 0, so we should use the parameterization $\boldsymbol{\gamma}^* = \boldsymbol{T_2}\boldsymbol{\tau}$ analogously to the other item parameters. The $\boldsymbol{T_2}$ is a transformation matrix of smaller order than $\boldsymbol{T}$, that is $m \times (m-1)$ while $\boldsymbol{\tau}$ is a $(m-1) \times 1$ vector of free parameters.

The free parameters $\boldsymbol{\alpha}^*$, $\boldsymbol{\delta}^*$ and $\boldsymbol{\tau}$ are respectively called "α-contrasts", "δ-contrasts" and "γ-contrasts". For example, the free parameters for a five-alternative multiple choice item are 5 "α-contrasts", 5 "δ-contrasts" and 4 "γ-contrasts", in total 14.

The model presents a further indeterminacy on the *sign* of $\boldsymbol{\alpha}^*$, due to the reflection of the latent variable θ, that is empirically avoided starting the estimation procedure with a positive α_h for the correct response.

2.3.2 Model interpretation

The MCM is principally used in the item calibration to perform a preliminary graphical analysis of the different response alternatives for each item. In fact, it can be usefully adopted to create a response curve for each option as a function of the θ ability, following the (2.6). As an example, consider the figure 2.1 which represents the response curves for the item 5 of the Psychology faculty[2].

[2] The item can be found in the appendix E, Psychology faculty test

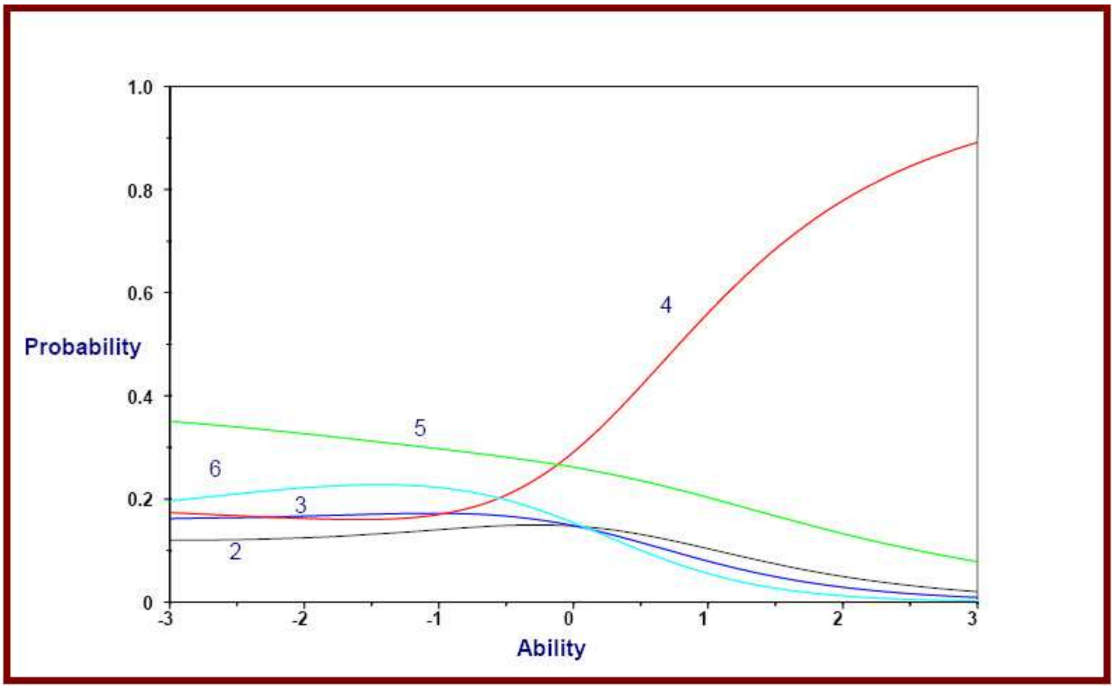


Figure 2.1. Response curves for the MCM - Item 5 Psychology

Each curve expresses the probability of selecting the single alternative as a function of ability. The response options are coded $\{a, b, c, d, e\}$ in the questionnaire and the correct one is c. The alternatives are numerically recoded as $\{2, 3, 4, 5, 6\}$ to preserve the model notation, so the correct one becomes the option 4 (the red curve). As we can see, the curve of the correct alternative has a monotonic increasing trend and a typical S-shape. This means that the probability of a correct response increases as ability increases, especially around intermediate values of abilities. The wrong alternatives $\{2, 3, 6\}$ all have a non-monotonic trend, slightly increasing for low abilities levels and decreasing on the rest of the domain. Finally, the wrong option 5 has a completely decreasing trend. The interpretation is straightforward: as ability decreases, also the probability of selecting a wrong option decreases. From the graphics we can also understand that the wrong options are not very effective distractors: the associated curves are quite low and the probability of a correct response is never zero. Also for very low abilities the distractors are not able to create serious response problems to the examinees.

The estimated values for the item parameters are not important for the interpretation. The α's represent the ordering between the options: for well calibrated items, the correct response has the biggest positive value that ensures a monotonic increasing behavior of the response curve, while the other options can have a decreasing or not monotone behavior with low and intermediate α_h, respectively. Particularly, what we expect, as the ability increases, is that the correct response has an increasing probability of been selected by the examinees while the wrong alternatives, named "distractors", have a decreasing probability to be chosen. We also allow a nonmonotonic trend for the incorrect response curves, that is increasing for low ability levels and decreasing for high abilities. Finally, the δ's reflect the selection relative frequency (for alternatives with similar values α_h, those with larger values of δ_h are chosen more frequently than the others) while the γ's are the proportions of guessing for each option.

2.3.3 Marginal maximum likelihood estimation

The marginal maximum likelihood (MML) is the most popular estimation method in IRT, next to the Bayesian methods. Its popularity depends on the advantages respect to other methods based on the likelihood maximization like the conditional maximum likelihood (CML) and joint maximum likelihood (JML), briefly discussed in the previous section. Given some specification of the population distribution of the latent variable, the parameters of the IRT models can be estimated by parallel marginal maximum likelihood procedures, following Bock and Lieberman (1970), Bock and Aitkin (1981) and Thissen and Steinberg (1984).

The focus of the MML method in the MCM is the estimation of the parameters ($\boldsymbol{\alpha}$*, $\boldsymbol{\gamma}$*, $\boldsymbol{\tau}$) for each item j, with $j = 1, ..., k$. The data are usually in the form of a $n \times k$ matrix containing the responses given by n examinees to a set of k items, each one with m response categories. Therefore each single response y_{ij}, with $i = 1, ..., n$ subjects, can take any value in the set $\{$1,2,..,$m\}$. The data are collected in response patterns $\boldsymbol{y}$ and summarized by their frequencies $r_{\boldsymbol{y}}$ through a m^k contingency table.

Under the assumption of local independence, the conditional probability of observing a response vector $\boldsymbol{y}$ is

$$P(\boldsymbol{y}|\theta) = \prod_{j=1}^{k} P(y_j|\theta), \tag{2.8}$$

where $P(y_j|\theta)$ is the (2.6) for the MCM. By the way, the (2.8) can be adapted to all the IRT models.

Under the assumption of θ as a random latent variable, one can obtain the joint unconditional or marginal probability of observing the data

$$P(\boldsymbol{y}) = \int\limits_{-\infty}^{\infty} \prod_{j=1}^{k} P(y_j|\theta)\phi(\theta)d\theta, \tag{2.9}$$

where $\phi(\theta)$ denotes the probability distribution of θ in the population, which is usually assumed to be standard normal, that is $\theta \sim N(0, 1)$. The (2.9) is known as the marginal likelihood and represents the starting point for the MML procedure. Simply, it is obtained integrating out θ from the joint distribution

$$P(\boldsymbol{y}) = \int\limits_{\Theta} P(\boldsymbol{y}, \theta)d\theta = \int\limits_{\Theta} P(\boldsymbol{y}|\theta)P(\theta)d\theta, \tag{2.10}$$

and then substituting $P(\boldsymbol{y}|\theta)$ with the (2.8) and $P(\theta)$ with $\phi(\theta)$. Because the ability has been integrated out, the marginal probability is a function of the item parameters only. Under the assumption of a standard normal distribution for θ, $\Theta = \mathbb{R}$.

The likelihood for the complete data set is

$$L = C \prod_{\boldsymbol{y}} P(\boldsymbol{y})^{r_{\boldsymbol{y}}}, \tag{2.11}$$

where the product runs over all the m^k response patterns and the C is a normalizing constant not dependent on the parameters.

The marginal log-likelihood, or better a function proportional to the log-likelihood, becomes

$$l_0 \sim \sum_{\boldsymbol{y}} r_{\boldsymbol{y}} \log P(\boldsymbol{y}). \tag{2.12}$$

Of course, as the number of items increases, many patterns are not observable because their associated frequency is zero. In these cases the authors suggest to include in the computation only the real observed patterns.

The basic idea in MML is to maximize the marginal log-likelihood to obtain the MML estimates of the item parameters using the first- and second-order derivatives. The observed frequencies of each response pattern are used instead of the single observations to reduce the number of computations. The full model cannot be estimated with 1 or 2 items because it is not identified. As an example, consider two multiple-choice items with 5 response alternatives. The number of parameters to be estimated is 14 for each item, that is 28. The number of possible response patterns is $5^2 = 25$, so the model is not identifiable. In case of 3 or 4 items, a direct maximization is possible through a Newton-type algorithm. In fact, the ML estimates are obtained directly maximizing the (2.12). In real dataset the number of items is larger than 4 and we need a different estimation method.

2.3.4 EM algorithm

The solution comes from the EM algorithm, an iterative two-step procedure created by Dempster et al. (1977) to solve a missing data problem. The EM algorithm was applied by Bock and Aitkin (1981) to obtain MML estimates for binary models and was extended by Thissen and Steinberg (1984) to multiple category models.

After a discretization of the continuous densities $\phi(\theta)$ and $P(\boldsymbol{y})$, the main idea of the EM procedure is to proceed with the estimation in two phases:

1. E-step: compute the expected values of the response patterns for each item, conditioned on the data and current parameter estimates;

2. M-step: maximize the log-likelihood with respect to the item parameters, using the expected values of the E-step as data.

The procedure is iterative, that is it starts with initial values for the item parameters and then repeats the two steps, using provisional estimates, until the convergence is reached.

The discretization is implemented by using Q discrete classes for the latent variable with values $\theta_1, \theta_2, ..., \theta_Q$ representing the individuals in the population.

The θ_q, with $q = 1, ..., Q$, are called *quadrature points* and allow the discrete representation of a continuous density. Of course, the larger Q is, the more precise the approximation is.

For each item j, a $m \times Q$ table $\boldsymbol{R}_j^*$, with the frequencies associated to the different responses of the individuals in the Q classes, would constitute a complete data sufficient statistic for the parameter estimation. Each element r_{jhq}^* of the table is the number of examinees in the θ_q class who choose the option h for the item j.

In detail, the E-steps computes the expected frequencies of responses for each alternative, using the current estimates of the item parameters and the θ_q, as follows

$$E(r_{jhq}^* | data; \{\boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*, \boldsymbol{\tau}\}) = \sum_{\{y_j=h\}} r_{\boldsymbol{y}} \left[P(\boldsymbol{y}; \theta_q) / \sum_q P(\boldsymbol{y}; \theta_q) \right], \tag{2.13}$$

where the sum is taken over all the examinees with $y_j = h$ and

$$P(\boldsymbol{y}; \theta_q) = \prod_{j=1}^{k} P(y_j | \theta_q) \phi(\theta_q). \tag{2.14}$$

We should note that the $P(y_j|\theta_q)$ is computed by using the (2.6) only for Q values of θ. In the end the E-step creates a $m \times Q$ table for each item with the artificial frequencies given by (2.13), used as data in the following step.

The M-step defines a set of k log-likelihoods as follows

$$l_j \sim \sum_{h=1}^{m} \sum_{q=1}^{Q} r_{jhq}^* \log P(y_j = h | \theta_q), \tag{2.15}$$

where the $P(y_j = h|\theta_q)$ is the (2.6) evaluated at θ_q. There are as many log-likelihoods terms as the items are and they should be maximized as a function of item parameters. This can be done using derivative-free methods or a Newton-Raphson approach as suggested in Thissen and Steinberg (1984). A detailed description of the equations involved in the M-step can be found in Thissen and Steinberg (1984, pg. 506-507). The sequence of the E-step and M-step iterations is repeated until the parameters become stable or a fixed number of cycles is reached.

The EM behavior depends mainly on the dataset; in some cases the change between cycles for all parameters becomes small while for some datasets most of the parameters remain stable after few cycles but few of them change indefinitely. Usually this happens for the α_h which rises toward a possibly infinite MLE. This problem is typical of the ML estimation and it is analogous to the so called "Heywood cases" for the factor analysis technique. Nevertheless we should note that in this case, even if the item parameter value is not well determined, the trace line fairly is. That's why the authors suggest to look more at the graphical representation of the response curves than at the punctual parameter estimate. Furthermore a ill-defined δ_h is caused by a associated estimated α_h close to zero.

Finally we should remark that the MCM is not completely identified. The parameter estimates computed maximizing the log-likelihood are not univocally determined, in fact if we change the sign of all the slope parameters α_h for each item the model fit remains unchanged. This operation is equivalent to reverse the direction of the latent variable θ and it represents the indeterminacy of reflections analogously to factor analysis. However this indetermination does not appear to affect the usefulness of the model, because the E-M algorithm only climbs local modes. In practical applications, choosing good starting values for the item parameters leads the algorithm to explore only the desired part of the likelihood surface and preserve the initial orientation of θ.

Fitting the algorithm is computationally intensive but using modern computers the problem is avoided. As an example, running the MCM with the program MULTILOG 7.0 edited by D. Thissen and implementing MML for a test of 20 items and a sample size of 1385 observations on a 1.73 GHz, 797 MHz Intel(R) Pentium(R) M processor takes only few seconds.

A major problem is the over-parameterization of the model leading to little information for some parameter estimation. In fact many multiple choice items include some distractors that are not attractive and may be selected only by few examinees. The parameters associated with these options need to be estimated fitting a nonlinear function giving the expected proportion of choosing that alternative. The resulting estimates would be unstable also in a hypothetical case of known θ. A possible solution would be to use prior distributions for the item parameters, restricting the estimates to a limited part of the space. A short discussion about the goodness of fit can be found in Section 2.4.3. Furthermore, an application of the MCM to the guidance test is presented in Section 2.5.

2.4 The three-parameter logistic model

In the previous section the use of the MCM for the item calibration has been described. However, the model presents several complexities due to the high number of parameters involved in the estimation process. Besides, in a practical context we are more interested on whether students could identify the correct answer. Therefore, a model for binary data, which is more stable and easy to interpret in terms of parameter estimates, can be used. Of course the data format conversion from polytomous to dichotomous causes loss of information. Nevertheless, in most cases many distractors are not so informative and the dichotomization process appears effective.

The three-parameter logistic (3PL) model (Birnbaum, 1968) is a member of the wide family of logistic models for binary data. The 3PL model cannot be directly included in the GLLAMM framework (Skrondal and Rabe-Hesketh, 2004). In fact, the model is not a generalized linear model, conditional to the latent variable, and can be included in the GLLAMM only if the guessing parameters are fixed. The same considerations are valid for the MCM, described in the previous section.

The IRT models for dichotomous responses have been proposed before the models for multiple categories and they are easier to interpret; usually the two contexts are described separately and the models for multiple response are introduced as generalizations of the models for binary data. However, coherently with the structure of this chapter, we would like to present the model showing that the 3PL model can be interpreted as a particular case of the MCM.

2.4.1 Model specification

When the item is binary, the possible responses are two: correct or incorrect. Usually the data are coded as "1" for a right answer and "0" for a wrong one. The number of categories is $m = 2$. Once more, the basic idea is the existence of a group of individuals who don't know (DK) the answer to the item j: a fraction γ_0 of them answers incorrectly while a fraction $\gamma_1 = 1 - \gamma_0$ gives the correct response. The probability of a correct response, by using the (2.6), becomes

$$P(y_j = 1|\theta) = \frac{\exp(\alpha_1\theta + \delta_1) + \gamma_1 \exp(\alpha_0\theta + \delta_0)}{\sum_{l=0}^{1} \exp(\alpha_l\theta + \delta_l)}. \tag{2.16}$$

The number of respondent given a correct response is composed of examinees who know the solution and answer correctly and of examinees who guess. The correspondent probability is expressed by the (2.16).

On the other hand, the probability of an incorrect response is

$$P(y_j = 0|\theta) = \frac{\gamma_0 \exp(\alpha_0\theta + \delta_0)}{\sum_{l=0}^{1} \exp(\alpha_l\theta + \delta_l)}, \tag{2.17}$$

reflecting only the DK people who answer incorrectly.

From the (2.16) we are able to reach an unconventional formulation of the 3PL model in few steps, see Thissen and Steinberg (1986). The probability of a correct response can be decomposed into two quantities, as follows

$$\begin{aligned} P(y_j = 1|\theta) =& \frac{\exp(\alpha_1\theta + \delta_1) + \gamma_1 \exp(\alpha_0\theta + \delta_0)}{\sum_{l=0}^{1} \exp(\alpha_l\theta + \delta_l)} = \frac{\exp(\alpha_1\theta + \delta_1) + \gamma_1 \exp(\alpha_0\theta + \delta_0)}{\exp(\alpha_0\theta + \delta_0) + \exp(\alpha_1\theta + \delta_1)} \\ =& \gamma_1 \frac{\exp(\alpha_0\theta + \delta_0)}{\exp(\alpha_0\theta + \delta_0) + \exp(\alpha_1\theta + \delta_1)} + \frac{\exp(\alpha_1\theta + \delta_1)}{\exp(\alpha_0\theta + \delta_0) + \exp(\alpha_1\theta + \delta_1)}. \end{aligned}$$

Then the constraints imposed with the MCM can be used, in fact $\alpha_0 = -\alpha_1$ and $\delta_0 = -\delta_1$ so that

$$P(y_j = 1|\theta) = \gamma_1 \frac{\exp[-(\alpha_1\theta + \delta_1)]}{\exp[-(\alpha_1\theta + \delta_1)] + \exp(\alpha_1\theta + \delta_1)} + \frac{\exp(\alpha_1\theta + \delta_1)}{\exp[-(\alpha_1\theta + \delta_1)] + \exp(\alpha_1\theta + \delta_1)}.$$

Finally, after some algebraic transformations, we find the formulation of the unconventional 3PL model

$$\begin{aligned} P(y_j = 1|\theta) =& \gamma_1 \left[1 - \frac{1}{1 + \exp[-2(\alpha_1\theta + \delta_1)]} \right] + \frac{1}{1 + \exp[-2(\alpha_1\theta + \delta_1)]} \\ =& \gamma_1 + (1 - \gamma_1) \frac{1}{1 + \exp[-2(\alpha_1\theta + \delta_1)]}. \end{aligned}$$

The 3PL model is written, in the conventional form, as

$$\begin{aligned} P(y_j = 1|\theta) =& \gamma_j + (1 - \gamma_j) \frac{1}{1 + \exp[-\mathbf{D}\alpha_j(\theta - \beta_j)]} \\ =& \gamma_j + (1 - \gamma_j) \frac{\exp[\mathbf{D}\alpha_j(\theta - \beta_j)]}{1 + \exp[\mathbf{D}\alpha_j(\theta - \beta_j)]}, \end{aligned} \tag{2.18}$$

where $\mathbf{D} = 1.702$ and $\delta_j = \alpha_j\beta_j$. This formulation of the model is coherent with the (2.3) equivalence and it is implemented in the software we are going to use for the analysis. The model is said to be in the *normal metric*. As we can see from (2.18), the item parameters are referred to the item j. In IRT models for binary data, the item characteristic curve (ICC) for an item j simply describes the relation between performance on the item j and the latent ability. Particularly, $P(y_j = 1|\theta)$ indicates the probability than an examinee with ability level θ answers correctly the item j. The set of item parameters for each item j is $\{\alpha_j, \beta_j \, \gamma_j\}$.

2.4.2 Parameter interpretation and identification

Again consider item 5 of the Psychology faculty. The ICC under the 3PL model is represented in Figure 2.2.

The item parameters interpretation is quite straightforward. The α_j is the *discrimination parameter* of the item, that is the capability of differentiating between the examinees with different ability levels. The higher α_j is, the more discriminant the item and steeper the ICC are. In fact, from a geometrical point of view, α_j is proportional to the slope of the ICC at the point $\theta = \beta_j$. The β_j represents the *difficulty parameter* for the item j and its values are collocated on the same scale of θ. The β_j is a location parameter because it defines the

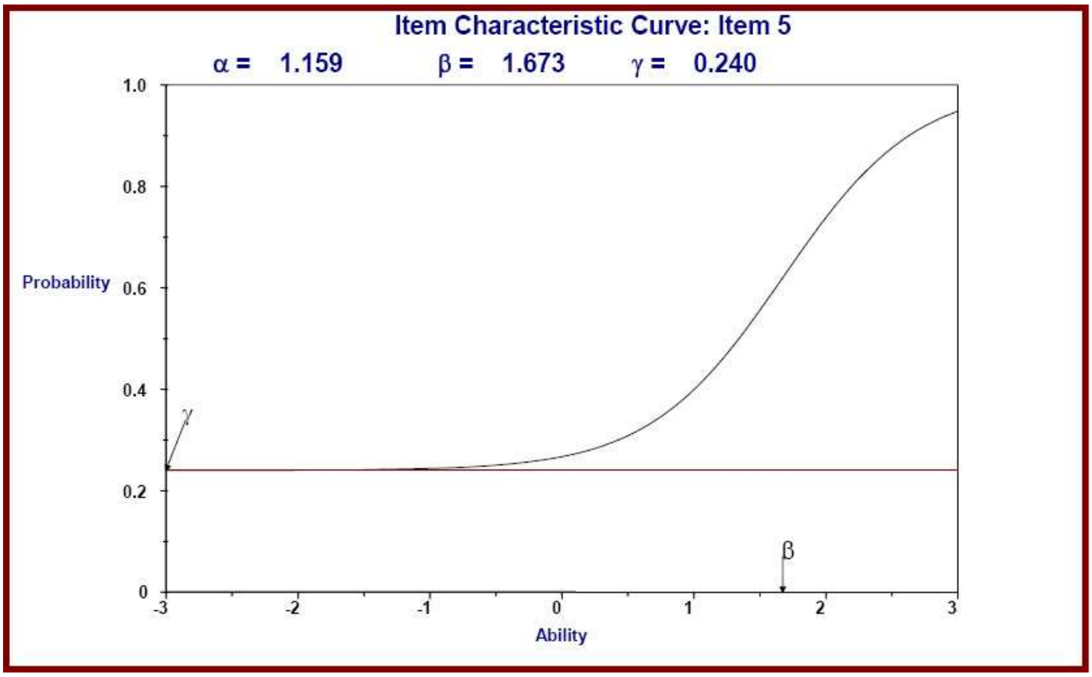


Figure 2.2. ICC for the item 5 of the Psychology faculty, according to the 3PL model

position of the ICC respect to the ability values. Particularly, as the difficulty parameter increases the ICC moves to the right, that is a higher level of ability is required to have the same probability of a correct answer. On the other hand, as the β_j decreases the ICC moves to the left side. Finally, the γ_j is called *guessing parameter* or, more precisely *pseudo-chance level parameter* (Hambleton and Swaminathan, 1985). Geometrically, it is the lower asymptote of the ICC and represents the probability of examinees with low ability to correctly answer the item j. According to the 3PL model, the probability of a correct response is never zero because a guessing factor is introduced. For this reason, the point on the horizontal axis where the difficulty β_j of the item is equal to the ability level θ correspond to $(1 + \gamma_j)/2$ on the vertical axis. In this case, the probability of a correct response is exactly the mean value between the highest and lowest probabilities of success. In the IRT applications to educational assessment and knowledge tests, the guessing parameter should never be excluded. In fact, the hypothesis of not guessing examinees is not reliable in this context while, for psychological tests, it may be likely and coherent. The information about the guessing parameter comes from data of people who respond incorrectly. In fact, referring to the individuals who don't know the correct answer, we can imagine that a proportion of them will guess the item while the remaining examinees will give the incorrect answer. The available data are the latter and they are used to estimate the fraction γ_j of the probability of really not knowing the answer (Thissen and Steinberg, 1986).

The ability domain is the whole set of real numbers and consequently this is also the range for the β's. Usually, only the interval $[-3; +3]$ is taken into account. The value of β_j for the item 5 in the figure 2.2 is equal to 1.673. The item is rather difficult. The discrimination parameter can also take any real value. In practice, a negative α_j would result in a decreasing ICC, and this is not coherent with the model interpretation. In fact, it would correspond to a decreasing probability as ability increases. If an item has a negative discrimination in the test, it probably is not coherent with the test specification, because it goes on the opposite direction respect to the latent ability. On the other hand, the α's should not be extremely high. This case would create a step function, that is the probability of a correct response would be equal to γ_j or equal to one, with no intermediate values. Usually we take the range $[0; 2]$ for the α's, even if values equal to zero mean lack of discrimination power. The item 5 has $\alpha_j = 1.673$ so the discrimination power is pretty high and this can be seen from the steepness of the curve. Obviously, the range interval for the guessing parameter is $[0; 1]$. The γ_j is the probability of guessing and should not have an extremely high value. For item 5, $\gamma_j = 0.24$.

From (2.18) we can show that the parameters are not identified but have two types of indeterminacies. In fact, if we multiply θ (and β_j because they are on the same scale) by a constant and divide α_j by the same constant the probability $P(y_j = 1|\theta)$ is preserved. Furthermore, if we add to θ (and to β_j) the same quantity, the probability doesn't change as well. Usually, the identification problem is solved by fixing the distribution of θ so that its mean value and standard deviation are 0 and 1, respectively.

The 3PL model can be estimated by the MML procedure with the implementation of the EM algorithm, analogously to the MCM. Therefore, the procedure is not described again.

2.4.3 Goodness of fit

Some general considerations about the goodness of fit for the IRT models are presented in this section. The goodness of fit is a crucial issue in IRT, because traditional tests for absolute fit loose their validity in case of sparse data. For multiple category models the fit to data against the general multinomial alternative may be evaluated with the conventional Pearson's or likelihood ratio test statistics. Unfortunately, they can be used only when the sample size is large and the number of items is small so that the m^k cross-classification of the responses produces a contingency table that may have expected values sufficiently large, where m is the number of alternatives for each item and k is the number of items. In this case, the statistics follow a Chi-square distribution and the overall goodness of fit of the model can be tested. Consider a test with multiple-choice items with 5 response alternatives. When the test is very short, from 3 to 5 items, the number of response patterns is quite moderate ($5^3 = 125$ for 3 items and $5^5 = 3125$ for 5 items) while with 6 items, the number of patterns becomes $5^6 = 15625$: this sample size is hardly reached in practice. Longer tests can be used in case of binary items where 10 items produce $2^{10} = 1024$ different possible response patterns, even if the number of patterns increases exponentially and it becomes very high with long tests (more than 15 items). The likelihood ratio test between hierarchically nested models may be used to evaluate the significance of different parametrizations. For large contingency tables, the likelihood ratio test between hierarchically nested models may be used to evaluate the significance of the additional parameters of the larger model. The likelihood ratio test G^2 compares the expected and observed pattern frequencies as follows

$$G^2 = 2 \sum_{\boldsymbol{y}} r_{\boldsymbol{y}} \ln \frac{r_{\boldsymbol{y}}}{N \bar{P}(\boldsymbol{y})}, \tag{2.19}$$

where the sum runs over all the 2^k response patterns in case of binary data, $r_{\boldsymbol{y}}$ is the frequency of pattern $\boldsymbol{y}$, N is the number of observations and $\bar{P}(\boldsymbol{y})$ is the marginal probability for $\boldsymbol{y}$, respectively.

The G^2, in presence of small contingency tables, follows a Chi-square distribution with $2^k - kp - 1$ degrees of freedom, where p is the number of item parameters in the model. The difference in the G^2 can be used to compare the fit of more versus less constrained models, even for sparse contingency tables. Proposals for sparse data have been developed by Glas (1988) for the Rash family models that do not allow differential slope or guessing parameters, Reiser (1996) and Bartholomew and Leung (2002). These solutions are based on the analysis of residuals from the marginal tables for pairs of variables. The problem of goodness of fit in IRT is still open, especially in terms of absolute fit of the model to data, and more research is certainly needed.

2.5 IRT calibration for the guidance test

The IRT unidimensional models described in this chapter are implemented in the calibration process of the item bank for the guidance tests. The multiple-choice model and the three-parameter logistic model are simultaneously used to find out the item properties and point out the main problems.

The first one is applied to the multiple-choice data to have a graphical representation of the single response curves and to understand the behavior of all the alternatives while the second one is implemented to estimate the item parameters, restricting the data to the binary case, that is "correct" and "wrong" responses.

The final objective of the item calibration is to create tests according to the following specifications:

- For each item, response alternatives with coherent behavior with respect to the latent trait;
- Items with high discrimination power;
- Items with moderate guessing parameter;
- For each test, items with different levels of difficulty.

The first condition reflects the considerations about the probability of selecting each single alternative as a function of the latent ability θ. We have already seen that, according to the MCM, the response curves should have a monotonic increasing trend for the correct alternative and a decreasing or non monotonic trend for the incorrect ones.

The second aspect concerns once more the relationship between the item and the trait. The discrimination is a slope parameter and its high value ensures a strong link between the item and the ability and is a symptom of unidimensionality. Furthermore, a test with high discriminating items is able to well differentiate between the examinees: small differences in the abilities correspond to substantial differences in the probabilities of a correct response.

Then, the probability of a correct response for the examinees with low abilities (that is, the guessing parameter) should be moderate; for example, with 5 alternatives multiple-choice items we can expect a probability of $1/5$ under the assumption of a complete random answer. In reality, we know that the response is hardly casual, because the examinees often have some knowledge which can help them to eliminate some of the alternatives. The wrong options don't have the same distractive power and this is the reason why, once more, the item development is an hard step.

Finally, each test should consist of easy, moderate and difficult items: this property is fundamental to obtain a rigorous evaluation of individuals, that are usually heterogeneous respect to their ability.

All these characteristics are difficult to obtain, especially with the first draft of the items. The item calibration turns out to be a continuous process implemented for all the duration of the guidance project.

A preliminary calibration, conducted on the first draft of the item bank, gave us indications about which items were completely not coherent with the project's objectives and needed to be excluded from the item bank. Furthermore, we had the possibility to modify the item specification and make it more clear to examinees. In the future, the calibration process will be used to test new items to insert in the test.

For the present work, the item calibration is used to find stable item parameter estimates. The ability estimation has not been performed yet. Students know the number of correct response for the general and the specific items. Future developments of the project will include the characterization of the θ's for examinees who produce a particular response pattern $\boldsymbol{y}$, taking the item parameters as known. The results of the ability estimation will be considered to create qualitative judgments about the students' performance.

Mainly, the success of the item calibration depends on two aspects:

- The relation of the responses to the latent ability;
- The sample size.

The first condition attains to the model choice and to the intensity of the relation between the items and the trait being measured. If the latent ability strongly influences the responses, the calibration will produce quite precise estimates with a small sample size.

With respect to the second aspect influencing the results, we should say that there is not a predefined sample size required to obtain a successful calibration. Clearly, the number of observations is dependent on the complexity of the selected model, on the number of parameters to be estimated and on the information provided by the data about each single parameter.

Thissen and Wainer (1982) noticed that few hundreds of examinees may serve to calibrate the items under the 1PL model while tens of thousands could be required for the 3PL model. On the other hand, following a recent paper by Rupp (2003) on the MML estimation of IRT models summarizing different applicative studies, the sample size seems to be smaller. With 15 to 50 items, about 250 observations are required for the 1PL and the 2PL while about 500-1000 may serve to calibrate the items under the 3PL and the MCM. No real guidelines are available for the standard errors of the estimates to be considered sufficiently precise.

As already discussed in the first chapter, each faculty-specific test consists of two different section: the general culture and the specific one. For each test, only 10 general culture items are submitted by using a block design created on an item bank of 30 items. The general culture items are common to all the faculties and they are calibrated separately respect to the 20 specific items to preserve unidimensionality and to gain information from the complete sample size.

Software The calibration phase of the item bank is conducted by using the software MULTILOG 7.0 (Thissen, 2003). The computer program can fit unidimensional IRT logistic models for polytomous and binary data. In the first category we find the graded response model (Samejima, 1969) for ordinal responses and the multiple-choice model (Thissen and Steinberg, 1984) for nominal data. Finally, the logistic models (1PL, 2PL, 3PL) can be applied to dichotomous items. All the models are estimated by using MML together with the EM algorithm.

Other IRT-specific software includes:

- BILOG-MG (Zimowski et al., 1996), for binary logistic IRT models;
- PARSCALE (Muraki and Bock, 1997), for binary and ordinal logistic IRT models;
- TESTFACT (Bock et al., 2002), for multidimensional binary probit models.

A complete review of the models and estimation procedure for the cited computer programs can be found in du Toit (2003).

2.5.1 The problem of incomplete design

Applications to real data typically encounter the problem of missing data. Missing data can be of three different types:

- Omitted items;
- Not presented items;
- Not reached items.

The first group consists of items voluntary skipped by the examinees, typically because they don't know or they are not sure about the correct answer. This type of missing can be simply treated as 'wrong response' or can be subject to imputation. The second group of missing occurs when the item is not administered to the respondents, that is the tester wants the examinees to answer to different blocks of items, which are a subset of the total item pool. The design of this test is called *incomplete*. The last type of missing is common in time-limited questionnaires, because not all the examinees are able to reach the last items of the test. In this case, the missing can be scored as 'wrong response', as 'correct response' or can be imputed. The missing data treatment depends on the context and on the specific test administered. The guidance test is without time limits but implements the incomplete design, as discussed in the first chapter. Therefore, it contains omitted and not presented items.

MULTILOG (Thissen, 2003) has been used for the unidimensional calibration. The computer program does not take into account different types of missing data, that is, it does not distinguish between not presented items and omitted responses. This is not a problem for the specific items, while it may introduce some bias in the estimation of the general culture item parameters. In fact, the specific items do not contain not presented items while the general culture items do, because of the block design. The missing data are treated as M.A.R. (missing at random) by MULTILOG.

The problem of incomplete design is especially encountered in the implementation of a multidimensional approach, because the items from the general culture and the specific sections are jointly analyzed. The incomplete design is taken into account by standard software for full-information factor analysis (see Chapter 3), as an example in TESTFACT (Bock et al., 2002). The incomplete design will be especially treated in case of Markov chain Monte Carlo estimation (see Chapter 4).

2.5.2 A specific case: the Psychology faculty

The aim of this section is to provide the results of the item calibration for the Psychology test. The 20 specific items are taken into account and the analysis has been conducted with a sample size of 1385 respondents. All the response are omitted for 3 examinees, so the actual sample size is 1382. First of all, some classical statistics are presented. Table 2.2 shows the number of respondents at each score level. The mean of the test score $T(i)$ is equal to 11.7 with a standard deviation of 3.0.

The percentages of correct, incorrect and omitted responses for the 20 items are shown in Table 2.3. The proportion of correct response, in the CTT context, is denominated item *facility* or *p-value.* From the table we can infer that the items present different levels of difficulties: in fact, the percentages of correct responses have a wide range of variation, from 21.66% of item 1 to 96.10% of item 14. Especially for item 14, the value seems to be rather extreme. Therefore, the results suggest that the total test score, in terms of sum of correct responses, is not a valid indicator of the student's performance.

Reliability is assessed with the *coefficient alpha* for binary items, that is the

score	obs.	score	obs.
0	0	11	168
1	2	12	205
2	1	13	169
3	6	14	138
4	10	15	107
5	18	16	68
6	27	17	31
7	52	18	18
8	79	19	7
9	127	20	7
10	142		

Table 2.2. Number of observations at each score, Psychology specific test

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	21.66	76.46	1.88	11	35.81	59.71	4.48
2	48.66	48.23	3.10	12	25.99	70.32	3.68
3	35.88	63.10	1.01	13	29.96	68.45	1.59
4	68.30	30.25	1.44	14	96.10	2.74	1.16
5	31.26	66.28	2.45	15	88.88	10.04	1.08
6	39.06	59.86	1.08	16	52.27	45.49	2.24
7	89.82	7.51	2.67	17	71.55	26.71	1.73
8	64.48	32.35	3.18	18	78.27	19.86	1.88
9	76.17	20.87	2.96	19	32.92	63.10	3.97
10	83.10	14.87	2.02	20	89.39	8.66	1.95

Table 2.3. Percentage of correct, incorrect and omitted responses for the Psychology specific items

Kuder-Richardson coefficient (Kuder and Richardson, 1937), calculated as follows

$$\mathrm{KR20} = \frac{k}{k-1} \frac{s^2 - \sum_{j=1}^{k} p_j(1-p_j)}{s^2}, \tag{2.20}$$

where k is the number of items, s^2 is the variance of the test scores and p_j is the item p-value. The KR20 indicator reflects the degree of agreement between all the items, in the way they measure the same theoretical construct. The estimated coefficient is equal to 0.64 and it is not particularly high.

Figure 2.3 shows the point biserial correlations plotted against the p-values. The point biserial correlation is the correlation between the test-takers's perfor-

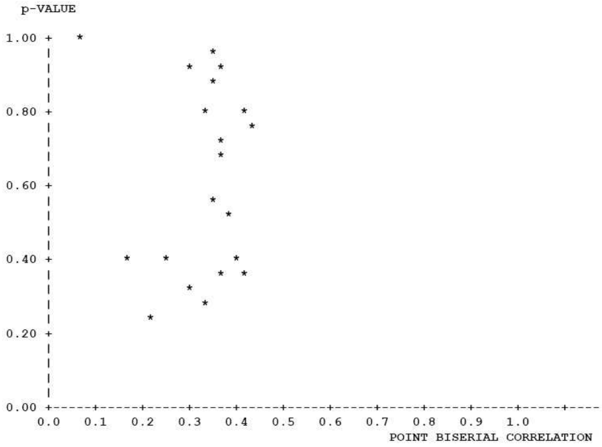


Figure 2.3. Plot of point biserial correlations vs. p-values

mance on one item compared to the test-takers' performance on the total test, calculated as follows

$$r_{pbis} = \frac{\mu_j - \mu}{s}\sqrt{p_j/(1-p_j)}, \tag{2.21}$$

where μ_j is the mean score on the test for those who get the item j correct, μ is the mean score on the test for the entire group, s is the standard deviation of the test score and p_j is the item p-value. The biserial correlation in (2.21) reflects the discrimination of the items. The plot shows that there is quite a large variation in these two indicators. In general, the plot of a measure of correlation and difficulty is useful to detect the presence of outliers, that is items with extreme values. The variation of the point biserial correlations indicates that there is a variation on how well the items discriminate. The results suggest that item discriminations and difficulties should not be excluded from a correct calibration of the test.

As a second step, we can look at the figure 2.4, representing the response category curves for each item according to the multiple-choice model obtained with MULTILOG.

Figure 2.4 shows the results for the 20 specific items, each one with 5 response category curves for the 5 multiple-choice alternatives. The horizontal axis represents the ability values, typically from -3 to 3, and the vertical axis the probability range $[0; 1]$. Each curve expresses the probability of selecting the correspondent alternative, as a function of ability. The convention of MULTILOG is

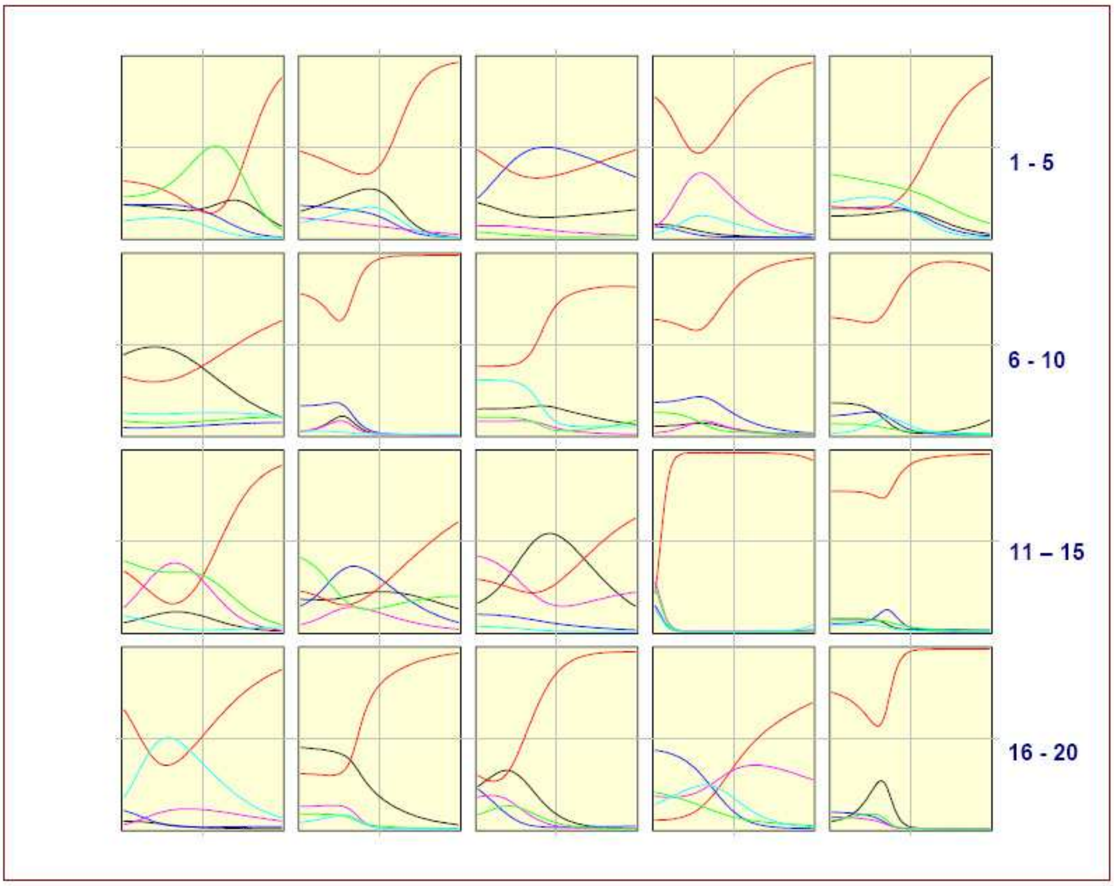


Figure 2.4. Matrix plot of MCM response category curves for the Psychology faculty, items 1-20

to highlight the correct option with a red line. The items which strictly respect the MCM features are the item 5 and the item 19. In the other cases the red curve slightly decreases for low ability levels and increases on the rest of the θ domain. Furthermore, for items (2,4,7,8,9,10,14,15,20), the correct option is always preferable, that is the distractors do not seem to be attractive for all the ability levels. The graphical representation for the MCM suggests, for several items, a decreasing trend for the correct alternative that is not coherent with the model interpretation, even for low ability levels. The model is not simply a multinomial logit model as (2.4) but includes the response probability of the latent category representing the totally undecided individuals as described by (2.6). Therefore, a large and positive α_h for the correct alternative h does not ensure an increasing trend for the response curve, for all the ability levels. The trend of the curve depends also on the estimates of γ_h and of the parameters for the latent category "0". As an example, an increasing trend in all the domain of θ is likely when the guessing parameter γ_h is very low or when the α_0 is not negative and large. It is very important to include in the item analysis the behavior of the incorrect options: the idea is that the item characteristics not only depend on the stem itself but also on the attractiveness of the different response alternatives. Nevertheless, we know that the MCM has identification problems and involves the estimation of a high number of parameters (for 20 multiple-choice items with 5 alternatives the number of parameters to be estimated is 280). MULTILOG fixes the item parameters according to the constraints discussed in section 2.3.1 but other identification rules are possible and may improve the interpretability of the results.

Restricting the item format to the binary case we can use the 3PL model to estimate the item parameters. The items have been calibrated by using MULTILOG. The number of quadrature points for MML estimation is 19, from -4.5 to 4.5 with intervals of 0.5. The EM-cycle convergence criterion is 0.001 and 25 cycles have been run. A Bayesian prior has been used for the logit of all the guessing parameters. The number of alternatives is 5, so we have chosen a Gaussian prior with mean equal to -1.4 (the logit of 0.2) and standard deviation equal to 1. The results of the parameter estimation in terms of discrimination α, difficulty β and guessing γ, all in the traditional normal metric, for the Psychology test are shown in Table 2.4.

Item	α	β	γ	Item	α	β	γ
1	2.76	1.93	0.19	11	1.24	1.03	0.20
2	0.94	0.89	0.32	12	0.75	1.84	0.15
3	1.90	2.25	0.35	13	1.13	1.94	0.25
4	0.81	0.20	0.44	14	0.10	-20.00	0.20
5	1.16	1.67	0.24	15	0.57	-2.26	0.24
6	0.65	2.23	0.32	16	0.62	0.93	0.33
7	1.05	-1.85	0.18	17	0.75	-0.82	0.12
8	0.42	-0.79	0.11	18	0.85	-1.16	0.11
9	0.51	-1.30	0.22	19	0.61	1.03	0.07
10	0.59	-1.75	0.18	20	1.97	-1.07	0.48

Table 2.4. Item parameter estimates for the 3PL model, Psychology specific items

The 3PL model is estimated as a binary form of the multiple-choice model: contrasts between the two slopes (correct and incorrect) and intercepts are estimated, as well as the logit of the lower asymptote. Table 2.5 shows the estimates with the associated standard errors.

Item	α-contrast	δ-contrast	logit of γ	Item	α-contrast	δ-contrast	logit of γ
1	4.70 (1.68)	-9.07 (3.19)	-1.43 (0.10)	11	2.11 (0.35)	-2.16 (0.41)	-1.37 (0.17)
2	1.60 (0.34)	-1.43 (0.43)	-0.77 (0.18)	12	1.27 (0.35)	-2.34 (0.60)	-1.71 (0.26)
3	3.24 (1.74)	-7.27 (3.66)	-0.64 (0.09)	13	1.92 (0.81)	-3.72 (1.37)	-1.12 (0.13)
4	1.38 (0.32)	-0.27 (0.45)	-0.24 (0.33)	14	0.17 (0.21)	3.34 (0.36)	-1.40 (1.00)
5	1.97 (0.49)	-3.29 (0.93)	-1.15 (0.13)	15	0.96 (0.17)	2.18 (0.38)	-1.15 (1.20)
6	1.10 (0.57)	-2.45 (1.05)	-0.77 (0.24)	16	1.05 (0.32)	-0.98 (0.54)	-0.73 (0.32)
7	1.79 (0.26)	3.30 (0.31)	-1.50 (0.96)	17	1.27 (0.16)	1.05 (0.21)	-2.00 (0.93)
8	0.71 (0.11)	0.56 (0.25)	-2.09 (1.21)	18	1.44 (0.15)	1.67 (0.20)	-2.10 (1.01)
9	0.86 (0.16)	1.12 (0.37)	-1.29 (0.97)	19	1.03 (0.18)	-1.06 (0.29)	-2.61 (0.77)
10	1.01 (0.16)	1.76 (0.30)	-1.48 (1.22)	20	3.36 (0.64)	3.59 (0.46)	-0.09 (0.37)

Table 2.5. Slope and intercept contrasts and logit of γ for the 3PL model, Psychology specific items (standard errors in brackets)

Referring to Table 2.4, we can notice that the discrimination estimates are quite good, except some low values on the items (8,9,10,14). Items with a value of α greater than 1 present a moderate/high capability of differentiating between the examinees and the correspondent ICC's result quite steep. The ICC's for all the items are shown in Figure 2.5.

No extreme high values are noticed, even if the discrimination for item 1 is quite large and makes the curve very steep. The tendency is to create a curve which does not differentiate between examinees with low ability values. In fact, as we can see from the corresponding ICC in Figure 2.5, examinees with low and medium ability levels all have a probability of success equal to the guessing parameter. The item is able to well discriminate only for a small subset of examinees, that is the individuals with high ability. This is also because the item is rather difficult (the position of the ICC is right-shifted).

The items are quite different in terms of difficulty parameters and this is a desirable property for the test. Only an extreme value is noticed corresponding to item 14.

In fact, item 14 shows the relevant problem of extreme easiness. The item has 96.1% of correct response, presents a discrimination power close to 0 (reflecting a flat ICC in Figure 2.5) and a not acceptable estimate for the difficulty parameter (-20.00). Generally, items with extreme values are not useful for student's evaluation since they provide no information about ability.

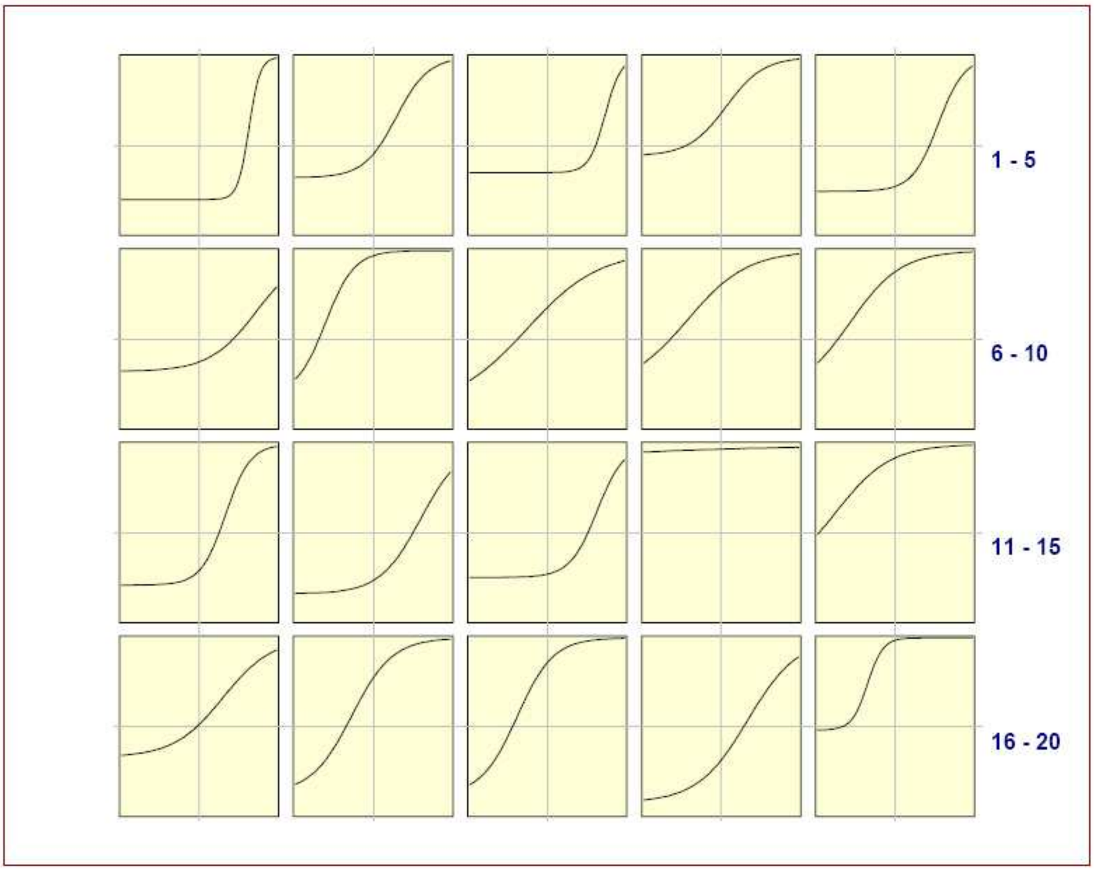


Figure 2.5. Matrix plot of the ICC's for the 3PL model, Psychology faculty, items 1-20

Finally, the guessing parameter seems to be quite moderate for about 2/3 of the items. Items (2,3,4,6,16,20) have a guessing parameter greater than 0.3. The estimation has been conducted also without imposing a prior on the γ_j's. In this case, the guessing is estimated equal to zero for items (8,17,18,19), which have low guessing according to Table 2.4. The guessing parameters for the other items are generally overestimated respect to the results obtained imposing a Gaussian prior.

The observed and expected proportions for a correct response are given in Table 2.6 for all the 20 specific items.

The observed and expected proportions are quite similar; nevertheless, as we stated in section 2.4.3, a deep residual analysis should be conducted on couple or triplets of items.

Item	Obs.	Exp.	Item	Obs.	Exp.
1	0.2208	0.2218	11	0.3749	0.3736
2	0.5022	0.5016	12	0.2699	0.2700
3	0.3625	0.3618	13	0.3045	0.3048
4	0.6930	0.6920	14	0.9722	0.9722
5	0.3205	0.3205	15	0.8985	0.8984
6	0.3949	0.3948	16	0.5347	0.5343
7	0.9228	0.9229	17	0.7281	0.7274
8	0.6659	0.6652	18	0.7976	0.7979
9	0.7850	0.7847	19	0.3429	0.3419
10	0.8482	0.8485	20	0.9116	0.9100

Table 2.6. Observed and expected correct proportion comparison for the 3PL model, Psychology specific items

The likelihood ratio test does not follow a Chi-square distribution because the contingency table is sparse (the number of possible response patterns is 2^20 while the sample size is only 1385). Therefore, we can only implement a test between nested models. We have estimated the 2PL model and the correspondent difference between the G^2 is equal to 84,3. The resulting p-value is < 0.000 so the test is significant. The 3PL model seems to fit the data better than the 2PL model.

2.5.3 General conclusions for the test

The test calibration has been performed for all the guidance tests. The general culture items have been analyzed by using the complete dataset of 9370 respondents while the specific items were calibrated with the sample sizes described in Table 1.2. Particularly, the faculties of Agriculture, Education Science and Statistics obtained a small number of respondents. We have already mentioned that the success of the calibration depends also on the sample size. It is difficult to obtain stable estimates with a small number of respondents and we hope to increase it in the future.

The percentages of correct, incorrect and missing responses for all the faculties on the specific items are shown in Appendix A (the results on the general culture items have been already reported in the first chapter, see Table 1.7). The percentages of correct responses are very different inside the same test, suggesting once more the idea that the number of correct responses cannot be used as test score. The percentage of omitted items is quite low for general culture items and the Political Science test respect to the other faculties.

The response category curves for the MCM are shown in Appendix B, both for general culture and faculty-specific items. The behavior of the single items should be evaluated. As a general comment, we can see that most items respect the MCM features for the general culture section and for the faculties of Arts and Humanities, Economics and Political Science. A deeper work on the behavior of the alternatives is needed.

Finally, Appendix C shows the results of the item parameter estimates according to the 3PL model for the general culture and the specific items. The items should be excluded from the test when they present a negative or close to zero discrimination. As examples, item 4 (Arts and Humanities, Table C.3), item 6 (Foreign Languages and Literature, Table C.6) and item 6 (Statistics, Table C.9). Very high discrimination estimates (greater than 2.5) should also be considered carefully. In fact, the ICC becomes a step function: the probability of a correct response is equal to the guessing parameter or to one for most ability levels. Clear examples can be found in item 19 (Agriculture, Table C.2), item 15 (Education Science, Table C.5), item 19 (Statistics, Table C.9). Difficulty estimates are quite different inside each single dataset: extreme easy items are noticed, as an example, in the Arts and Humanities test (items 4 and 14, Table C.3). The guessing parameters are quite moderate except few cases.

Chapter 3

Multidimensional approach and Full Information Factor Analysis

3.1 Towards a multidimensional approach

The structure of the guidance test has been a source of wide debate in terms of dimensionality. So far we have described the phase of calibration for the tests considering the general culture and the faculty-specific parts as separated. This approach is strongly confirmative and based on the assumption of unidimensionality. In fact, we have supposed that a single latent trait is underlying the general culture responses. Analogously, the existence of a single latent ability has been assumed for the specific section of all the guidance tests. On the other hand, a multidimensional approach can be introduced to allow the simultaneous existence of more than one latent ability.

3.1.1 The importance of dimensionality

The detection of dimensionality is a crucial issue in IRT and in latent variable modelling. Recently, a work by Tate (2003) reviewed methods for empirically assessing the structure of tests with dichotomous items. In particular, the author pointed out the importance of assessing the test statistical structure, resulting from the interaction of examinees with items. This aspect should be an important part of the development, evaluation, and maintenance of large-scale tests.

Many IRT models are based on the unidimensionality assumption. Nevertheless, the assumption of local independence is valid only when the complete latent space has been specified. For this reason, many researchers tried to define the concept of dimensionality and to develop methods for its detection. Our idea is that still a precise and unambiguous definition of dimensionality does not exist yet. This is a direct consequence of the latent nature of the phenomenon and of the impossibility of comparison with observed results.

Hambleton and Swaminathan (1985) justified the assumption of unidimensionality for IRT models with the existence of a dominant trait able to account for the responses. In this sense, we can imagine that a single trait always exists but the real point is how dominant the trait is. On the other hand, an opposite approach comes from Traub (1983), who argued that unidimensionality is probably more the exception than the rule, considering all the skills necessary to solve the items on most cognitive tests.

Adams et al. (1997) reviewed some weak points of unidimensionality assumption in order to propose a multidimensional IRT model. First of all, unidimensionality may be inappropriate for tests deliberately constructed from subcomponents that are assumed to measure different traits. IRT models seemed to be robust to these violations of unidimensionality, especially with highly correlated traits. In fact, if a single latent trait is assumed, it can be considered the dominant factor reflecting the different composition of the items. Nevertheless, when a test contains mutually exclusive subsets of items or when the underlying dimensions are not highly correlated, the use of a unidimensional model can bias the parameter estimation, adaptive item selection and trait estimation. The problem is highlighted especially in adaptive testing, when the examinees are administered different combinations of items and the traits underlying the performance may reflect the different composition of the items.

Secondly, the assessment of knowledge, competencies and achievement is more and more going toward a multidimensional evaluation. As an example, the evaluation of examinees should be conducted not only in terms of degree of correctness of the responses but also in terms of strategy and methods applied in the performances. In the University context, the student's evaluation is typically multidimensional at each level: inside a single course and during all the University career, students are evaluated on the basis of multiple competencies.

Finally, the multidimensional approach turns out to be more efficient when calibrating several unidimensional tests (test batteries) because it takes the correlations between the latent abilities into account (Wang et al., 2004).

What we would like to investigate is the possibility of implementing a multidimensional approach for the guidance test by using Multidimensional Item Response Theory (MIRT) models. In fact, multidimensionality seems to be within the framework of our guidance test: each test is made up of two sections and each student can try different faculty-specific tests. Ackerman (1994) highlighted the use of a two-way classification scheme for standardized achievement tests: items should be classified according to the content and to the type of skill required to succeed. In this sense, MIRT represents a useful mean to validate the test specification, helping test developers to understand which composite skills are being measured at item and test level.

3.1.2 Main features

Mainly, the multidimensional investigation can be conducted by using two different approaches: the *exploratory* approach and the *confirmatory* one. In the first one no prior knowledge is included in the model, in terms of relationship between the items and the latent traits. The number of abilities may be specified in advance: in this case the method is not purely exploratory. According to the confirmatory approach, not only the number of latent variables is pre-specified but also their relationships with the items. In fact, the researcher can use prior knowledge to define which items load on which factors.

Furthermore, *between-item* or *within-item* multidimensionality can be specified. In the first case, each subset of items is designed to measure a specific latent trait, that is each item is allowed to load only on a single ability. On the other hand, in the within-items solution some items are designed to simultaneously measure more than one single latent trait. This approach becomes exploratory when all the items are allowed to measure all the latent traits.

Since the concept of dimensionality is very wide and complex, a good choice should always take into account the aim of the test. In fact, the exploration of a test internal structure may be affected by the purpose of assessment (Reckase, 2004). In this sense, a deep knowledge of the test battery structure or test structure may require a high dimensionality while student's evaluation should be simplified as much as possible with a low dimensionality.

If we go back to the guidance project and consider the faculty-test level, an example of multidimensional approach is presented in Figure 3.1.

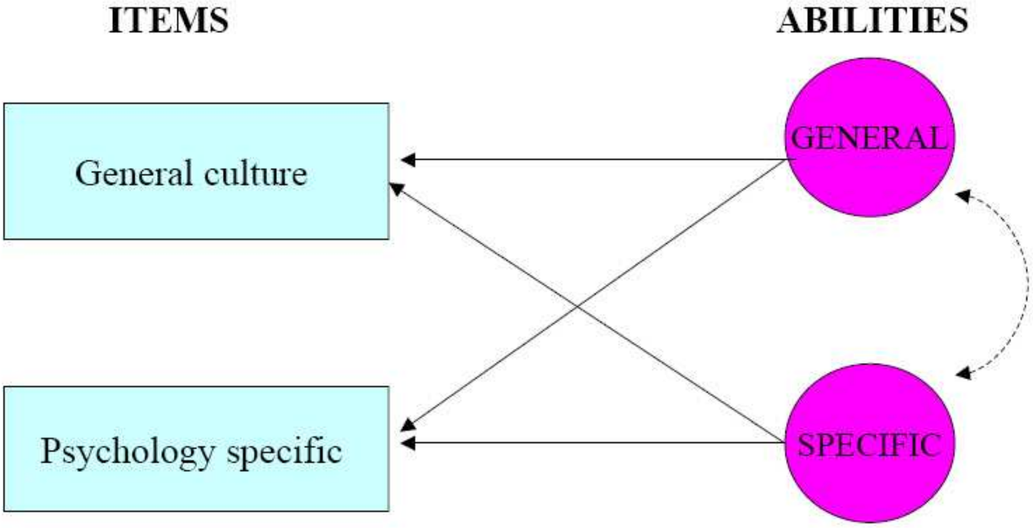


Figure 3.1. Dimensionality for student's evaluation - Faculty of Psychology

Figure 3.1 is referred to the Psychology faculty test. We suppose the existence of two latent abilities, the general and the specific, which should be able to account for all the variability between the items. The approach is exploratory (but we assume two latent traits) and complete within item multidimensionality has been specified.

The main estimation methods for item factor analysis are:

- Factor analysis of tetrachoric correlations (Muthén and Christoffersson, 1981; Joreskog, 1990);

- Full-information factor analysis (Bock et al., 1988);

- Markov chain Monte Carlo estimation (Albert, 1992; Béguin and Glas, 2001).

The first method will not be considered, while the Full-information factor analysis will be discussed in the following section. Finally, a Bayesian approach will be described in Chapter 4.

3.2 Full-information factor analysis

Item factor analysis has been proposed all along as a method to investigate the dimensionality of a set of items. The attractive feature of factor analysis is the possibility of analyzing simultaneously a set of items and assigning them to particular dimensions. Furthermore, the interpretation of common item sets is possible in terms of content and format, respect to multiple factors. Full-information factor analysis (FIFA) is based on marginal maximum likelihood (MML) estimation via the EM algorithm (Bock and Aitkin, 1981) and can be distinguished from the partial information methods due to the different use of the available data. In fact, the full-information method uses the frequencies of all distinct response patterns, i.e. the information on the joint frequencies of all orders, while the limited information methods based on tetrachoric correlations (Divgi, 1979) and generalized least squares (Christoffersson, 1975; Muthén, 1978, 1984) use low-order joint occurrence frequencies. The most important work on FIFA is the paper by Bock et al. (1988), which provided a detailed extension of Bock and Aitkin (1981) and discussed technical problems of implementation.

Consider a set of k items and n observations: the continuous unobservable response process x_{ij}, with $i = 1, ..., n$ and $j = 1, ..., k$, is expressed as a linear combination of Q common factors $(\theta_1, \theta_2, ..., \theta_Q)$ according to the Thurstone's multiple-factor model (Thurstone, 1947)[1], as follows

$$x_{ij} = a_{j1}\theta_{i1} + a_{j2}\theta_{i2} + ... + a_{jQ}\theta_{iQ} + \varepsilon_{ij}, \tag{3.1}$$

where:

[1] The original model notation is not preserved to provide a coherent and unified treatment of all the models in this work.

- $(a_{j1}, a_{j2}, ..., a_{jQ})$ are the Q factor loadings for the item j;
- $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iq}, ..., \theta_{iQ})$ is the vector of abilities for the individual i;
- ε_{ij} is assumed to be an unobservable random variable normally distributed as $\varepsilon_{ij} \sim N(0; \sigma_j^2)$.

The continuous response variable can be dichotomized in order to obtain a binary response variable y_{ij} by using a threshold d_j, as follows

$$\begin{cases} y_{ij} = 1 & \text{if } x_{ij} \geq d_j, \\ y_{ij} = 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

Therefore, a correct response is obtained when the latent x_{ij} equals or exceeds the threshold d_j. On the assumption for ε_{ij}, the probability of a correct response to item j from person i, with abilities $\boldsymbol{\theta}_i$, can be expressed as

$$P(y_{ij} = 1|\boldsymbol{\theta}_i) = P(x_{ij} > d_j|\boldsymbol{\theta}_i) = \Phi\left(\frac{\sum_{q=1}^{Q} a_{jq}\theta_{iq} - d_j}{\sigma_j}\right), \tag{3.3}$$

where Φ is the standard normal cumulative distribution function. The conditional response probability is a normal ogive model. To simplify the estimation procedure, it is convenient to express the (3.3) in terms of slope and intercept parameters through the following substitution

$$\frac{\sum_{q=1}^{Q} a_{jq}\theta_{iq} - d_j}{\sigma_j} = \sum_{q=1}^{Q} \alpha_{jq}\theta_{iq} + \delta_j,$$

where $\alpha_{jq} = a_{jq}/\sigma_j$ and $\delta_j = -d_j/\sigma_j$, for each item $j = 1, ..., k$ and dimension $q = 1, ..., Q$. The model is now expressed with the conventional IRT parameterization, in fact the slopes can be interpreted as the multidimensional discrimination parameters (one for each latent ability) and the intercept is proportional to the item difficulty. Therefore, the model becomes

$$P(y_{ij} = 1|\boldsymbol{\theta}_i) = \Phi\left(\sum_{q=1}^{Q} \alpha_{jq}\theta_{iq} + \delta_j\right). \tag{3.4}$$

The (3.4) is the typical formulation of the multidimensional two-parameter normal ogive (2PNO) model (Lord, 1952; Lord and Novick, 1968) in IRT. The model is referred as a *compensatory* model, because a low value on one ability can be compensated by a higher one on another dimension. The 2PNO model can be perfectly included in the latent variable models (Bartholomew and Knott, 1999) and in the more general framework of GLLVMM (Skrondal and Rabe-Hesketh, 2004).

The implementation of the MML estimation for item parameters following Bock and Aitkin (1981) assumes the existence of a multivariate population distribution for the vector of abilities. We assume that this distribution is a multivariate normal, that is $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{I})$. The expected value is equal to a $(Q \times 1)$ vector of zeros while the variance-covariance matrix is the identity matrix of order Q. Furthermore, in factor analysis the continuous response process x_j is distributed as a standard normal, that is $x_j \sim N(0,1)$, for $j = 1, .., k$ items. Therefore, the variance σ_j^2 of the error term becomes:

$$\sigma_j^2 = 1 - \sum_{q=1}^{Q} a_{jq}^2. \tag{3.5}$$

MML estimates for a_{jq} and d_j may be obtained, from MML estimates of α_{jq} and δ_j, by

$$\hat{a}_{jq} = \frac{\hat{\alpha}_{jq}}{\sqrt{1 + \sum_{q=1}^{Q} \hat{\alpha}_{jq}^2}} \tag{3.6}$$

and

$$\hat{d}_j = -\frac{\hat{\delta}_j}{\sqrt{1 + \sum_{q=1}^{Q} \hat{\alpha}_{jq}^2}}. \tag{3.7}$$

Each parameter $\hat{a}_{jq}$ is also called *standardized* $\hat{\alpha}_{jq}$ while the $\hat{d}_j$ is called *standardized difficulty.* The MML estimation procedure with the EM algorithm is analogous to the one implemented for the multiple-choice model (see Chapter 2), restricting the response to the binary case and introducing the possibility of more than one latent trait. Details can be found in Bock et al. (1988). The formal equivalence of marginal likelihood for the 2PNO model in IRT and the factor analysis of dichotomized variables was formally proved by Takane and de Leeuw

(1987).

3.2.1 Test for choosing the number of factors

The problem of sparse data has been raised in the discussion of goodness of fit for unidimensional models (Section 2.4.3). The likelihood ratio test of fit of the model relative to the general multinomial alternative is

$$G^2 = 2\sum_{l=1}^{2^k} r_l \ln \frac{r_l}{N\bar{P}_l}, \tag{3.8}$$

where 2^k is the number of possible patterns for binary data and k items, r_l is the frequency of pattern l, N is the number of observations ($N = \sum_l r_l$) and $\bar{P}_l$ is the expected probability computed from the ML estimates of the item parameters.

When the contingency table is not sparse, the G^2 follows a Chi-square distribution with $2^k - 1 - k(Q+1) + Q(Q+1)/2$ degrees of freedom, where Q is the number of latent abilities. Therefore, the statistical test may be used to decide the number of factors, performing repeated analysis, adding one factor at a time. When the G^2 falls into insignificance, no further factors are needed.

In case of sparse data, the (3.8) is not distributed as Chi-square. Nevertheless, Haberman (1977) has proved, for large samples, that the difference of the G^2's for nested models follows a Chi-square distribution, with degrees of freedom equal to the difference into the respective degrees of freedom. Therefore, the contribution of the last factor added to the model is significant if the corresponding change of the G^2 for the previous and the current model is statistically significant.

3.2.2 FIFA: an application to the Psychology faculty

This section presents the application of FIFA on the Psychology complete test. The general culture and the faculty-specific items are jointly considered and two different abilities are specified in the model (see Figure 3.1). The aim of the analysis is to understand the relationship between items and abilities, so that individuals could be evaluated on the basis of the results. The 2PNO model, with two latent variables, becomes

$$P(y_{ij} = 1|\boldsymbol{\theta}_i) = \Phi(\alpha_{j1}\theta_{i1}\alpha_{j2}\theta_{i2} + \delta_j). \tag{3.9}$$

According to this specification, two discrimination parameters (α_{j1}, α_{j2}) and an intercept term δ_j are estimated for each item j.

The model in (3.9) has been implemented in TESTFACT 4.0 (Bock et al., 2002), a software designed for factor analysis based on item response theory. TESTFACT is able to distinguish between two types of missing data: the omitted and the not presented items. The omitted items have been recoded as wrong responses while the not presented items are not included in the analysis. The MML estimation with the EM algorithm has been implemented with adaptive quadrature, so that the placement of the points is adapted to the region of the factor space occupied by the posterior distribution corresponding to each pattern. When the number of items is large (> 30), this method is suggested. In case of a bidimensional model, 5 quadrature points are specified. We specified 25 cycles for the EM algorithm.

Table 3.1 shows the slope estimates from the analysis on the complete test of Psychology containing the general culture items (1-30) and the Psychology specific items (31-50). The promax method (Hendrickson and White, 1964) for oblique rotation of factor pattern has been selected. In fact, the promax solution is particularly useful for identifying one-dimensional subsets of items from a multidimensional set.

The structure of the test is quite clear: the general culture items load on the first factor while the specific items load on the second factor, with few exceptions. The first latent variable may be interpreted as the general ability while the second one as the Psychology-specific ability. In particular, the general culture items (18,21,27,29) have positive and larger slope on the second dimension instead of the first one. Items (18,27,29) are technical-scientific while item 21 is a humanistic one. The Psychology specific items are based on Italian language, logic, natural science and English contents and the scientific area is predominant. Therefore, the results on items (18,27,29) are not surprisingly while the behavior of item 21 is more difficult to justify. The item slopes may be transformed into the factor loading by using the (3.6). The slopes are interpreted as discrimination parameters, i.e. they measure the capability of the item to differentiate between the examinees, respect to the latent ability. The estimated correlation between

Item	α_{j1}	α_{j2}	Item	α_{j1}	α_{j2}
1	**0.837**	-0.266	26	**0.640**	-0.004
2	**0.354**	0.107	27	0.215	**0.480**
3	**1.009**	-0.191	28	**1.185**	-0.165
4	**0.350**	-0.004	29	0.273	**0.765**
5	**0.700**	0.049	30	**0.613**	0.100
6	**0.983**	-0.251	31	0.025	0.120
7	**0.401**	0.169	32	0.146	**0.309**
8	**0.236**	0.070	33	0.005	0.045
9	**0.650**	0.024	34	0.170	**0.340**
10	**0.659**	0.353	35	-0.130	**0.634**
11	**0.811**	-0.123	36	-0.011	0.186
12	**0.653**	-0.008	37	-0.397	**1.386**
13	**0.430**	-0.071	38	-0.218	**0.650**
14	**0.706**	-0.066	39	-0.064	**0.521**
15	**0.500**	0.006	40	-0.074	**0.638**
16	**0.331**	0.095	41	-0.079	**0.555**
17	**0.355**	0.268	42	0.051	**0.345**
18	0.327	**0.360**	43	-0.101	**0.301**
19	**0.606**	0.052	44	0.057	**0.377**
20	**0.391**	0.181	45	0.100	**0.479**
21	0.128	**0.368**	46	0.102	**0.279**
22	**0.609**	-0.048	47	-0.100	**0.734**
23	**1.174**	-0.093	48	-0.019	**0.820**
24	**0.991**	-0.006	49	-0.105	**0.532**
25	**0.451**	0.083	50	-0.145	**1.095**

Table 3.1. Item slopes for the two-dimensional solution of FIFA (promax rotation) on the Psychology complete test

the two latent abilities, with the promax method, is 0.549.

Table (3.2) shows the estimates of the intercepts δ_j and the standardized difficulties d_j. As we can see, the two different parameters have opposite sign. In fact, the intercepts may be interpreted as facility parameters for the item. Considering $\boldsymbol{\theta} = \mathbf{0}$, the probability of a correct response becomes the standard normal cumulative distribution function in δ_j. Therefore, the intercept determines the probability of success for the median individual. The d_j is an indicator of the level of difficulty of the item: the higher the parameter, the more difficult the item.

When more than one dimension is involved in the model, it is not possible

Item	$\boldsymbol{\delta_j}$	$\boldsymbol{d_j}$	**Item**	$\boldsymbol{\delta_j}$	$\boldsymbol{d_j}$
1	-0.213	0.178	26	-0.456	0.385
2	0.844	-0.775	27	1.349	-1.124
3	1.964	-1.508	28	-0.549	0.388
4	1.842	-1.740	29	3.031	-2.066
5	1.089	-0.875	30	0.329	-0.270
6	-0.620	0.491	31	-0.786	0.779
7	0.710	-0.627	32	-0.021	0.019
8	1.308	-1.259	33	-0.359	0.359
9	0.034	-0.028	34	0.546	-0.494
10	1.930	-1.334	35	-0.445	0.390
11	-0.111	0.090	36	-0.275	0.271
12	0.942	-0.791	37	1.809	-1.306
13	0.557	-0.520	38	0.438	-0.388
14	1.530	-1.280	39	0.811	-0.732
1	0.106	-0.095	40	1.138	-0.982
16	0.517	-0.480	41	-0.389	0.347
17	0.245	-0.212	42	-0.678	0.634
18	-0.309	0.260	43	-0.534	0.518
19	0.533	-0.447	44	2.270	-2.096
20	-0.167	0.147	45	1.425	-1.246
21	1.737	-1.576	46	0.073	-0.069
22	0.671	-0.582	47	0.708	-0.592
23	0.178	-0.122	48	1.037	-0.809
24	0.218	-0.155	49	-0.474	0.429
25	1.455	-1.295	50	1.766	-1.284

Table 3.2. Item intercepts and standardized difficulties for the two-dimensional solution of FIFA on the Psychology complete test

to graphically represent the probability of success as a function of ability in the Cartesian plan through the ICC. In case of a two-dimensional solution, a probability surface should be considered: the item response surface (IRS).

Figure 3.2 shows the IRS for item 35 of the Psychology complete test (that is, item 5 for the Psychology-specific test, considered in Chapter 2). The surface represents the probability of a correct response for the item given the two traits. The estimated item parameters from the unrotated solution are considered so that orthogonal axes for the abilities are assumed. The surface reflects the increasing trend of the probability on the second dimension θ_2 and the slightly increasing trend on the first dimension θ_1.

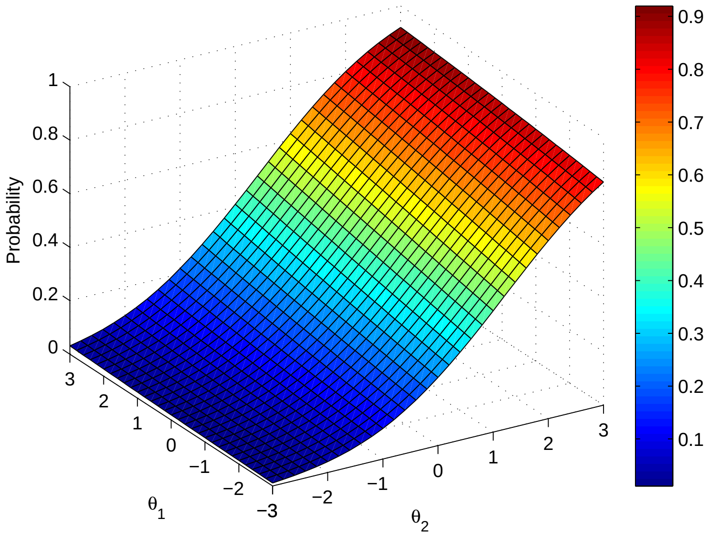


Figure 3.2. IRS for item 35, Psychology complete test. $\alpha_{35,1} = 0.072$, $\alpha_{35,1} = 0.544$, $\delta_{35} = -0.445$

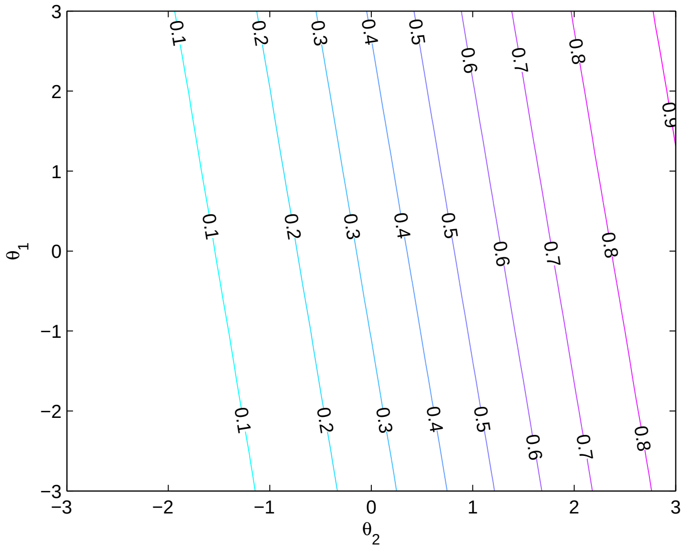


Figure 3.3. Contour plot for item 35, Psychology complete test

A more informative graphical representation is the contour plot, describing the probability iso-lines for combinations of the two abilities. The contour plot for item 35 is presented in Figure 3.3. The equiprobability contours are always parallel for this type of model. Examinees whose combination of θ_1 and θ_2 places them in the same contour, all have the same probability to answer positively the item.

FIFA was performed also for the unidimensional and three-dimensional models. Table 3.3 shows the comparison in terms of percentage of explained variance.

	1DIM	2DIM	3DIM
% Variance	18.4	25.8	28.7

Table 3.3. Percentage of explained variance for one-, two- and three-dimensional solutions

The percentage of explained variance is quite low for all the solutions. Nevertheless, this is not unusual in IRT models and the increment of the number of factors does not lead to a significant improvement. From the one- to the two-dimensional solution the explained variance goes from 18.4 to 25.8 % while for the three-dimensional solution it is 28.7 %. Therefore, there is a meaningful increment from the first to the second solution while there is not a big improvement with the third one.

The G^2 statistics (see (3.8)) are shown in Table 3.4.

	1DIM	2DIM	3DIM
G^2	22646.94	22233.7	22094.97
G^2 difference		413.24 (df=49)	138.73 (df=48)
p-value		< 0.000	< 0.000

Table 3.4. G^2 and G^2 difference test for one, two and three-dimensional solutions

The difference between the likelihood ratio tests of fit G^2 may be considered as a test for the number of factors. The difference between the statistics for the one- and the two-dimensional solutions follows a Chi-square distribution with 49 degrees of freedom. The correspondent p-value is < 0.000: the test is significant and the contribution of the second factor improves the model fit. The same conclusion can be inferred from the comparison between the two- and the three-dimensional models. The difference in the statistics is distributed as a Chi-square with 48 degrees of freedom. The test is significant (p-value < 0.000) and the contribution of the third factor is considered relevant. Nevertheless, the G^2 statistics should be considered carefully. Sometimes, they may be inflated from cluster effects of respondents. Because cluster effects have not been studied yet and respondents come from different sites (respect to gender, school and district), we prefer to be conservative about the value of the likelihood ratio test and to divide it by a design factor of 2. In this case, the difference between the first two solutions becomes 206.62 and the test is still significant (p-value< 0.000). On the other hand, the G^2 difference between the two and three factor models becomes 69.365 (p-value$= 0.0234$). The test is not significant at 1% level. Furthermore, the interpretation of the test structure becomes difficult with three abilities.

The results have shown a significant improvement in the fit from a unidimensional to a bidimensional model while three dimensions do not lead to a meaningful improvement. Furthermore, the choice of dimensionality depends on the purpose of assessment and we believe that the two-dimensional solution is able to explain the data and also to give a simple and interpretable instrument for students' evaluation.

The implementation of FIFA in TESTFACT has disclosed several limits. First of all, it is not possible to implement a confirmatory approach and to identify the model imposing constraints on the item parameters. Furthermore, no standard errors of estimates are given and the goodness of fit tools are limited. Finally, the number of latent variables which may be included in the analysis is restricted to 5 for not adaptive quadrature and 10 for adaptive quadrature. On the other hand, some advantages should be highlighted. Firstly, the software is able to deal with different types of missing data: omitted and not presented items. Secondly, the MML estimation with the EM algorithm is computationally fast. Results may be obtained for a bidimensional model in less than one minute on a 1.73 GHz, 797 MHz Intel(R) Pentium(R) M processor. Finally, TESTFACT represents a flexible and easy tool for exploratory analysis.

Chapter 4

An alternative approach: Gibbs sampler for the IRT model estimation

4.1 Introduction

The Bayesian approach using Markov chain Monte Carlo (MCMC) has recently become very popular in the estimation of item response models. It can be seen as an alternative and at the same time as a compensatory solution to the classical EM algorithm implemented in the marginal maximum likelihood (MML) estimation. Alternative, since it works with simulation and introduces an informative prior distribution in the estimation process. Unlike the MML method, the Bayesian approach regards both the latent variables and the item parameters as random. Compensatory, since the posterior distribution generated with MCMC can be used to assess the suitability of the normal approximations in the MML and the two methods can be compared in terms of accuracy of parameter recovery.

As we will see in this section, MCMC is very useful to make inference when the model is very complex and it is difficult to sample or directly simulate from the posterior distribution, for example with multidimensional and multilevel models. Particularly, the Gibbs sampler seems to be a very accurate strategy in creating suitable samples from the posterior density. This method is also not very constraining and relatively easy to implement, compared with other methods. For these reasons, several researchers have recently decided to implement the MCMC strategies in the item response theory framework, studying the properties of the methods, the parameter recovery respect to the classical methods and the application to more and more complex models. The main advantages of the MCMC approach respect to the classical MML estimation are the flexibility in terms of modelling all the dependencies between the latent and the observed variables and the suitability for more complex models. Furthermore, the Bayesian algorithm is not sensitive to the choice of starting values unlike the EM algorithm. Finally, more advantages have been noticed respect to the software TESTFACT used for FIFA: the model identification is possible and the number of latent variables to be included in the analysis is not limited.

Our aim is to perform a Bayesian estimation through the Gibbs sampler for the unidimensional and multidimensional two-parameter normal ogive (2PNO) model (Lord, 1952; Lord and Novick, 1968) including the incomplete design. The two main references for this purpose are the works of Albert (1992) and Béguin and Glas (2001). The first one implements the Gibbs sampler to the unidimensional 2PNO model and consists of a comparison between the MCMC algorithm and the EM method through a real educational application. The second work extends the work of Albert to the presence of the guessing parameter and to the multidimensional case. It also discusses incomplete design and multiple groups applications with several examples. Other item response applications of MCMC can be found in Fox and Glas (2001); Patz and Junker (1999a,b).

This section includes some review of both the Bayesian estimation and the MCMC principal algorithms and the implementation of the Gibbs sampler to the unidimensional and multidimensional 2PNO models with incomplete design. Applications of the algorithm will be discussed as well.

4.2 Markov chain Monte Carlo methods

Adopting a Bayesian perspective, the main aim of the researcher is to investigate the properties of the posterior distribution, say $P(\rho|y)$, where y represents the observed data and ρ the parameters of interest. Therefore, two different distributions are taken into account: the prior, which reflects the opinion or the knowledge of the researcher about ρ, and the posterior, which expresses the distribution after the data have been observed. According to the Bayes' theorem, the posterior distribution is defined as

$$P(\rho|y) = \frac{f(y|\rho)P(\rho)}{f(y)}, \tag{4.1}$$

where $f(y) = \int f(y|\rho)P(\rho)d\rho$, in case of continuous quantities. Following (4.1), the posterior distribution of ρ is proportional to the product of the likelihood and the prior distribution

$$P(\rho|y) \propto f(y|\rho)P(\rho). \tag{4.2}$$

When the posterior density does not have a familiar functional form and/or direct simulation is not applicable because of the complexity of the model, the Markov chain simulation seems to be the easiest way to obtain samples from the posterior distribution $P(\rho|y)$.

The Markov chain Monte Carlo (MCMC) is a general class of methods based on the idea of reproducing the target distribution, in our context the posterior $P(\rho|y)$, through the simulation of one or more sequences of correlated random variables. The algorithm simulates a random walk in the ρ space, where each value $\rho^{(s)}$ (for $s = 1, .., S$ iterations) is drawn from a probability distribution dependent on the previous value $\rho^{(s-1)}$. The basic idea is that the random walk visits the regions of the state space in proportion to their posterior probabilities and, when the number of iterations is sufficiently large, it can approximate the target distribution. The main difference with Monte Carlo methods is that the sampled values are correlated, instead of being statistically independent. The proof of convergence (Gelman et al., 1995, pg. 325-326) consists of the demonstration that the generated Markov chain has a stationary unique distribution and that this distribution coincides with the target distribution. The method is rather complicated and should be used carefully; nevertheless it seems to give reliable results in reproducing the marginal posterior distributions of item response parameters that cannot be derived analytically.

The main difficulty with MCMC is to create a chain long enough to best approximate the target distribution and whose convergence is not slow. Not all the sampled values should be included: a burn-in period (a sequence containing the first iterations) should be fixed and removed from the analysis. Its length can vary according to the complexity of the posterior distribution, the convergence speed and the starting values; a general indication given by Gelman et al. (1995) suggests half the sampled values for each generated sequence while several researchers support the theory of not excluding iterations. In practice, what we suggest is to check the plots of the sampled parameters in the progression of iterations and decide afterwards; for example, in one of the studies presented in Béguin and Glas (2001), the burn-in phase is of 1000 values against a run length of 30000.

Furthermore, the discussion about the number of chains needed for the algorithm is open, as stated in Gilks et al. (1996). Suggestions are to create either one long chain or several quite long chains or many short chains. The first case is supported by the idea that the longer the chain is, the higher the possibility to find new modes is. On the other hand, multiple chains allow the comparison between the results which may reveal some relevant differences, symptom of not reached stationarity. The use of short chains, motivated by the creation of independent samples, is not recommended because convergence may take a long time to be reached and independent samples are not required.

Depending on the characteristics of the problem and on the desired properties of the Markov chain, several MCMC algorithms have been developed which specify a transition distribution $P(\rho^{(s)}|\rho^{(s-1)})$ and a rule for moving from each value of the sequence to the following, beginning from proper starting values $\rho^{(0)}$.

The application on real data consists of the implementation of the Gibbs sampler, which is a particular case of the Metropolis-Hastings algorithm. Therefore, we provide a brief description of both the methods. More details about MCMC can be found in Gelman et al. (1995), Gamerman (1997) and Gilks et al. (1996).

4.2.1 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm (Hastings (1970)) simply consists of three main steps:

1. Generation of a candidate point ρ^* through a *proposal density*,
2. Computation of the ratio r of importance ratios,
3. Acceptation or rejection of the ρ^*.

Assuming that ρ is a multi-component vector of continuous-valued parameters, that we have provided suitable starting values and that the current state of the chain is $\rho^{(s-1)}$, the first step of the algorithm consists of sampling a candidate point ρ^* by using a *proposal density* $\alpha(\rho^*|\rho^{(s-1)})$. The density α is also called *jumping distribution* to better express the idea of moving the chain from the current value to the following. Analogously, it is possible to define the probability of "jumping" in the opposite direction from the candidate point to the current value, that is $\alpha(\rho^{(s-1)}|\rho^*)$. The *proposal density* does not need to be symmetric, but the more restricted Metropolis algorithm requires this property too. Next, we should define $\alpha(\cdot)$ in order to satisfy a set of specific characteristics of the chain, i.e. irreducibility, aperiodicity and not transitoriness. A chain is irreducible if it is possible to move from one state to any other state in a finite number of steps with positive probability, aperiodic if all the states are acyclic, and not transient if all the states are recurrent (the probability to return to a state from the same state is equal to one). Furthermore, it is required that the ratio $\frac{\alpha(\rho^*|\rho^{(s-1)})}{\alpha(\rho^{(s-1)}|\rho^*)}$ is strictly positive, for all the ρ-values for which both the numerator and the denominator are nonzero. The second step is to compute the ratio r as

$$r = \frac{P(\rho^*|y)\alpha(\rho^{(s-1)}|\rho^*)}{P(\rho^{(s-1)}|y)\alpha(\rho^*|\rho^{(s-1)})},$$

and the *acceptance probability*, defined as $prob = min(1, r)$. The higher the $prob$ is, the more probable the acceptation of the candidate value ρ^* will be. The indicator r is made up of the ratio of the posterior probabilities, which makes the algorithm moving to the ρ-value with higher posterior density, and the ratio of the *proposal densities*, which can also determine the direction towards one or another parameter value. The third step of the M-H algorithm is to decide wether to accept or refuse the candidate value ρ^*. Therefore, we should draw a random number u from the uniform distribution in the [0,1] interval and then set $\rho^{(s)}$ equal to ρ^* if $u < prob$ and equal to $\rho^{(s-1)}$ otherwise. Hence, the candidate value ρ^* is accepted with probability $prob = min(1, r)$ and rejected in favor of $\rho^{(s-1)}$ in case of $u \geq prob$. However, the rejection of the jump implies the progression of the iterations. The M-H algorithm can be applied in case of discrete-values parameters too and the α density becomes the probability mass function used to generate candidate points.

4.2.2 Gibbs sampler

The Gibbs sampler, named by Geman and Geman (1984) and formalized by Gelfand and Smith (1990) is a special case of the Metropolis-Hastings algorithm, when the ratio r is equal to one (proof in Gelman et al. 1995, pg. 328). This means that the *acceptance probability* of the new value is one and the jump is made at each iteration. The Gibbs sampler is based on iterative sampling of the conditional distributions resulting from the decomposition of the full posterior density. The algorithm is particularly useful in order to estimate the parameters of multidimensional or hierarchical models and it overcomes the major disadvantage of the MH method: the choice of the proposal density. An unreasonable choice of the $\alpha(\cdot)$ may create a very slow algorithm, which might not converge after thousands of iterations. In fact, if the proposal density is too constraining, the method will hardly visit distant states of the space in few iterations and furthermore, there will probably be a high correlation between the iterates; on the other hand, if the *proposal densities* is too broad it will take a long time to reach the equilibrium distribution.

The main idea of the Gibbs sampler is very simple and does not depend on the choice of the *jump distribution*. If it is difficult to sample from the high-parameterized posterior distribution and it is possible to subdivide the parameter vector, we should sequentially generate the parameter values from the single conditional distributions. In the case we have a multidimensional or a hierarchical model, we may deal with a multi-component vector of parameters, say $\boldsymbol{\rho} = (\rho_1, \rho_2, ..., \rho_P)$ with P elements. The full posterior distribution of interest is

$P(\boldsymbol{\rho}|y) = P(\rho_1, \rho_2, ..., \rho_P|y)$, that we can decompose into the conditional distributions $P_1(\rho_1|\rho_2, ..., \rho_P, y)$, $P_2(\rho_2|\rho_1, \rho_3, ..., \rho_P, y)$, $\ldots$, $P_P(\rho_P|\rho_1, \rho_2, ..., \rho_{P-1}, y)$, where each component of the ρ vector is conditional to all the other elements and to the data. Suppose that the algorithm is in the iteration $(s-1)$ and that we have simulated the values $\rho_1^{(s-1)}$, $\rho_2^{(s-1)}$, $\ldots$, $\rho_P^{(s-1)}$; the Gibbs sampler works as follows:

- Sample $\rho_1^{(s)}$ from $P_1(\rho_1|\rho_2^{(s-1)}, ..., \rho_P^{(s-1)}, y)$;
- Sample $\rho_2^{(s)}$ from $P_2(\rho_2|\rho_1^{(s)}, \rho_3^{(s-1)}, ..., \rho_P^{(s-1)}, y)$;
- ...
- Sample $\rho_P^{(s)}$ from $P_P(\rho_P|\rho_1^{(s)}, \rho_2^{(s)}, ..., \rho_{P-1}^{(s)}, y)$.

At each iteration the values of the parameter vector $\boldsymbol{\rho}$ are sequentially updated, each value from its own conditional distribution depending on the other most recent sampled values. The sequence of the sampler and the starting values for the algorithm are decided by the researcher. Under suitable regularity conditions, the distribution of $\boldsymbol{\rho}^{(s)}$ will converge to the posterior distribution of interest. Since in many examples the convergence is very fast, we can consider the complete sequence $\{\boldsymbol{\rho}^{(s)}\}$ as a simulated sample from the target distribution.

4.3 Gibbs sampler for the unidimensional two-parameter normal ogive model

4.3.1 The model

The unidimensional 2PNO model (Lord, 1952) specifies the probability of a correct response for the individual i on item j, with $i = 1, ..., n$ examinees and $j = 1, ..., k$ items, as a function of person and item parameters, as follows

$$P(y_{ij} = 1|\theta_i, \alpha_j, \delta_j) = \Phi(\alpha_j\theta_i - \delta_j) = \int\limits_{-\infty}^{\alpha_j\theta_i-\delta_j} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \tag{4.3}$$

with:

- y_{ij} = binary response of the examinee i to the item j, taking the value 1 for a correct response and 0 for an incorrect one,
- θ_i = latent ability of person i,
- α_j = discrimination power of item j,
- δ_j = difficulty of item j,
- Φ = standard normal cumulative distribution function.

The (4.3) has been introduced in Chapter 2 (Section 2.1.1) by using the linear function $\alpha_j(\theta_i - \beta_j)$ instead of $\alpha_j\theta_i - \delta_j$, where $\delta_j = \alpha_j\beta_j$. The two specifications are equivalent: here we prefer the second one to facilitate the implementation of the Gibbs sampler. Furthermore, the (4.3) has been used for FIFA (Chapter 3) in the multidimensional context. The intercept term had opposite sign, in fact the δ_j parameter in FIFA is interpreted as an easiness parameter while in this formulation it means difficulty.

The probability of observing an individual response pattern (the complete sequence of responses) can be expressed, by using the assumption of local independence, as

$$\begin{aligned} P(y_{i1}, ..., y_{ik}|\theta_i, \boldsymbol{\xi}) &= \prod_{j=1}^{k} P(y_{ij}|\theta_i, \boldsymbol{\xi}) \\ &= \prod_{j=1}^{k} [P(y_{ij} = 1|\theta_i, \boldsymbol{\xi})]^{y_{ij}} [1 - P(y_{ij} = 1|\theta_i, \boldsymbol{\xi})]^{1-y_{ij}} \\ &= \prod_{j=1}^{k} [\Phi(y_{ij} = 1|\theta_i, \boldsymbol{\xi})]^{y_{ij}} [1 - \Phi(y_{ij} = 1|\theta_i, \boldsymbol{\xi})]^{1-y_{ij}}, \end{aligned} \tag{4.4}$$

where $\boldsymbol{\xi}$ is the vector of item parameters for all the k items.

Finally, the likelihood function for all the data is obtained multiplying over all the examinees, thanks to the assumption of experimental independence, such as

$$P(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\xi}) = \prod_{i=1}^{n}\prod_{j=1}^{k}[\Phi(y_{ij}=1|\theta_i,\boldsymbol{\xi})]^{y_{ij}}[1-\Phi(y_{ij}=1|\theta_i,\boldsymbol{\xi})]^{1-y_{ij}}. \tag{4.5}$$

Usually we assume that $\theta_1, ..., \theta_n$ is a random sample from a normal distribution. Clearly, from (4.3) we can see that the model is not univocally determined: if we multiply θ_i by a constant and divide α_j by the same constant or if we add to θ_i and to δ_j/α_j the same quantity, the model does not change. Constraints on item or person parameters allow to solve these two indeterminacies. Usually, in the case of unidimensionality, we are used to fix the mean value and the standard deviation of the ability distribution to 0 and 1, respectively. As we will see later, in multidimensional models the interpretation of the ability is more complicated and it is convenient to impose constraints on the item parameters, rather than on the person parameters, for identification purpose.

4.3.2 The underlying variable approach

The presence of a dichotomous variable y_{ij}, indicating correct or incorrect response of person i to item j, can be modeled introducing an additional underlying variable, say Z_{ij}. The probability of success for the dichotomous variable is generally expressed as a function F of a linear predictor $\eta_{ij} = \mathbf{x}_i^{'}\boldsymbol{\beta_j}$, as follows

$$P(y_{ij}=1) = F(\eta_{ij}) = F(\mathbf{x}_i^{'}\boldsymbol{\beta_j}). \tag{4.6}$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{ij}, ..., x_{iQ})'$ is the Q-dimensional vector of the covariates referring to the observation i, with $i = 1, ..., n$ and $q = 1, ..., Q$, and $\boldsymbol{\beta_j} = (\beta_{j1}, \beta_{j2}, ..., \beta_{jQ})'$ is the Q-dimensional vector of regression coefficients.

The linear regression of a binary variable on the covariates is theoretically not possible: the problem is solved by using a probit or logit link function, so that the interval [0,1] is converted into the real line. The binary regression is very common in many contexts as medical, biological, economical, social and of course the educational one. Then, the dichotomous variable (yes or no, success or failure, agreement or disagreement) is able to model, for example, the patient recovery, the efficacy of a treatment or simply the agreement to a sentence. In the educational field, the main interest is on student knowledge and a competence test made of several items is submitted. Therefore, the y_{ij} indicates, for each item i and observation j, correct answer ($y_{ij} = 1$) or wrong answer ($y_{ij} = 0$). The covariates are represented by the different Q abilities needed to solve the items while the regression coefficients are the discrimination parameters. Thus, for each observation, we can compute the regression of a underlying variable Z_{ij} on $\mathbf{x}_i$ adding a random error term ϵ_{ij}, independent and identically distributed according to the function F

$$Z_{ij} = \mathbf{x}_i^{'}\boldsymbol{\beta_j} + \epsilon_{ij}. \tag{4.7}$$

where we assume the following relation between the observed and the underlying variables

$$y_{ij} = \begin{cases} 1 & \text{if } Z_{ij} > 0, \\ 0 & \text{if } Z_{ij} \leq 0. \end{cases} \tag{4.8}$$

Accordingly, $Z_{ij} > 0$ if and only if the corresponding observed response y_{ij} is a success, that is $P(y_{ij} = 1) = P(Z_{ij} > 0)$. This method, called *underlying variable* approach (see Bartholomew and Knott (1999)), introduces a partition of the continuous underlying variable to represent the dichotomy and has been developed also for ordinal data, where the *underlying variable* is segmented according to threshold levels. Furthermore, using the standard cumulative distribution function Φ so that $P(y_{ij} = 1) = P(Z_{ij} > 0) = \Phi(\eta_{ij}) = \Phi(\mathbf{x}_i^{'}\boldsymbol{\beta_j})$ and a noninformative uniform prior for $\boldsymbol{\beta_j}$, the resulting posterior conditional distribution of $\boldsymbol{\beta_j}$ given $\boldsymbol{Z_j} = (Z_{1j}, Z_{2j}..., Z_{nj})'$ is normal:

$$\boldsymbol{\beta_j}|\boldsymbol{Z_j} \sim N(\hat{\boldsymbol{\beta_j}}, (\mathbf{X}'\mathbf{X})^{-1}), \tag{4.9}$$

where $\hat{\boldsymbol{\beta_j}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Z_j}$ is the least squares estimate of $\boldsymbol{\beta_j}$ and $\mathbf{X}$ is the $n \times Q$ explanatory variable matrix.

The proof of this result is very simple. The Z_{ij} are normally distributed, in fact $P(Z_{ij} \leq 0) = \Phi(-\mathbf{x}_i^{'}\boldsymbol{\beta_j})$ implies that $Z_{ij}|\boldsymbol{\beta_j} \sim N(\mathbf{x}_i^{'}\boldsymbol{\beta_j}; 1)$, in vector notation $\boldsymbol{Z_j} \sim N(\mathbf{X}\boldsymbol{\beta_j}; I)$. Since we assume a uniform prior for $\boldsymbol{\beta_j}$, the conditional distribution $P(\boldsymbol{\beta_j}|\boldsymbol{Z_j})$ is proportional to $P(\boldsymbol{Z_j}|\boldsymbol{\beta_j})$, that is, $\exp\big(-\frac{1}{2}(\boldsymbol{Z_j} - \mathbf{X}\boldsymbol{\beta_j})'(\boldsymbol{Z_j} - \mathbf{X}\boldsymbol{\beta_j})\big)$. Now we prove that the mean is $\hat{\boldsymbol{\beta_j}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Z_j}$ and the variance is equal to $(\mathbf{X}'\mathbf{X})^{-1}$. To this aim we show that the quantities $(\boldsymbol{Z_j}-\mathbf{X}\boldsymbol{\beta_j})'(\boldsymbol{Z_j}-\mathbf{X}\boldsymbol{\beta_j})$ and $(\boldsymbol{\beta_j}-\hat{\boldsymbol{\beta}}_{\boldsymbol{j}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta_j}-\hat{\boldsymbol{\beta}}_{\boldsymbol{j}})$ are equal, up to additive constants not depending on $\boldsymbol{\beta_j}$ that are multiplicative terms in the posterior density function. The expansion of the two quadratics leads, in the first case, to

$$\begin{aligned}&(\boldsymbol{Z_j}-\mathbf{X}\boldsymbol{\beta_j})'(\boldsymbol{Z_j}-\mathbf{X}\boldsymbol{\beta_j}) =\\ &= \boldsymbol{Z_j}'\boldsymbol{Z_j} - \boldsymbol{Z_j}'\mathbf{X}\boldsymbol{\beta_j} - \boldsymbol{\beta_j}'\mathbf{X}'\boldsymbol{Z_j} + \boldsymbol{\beta_j}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta_j} =\\ &= constant - \boldsymbol{Z_j}'\mathbf{X}\boldsymbol{\beta_j} - \boldsymbol{\beta_j}'\mathbf{X}'\boldsymbol{Z_j} + \boldsymbol{\beta_j}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta_j},\end{aligned}$$

while in the second case to

$$\begin{aligned}&(\boldsymbol{\beta_j}-\hat{\boldsymbol{\beta}}_{\boldsymbol{j}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta_j}-\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}) =\\ &= \boldsymbol{\beta_j}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta_j} - \boldsymbol{\beta_j}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Z_j} - \boldsymbol{Z_j}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta_j} + constant =\\ &= \boldsymbol{\beta_j}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta_j} - \boldsymbol{\beta_j}'\mathbf{X}'\boldsymbol{Z_j} - \boldsymbol{Z_j}'\mathbf{X}\boldsymbol{\beta_j} + constant.\end{aligned}$$

The two quantities are equal (proof adapted from the online solution of the exercises in Gelman et al. (1995)).

Another important result (see for example Johnson and Albert (1999), Chap. 3) is that the distribution of each Z_{ij} given y_{ij} and $\boldsymbol{\beta_j}$ is normal, truncated by 0 to the left if $y_{ij} = 1$ and to the right if $y_{ij} = 0$

$$P(Z_{ij}|y_{ij},\boldsymbol{\beta_j}) \propto \begin{cases} \phi(z;\eta_{ij},1)I(z>0) & \text{if } y_{ij}=1,\\ \phi(z;\eta_{ij},1)I(z\leq 0) & \text{if } y_{ij}=0.\end{cases} \tag{4.10}$$

where $\phi(z;\eta_{ij},1)$ is a normal density with expected value $\eta_{ij} = \mathbf{x}_i^{'}\boldsymbol{\beta_j}$ and variance equal to 1 and $I(\cdot)$ is the indicator function, taking value 1 if the argument is true or 0 otherwise.

The underlying variable approach can be applied also to the normal ogive model: the $\{Z_{ij}\}$ are independent and identically normally distributed random variables with expected value $\eta_{ij} = \alpha_j\theta_i - \delta_j$ and variance equal to 1, for $i = 1, ..., n$ and $j = 1, ..., k$. The relation between the binary response variables y_{ij} and the continuous unobserved variables Z_{ij} becomes

$$y_{ij} = \begin{cases} 1 & \text{if } Z_{ij} > 0, \\ 0 & \text{if } Z_{ij} \leq 0. \end{cases} \tag{4.11}$$

Consequently the conditional distribution of each Z_{ij} given y_{ij} and η_{ij} can be expressed as

$$P(Z_{ij}|y_{ij}, \eta_{ij}) \propto \phi(Z_{ij}; \eta_{ij}, 1)[I(Z_{ij} > 0)I(y_{ij} = 1)+I(Z_{ij} \leq 0)I(y_{ij} = 0)], \tag{4.12}$$

where $\phi(Z_{ij}; \eta_i, 1)$ is the normal density with expected value η_i and variance equal to 1 and $I(\cdot)$ is the indicator function, taking value 1 when the argument is true or 0 otherwise.

4.3.3 Gibbs sampler implementation

To perform a Bayesian estimation of item and person parameters of the 2PNO model we should be able to simulate from the joint posterior distribution of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi})$ introducing the vector $\mathbf{Z} = (Z_{11}, ..., Z_{nk})$ and using the following assumptions:

- $\{Z_{ij}\}$ i.i.d.$\sim N(\eta_{ij}, 1)$, with $\eta_{ij} = \alpha_j\theta_i - \delta_j$,
- $\{y_{ij}\}$ indicators of values of $\{Z_{ij}\}$,
- Standard normal prior distribution on $\{\theta_i\}$: $\{\theta_i\}$ i.i.d.$\sim N(0, 1)$,
- Prior distribution on item parameters $\boldsymbol{\xi}$: $P(\boldsymbol{\xi}) = \prod_{j=1}^{k} I(\alpha_j > 0)$.

The last assumption insures that the discrimination parameters are positive to preserve the increasing monotonic trend of the item characteristic curve. Thus, the joint posterior distribution is given by

$$\begin{aligned} P(\mathbf{Z},\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y}) =& P(\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\xi},\mathbf{y})P(\boldsymbol{\theta})P(\boldsymbol{\xi}) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{k}\{\phi(Z_{ij};\eta_{ij},1)[I(Z_{ij}>0)I(y_{ij}=1)+I(Z_{ij}\leq 0)I(y_{ij}=0)]\} \\ &\prod_{i=1}^{n}\phi(\theta_i;0,1)\prod_{j=1}^{k}I(\alpha_j>0). \end{aligned} \tag{4.13}$$

Because of the intractable form of (4.13) we can resort to the Gibbs sampler using the conditional distributions of $\mathbf{Z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, respectively $P(\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\xi},\mathbf{y})$, $P(\boldsymbol{\theta}|\mathbf{Z},\boldsymbol{\xi},\mathbf{y})$ and $P(\boldsymbol{\xi}|\mathbf{Z},\boldsymbol{\theta},\mathbf{y})$, which are tractable and easy to draw samples from. The conditional distribution of the independent Z_{ij} is normal, with expected value $\eta_{ij}=\alpha_j\theta_i-\delta_j$ and variance equal to 1, truncated by 0 to the left if $y_{ij}=1$ and to the right if $y_{ij}=0$ as similarly stated in (4.10)

$$Z_{ij}|\boldsymbol{\theta},\boldsymbol{\xi},\mathbf{y}\sim\begin{cases} N(\eta_{ij},1) & \text{with } Z_{ij}>0 \text{ if } y_{ij}=1, \\ N(\eta_{ij},1) & \text{with } Z_{ij}\leq 0 \text{ if } y_{ij}=0. \end{cases} \tag{4.14}$$

The conditional distribution of $\boldsymbol{\theta}$ is also normal: the person parameters $\theta_1,...,\theta_n$ are independent with the following conditional posterior distribution

$$P(\theta_i|\mathbf{Z},\boldsymbol{\xi},\mathbf{y})\propto\prod_{j=1}^{k}\phi(Z_{ij};\eta_{ij},1)\phi(\theta_i;0,1). \tag{4.15}$$

Similarly to (4.7) the normal regression model for observation i is given by

$$\begin{aligned} Z_{ij}&=\alpha_j\theta_i-\delta_j+\epsilon_{ij} \\ Z_{ij}+\delta_j&=\alpha_j\theta_i+\epsilon_{ij}, \end{aligned} \tag{4.16}$$

where ϵ_{ij} i.i.d.$\sim N(0,1)$. The second formulation can be interpreted as the multiple regression of $(Z_{ij}+\delta_j)$ on the regressors α_j, with $j=1,...,k$, considering the θ_i as regression coefficients and following the vector notation

$$\begin{pmatrix} Z_{i1}+\delta_1 \\ Z_{i2}+\delta_2 \\ \vdots \\ Z_{ik}+\delta_k \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} \theta_i + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{ik} \end{pmatrix}. \tag{4.17}$$

With a noninformative prior distribution on θ_i, the posterior distribution would be analogous to (4.9); because our assumptions provide a standard normal prior distribution for θ_i, we need to combine the likelihood and prior distribution information together in a normal model. The likelihood function of θ_i follows the normal distribution given by (4.9) with mean equal to the least square estimate of θ_i, specifically $\hat{\theta}_i = (\alpha'_j\alpha_j)^{-1}\alpha'_j(Z_{ij}+\delta_j)$ and variance $v = (\alpha'_j\alpha_j)^{-1}$. Practically $\hat{\theta}_i$ and v can be calculated by

$$\begin{aligned} \hat{\theta}_i =& [\alpha_1^2 + ... + \alpha_k^2]^{-1} \sum_{j=1}^{k} \alpha_j(Z_{ij}+\delta_j) = \frac{\sum_{j=1}^{k} \alpha_j(Z_{ij}+\delta_j)}{\sum_{j=1}^{k} \alpha_j^2} \\ v =& \frac{1}{\sum_{j=1}^{k} \alpha_j^2}. \end{aligned}$$

Generally, a prior distribution as $\theta_i \sim N(\mu, \sigma^2)$ combined with a normal likelihood with expected value $\hat{\theta}_i$ and variance v yields to the following posterior distribution

$$\theta_i|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y} \sim N\left(\frac{\hat{\theta}_i/v + \mu/\sigma^2}{1/v + 1/\sigma^2}; \frac{1}{1/v + 1/\sigma^2}\right). \tag{4.18}$$

Therefore, the combination of our standard normal prior distribution and the likelihood results, leads to the following normal posterior distribution for θ_i

$$\theta_i|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y} \sim N\left(\frac{\hat{\theta}_i/v}{1/v + 1}; \frac{1}{1/v + 1}\right), \tag{4.19}$$

with expected value and variance equal to $\sum_{j=1}^{k} \alpha_j(Z_{ij}+\delta_j)/(\sum_{j=1}^{k} \alpha_j^2 + 1)$ and $1/(\sum_{j=1}^{k} \alpha_j^2 + 1)$, respectively.

The third conditional distribution $P(\boldsymbol{\xi}|\mathbf{Z},\boldsymbol{\theta},\mathbf{y})$ can be computed by using the same approach applied to the fully conditional distribution of θ_i. Consider the k item parameters $\boldsymbol{\xi}_1,...,\boldsymbol{\xi}_k$, with $\boldsymbol{\xi}_j' = [\alpha_j;\delta_j]$, independent with the following posterior distribution

$$P(\boldsymbol{\xi}_j|\boldsymbol{\theta},\mathbf{Z},\mathbf{y}) \propto \prod_{i=1}^{n} \phi(Z_{ij};\eta_{ij},1)I(\alpha_j > 0). \tag{4.20}$$

The normal regression model for each item j, with $j = 1,...,k$, is

$$\mathbf{Z}_j = [\boldsymbol{\theta} \ -\mathbf{1}]\boldsymbol{\xi}_j + \boldsymbol{\epsilon}_j, \tag{4.21}$$

where $\boldsymbol{\theta}$ is the n-dimensional vector of individual abilities, -**1** is a n-dimensional vector with entries equal to -1 and $\boldsymbol{\epsilon}_j = (\epsilon_{1j},...,\epsilon_{nj})$ is a random sample from a standard normal distribution. The model can be interpreted as the regression of $\mathbf{Z}_j$ on the explanatory variables $\mathbf{X} = [\boldsymbol{\theta} \ -\mathbf{1}]$, considering the $\boldsymbol{\xi}_j$ as regression coefficients as follows

$$\begin{pmatrix} Z_{1j} \\ Z_{2j} \\ \vdots \\ Z_{nj} \end{pmatrix} = \begin{pmatrix} \theta_1 & -1 \\ \theta_2 & -1 \\ \vdots & \vdots \\ \theta_n & -1 \end{pmatrix} \begin{pmatrix} \alpha_j \\ \delta_j \end{pmatrix} + \begin{pmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \vdots \\ \epsilon_{nj} \end{pmatrix} \tag{4.22}$$

Similarly to the previous case, the likelihood function of $\boldsymbol{\xi}_j$ follows the normal distribution with mean equal to the usual least squares estimate $\hat{\boldsymbol{\xi}}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_j$ and variance equal to $(\mathbf{X}'\mathbf{X})^{-1}$.

Finally the posterior distribution obtained combining the likelihood function and the prior distribution on item parameters $\boldsymbol{\xi}_j$ is given by

$$\boldsymbol{\xi}_j|\mathbf{Z},\boldsymbol{\theta},\mathbf{y} \sim N(\hat{\boldsymbol{\xi}}_j;(\mathbf{X}'\mathbf{X})^{-1})I(\alpha_j > 0). \tag{4.23}$$

Another possible solution for computing the posterior density is to choose a prior covariance matrix for the item parameters denoted by

$$\mathbf{\Sigma_0} = \begin{pmatrix} s_\alpha^2 & 0 \\ 0 & s_\delta^2 \end{pmatrix},$$

where s_α and s_δ are the prior standard deviations for α_j and δ_j. Therefore, the conditional posterior distribution of $\boldsymbol{\xi}_j$ is a multivariate normal with mean vector equal to $(\mathbf{X}'\mathbf{X} + \mathbf{\Sigma_0}^{-1})^{-1}\mathbf{X}'\mathbf{Z}_j$ and covariance matrix equal to $(\mathbf{X}'\mathbf{X} + \mathbf{\Sigma_0}^{-1})^{-1}$.

We have reached a quite simple formulation of the conditional distributions of $\mathbf{Z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ and we are able to implement the Gibbs sampler to generate a sequence of drawings from these distributions in three steps:

1. Start with initial values $\boldsymbol{\xi}^{(0)}$, $\boldsymbol{\theta}^{(0)}$ and sample $\boldsymbol{Z}^{(0)}$ from $P(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{y})$;

2. Use $\boldsymbol{Z}^{(0)}$, $\boldsymbol{\xi}^{(0)}$ and sample $\boldsymbol{\theta}^{(1)}$ from $P(\boldsymbol{\theta}|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y})$;

3. Use $\boldsymbol{Z}^{(0)}$, $\boldsymbol{\theta}^{(1)}$ and sample $\boldsymbol{\xi}^{(1)}$ from $P(\boldsymbol{\xi}|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{y})$.

The steps are repeated until convergence.

The MCMC sampling procedure does not seem to be too much sensitive to the choice of starting values; however, reasonable initial values can reduce the time of convergence. Coherently with the prior assumptions about the $\theta's$, a possible solution is to initialize the ability parameters to their prior mean, which is equal to 0. According to Albert (1992), starting values for the item parameters α_j and δ_j, can be respectively set to 2 and $-\Phi^{-1}[(\hat{p}_j)\sqrt{5}]$, for each item j, where $\hat{p}_j = \sum_i y_{ij}/n$. This is equivalent to estimate the probability of a correct response of item j by the corresponding sample proportion. The unconditional probability of a correct response can be expressed as $P(Y_j = 1) = \int P(Y_j = 1|\theta)P(\theta)d\theta$, which turns out to be $P(Y_j = 1) = \Phi(-b_j/\sqrt{1 + a_j^2})$ in the case of a standard normal distribution for the latent trait. If we assume a prior mean for each discrimination parameter, we can use that value to initialize the α_j and compute the starting values for the $\delta's$. However one can decide to initialize the discrimination parameters to suitable values, according to prior knowledge. For example, because we expect that the discrimination parameters vary between 0 and 2 and since the difficulty parameters are on the real line, we may decide to set all the initial $\alpha's$ to 1 and all the $\delta's$ to 0 (see Béguin and Glas (2001)). Another possibility is to use the marginal maximum likelihood (MML) parameter estimates, as described in Chapter 3, but this procedure requires the implementation of the EM algorithm.

So far we have supposed that all the data are available, that is the $\mathbf{y}$ data matrix is complete and consists of correct and wrong responses coded by 1 and 0, respectively. On the other hand, we know that the data available from the guidance project present omitted and not presented items. Therefore, we are especially interested in developing a method which can be used in presence of these two types of missing data. Exploiting the absence of time limit and the fact that we have a competence test (and not a psychological test), we can code the omitted data as 'wrong responses', without lack of information. The incomplete design can be implemented in the Gibbs sampler, imposing the algorithm to skip the missing data, as suggested in Béguin and Glas (2001). Next to the data matrix $\mathbf{y}$, which contains correct, incorrect and missing responses corresponding to n examinees and k items, we can create a new matrix $\mathbf{D}$ as indicator of the incomplete design. Particularly, we have

$$d_{ij} = \begin{cases} 1 & \text{if the item } j \text{ is administered to the respondent } i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.24}$$

Therefore, the Gibbs sampler works as follows:

1. Start with initial values $\boldsymbol{\xi}^{(0)}$, $\boldsymbol{\theta}^{(0)}$ and sample $\boldsymbol{Z}^{(0)}$ from $P(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{y})$ only for the elements equal to 1 of the $\mathbf{D}$ matrix;

2. Use $\boldsymbol{Z}^{(0)}$, $\boldsymbol{\xi}^{(0)}$ and sample $\boldsymbol{\theta}^{(1)}$ from $P(\boldsymbol{\theta}|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y})$ conditionally on $\mathbf{D}$;

3. Use $\boldsymbol{Z}^{(0)}$, $\boldsymbol{\theta}^{(1)}$ and sample $\boldsymbol{\xi}^{(1)}$ from $P(\boldsymbol{\xi}|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{y})$ conditionally on $\mathbf{D}$.

4.4 Gibbs sampler for the multidimensional two-parameter normal ogive model

The multidimensional approach extends the model to the contemporaneous existence of more than one latent trait, specifying the probability of a correct response of person i to item j as follows

$$P(y_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\xi}_j) = \Phi(\eta_{ij}) = \Phi\bigg(\sum_{q=1}^{Q} \alpha_{jq}\theta_{iq} - \delta_j\bigg), \tag{4.25}$$

where:

- $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iq}, ..., \theta_{iQ})'$ is the vector of ability parameters for the individual i, with $q = 1, .., Q$ dimensions,
- $\boldsymbol{\xi}_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jQ}, \delta_j)'$ is the parameter vector for the item j, containing the Q discriminations, one for each dimension, and the difficulty,
- $\eta_{ij} = \sum_{q=1}^{Q} \alpha_{jq}\theta_{iq} - \delta_j = \alpha_{j1}\theta_{i1} + ... + \alpha_{jQ}\theta_{iQ} - \delta_j$.

The multidimensional model was formalized for the first time by Lord and Novick (1968). It is assumed that the $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ are independent and multivariate normally distributed as $\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta)$, where $\boldsymbol{\mu}$ is the $Q-$dimensional mean vector and $\boldsymbol{\Sigma}_\theta$ is the $Q \times Q$ variance-covariance matrix for the Q latent traits. The prior distributions $P(\boldsymbol{\mu}|\boldsymbol{\Sigma}_\theta)$ and $P(\boldsymbol{\Sigma}_\theta)$ are normal and inverse-Wishart, respectively. Furthermore, normal distributions are assumed for item parameters, that is $\alpha \sim N(\mu_\alpha, \sigma^2_\alpha)$ and $\delta \sim N(\mu_\delta, \sigma^2_\delta)$.

The model in (4.25) has been introduced in Chapter 3 and it is not identified. A necessary condition for identification is that each latent variable must be assigned a scale. Another issue is latent variable indeterminacy, since different rotations for the discrimination parameters are possible. Mainly, there are two approaches to identify the model: the first one is to constrain the ability parameters while the other one is to act on the item parameters. The first solution consists of setting $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\theta$ equal to a vector of zeros and to the identity matrix, respectively. Furthermore one has to fix the $\alpha_{jq} = 0$ for $j = 1, ..., Q-1$ and $q = j+1, ..., Q$. This approach is used for example in Fraser (1988), who implemented the harmonic analysis robust method for the normal ogive model in the computer program NOHARM. Another way to identify the model is to set some restrictions only on item parameters, that is

- Impose $\alpha_{jq} = 1$ if $j = q$ and $\alpha_{jq} = 0$ if $j \neq q$, for $j = 1, ..., Q$ and $q = 1, .., Q$;
- Set Q item difficulties δ_j equal to 0.

Béguin and Glas (2001) proved that the two identification methods are interchangeable.

As in the unidimensional model we can express the joint posterior distribution for $\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\theta$, where

$$\begin{aligned} P(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\mu}, \Sigma_\theta|\mathbf{y}) =& P(\mathbf{Z}|\mathbf{y}, \boldsymbol{\xi}, \boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\mu}, \Sigma_\theta)P(\boldsymbol{\mu}|\Sigma_\theta)P(\Sigma_\theta)P(\boldsymbol{\xi}) \\ & \propto \prod_{i=1}^{n}\prod_{j=1}^{k}\{\phi(Z_{ij}; \eta_{ij}, 1)[I(Z_{ij} > 0)I(y_{ij} = 1) + I(Z_{ij} \leq 0)I(y_{ij} = 0)]\} \\ & \prod_{i=1}^{n}\phi(\boldsymbol{\theta}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta)P(\boldsymbol{\mu}|\boldsymbol{\Sigma}_\theta)P(\boldsymbol{\Sigma}_\theta)P(\boldsymbol{\xi}). \end{aligned} \tag{4.26}$$

Assuming that initial appropriate estimate of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are available, the Gibbs sampler works with the following conditional densities:

- $P(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta|\boldsymbol{\theta})$;
- $P(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{y})$;
- $P(\boldsymbol{\theta}|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta)$;
- $P(\boldsymbol{\xi}|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{y})$.

In order to express the conditional distribution $P(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta|\boldsymbol{\theta})$, we should combine the information of the prior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\theta$ and the likelihood function $P(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta)$. First of all, we assume that $P(\boldsymbol{\mu}|\boldsymbol{\Sigma}_\theta)$ is a multivariate normal density, that is $\boldsymbol{\mu}|\boldsymbol{\Sigma}_\theta \sim N(\boldsymbol{\mu}_0, \frac{\boldsymbol{\Sigma}_\theta}{\kappa_0})$, where $\boldsymbol{\mu}_0$ is the prior mean vector and κ_0 is the number of prior measurements on the $\boldsymbol{\Sigma}_\theta$ scale, explicitly

$$P(\boldsymbol{\mu}|\boldsymbol{\Sigma}_\theta) = (2\pi)^{-Q/2}\left|\frac{\boldsymbol{\Sigma}_\theta}{\kappa_0}\right|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\left(\frac{\boldsymbol{\Sigma}_\theta}{\kappa_0}\right)^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right\}. \tag{4.27}$$

Furthermore, we assume $\boldsymbol{\Sigma}_\theta$ having an Inverse-Wishart density, that is $\boldsymbol{\Sigma}_\theta \sim$ Inv-Wishart$_{\nu_0}(\boldsymbol{\Lambda}_0^{-1})$, where ν_0 describes the degrees of freedom and $\boldsymbol{\Lambda}_0$ is the symmetric and positive definite scale matrix for the distribution on $\mathbf{\Sigma}_\theta$, also positive definite, as follows

$$\begin{aligned} P(\mathbf{\Sigma}_\theta) =&\left(2^{\nu_0 Q/2}\pi^{Q(Q-1)/4}\prod_{q=1}^{Q}\Gamma\left(\frac{\nu_0+1-q}{2}\right)\right)^{-1} \\ &\times|\mathbf{\Lambda}_0|^{\nu_0/2}|\mathbf{\Sigma}_\theta|^{-(\frac{\nu_0+Q+1}{2})}\times\exp\left(-\frac{1}{2}tr(\mathbf{\Lambda}_0\mathbf{\Sigma}_\theta^{-1})\right). \end{aligned} \tag{4.28}$$

Therefore, the conjugate prior distribution $P(\boldsymbol{\mu}, \mathbf{\Sigma}_\theta)$ is a normal-inverse-Wishart (see Gelman et al. (1995)), parameterized in terms of hyperparameters $(\boldsymbol{\mu}_0, \mathbf{\Sigma}_\theta/\kappa_0; \nu_0, \mathbf{\Lambda}_0)$ as follows

$$P(\boldsymbol{\mu}, \mathbf{\Sigma}_\theta) \propto |\mathbf{\Sigma}_\theta|^{-((\nu_0+Q)/2+1)}\exp\left(-\frac{1}{2}tr(\mathbf{\Lambda}_0\mathbf{\Sigma}_\theta^{-1}) - \frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\mathbf{\Sigma}_\theta^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right). \tag{4.29}$$

The likelihood function can be computed starting from the assumption of independent and multivariate normally distributed $\boldsymbol{\theta}_i$ with Q-dimensional mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{\Sigma}_\theta$, as follows

$$\begin{aligned} P(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{\Sigma}_\theta) =&\prod_{i=1}^{n} P(\boldsymbol{\theta}_i|\boldsymbol{\mu}, \mathbf{\Sigma}_\theta) \\ \propto&\prod_{i=1}^{n}|\mathbf{\Sigma}_\theta|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_i-\boldsymbol{\mu})'(\mathbf{\Sigma}_\theta)^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu})\right\} \\ =&|\mathbf{\Sigma}_\theta|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\theta}_i-\boldsymbol{\mu})'(\mathbf{\Sigma}_\theta)^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu})\right\}. \end{aligned} \tag{4.30}$$

Consequently, the posterior conditional distribution $P(\boldsymbol{\mu}, \mathbf{\Sigma}_\theta|\boldsymbol{\theta})$ will be a normal-inverse-Wishart (see, for example, Gelman et al. (1995)) with parameters $\boldsymbol{\mu}_n$, $\mathbf{\Sigma}_\theta/\kappa_n$, ν_n and $\mathbf{\Lambda}_n$, where:

$$\begin{aligned}
\boldsymbol{\mu}_n &= \frac{\kappa_0}{\kappa_0 + n}\boldsymbol{\mu}_0 + \frac{n}{\kappa_0 + n}\bar{\boldsymbol{\theta}} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\boldsymbol{\Lambda}_n &= \boldsymbol{\Lambda}_0 + \mathbf{S} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\boldsymbol{\theta}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{\theta}} - \boldsymbol{\mu}_0)'.
\end{aligned}$$

The matrix $\mathbf{S}$ is the $Q \times Q$ sum of squares matrix relative to the sample mean $\bar{\boldsymbol{\theta}}$, that is

$$\mathbf{S} = \sum_{i=1}^{n}(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})'.$$

The vector $\boldsymbol{\theta}_i$ was previously discussed while the elements of $\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \bar{\theta}_2, ..., \bar{\theta}_Q)'$ are the sample means of the Q dimensions. The matrix $\mathbf{S}$ is square and symmetric with elements $S_{pq} = \sum_{i=1}^{n}(\theta_{ip} - \bar{\theta}_p)(\theta_{iq} - \bar{\theta}_q)$, where $\bar{\theta}_p$ and $\bar{\theta}_q$ are the means of the ability parameters of dimensions p and q, with $p, q = 1, ..., n$.

The second conditional distribution of interest, $P(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{y})$, is a truncated normal, following the same formulation of the unidimensional model with proper substitutions of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, particularly

$$Z_{ij}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{y} \sim \begin{cases} N(\eta_{ij}, 1) & \text{with } Z_{ij} > 0 \text{ if } y_{ij} = 1, \\ N(\eta_{ij}, 1) & \text{with } Z_{ij} \leq 0 \text{ if } y_{ij} = 0, \end{cases} \tag{4.31}$$

where $\eta_{ij} = \sum_{q=1}^{Q} \alpha_{jq}\theta_{iq} - \delta_j$.

The conditional distribution of $P(\boldsymbol{\theta}|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta)$ is obtained with a transformation of the $\boldsymbol{\theta}_i$ vector of ability parameters and, similarly to the unidimensional case, with a normal regression interpretation of the model. Generally, a random draw from a multivariate normal distribution with expected value $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ can be performed using the Cholesky decomposition of $\boldsymbol{\Sigma}$, so that $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$, and a vector of d independent univariate normal random variables $\mathbf{Z} = (Z_1, ..., Z_d)$. A random draw from the multivariate distribution is $\boldsymbol{\theta} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$, (see Gelman et al. (1995)). In our case, we know that the vector $\boldsymbol{\theta}_i$ of ability parameters for the individual i has a multivariate normal distribution with expected value $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}_\theta$; therefore, we can apply the Cholesky decomposition to the $\boldsymbol{\Sigma}_\theta$ matrix so that $\boldsymbol{\Sigma}_\theta = \mathbf{LL}'$, where $\mathbf{L}$ is a lower-triangular matrix called "Cholesky factor", and define an orthogonally standardized proficiency variable $\boldsymbol{\theta}^o$. The vector $\boldsymbol{\theta}_i^o = (\theta_{i1}^o, ..., \theta_{iQ}^o)$ is defined with all independent and standard normally distributed elements. Consequently, $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\theta}_i^o$ represents a random draw from the multivariate normal distribution and it is possible to define $\boldsymbol{\theta}_i^o$ in respect to $\boldsymbol{\theta}_i$, i.e. $\boldsymbol{\theta}_i^o = \mathbf{L}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu})$. The combination of item and person parameters η_{ij} becomes

$$\eta_{ij} = \sum_{q=1}^{Q} \alpha_{jq}\theta_{iq} - \delta_j = \sum_{q=1}^{Q}\left(\alpha_{jq}\sum_{h=1}^{Q} L_{hq}\theta_{iq}^o + \mu_q\right) - \delta_j.$$

Using a matrix notation, $\boldsymbol{\eta}_i$ can be written as

$$\boldsymbol{\eta}_i = \mathbf{ALL}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu} + \boldsymbol{\mu}) - \boldsymbol{\delta} = \mathbf{A}(\mathbf{L}\boldsymbol{\theta}_i^o + \boldsymbol{\mu}) - \boldsymbol{\delta},$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{ik})'$, $\boldsymbol{\delta} = (\delta_1, ..., \delta_k)'$ and the matrix $\mathbf{A}$ expressed as follows

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1q} & \dots & \alpha_{1Q} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2q} & \dots & \alpha_{2Q} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \alpha_{j1} & \alpha_{j2} & \dots & \alpha_{jq} & \dots & \alpha_{jQ} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \alpha_{k1} & \alpha_{k2} & \dots & \alpha_{kq} & \dots & \alpha_{kQ} \end{pmatrix}$$

The posterior distribution of the $\boldsymbol{\theta}_i^o$ variables becomes

$$P(\boldsymbol{\theta}_i^o|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y}) \propto \phi(\boldsymbol{\theta}_i^o; \mathbf{0}, \mathbf{I}) \prod_{j=1}^{k} \phi(z_{ij}; \eta_{ij}, 1), \tag{4.32}$$

where $\mathbf{0}$ is a Q-dimensional vector of zeros and $\mathbf{I}$ is the $Q \times Q$ identity matrix. Similarly to the unidimensional case, since $E(\mathbf{Z}_i) = \boldsymbol{\eta}_i = \mathbf{A}(\mathbf{L}\boldsymbol{\theta}_i^o + \boldsymbol{\mu}) - \boldsymbol{\delta}$, we can define a regression model in the following steps

$$
\begin{aligned}
\mathbf{Z}_i &= \mathbf{A}(\mathbf{L}\boldsymbol{\theta}_i^o + \boldsymbol{\mu}) - \boldsymbol{\delta} + \boldsymbol{\epsilon}_i \\
\mathbf{Z}_i + \boldsymbol{\delta} - \mathbf{A}\boldsymbol{\mu} &= \mathbf{A}\mathbf{L}\boldsymbol{\theta}_i^o + \boldsymbol{\epsilon}_i \\
\mathbf{Z}_i + \boldsymbol{\delta} - \mathbf{A}\boldsymbol{\mu} &= \mathbf{B}\boldsymbol{\theta}_i^o + \boldsymbol{\epsilon}_i,
\end{aligned} \tag{4.33}
$$

where $\boldsymbol{\epsilon}_i$ is the vector of the k independent and standard normally distributed error terms ϵ_{ij} and $\mathbf{B} = \mathbf{AL}$. The third formulation of (4.33) expresses the regression of the $(\mathbf{Z}_i+\boldsymbol{\delta}-\mathbf{A}\boldsymbol{\mu})$ on $\mathbf{B}$, with regression coefficients $\boldsymbol{\theta}_i^o$, for observation i. The likelihood function for $\boldsymbol{\theta}_i^o$ is normal with expected value equal to the least square estimate $\hat{\boldsymbol{\theta}}_i^o = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'(\mathbf{Z}_i + \boldsymbol{\delta} - \mathbf{A}\boldsymbol{\mu})$ and variance-covariance matrix equal to $\boldsymbol{\Sigma} = (\mathbf{B}'\mathbf{B})^{-1}$. Combining the prior information about the standardized ability variables, the posterior distribution of $\boldsymbol{\theta}_i^o$ can be written as

$$
\boldsymbol{\theta}_i^o|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y} \sim N\big((\mathbf{I} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}_i^o; (\mathbf{I} + \boldsymbol{\Sigma}^{-1})^{-1}\big). \tag{4.34}
$$

The transformation $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\theta}_i^o$ can be easily used to obtain $\boldsymbol{\theta}_i$ from each $\boldsymbol{\theta}_i^o$.

Finally, we need to examine the conditional distribution $P(\boldsymbol{\xi}|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{y})$ of item parameters. We have already assumed that the item discriminations and difficulties follow a normal distribution, that is $\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$ and $\delta \sim N(\mu_\delta, \sigma_\delta^2)$; therefore, it is possible to express the prior distribution on item parameters for the item j, with $j = 1, ..., k$, as a multivariate normal distribution. Particularly, the vector of item parameters $\boldsymbol{\xi}_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jQ}, \delta_j)'$ has a multivariate normal distribution with a mean vector equal to $\boldsymbol{\mu}_{\xi_0} = (\mu_{\alpha 1}, ..., \mu_{\alpha Q}, \mu_\delta)'$ and variance $\boldsymbol{\Sigma}_{\xi_0} = diag(\sigma_{\alpha 1}, ..., \sigma_{\alpha Q}, \sigma_\delta)$. Conditional to $\mathbf{Z}$ and $\boldsymbol{\theta}$, the posterior distributions of $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k$ are independent with density given by

$$
P(\boldsymbol{\xi}_j|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{y}) \propto \prod_{i=1}^{n} \phi(Z_{ij}; \eta_{ij}, 1)\phi(\delta_j; \mu_{\delta_j}, \sigma_{\delta_j}) \prod_{q=1}^{Q} \phi(\alpha_{jq}; \mu_{\alpha_{jq}}, \sigma_{\alpha_{jq}}). \tag{4.35}
$$

After the definition of a $n \times (Q+1)$ matrix $\mathbf{X}$ containing the ability parameters θ_{iq} for each observation relative to all the Q latent dimension and a column with all the elements equal to -1, it is possible to express the regression of the $\mathbf{Z}_j$ on the covariates $\mathbf{X}$, with $\boldsymbol{\xi}_j$ as regression coefficients

$$\mathbf{Z}_j = \mathbf{X}\boldsymbol{\xi}_j + \boldsymbol{\epsilon}_j, \tag{4.36}$$

or, in explicit matrix notation

$$\begin{pmatrix} Z_{1j} \\ Z_{2j} \\ \vdots \\ Z_{nj} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \dots & \theta_{1q} & \dots & \theta_{1Q} & -1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \theta_{i1} & \dots & \theta_{iq} & \dots & \theta_{iQ} & -1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \theta_{n1} & \dots & \theta_{nq} & \dots & \theta_{nQ} & -1 \end{pmatrix} \begin{pmatrix} \alpha_{1j} \\ \alpha_{2j} \\ \vdots \\ \alpha_{Qj} \\ \delta_j \end{pmatrix} + \begin{pmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \vdots \\ \epsilon_{nj} \end{pmatrix}$$

where the ϵ_{ij} are the independent and standard normally distributed error terms. The likelihood function of $\boldsymbol{\xi}$ is normal with expected value equal to the usual least square estimate $\hat{\boldsymbol{\xi}}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_j$ and variance $\boldsymbol{v} = (\mathbf{X}'\mathbf{X})^{-1}$. Adding the prior information about the item parameters, the posterior distribution is also normal

$$\boldsymbol{\xi}_j|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{y} \sim N\big(\boldsymbol{\mu}_{\boldsymbol{\xi}_j}; ({\boldsymbol{\Sigma}_{\boldsymbol{\xi}_0}}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\big), \tag{4.37}$$

where $\boldsymbol{\mu}_{\boldsymbol{\xi}_j} = ({\boldsymbol{\Sigma}_{\boldsymbol{\xi}_0}}^{-1} + \mathbf{X}'\mathbf{X})^{-1}({\boldsymbol{\Sigma}_{\boldsymbol{\xi}_0}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\xi}_0} + \mathbf{X}'\mathbf{Z}_j)$.

The Gibbs sampler can be implemented to generate a sequence of drawings from the conditional distributions in the following four steps:

1. Starting with an initial value $\boldsymbol{\theta}^{(0)}$, sample $\boldsymbol{\Sigma}_\theta^{(0)}$ from $\boldsymbol{\Sigma}_\theta \sim \text{Inv-Wishart}_{\nu_n}(\boldsymbol{\Lambda}_n^{-1})$ and then sample $\boldsymbol{\mu}^{(0)}$ from $\boldsymbol{\mu}|\boldsymbol{\Sigma}_\theta, \boldsymbol{\theta} \sim N(\boldsymbol{\mu}_n, \frac{\boldsymbol{\Sigma}_\theta}{\kappa_n})$;

2. Start with initial values of $\boldsymbol{\xi}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ and sample $\boldsymbol{Z}^{(0)}$ from $P(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{y})$;

3. Use $\boldsymbol{Z}^{(0)}$, $\boldsymbol{\xi}^{(0)}$, $\boldsymbol{\Sigma}_\theta^{(0)}$ and $\boldsymbol{\mu}^{(0)}$ to sample $\boldsymbol{\theta}^{(1)}$ from $P(\boldsymbol{\theta}|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta)$;

4. Use $\boldsymbol{Z}^{(0)}$ and $\boldsymbol{\theta}^{(1)}$ to sample $\boldsymbol{\xi}^{(1)}$ from $P(\boldsymbol{\xi}|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{y})$.

A possible solution for the starting values is to use the MML estimates or the parameters obtained with NOHARM. Finally, the problem of missing data for not presented items can be handled in the estimation procedure including the **D** matrix of incomplete design, as in the unidimensional model. The sequence of drawings for the Gibbs sampler modifies as follows:

1. Starting with an initial value $\boldsymbol{\theta}^{(0)}$, sample $\boldsymbol{\Sigma}_\theta^{(0)}$ from $\boldsymbol{\Sigma}_\theta \sim \text{Inv-Wishart}_{\nu_n}(\boldsymbol{\Lambda}_n^{-1})$ and then sample $\boldsymbol{\mu}^{(0)}$ from $\boldsymbol{\mu}|\boldsymbol{\Sigma}_\theta, \boldsymbol{\theta} \sim N(\boldsymbol{\mu}_n, \frac{\boldsymbol{\Sigma}_\theta}{\kappa_n})$;

2. Start with initial values of $\boldsymbol{\xi}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ and sample $\boldsymbol{Z}^{(0)}$ from $P(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{y})$, only for the elements corresponding to $d_{ij} = 1$;

3. Use $\boldsymbol{Z}^{(0)}$, $\boldsymbol{\xi}^{(0)}$, $\boldsymbol{\Sigma}_\theta^{(0)}$ and $\boldsymbol{\mu}^{(0)}$ to sample $\boldsymbol{\theta}^{(1)}$ from $P(\boldsymbol{\theta}|\mathbf{Z}, \boldsymbol{\xi}, \mathbf{y}, \boldsymbol{\mu}, \Sigma_\theta)$, conditionally to the **D** matrix;

4. Use $\boldsymbol{Z}^{(0)}$ and $\boldsymbol{\theta}^{(1)}$ to sample $\boldsymbol{\xi}^{(1)}$ from $P(\boldsymbol{\xi}|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{y})$, conditionally to the **D** matrix.

Applications of the algorithm will be presented in the following section.

4.5 Applications of the Gibbs sampler

General software for MCMC in IRT models has been implemented in R (Martin and Quinn, 2007) and in S-plus (Patz and Junker, 1999).

Specifically, the Gibbs sampler has been applied to the unidimensional 2PNO model, without incomplete design, by Albert (1992). The syntax is available. Béguin and Glas (2001) implemented the algorithm introducing the guessing parameter and the incomplete design in the model. Unfortunately, no syntax is available.

Therefore, we have decided to write a specific code for the application of Gibbs sampler for the unidimensional and bidimensional 2PNO models, including the incomplete design. The syntax has been written by using the software Mathlab 7.1 (Mat, 2005). The use of a specific code allows flexibility in the estimation process and increases the capability of modelling all the dependencies. The developed code can be found in Appendix D. This section will show several results of the Gibbs sampler application, both for simulated and real data.

First of all, the capability of parameter recovery for the MCMC procedure is tested with simulated data, considering the unidimensional 2PNO model. Responses to 10 binary items have been simulated for 1500 observations, drawn from a standard normal ability distribution. The item parameters α_j and δ_j for the 10 items are presented in columns 2 to 3 of Table 4.1 (True values). The model has been identified by imposing the mean and the variance for the ability distribution equal to zero and one, respectively. The starting values for the item parameters have been fixed to one for all the slopes and zero for all the difficulties. The ability vector has been initialized to a vector of zeros. A prior covariance matrix Σ_0 has been chosen for the item parameters, with prior standard deviations s_α and s_δ equal to 1. No significative differences in the results have been noticed, choosing different values for the prior standard deviations. The MCMC procedure had a run length of 5000 iterations and took only few seconds on a 2.21 Ghz, AMD Athlon 64, 3500+ processor. The plots of sampled values suggested a limited burn-in period of 100 iterations. The simulation has been conducted also with a run length of 30000 iterations, but no improvement has been noticed. Table 4.1 shows the results of the simulation. The estimated item parameters obtained with the Gibbs sampler procedure are presented, both for the slopes $\hat{\alpha_j}$ and the intercepts $\hat{\delta_j}$, together with the correspondent standard deviations. The estimates are also called EAP (Expected A Posteriori). The last two columns show the absolute value of the difference between the true and the estimated values, for all the item parameters. Figure 4.1 shows the plot of the true values (x-axis) against the EAP estimates (y-axis) for all the item parameters. The residuals are low as well as the standard deviations of the simulated

	True values		Gibbs sampler				Abs.difference	
Item	α_j	δ_j	$\hat{\alpha_j}$	sd($\hat{\alpha_j}$)	$\hat{\delta_j}$	sd($\hat{\delta_j}$)	$\lvert\alpha_j - \hat{\alpha_j}\rvert$	$\lvert\delta_j - \hat{\delta_j}\rvert$
1	0.675	-1.041	0.600	0.06	-1.036	0.05	0.075	0.005
2	0.585	0.480	0.733	0.06	0.496	0.04	0.148	0.016
3	0.240	0.868	0.260	0.05	0.844	0.04	0.020	0.024
4	0.662	0.688	0.753	0.06	0.745	0.04	0.091	0.057
5	0.143	-0.086	0.181	0.04	-0.113	0.03	0.038	0.027
6	1.272	0.093	1.175	0.09	0.137	0.04	0.097	0.044
7	0.369	-0.031	0.368	0.04	-0.060	0.03	0.001	0.029
8	0.644	-0.277	0.646	0.06	-0.266	0.03	0.002	0.011
9	0.681	-1.079	0.736	0.07	-1.136	0.06	0.055	0.057
10	0.621	-0.161	0.584	0.05	-0.128	0.03	0.037	0.033

Table 4.1. MCMC parameter recovery of simulated data, unidimensional 2PNO model

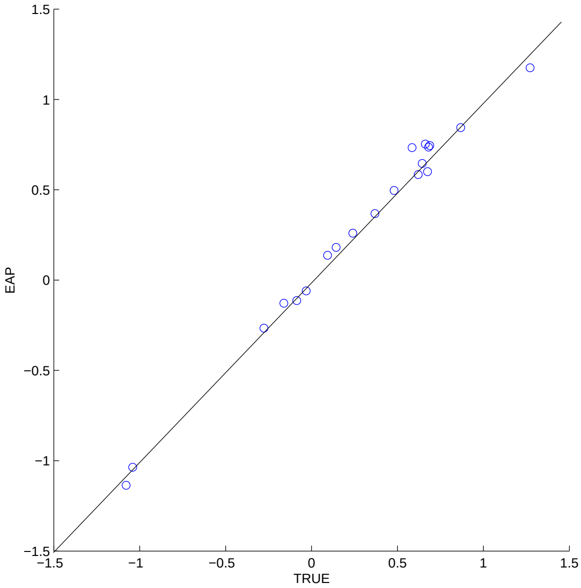


Figure 4.1. Plot of true and EAP estimates with the MCMC procedure

values. Therefore, the Gibbs sampler seems to converge and to recover the original parameters quite well. As example, the plot of the simulated values for item 7 is presented in Figure 4.2. The discrimination parameter is considered in the left-side graphic while the difficulty term in the right-side one.

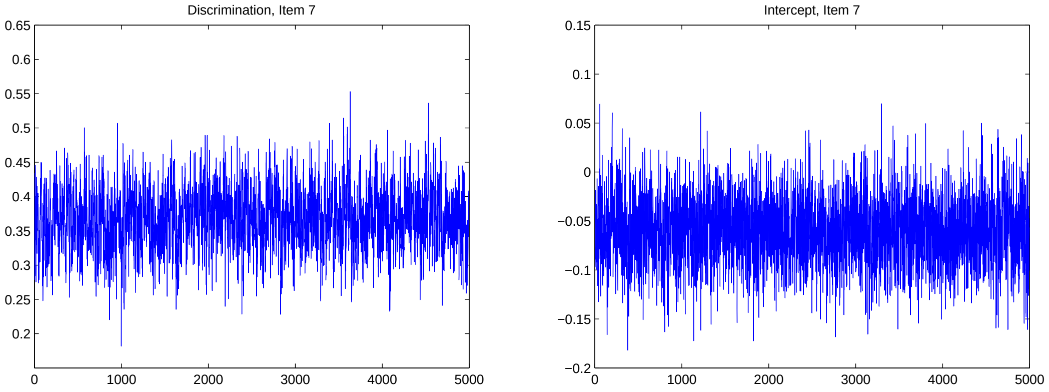


Figure 4.2. Plot of simulated item parameters for item 7

The second study is conducted on real data. The Psychology specific test is considered and unidimensionality is assumed. The data have been analyzed using the Gibbs sampler for the 2PNO model. The main purpose is to estimate the item parameters for the 20 specific items. The missing data from omitted items have been recoded as "wrong" responses. We have generated 30000 iterations and 5000 iterations have been excluded from the computation as burn-in phase. The starting values used in the MCMC algorithm are $\alpha_j = 1$ and $\delta_j = 0$ for each item j and $\theta_i = 0$ for all individuals. As in the previous study, a prior covariance matrix has been chosen for the item parameters, with prior standard deviations s_α and s_δ equal to 0.5. The results of the EAP estimates for the item parameters are presented in Table 4.2. The discrimination parameters are estimated to be all positive, with values ranging from 0.058 to 0.931. This is a good result, considering that no prior on the discrimination parameter has been imposed to constrain the estimates to be positive. The difficulty levels vary from -1.783 of item 14, estimated to be the easiest item, to 0.789 of item 1, the most difficult item in the set. The standard deviations are moderate. The results are coherent with the calibration conducted on the test in Section 2.5.2.

Item	α_j	sd(α_j)	δ_j	sd(δ_j)	Item	α_j	sd(α_j)	δ_j	sd(δ_j)
1	0.158	0.047	0.789	0.039	11	0.516	0.053	0.405	0.037
2	0.401	0.048	0.035	0.034	12	0.389	0.052	0.686	0.039
3	0.058	0.042	0.361	0.034	13	0.265	0.046	0.541	0.036
4	0.440	0.051	-0.518	0.038	14	0.226	0.072	-1.783	0.065
5	0.296	0.046	0.506	0.036	15	0.540	0.069	-1.376	0.059
6	0.181	0.043	0.282	0.034	16	0.355	0.046	-0.061	0.035
7	0.847	0.093	-1.647	0.086	17	0.702	0.065	-0.690	0.044
8	0.474	0.050	-0.409	0.037	18	0.789	0.075	-0.985	0.054
9	0.485	0.056	-0.787	0.041	19	0.511	0.054	0.493	0.038
10	0.591	0.066	-1.106	0.052	20	0.931	0.096	-1.686	0.091

Table 4.2. Item parameter estimates with the Gibbs sampler algorithm, Psychology specific test

To make a comparison with FIFA, the estimates obtained with TESTFACT are presented in Table 4.3[1].

As we can note, no standard errors are provided by TESTFACT. This limit

[1] The δ_j parameters have opposite sign to make the comparison possible with the estimates of the Gibbs sampler.

Item	α_j	δ_j	Item	α_j	δ_j
1	0.138	0.786	11	0.511	0.390
2	0.391	0.021	12	0.373	0.676
3	0.043	0.359	13	0.251	0.533
4	0.421	-0.535	14	0.367	-2.246
5	0.537	0.444	15	0.514	-1.402
6	0.175	0.275	16	0.323	-0.072
7	0.884	-1.744	17	0.677	-0.716
8	0.480	-0.431	18	0.792	-1.031
9	0.466	-0.807	19	0.482	0.476
10	0.570	-1.131	20	0.975	-1.797

Table 4.3. Item parameter estimates with FIFA, Psychology specific test

has been already mentioned in the discussion about FIFA. The results of the two methods are coherent and the estimates both for discrimination and difficulty parameters are quite close for most of the items. Larger discrepancies can be noted for items 5 and 14, for both parameters.

Finally the two-dimensional 2PNO model is considered in the application on the complete Psychology test. The item parameters of the 30 general culture and the 20 specific items are jointly estimated with the Gibbs sampler algorithm for the multidimensional model with incomplete design. The starting values for the algorithm are $\alpha_{j1} = \alpha_{j2} = 1$ and $\delta_j = 0$ for each item j and $\theta_{i1} = \theta_{i2} = 0$ for all individuals i. Normal priors have been employed for the item parameters ($\alpha_1 \sim N(1,1)$, $\alpha_2 \sim N(1,1)$ and $\delta \sim N(0,1)$). We have performed 20000 iterations of the algorithm and we have chosen a burn-in period of 5000 to be conservative. The estimation process for the Psychology dataset took nearly 3 hours. The model has been identified imposing the mean vector of abilities equal to a vector of zeros and fixing 4 discrimination parameters. In particular, prior knowledge about the composition of the items was used to subdivide the items into the general and specific sections. Therefore, we have fixed $\alpha_{1,1} = 1$, $\alpha_{1,2} = 0$, $\alpha_{31,1} = 0$ and $\alpha_{31,2} = 1$, where the first index refers to the item and the second one to the ability θ. The identification reflects the belief that the first general culture item (item 1) does not load on the second dimension and the first specific item (item 31) does not load on the first dimension. The results of the discrimination estimates are presented in Table 4.4.

The results show that the data support a two-dimensional interpretation, as

Item	α_{j1}	sd(α_{j1})	α_{j2}	sd(α_{j2})	Item	α_{j1}	sd(α_{j1})	α_{j2}	sd(α_{j2})
1	**1.000**	0.000	**0.000**	0.000	26	**0.725**	1.084	0.257	0.455
2	**0.767**	1.067	-0.064	0.502	27	**0.721**	1.062	0.169	0.461
3	**0.737**	1.032	-0.230	0.546	28	**0.743**	1.064	0.310	0.457
4	**0.748**	1.039	-0.269	0.560	29	**0.709**	1.106	-0.180	0.537
5	**0.745**	1.054	-0.023	0.497	30	**0.731**	1.075	0.318	0.452
6	**0.738**	1.014	-0.014	0.491	31	**0.000**	0.000	**1.000**	0.000
7	**0.741**	1.097	0.048	0.472	32	0.298	1.660	**2.088**	0.996
8	**0.766**	1.058	-0.263	0.552	33	**0.543**	1.229	0.249	0.546
9	**0.746**	1.013	0.120	0.471	34	0.335	1.654	**2.082**	1.104
10	**0.776**	1.045	-0.053	0.496	35	0.429	1.524	**1.765**	0.779
11	**0.715**	1.106	0.032	0.478	36	0.534	1.304	**1.220**	0.655
12	**0.750**	1.068	0.033	0.487	37	0.214	1.589	**1.061**	1.166
13	**0.748**	1.108	-0.068	0.501	38	0.638	1.409	**2.061**	1.150
14	**0.744**	1.095	-0.065	0.505	39	0.141	1.412	**2.033**	1.150
15	**0.753**	1.040	0.151	0.463	40	0.140	1.227	**1.767**	1.326
16	**0.723**	1.068	0.228	0.464	41	0.235	1.698	**2.052**	1.203
17	**0.722**	1.128	0.422	0.466	42	0.497	1.354	**2.080**	0.992
18	**0.670**	1.238	0.357	0.458	43	0.330	1.648	**1.589**	0.752
19	**0.712**	1.123	0.255	0.464	44	0.562	1.471	**1.482**	0.793
20	**0.656**	1.195	0.244	0.467	45	0.431	1.461	**1.910**	1.297
21	**0.737**	1.055	0.026	0.484	46	0.389	1.544	**2.001**	0.893
22	**0.683**	1.153	0.158	0.468	47	0.339	1.543	**0.949**	1.078
23	**0.698**	1.123	0.427	0.454	48	0.264	1.459	**0.714**	0.790
24	**0.707**	1.111	0.472	0.467	49	0.490	1.661	**1.841**	1.344
25	**0.717**	1.085	0.025	0.484	50	0.459	1.654	**0.677**	0.721

Table 4.4. Discrimination estimates with the Gibbs sampler algorithm, Psychology complete test

we have pointed out in the FIFA estimation of the same dataset. In fact, the general culture items (1-30) load on the first dimension, that is they have an higher discrimination parameter α_{j1} respect to α_{j2}. For general culture items, several discrimination estimates are negative and close to zero on the second dimension. On the other hand, the specific items (31-50) load on the second latent variable, with the only exception of item 33. The standard deviations are quite high respect to the results obtained in the unidimensional estimation. The estimated correlation between θ_1 and θ_2 is equal to 0.216.

Table 4.5 refers to the difficulty estimates.

Item	δ_j	sd(δ_j)	Item	δ_j	sd(δ_j)
1	0.046	0.021	26	0.093	0.022
2	-0.165	0.039	27	-0.209	0.022
3	-0.260	0.024	28	0.095	0.027
4	-0.276	0.023	29	-0.280	0.026
5	-0.181	0.029	30	-0.054	0.032
6	0.122	0.029	31	0.795	0.039
7	-0.137	0.025	32	0.035	0.035
8	-0.242	0.033	33	0.364	0.035
9	0.003	0.069	34	-0.527	0.038
10	-0.248	0.029	35	0.510	0.036
11	0.028	0.029	36	0.283	0.034
12	-0.172	0.028	37	-1.668	0.088
13	-0.119	0.024	38	-0.408	0.037
14	-0.246	0.023	39	-0.841	0.049
15	-0.019	0.039	40	-1.219	0.076
16	-0.117	0.021	41	0.411	0.037
17	-0.057	0.041	42	0.693	0.039
18	0.057	0.022	43	0.549	0.037
19	-0.111	0.023	44	-1.819	0.068
20	0.032	0.022	45	-1.391	0.059
21	-0.266	0.027	46	-0.062	0.034
22	-0.130	0.029	47	-0.746	0.052
23	-0.026	0.028	48	-1.087	0.068
24	-0.034	0.028	49	0.511	0.039
25	-0.242	0.034	50	-1.729	0.096

Table 4.5. Difficulty estimates with the Gibbs sampler algorithm, Psychology complete test

The difficulty parameters vary between the items and lower standard deviations are noticed respect to the discrimination estimates.

The Gibbs sampler provided an alternative and useful method for the parameter estimation. Nevertheless, the approach is not without problems. The most relevant one is the computational time, which is sensibly higher respect to the MML estimation in case of a two-dimensional model. Furthermore, no improvement in the precision of estimates has been noticed increasing the number of iterations. More research is certainly needed in this sense.

Chapter 5

Conclusions

This work has considered the entrance guidance in the University education context. A new guidance test has been created in order to give to students an effective and simple instrument for the University faculty choice. In this regard, items able to verify knowledge, interest and reasoning ability, instead of aptitude, have been preferred. The items have been divided into two sections: general culture and faculty-specific.

In order to perform the item calibration, preliminary classical analyses have been conducted which showed that the items are very different in terms of p-values and point biserial correlations. Therefore, Item Response Theory (IRT) has been used to express the relation between responses and latent abilities through probabilistic models which take into account discrimination, difficulty and guessing of the items. The calibration for the Psychology specific items has been conducted under the unidimensionality assumption by using the Multiple-choice model and the 3PL model. The results have proved that not all the items have a correct trend, in terms of response category curves, and that distractors are often not so attractive. Nevertheless, the items are quite discriminating, have different levels of difficulty and quite moderate guessing. Relevant problems have been encountered for item 14, due to extreme easiness and very low discrimination.

Furthermore, multidimensionality has been investigated. Multiple abilities have been considered to estimate the item parameters of the complete test. In particular, the estimation of the multidimensional 2PNO model has been described through two different approaches: the classical FIFA and the MCMC in the Bayesian framework. A two-dimensional solution is considered. Furthermore, the problem of incomplete design has been introduced. The model has been estimated with MML by using the software TESTFACT (Bock et al., 2002). The FIFA results have shown that a two-dimensional model is supported by the data and the test structure is quite clear: the general culture items load on the first ability while the specific items load on the second one. This solution fits better respect to the unidimensional one.

Both unidimensional and two-dimensional models have been considered in the MCMC estimation. Particularly, the Gibbs sampler has been implemented by writing a specific code in Mathlab (Mat, 2005) and results from simulated and real data have been described. The application on the Psychology dataset confirms the two-dimensional interpretation of the data. The Gibbs sampler has been chosen for the capability of modelling all the dependencies between the items and the latent variables and for overcoming the limits of standard software for MML estimation, like the identification of the model and the limited number of latent variables. The Gibbs sampler is also relative easy to implement, respect to the EM algorithm. On the other hand, the MCMC estimation needs a higher computational time respect to MML estimation with more than one ability. Further research in order to improve the efficiency of the method and to deeply study its capabilities is certainly needed.

The analysis conducted on the guidance data revealed a potentially useful but still provisional instrument. In order to improve the test, future possible developments are described.

First of all, the item calibration should be completed to proceed with the students' evaluation. We have experimented that obtaining items with the desired characteristics is very difficult in practice. The item parameter estimates should be stable to calibrate correctly the test. To this end, a larger sample size is required, especially for the faculties with low number of respondents, and a deep study of differential item functioning (DIF) is needed. The DIF is useful to understand if items behave differently for different groups of individuals. Particularly, an item shows DIF if the item response functions across different subgroups are not identical (Hambleton et al., 1991). In our test, different groups of individuals may be characterized by sex, attended school or place of origin.

Secondly, the ability estimation needs to be performed. Students should be provided with a qualitative and informative comment on their performance, instead of the simple total test score. The estimation of an ability score for each student may be obtained in IRT by using ML or Bayesian methods. An alternative solution may be found in the latent class modelling (Lazarsfeld and Henry, 1968), assuming categorical latent variables. Therefore, a qualitative judgment on the performance may be based on a class of ability. Furthermore, the incomplete design should be treated to make the tests comparable. In fact, different tests for the same faculty are generated by the block design. The IRT test equating may provide an effective solution in order to make the scores comparable.

Furthermore, the item and test information should be taken into account. In IRT, item information is used in order to understand the single contribution of the items in the test and consequently, for item selection. Item selection in case of multiple abilities and computerized adaptive testing has been deeply described in Veldkamp (2001). Applications to the guidance test are possible, so that items with low or high informative capability respect to the latent construct may be detected.

Finally, a complete guidance process is needed. The competence test should represent only a single step of a more complex guidance path. Therefore, the purpose is create several different and complementary instruments to guide students into the University choice. The first one will be a psychological and aptitude evaluation, followed by the competence test. Furthermore, students will have the possibility to use e-learning instruments and finally to know the working possibilities of the single faculties.

Bibliography

T.A. Ackerman. Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4): 255–278, 1994.

A.J. Adams, M. Wilson, and W.-C. Wang. The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1): 1–23, 1997.

J.H. Albert. Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17(3):251–269, 1992.

D.J. Bartholomew and M. Knott. *Latent Variable Models and Factor Analysis.* Arnold Publishers, London, 1999.

D.J. Bartholomew and S.O. Leung. A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, 55:1–15, 2002.

A.A. Béguin and C.A.W. Glas. MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66(4):541–562, 2001.

A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, editors, *Statistical theories of mental test scores.* Addison-Wesley, Reading MA, 1968.

R.D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972.

R.D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, 46(4):443–459, 1981.

R.D. Bock and M. Lieberman. Fitting a response model for n dichotomously scored items. *Psychometrika*, 35(2):179–197, 1970.

R.D. Bock, R.Gibbons, and E. Muraki. Full-information item factor analysis. *Applied Psychological Measurement*, 12(3):261–280, 1988.

R.D. Bock, R. Gibbons, S.G. Schilling, E. Muraki, D.T. Wilson, and R. Wood. *TESTFACT 4*. Scientific Software International, Lincolnwood, IL, 2002.

D.M. Bolt and V.F. Lall. Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, 27(6):395–414, 2003.

A. Christoffersson. Factor analysis of dichotomized variables. *Psychometrika*, 40: 5–32, 1975.

J. de la Torre and R.J. Patz. A multidimensional item response theory approach to simultaneous ability estimation. Paper presented to the *National Council on Measurement in Education* in New Orleans, LA, 2002.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm (with Discussion). *Journal of the Royal Statistical Society*, 39(Series B):1–38, 1977.

D.R. Divgi. Calculation of the tetrachoric correlation coefficient. *Psychometrika*, 44:169–172, 1979.

M. du Toit. *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Scientific Software International, Lincolnwood, IL, 2003.

S.E. Embretson and S.P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah-New Jersey, 2000.

A. Evers, N. Anderson, and O. Oskuijil. *The Blackwell handbook of personnel selection*. Blackwell, Malden, MA, 2005.

G.P. Fox and C.A.W. Glas. Bayesian estimation of a multilevel of an IRT model using Gibbs sampling. *Psychometrika*, 66(2):271–288, 2001.

C. Fraser. *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. University of New England, Armidale, Australia, 1988.

D. Gamerman. *Markov Chain Monte Carlo*. Chapman and Hall, London, 1997.

A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85:398–409, 1990.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, 1995.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactons on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

C.A.W. Glas. The derivation of some tests for the rasch model from the multinomial distribution. *Psychometrika*, 53:525–546, 1988.

J.S. Haberman. Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, 5:1148–1169, 1977.

D.C. Haley. Estimation of the dosage mortality relationship when the dose is subject to error. Technical Report 15 (Office of Naval Research Contract No. 25140, NR-342-022), Stanford University: Applied Mathematics and Statistics Laboratory, 1952.

R.K. Hambleton and H. Swaminathan. *Item Response Theory: Principles and applications*. Kluwer Nijhoff Publishing, Boston, 1985.

R.K. Hambleton, H. Swaminathan, and H.J. Rogers. *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, CA, 1991.

W.K. Hastings. Monte Carlo simulation methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

A.E. Hendrickson and P.O. White. Promax: a quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology*, 17:65–70, 1964.

V.E. Johnson and J.H. Albert. *Ordinal Data Modeling.* Springer-Verlag, New York, 1999.

K. Joreskog. New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, 24: 387–404, 1990.

G.F. Kuder and M.W. Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2:151–160, 1937.

P.F. Lazarsfeld and N.W. Henry. *Latent Structure Analysis.* Houghton Mill, Boston, 1968.

F.M. Lord. A theory of test scores. *Psychometric Monograph No. 7*, 1952.

F.M. Lord and M.R. Novick. *Statistical theories of mental test scores.* Addison-Wesley, Reading, MA, 1968.

A.D. Martin and K.M. Quinn. *Markov chain Monte Carlo (MCMC) package, version 0.8-1.* 2007.

Mathlab 7.1. The MathWorks, Inc., Natick, MA, 2005.

E. Muraki and R.D. Bock. *PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales.* Scientific Software International, Chicago, 1997.

E. Muraki and G.Jr. Engelhard. Full-information item factor analysis: applications of eap scores. *Applied Psychological Measurement*, 9(4):417–430, 1985.

B. Muthén. Contributions to factor analysis of dichotomized variables. *Psychometrika*, 43:551–560, 1978.

B. Muthén. A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, 49:115–132, 1984.

B. Muthén and A. Christoffersson. Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, 46:407–419, 1981.

R.J. Patz and B.W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24:342–366, 1999a.

R.J. Patz and B.W. Junker. *STATLIB package mcmcirt, release 1.0.* 1999.

R.J. Patz and B.W. Junker. A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24:146–178, 1999b.

G. Rasch. *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen, 1960.

M.D. Reckase. A linear logistic multidimensional model for dichotomous item response data. In W.J. Van der Linden and R.K. Hambleton, editors, *Handbook of Modern Item Response Theory*, chapter 16, pages 271–296. Springer-Verlag, New York, 1997.

M.D. Reckase. What is the "correct" dimensionality for a set of item response data? In D. Laveault, B. D. Zumbo, M. E. Gessaroli, and M.W. Boss, editors, *Modern theories of measurement: Problems andissues*, pages 87–92. University of Ottawa, (Ontario, Canada), 1994.

M. Reiser. Analysis of residuals for the multinomial item response model. *Psychometrika*, 61:509–528, 1996.

A.A. Rupp. Item Response Modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing*, 3(4):365–384, 2003.

S.K. Sahu. Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, 72(3):217–232, 2002.

F. Samejima. Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17*, 1969.

F. Samejima. A new family of models for the multiple choice items. Research Report 79-4, University of Tennessee, Department of Psychology, Knoxville, 1979.

F. Samejima. Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 39(1):111–121, 1974.

A. Skrondal and S. Rabe-Hesketh. *Generalized latent variable modeling : multilevel, longitudinal, and structural equation models.* Chapman and Hall/CRC, Boca Raton, Florida, 2004.

Y. Takane and J. de Leeuw. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3):393–408, 1987.

R. Tate. A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27:159–203, 2003.

D. Thissen. *MULTILOG 7.0. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory.* Scientific Software International, Lincolnwood, IL, 2003.

D. Thissen and M.C. Edwards. Diagnostic scores augmented using multidimensional item response theory. Paper presented to the *National Council on Measurement in Education*in Montreal, Canada, 2005.

D. Thissen and L. Steinberg. A response model for multiple choice items. *Psychometrika*, 49(4):501–519, 1984.

D. Thissen and L. Steinberg. A taxonomy of item response models. *Psychometrika*, 51(4):567–577, 1986.

D. Thissen and H. Wainer. Some standard errors in item response theory. *Psychometrika*, 47:397–412, 1982.

L.L. Thurstone. *Multiple-factor analysis.* University of Chicago press, Chicago, 1947.

R.E. Traub. A priori considerations in choosing an item response model. In R.K. Hambleton, editor, *Applications of item response theory*, pages 57–70. Educational Research Institute of British Columbia, Vancouver, BC, 1983.

W.J. van der Linden and R.K. Hambleton. *Handbook of Modern Item Response Theory*. Springer-Verlag, New York, 1997.

B.P. Veldkamp. *Principles and Methods of Constrained Test Assembly*. Unpublished doctoral dissertation, University of Twente, The Netherlands, 2001.

W.-C. Wang, P.-H. Chen, and Y.-Y. Cheng. Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9(1):116–136, 2004.

M.F. Zimowski, E. Muraki, R.J. Mislevy, and R.D. Bock. *BILOG-MG: Multiple group IRT analysis and test maintenance for binary items*. Scientific Software International, Chicago, 1996.

Appendix A

Percentage of correct, incorrect and omitted responses for the specific items

A.1 Agriculture

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	41.70	54.26	4.04	11	39.46	55.16	5.38
2	78.48	19.73	1.79	12	69.96	27.80	2.24
3	44.39	50.22	5.38	13	56.95	39.91	3.14
4	59.19	39.01	1.79	14	48.88	46.19	4.93
5	83.41	11.66	4.93	15	21.97	75.78	2.24
6	60.99	35.87	3.14	16	76.23	21.08	2.69
7	45.29	47.09	7.62	17	11.66	83.86	4.48
8	56.05	36.77	7.17	18	62.33	34.98	2.69
9	56.95	38.57	4.48	19	13.45	72.20	14.35
10	78.03	16.14	5.83	20	47.09	47.98	4.93

Table A.1. Percentage of correct, incorrect and omitted responses - Agriculture

A.2 Arts and Humanities

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	47.31	50.85	1.84	11	55.88	40.33	3.79
2	22.48	74.53	2.99	12	32.30	62.61	5.08
3	28.61	69.49	1.89	13	75.62	22.33	2.04
4	87.79	11.27	0.95	14	98.31	0.60	1.10
5	51.00	47.31	1.69	15	26.47	67.50	6.03
6	29.06	68.69	2.24	16	80.11	18.54	1.35
7	87.69	10.92	1.40	17	54.64	43.47	1.89
8	86.59	12.11	1.30	18	41.48	56.58	1.94
9	76.62	21.19	2.19	19	53.94	41.72	4.34
10	98.31	0.95	0.75	20	36.99	61.32	1.69

Table A.2. Percentage of correct, incorrect and omitted responses - Arts and Humanities

A.3 Economics

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	53.49	42.17	4.34	11	84.81	13.21	1.98
2	45.47	49.81	4.72	12	73.96	22.74	3.30
3	66.98	28.11	4.91	13	65.19	31.89	2.92
4	55.19	40.57	4.25	14	55.38	42.26	2.36
5	25.85	68.21	5.94	15	75.66	20.85	3.49
6	51.60	44.91	3.49	16	75.28	20.57	4.15
7	92.45	5.19	2.36	17	56.04	39.25	4.72
8	78.11	18.58	3.30	18	59.43	37.64	2.92
9	75.57	20.38	4.06	19	88.68	9.53	1.79
10	55.57	42.17	2.26	20	34.62	62.36	3.02

Table A.3. Percentage of correct, incorrect and omitted responses - Economics

A.4 Education Science

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	67.23	30.67	2.10	11	42.44	53.36	4.20
2	87.39	9.66	2.94	12	73.53	21.01	5.46
3	78.15	19.33	2.52	13	92.02	6.30	1.68
4	91.18	6.72	2.10	14	53.78	39.92	6.30
5	84.45	12.61	2.94	15	15.13	74.79	10.08
6	58.82	36.13	5.04	16	27.73	62.18	10.08
7	84.87	11.76	3.36	17	41.60	54.20	4.20
8	23.11	65.55	11.34	18	28.57	68.91	2.52
9	47.06	48.32	4.62	19	41.60	56.30	2.10
10	70.17	26.89	2.94	20	23.95	72.69	3.36

Table A.4. Percentage of correct, incorrect and omitted responses - Education Science

A.5 Foreign Languages and Literature

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	45.88	49.84	4.28	11	45.78	50.80	3.42
2	22.99	70.05	6.95	12	50.48	45.67	3.85
3	57.65	36.79	5.56	13	73.37	23.42	3.21
4	36.90	58.50	4.60	14	89.20	6.84	3.96
5	17.33	73.37	9.30	15	76.15	20.32	3.53
6	60.53	37.01	2.46	16	35.83	57.86	6.31
7	52.41	41.60	5.99	17	54.22	41.93	3.85
8	21.07	76.47	2.46	18	71.76	23.42	4.81
9	24.17	68.13	7.70	19	43.64	49.73	6.63
10	47.27	49.73	2.99	20	58.82	38.18	2.99

Table A.5. Percentage of correct, incorrect and omitted responses - Foreign Languages and Literature

A.6 Pharmacy

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	67.41	28.01	4.58	11	33.12	60.17	6.71
2	81.47	14.06	4.47	12	45.79	46.75	7.45
3	37.27	58.04	4.69	13	67.31	27.69	5.01
4	56.12	37.81	6.07	14	29.82	55.70	14.48
5	71.57	24.49	3.94	15	23.43	62.41	14.16
6	48.03	46.33	5.64	16	82.32	13.53	4.15
7	53.99	42.07	3.94	17	59.96	31.10	8.95
8	56.98	38.76	4.26	18	91.48	4.26	4.26
9	81.15	14.06	4.79	19	60.81	34.19	5.01
10	42.28	49.31	8.41	20	79.77	15.87	4.37

Table A.6. Percentage of correct, incorrect and omitted responses - Pharmacy

A.7 Political Science

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	92.51	5.83	1.66	11	96.71	2.25	1.04
2	43.45	53.47	3.08	12	82.19	15.86	1.96
3	79.78	18.14	2.08	13	97.63	0.87	1.50
4	88.51	10.07	1.41	14	63.21	35.54	1.25
5	72.78	25.76	1.46	15	96.88	1.83	1.29
6	53.39	43.99	2.62	16	60.22	38.12	1.66
7	85.68	13.28	1.04	17	95.96	2.91	1.12
8	84.10	14.48	1.41	18	72.66	25.22	2.12
9	84.60	13.90	1.50	19	49.27	47.77	2.95
10	75.61	22.39	2.00	20	66.17	32.13	1.71

Table A.7. Percentage of correct, incorrect and omitted responses - Political Science

A.8 Statistics

Item	Correct	Incorrect	Omitted	Item	Correct	Incorrect	Omitted
1	75.69	18.78	5.52	11	88.40	7.73	3.87
2	74.59	19.34	6.08	12	43.65	50.28	6.08
3	66.30	29.28	4.42	13	56.91	38.12	4.97
4	80.11	12.15	7.73	14	62.98	32.60	4.42
5	53.59	41.99	4.42	15	40.33	55.25	4.42
6	38.67	56.35	4.97	16	48.07	46.96	4.97
7	56.91	37.57	5.52	17	91.16	4.42	4.42
8	54.70	40.88	4.42	18	87.29	6.08	6.63
9	63.54	32.04	4.42	19	61.33	34.25	4.42
10	77.35	17.13	5.52	20	29.83	66.30	3.87

Table A.8. Percentage of correct, incorrect and omitted responses - Statistics

Appendix B

Response category curves according to the Multiple Choice Model

B.1 General culture items

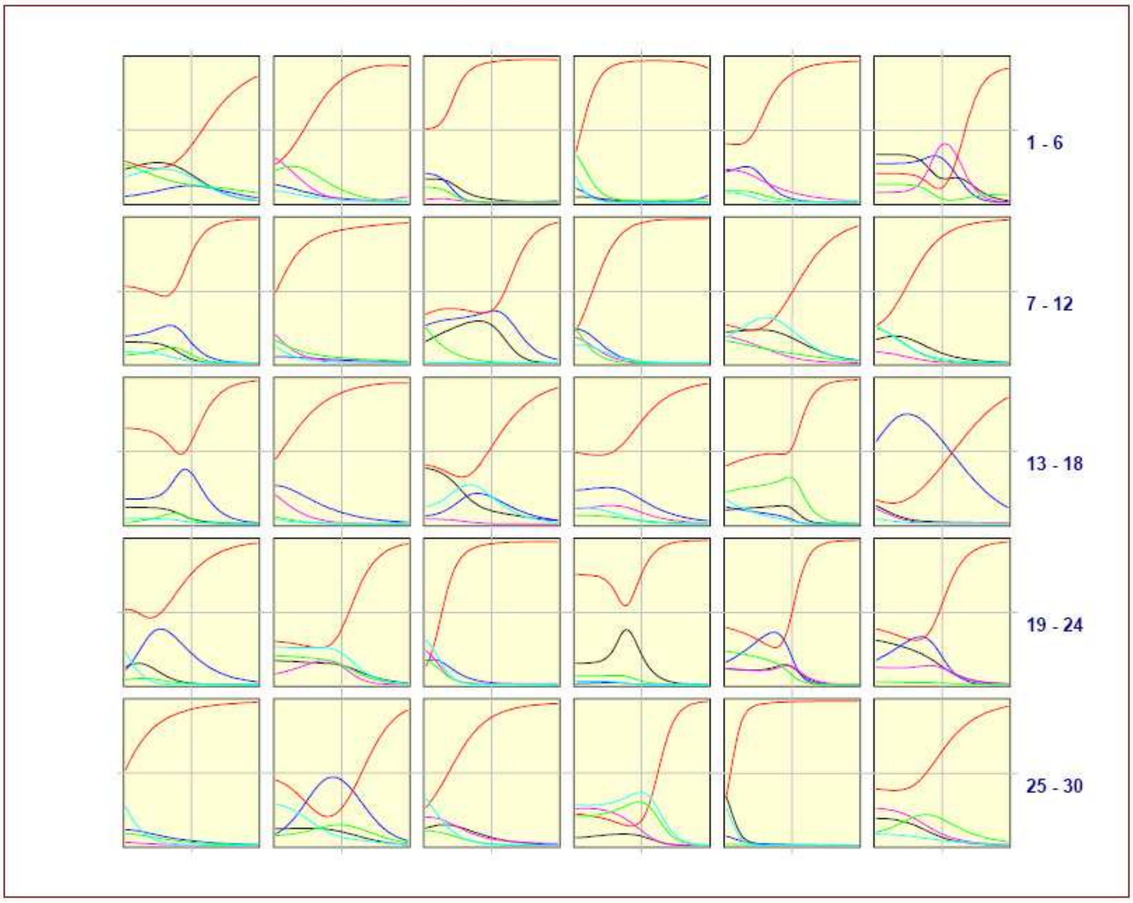


Figure B.1. Matrix plot of MCM response category curves for the 30 general culture items, calibrated for the complete dataset

B.2 Agriculture

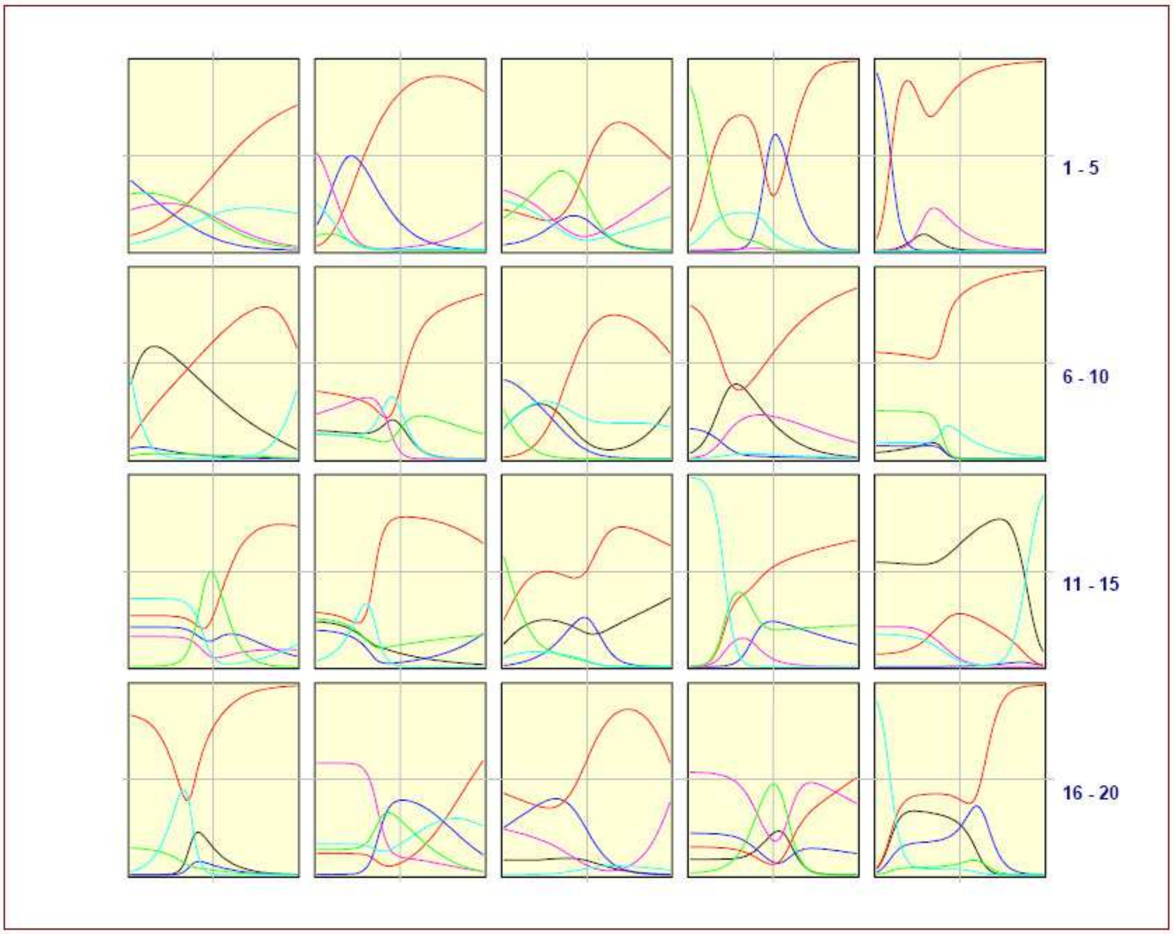


Figure B.2. Matrix plot of MCM response category curves for the Agriculture faculty, items 1-20

B.3 Arts and Humanities

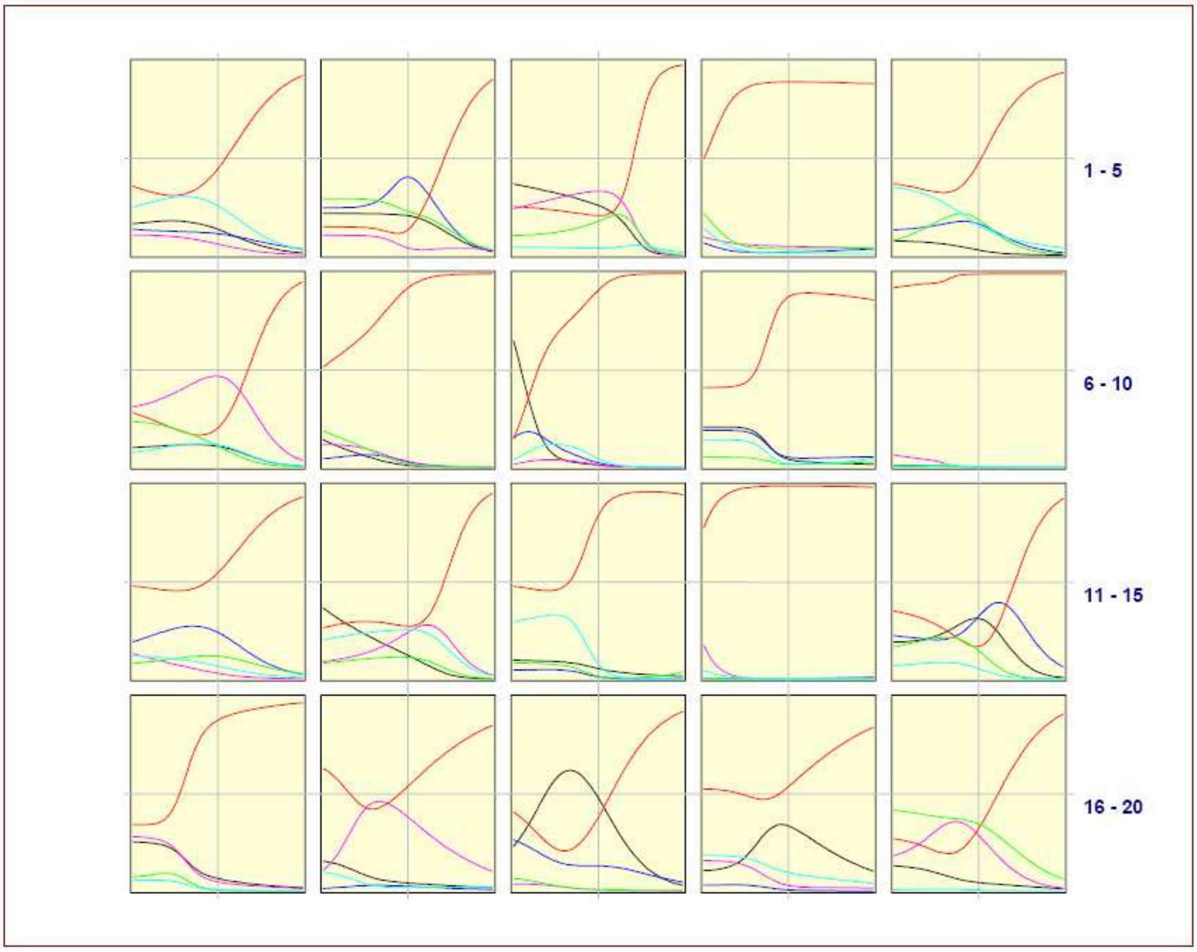


Figure B.3. Matrix plot of MCM response category curves for the Arts and Humanities faculty, items 1-20

B.4 Economics

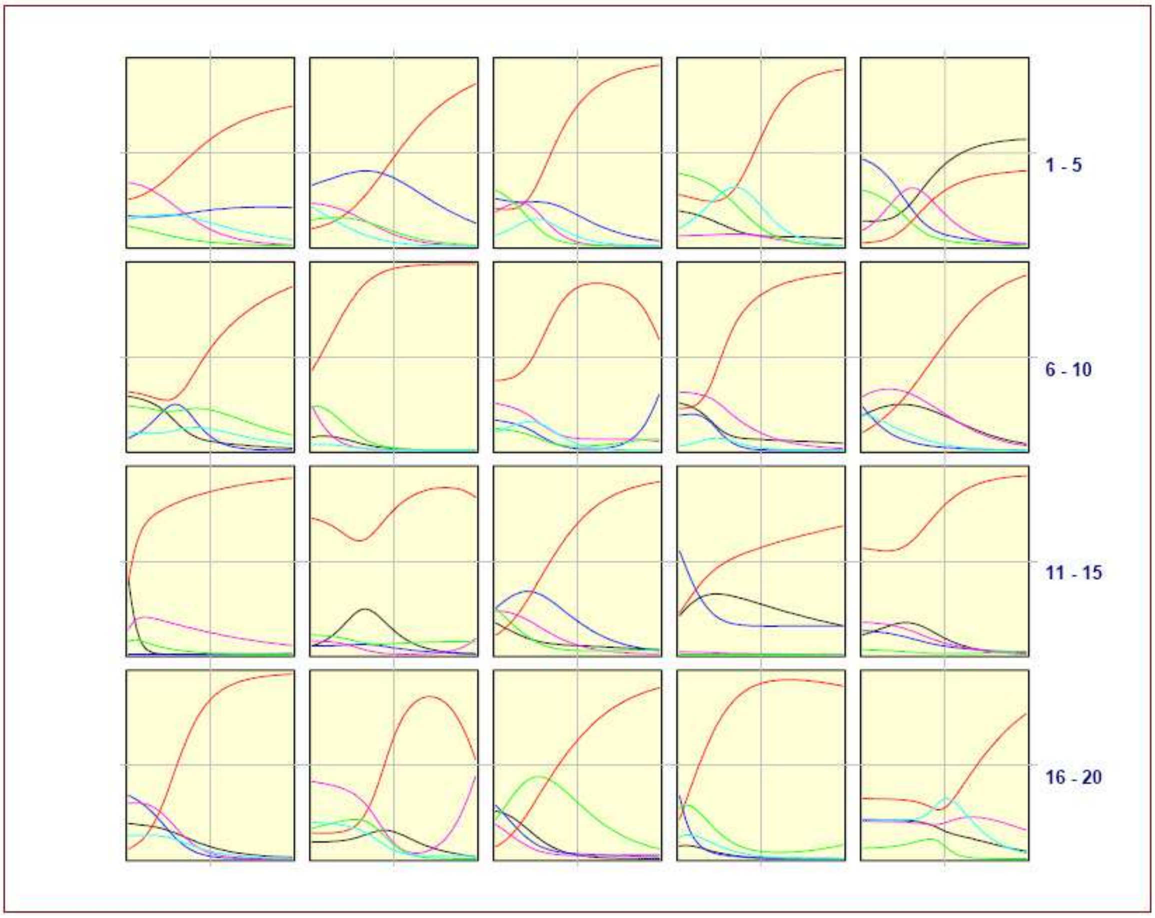


Figure B.4. Matrix plot of MCM response category curves for the Economics faculty, items 1-20

B.5 Education Science

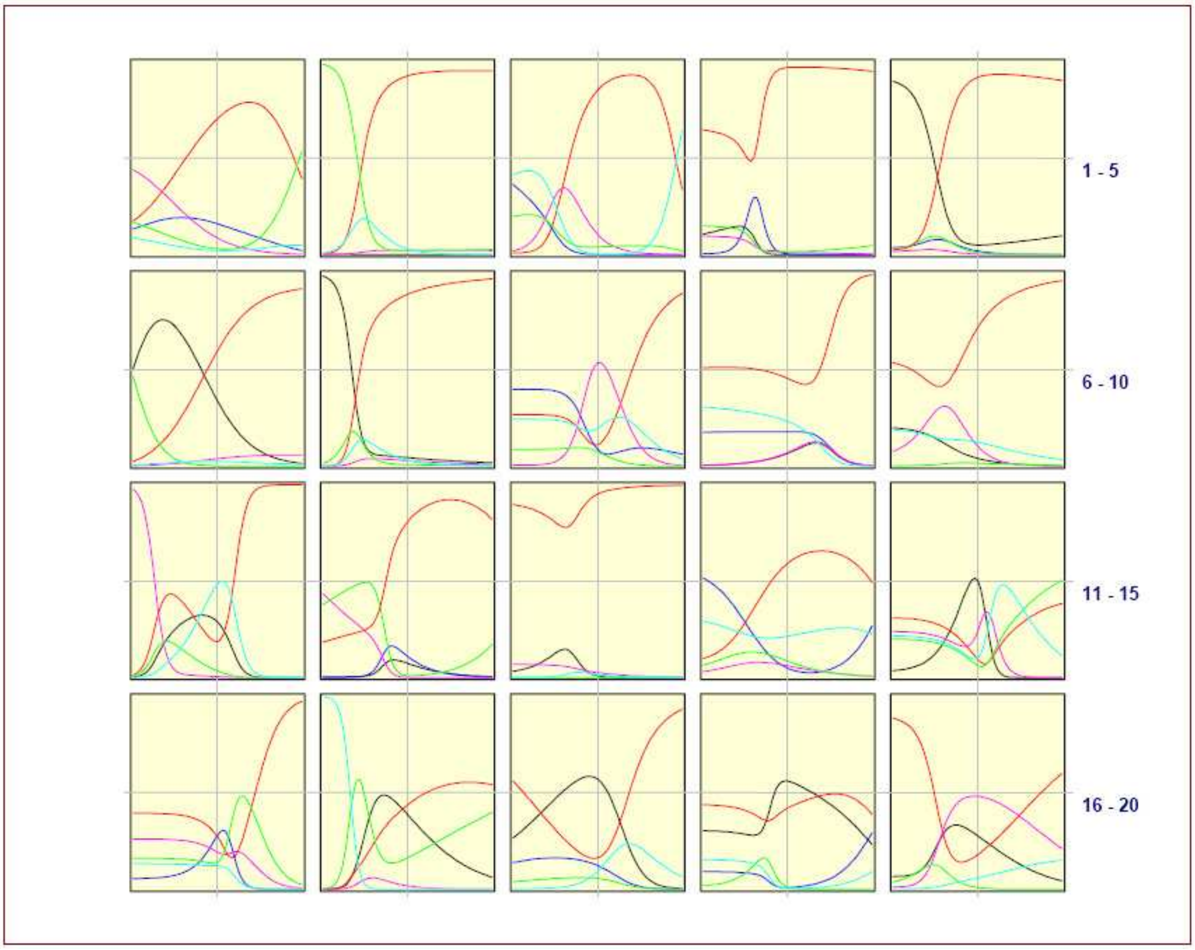


Figure B.5. Matrix plot of MCM response category curves for the Education Science faculty, items 1-20

B.6 Foreign Languages and Literature

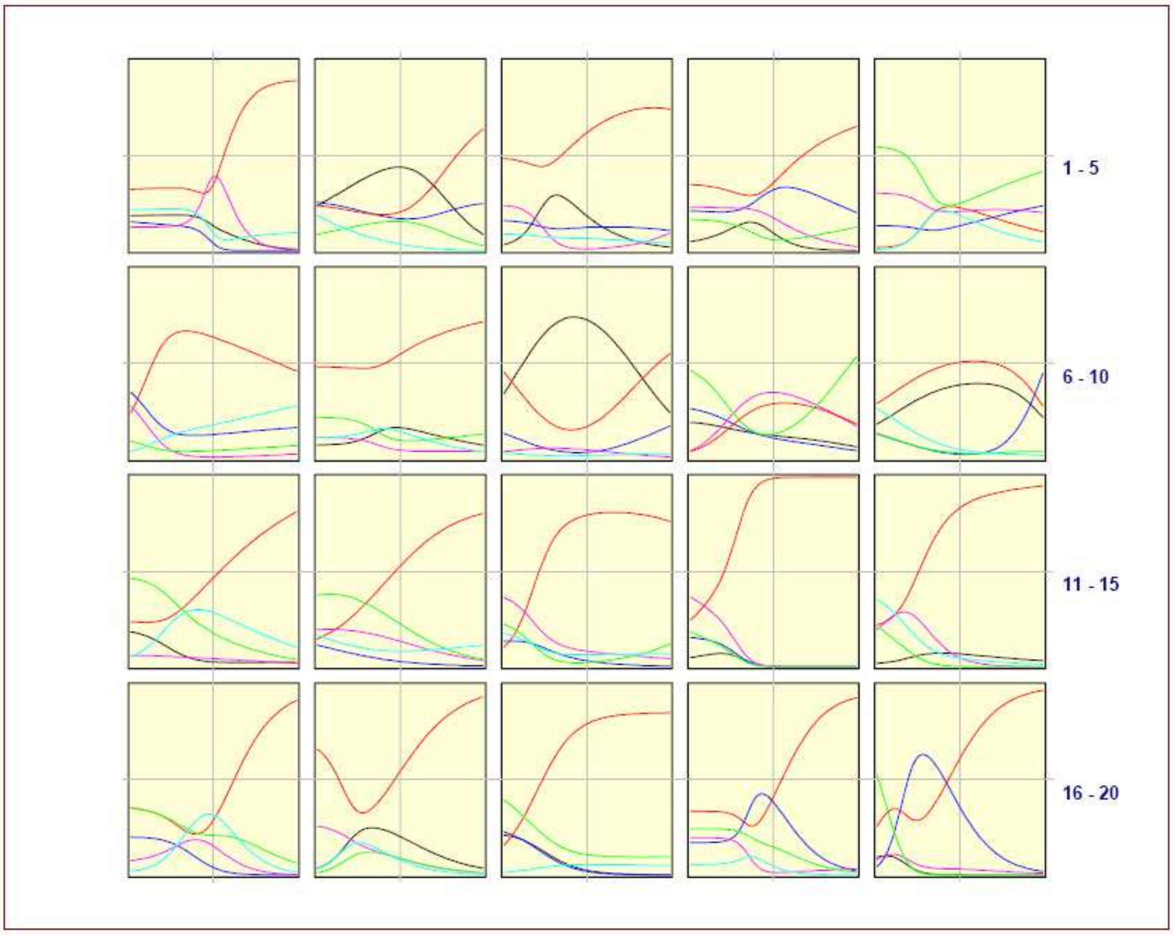


Figure B.6. Matrix plot of MCM response category curves for the Foreign Languages and Literature faculty, items 1-20

B.7 Pharmacy

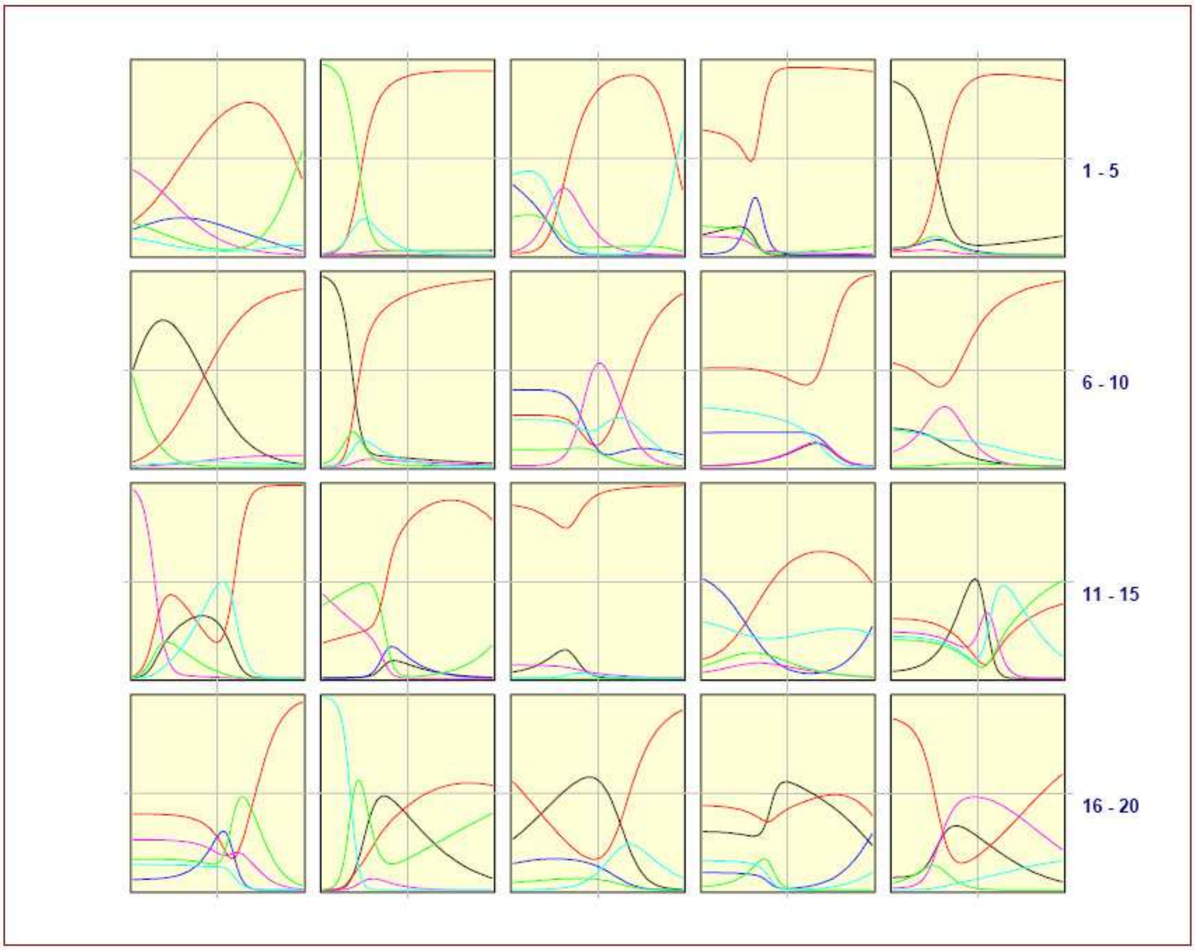


Figure B.7. Matrix plot of MCM response category curves for the Pharmacy faculty, items 1-20

B.8 Political Science

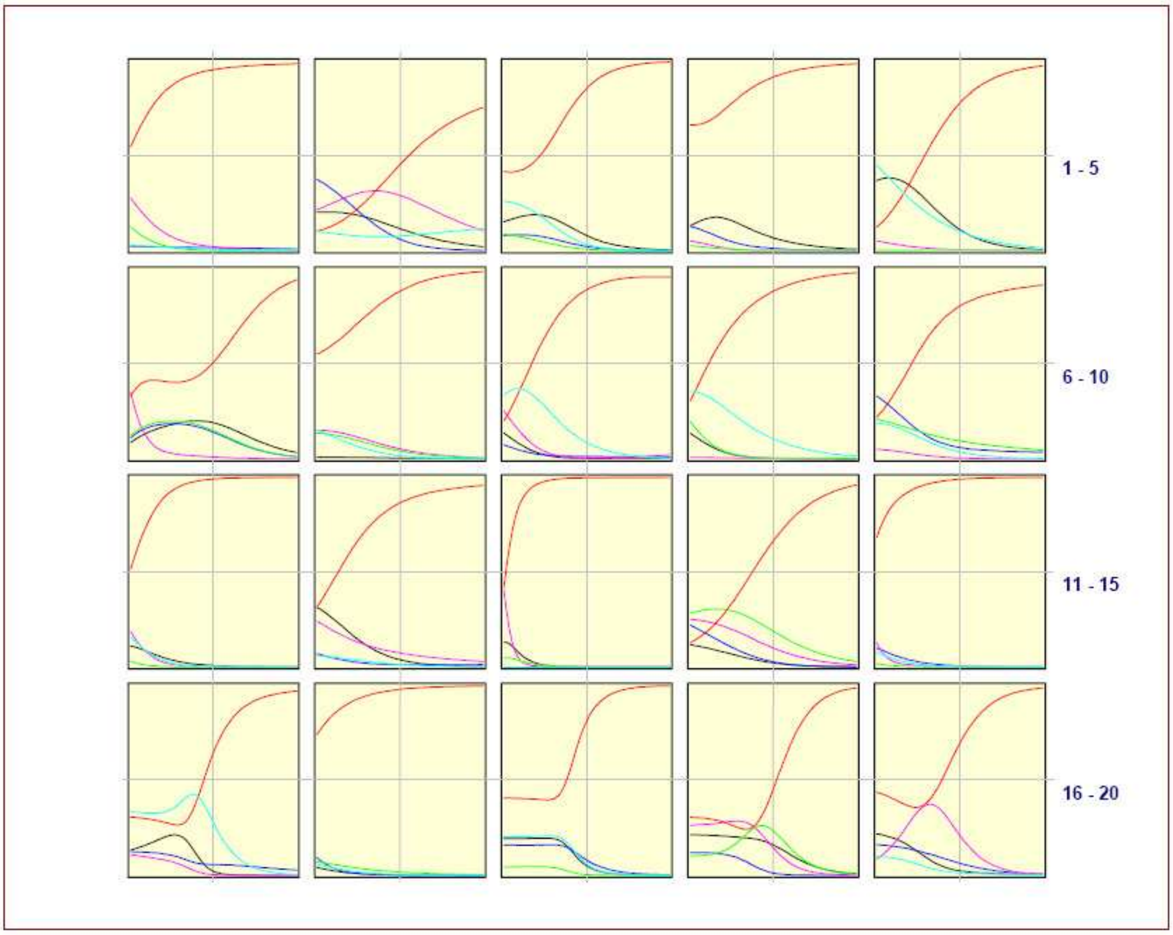


Figure B.8. Matrix plot of MCM response category curves for the Political Science faculty, items 1-20

B.9 Statistics

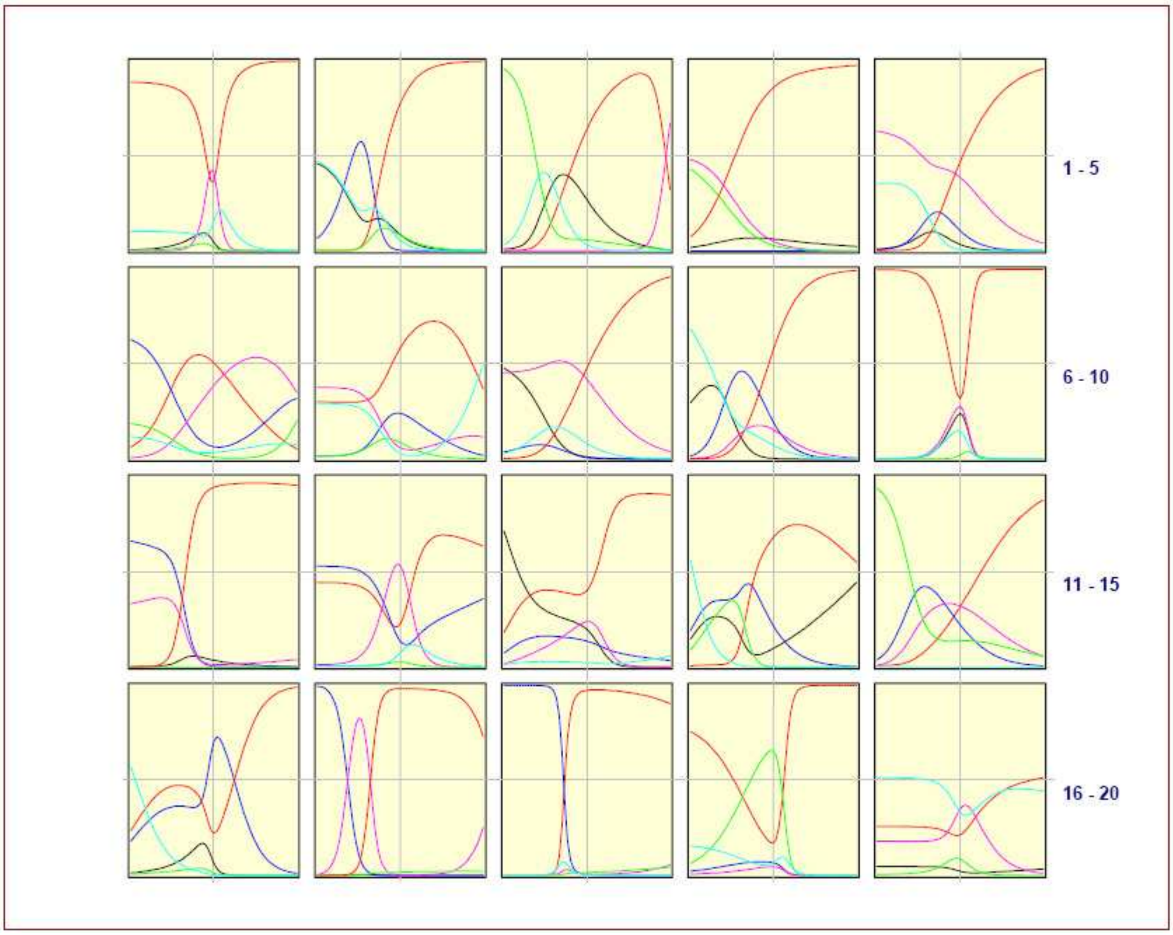


Figure B.9. Matrix plot of MCM response category curves for the Statistics faculty, items 1-20

Appendix C

Item parameter estimates for the Three-parameter logistic model

C.1 General culture items

Item	α	β	γ	Item	α	β	γ
1	0.66	0.96	0.19	16	0.66	-0.16	0.37
2	0.58	-1.51	0.15	17	1.56	0.44	0.45
3	0.77	-2.49	0.20	18	0.61	0.86	0.14
4	0.52	-3.92	0.19	19	0.55	-0.94	0.20
5	0.63	-1.65	0.16	20	1.07	0.65	0.22
6	2.13	1.31	0.16	21	0.70	-2.75	0.16
7	1.65	0.11	0.50	22	2.41	0.25	0.62
8	0.38	-3.48	0.18	23	1.97	0.15	0.23
9	1.88	1.18	0.37	24	1.13	0.18	0.25
10	1.20	-1.89	0.16	25	0.49	-3.07	0.21
11	0.74	0.39	0.16	26	1.79	1.29	0.25
12	0.68	-1.44	0.19	27	0.59	-1.81	0.15
13	1.51	0.74	0.53	28	1.50	0.92	0.14
14	0.46	-2.51	0.23	29	0.89	-3.45	0.18
15	0.92	0.71	0.34	30	0.56	-0.34	0.18

Table C.1. Item parameter estimates for the 3PL model, general culture items, calibrated on the complete dataset

C.2 Agriculture

Item	α	β	γ	Item	α	β	γ
1	0.45	1.03	0.15	11	3.05	0.95	0.28
2	0.52	-1.45	0.18	12	0.19	-1.96	0.19
3	0.77	0.47	0.13	13	0.31	0.14	0.20
4	0.30	0.20	0.24	14	0.35	0.61	0.17
5	0.82	-1.58	0.22	15	0.55	4.86	0.21
6	0.19	-0.47	0.20	16	0.77	-0.99	0.19
7	0.86	0.57	0.20	17	2.03	2.75	0.12
8	0.62	-0.22	0.12	18	0.68	-0.21	0.20
9	0.59	0.46	0.31	19	4.21	1.80	0.13
10	1.00	-1.07	0.20	20	1.52	0.95	0.35

Table C.2. Item parameter estimates for the 3PL model, Agriculture specific items

C.3 Arts and Humanities

Item	α	β	γ	Item	α	β	γ
1	1.17	1.13	0.35	11	0.93	1.09	0.45
2	1.19	1.48	0.12	12	1.95	1.62	0.29
3	3.13	1.40	0.22	13	0.84	-0.92	0.17
4	0.09	-11.33	0.20	14	0.47	-6.51	0.20
5	0.83	0.60	0.26	15	2.04	1.59	0.22
6	1.26	1.30	0.17	16	0.61	-1.45	0.16
7	1.05	-1.19	0.43	17	0.36	0.69	0.26
8	0.91	-1.56	0.17	18	0.90	0.91	0.21
9	0.49	-1.50	0.13	19	0.76	1.56	0.47
10	1.10	-3.21	0.20	20	0.90	1.11	0.19

Table C.3. Item parameter estimates for the 3PL model, Arts and Humanities specific items

C.4 Economics

Item	α	β	γ	Item	α	β	γ
1	0.30	0.16	0.15	11	0.26	-3.71	0.19
2	0.55	0.53	0.13	12	0.35	-1.64	0.18
3	0.81	-0.43	0.24	13	0.52	-0.60	0.15
4	0.68	0.05	0.17	14	0.97	1.59	0.50
5	0.40	2.14	0.08	15	0.52	-1.18	0.25
6	0.49	0.41	0.19	16	1.06	-0.85	0.19
7	0.85	-2.41	0.18	17	0.94	-0.15	0.10
8	0.45	-1.75	0.16	18	0.55	-0.25	0.14
9	0.59	-1.32	0.14	19	0.63	-2.24	0.18
10	0.53	-0.02	0.13	20	1.09	1.70	0.28

Table C.4. Item parameter estimates for the 3PL model, Economics specific items

C.5 Education Science

Item	α	β	γ	Item	α	β	γ
1	0.31	-0.93	0.19	11	2.84	0.87	0.29
2	0.92	-1.74	0.18	12	0.59	-1.20	0.16
3	0.56	-1.48	0.15	13	0.35	-4.29	0.20
4	0.59	-2.74	0.20	14	0.20	0.30	0.19
5	0.55	-2.08	0.18	15	3.13	2.13	0.15
6	0.52	-0.12	0.19	16	1.71	1.76	0.26
7	0.48	-2.36	0.20	17	0.34	1.26	0.15
8	1.55	1.42	0.16	18	1.47	1.40	0.19
9	0.31	1.23	0.21	19	0.58	1.94	0.31
10	0.38	-1.05	0.20	20	2.94	1.94	0.22

Table C.5. Item parameter estimates for the 3PL model, Education Science specific items

C.6 Foreign Languages and Literature

Item	α	β	γ	Item	α	β	γ
1	0.80	0.72	0.23	11	0.50	0.71	0.16
2	3.01	2.05	0.23	12	0.69	0.94	0.32
3	0.34	0.02	0.22	13	0.26	-2.01	0.20
4	0.68	1.77	0.27	14	2.51	-1.48	0.16
5	3.26	2.36	0.18	15	0.53	-1.44	0.14
6	-0.11	0.60	0.20	16	0.78	1.11	0.18
7	0.23	0.75	0.22	17	0.56	0.38	0.24
8	1.70	2.14	0.19	18	0.34	-1.58	0.17
9	2.39	2.18	0.24	19	0.85	0.65	0.20
10	1.48	2.34	0.47	20	0.63	-0.18	0.15

Table C.6. Item parameter estimates for the 3PL model, Foreign Languages and Literature specific items

C.7 Pharmacy

Item	α	β	γ	Item	α	β	γ
1	1.01	-0.37	0.25	11	1.46	1.43	0.27
2	0.82	-1.48	0.16	12	0.81	0.58	0.21
3	0.69	0.86	0.11	13	1.00	-0.45	0.22
4	0.73	-0.13	0.14	14	1.06	1.00	0.14
5	0.86	-0.76	0.17	15	0.91	1.56	0.15
6	0.65	0.37	0.15	16	0.38	-2.53	0.21
7	0.31	0.13	0.15	17	0.86	-0.04	0.29
8	0.90	0.12	0.23	18	0.80	-2.64	0.19
9	0.94	-1.31	0.20	19	0.99	0.09	0.31
10	0.71	0.52	0.13	20	0.36	-2.35	0.21

Table C.7. Item parameter estimates for the 3PL model, Pharmacy specific items

C.8 Political Science

Item	α	β	γ	Item	α	β	γ
1	0.44	-3.69	0.20	11	0.84	-3.11	0.19
2	0.44	0.70	0.10	12	0.46	-2.07	0.16
3	0.76	-0.99	0.32	13	1.15	-3.21	0.18
4	0.41	-2.99	0.20	14	0.64	-0.32	0.16
5	0.58	-1.08	0.10	15	0.74	-3.57	0.18
6	0.75	0.88	0.35	16	0.94	-0.05	0.20
7	0.48	-2.10	0.25	17	0.47	-4.42	0.20
8	0.55	-1.92	0.16	18	1.23	-0.35	0.34
9	0.48	-2.22	0.15	19	1.07	0.43	0.20
10	0.42	-1.54	0.15	20	0.86	-0.15	0.29

Table C.8. Item parameter estimates for the 3PL model, Political Science specific items

C.9 Statistics

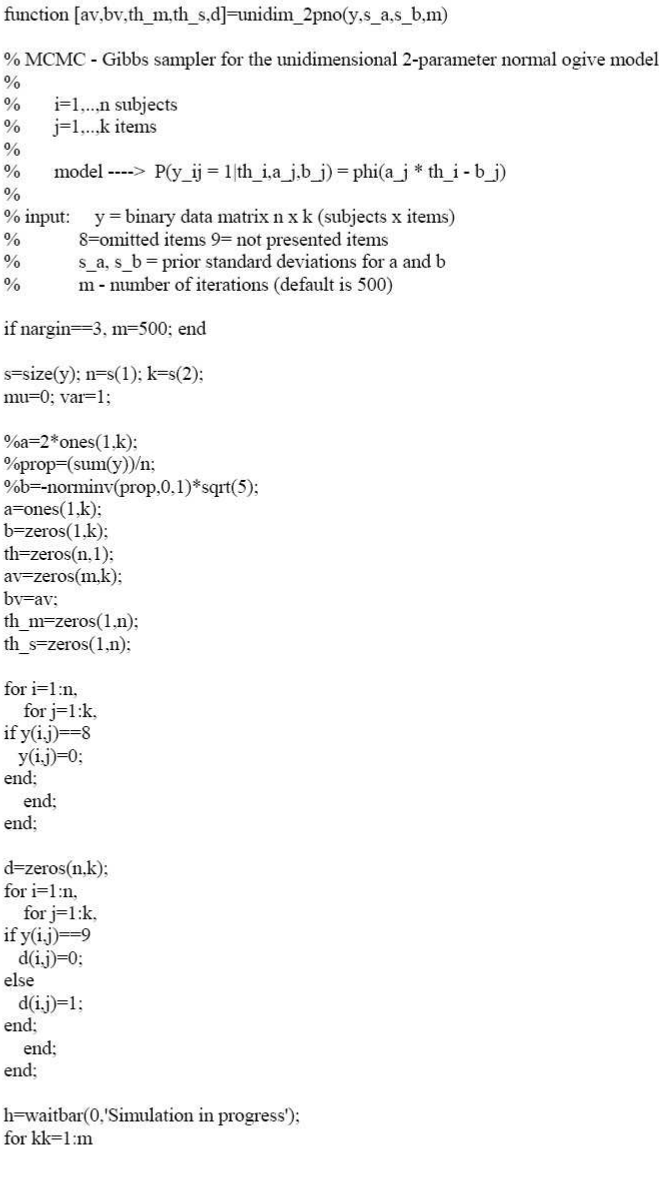
Item	α	β	γ	Item	α	β	γ
1	0,73	-1,10	0,22	11	0,77	-2,18	0,16
2	1,17	-0,90	0,15	12	0,81	0,85	0,23
3	0,57	-0,72	0,14	13	1,88	0,59	0,41
4	0,53	-2,03	0,20	14	0,37	-0,60	0,18
5	0,76	0,03	0,13	15	2,55	0,87	0,26
6	-0,29	-1,84	0,16	16	0,57	0,41	0,15
7	0,40	-0,10	0,18	17	0,54	-3,44	0,19
8	0,78	-0,01	0,13	18	0,70	-2,48	0,18
9	0,96	-0,33	0,17	19	6,69	0,02	0,27
10	0,83	-0,82	0,36	20	0,71	1,80	0,19

Table C.9. Item parameter estimates for the 3PL model, Statistics specific items

Appendix D

Gibbs sampler algorithm syntax

D.1 Unidimensional model

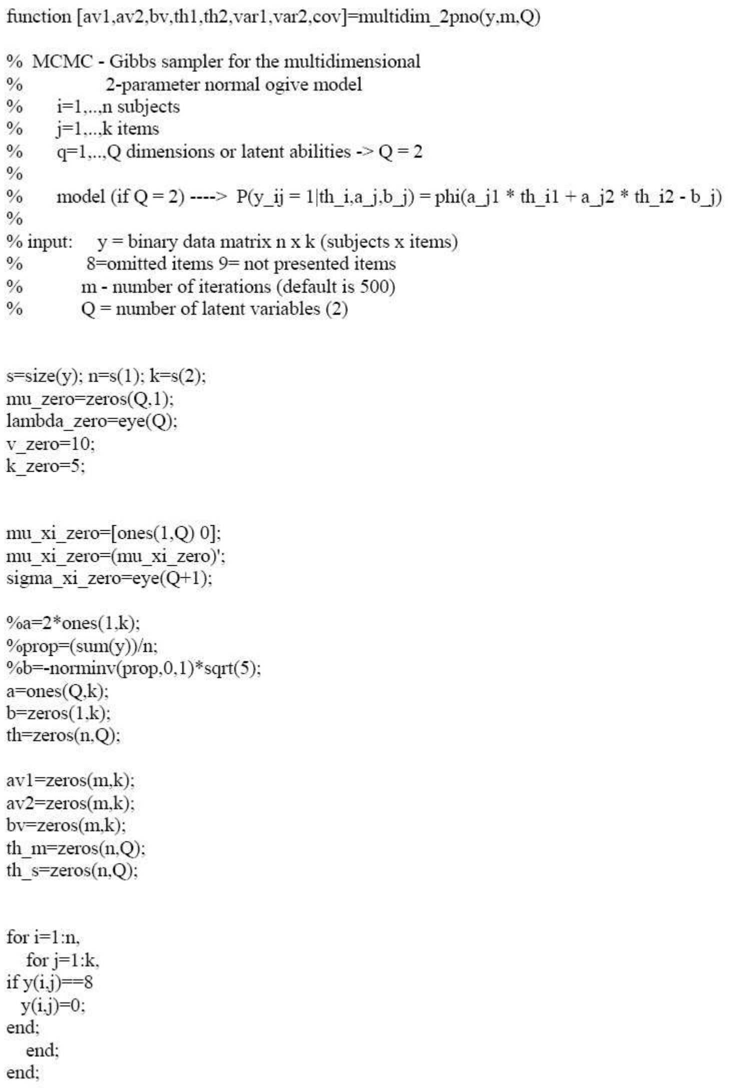


```
function [av,bv,th_m,th_s,d]=unidim_2pno(y,s_a,s_b,m)

% MCMC - Gibbs sampler for the unidimensional 2-parameter normal ogive model
%
%       i=1,..,n subjects
%       j=1,...,k items
%
%       model ----> P(y_ij = 1|th_i,a_j,b_j) = phi(a_j * th_i - b_j)
%
% input:     y = binary data matrix n x k (subjects x items)
%            8=omitted items 9= not presented items
%            s_a, s_b = prior standard deviations for a and b
%            m - number of iterations (default is 500)

if nargin==3, m=500; end

s=size(y); n=s(1); k=s(2);
mu=0; var=1;

%a=2*ones(1,k);
%prop=(sum(y))/n;
%b=-norminv(prop,0,1)*sqrt(5);
a=ones(1,k);
b=zeros(1,k);
th=zeros(n,1);
av=zeros(m,k);
bv=av;
th_m=zeros(1,n);
th_s=zeros(1,n);

for i=1:n,
   for j=1:k,
if y(i,j)==8
  y(i,j)=0;
end;
   end;
end;

d=zeros(n,k);
for i=1:n,
   for j=1:k,
if y(i,j)==9
  d(i,j)=0;
else
  d(i,j)=1;
end;
   end;
end;

h=waitbar(0,'Simulation in progress');
for kk=1:m
```

```
      % SIMULATE ZETA

         lp=th*a-ones(n,1)*b;
         bb=normcdf(-lp,0,1)   ;
         u=rand(n,k);
         tt=(bb.*(1-y)+(1-bb).*y).*u+bb.*y;
         z=norminv(tt,0,1)+lp;


   for i=1:n,
      for j=1:k,
if d(i,j)==0
  z(i,j)=0;
end;
   end;
end;


     % SIMULATE THETA

         v=1/sum(a.^2);
         pvar=1/(1/v+1/var);
         mn=sum(((ones(n,1)*a.*d).*(z+ones(n,1)*b))')';
         pmean=(mn+mu/var)*pvar;
         th=randn(n,1)*sqrt(pvar)+pmean;


   th_mean=mean(th);
   th_sd=std(th,1); % not correct standard deviation
   th=(th-th_mean)/th_sd;

     % SIMULATE ITEM PARAMETERS

         x=[th -ones(n,1)];
  pp=[1/s_a^2 0;0 1/s_b^2];
         amat=chol(inv(x'*x+pp));
         bz=(x'*x+pp)\(x'*z);
  beta=amat'*randn(2,k)+bz;
  a=beta(1,:); b=beta(2,:);

         av(kk,:)=a;
         bv(kk,:)=b;
         th_m=th_m+th';
         th_s=th_s+th'.^2;
   z=z;

  waitbar(kk/m)

end
```

```
close(h)

th_m=th_m/m;
th_s=sqrt(th_s/m-th_m.^2);


t='1';
for i=2:k
t=str2mat(t,num2str(i));
end
figure
bm=mean(bv); am=mean(av);
br=max(bm)-min(bm); ar=max(am)-min(am);
ax=[min(bm)-.1*br max(bm)+.1*br min(am)-.1*ar max(am)+.1*ar];
text(mean(bv),mean(av),t)
axis(ax)
xlabel('DIFFICULTY');ylabel('DISCRIMINATION')
```

D.2 Two-dimensional model



```
function [av1,av2,bv,th1,th2,var1,var2,cov]=multidim_2pno(y,m,Q)

%  MCMC - Gibbs sampler for the multidimensional
%              2-parameter normal ogive model
%      i=1,..,n subjects
%      j=1,...,k items
%      q=1,...,Q dimensions or latent abilities -> Q = 2
%
%      model (if Q = 2) ----> P(y_ij = 1|th_i,a_j,b_j) = phi(a_j1 * th_i1 + a_j2 * th_i2 - b_j)
%
% input:     y = binary data matrix n x k (subjects x items)
%           8=omitted items 9= not presented items
%          m - number of iterations (default is 500)
%          Q = number of latent variables (2)


s=size(y); n=s(1); k=s(2);
mu_zero=zeros(Q,1);
lambda_zero=eye(Q);
v_zero=10;
k_zero=5;


mu_xi_zero=[ones(1,Q) 0];
mu_xi_zero=(mu_xi_zero)';
sigma_xi_zero=eye(Q+1);

%a=2*ones(1,k);
%prop=(sum(y))/n;
%b=-norminv(prop,0,1)*sqrt(5);
a=ones(Q,k);
b=zeros(1,k);
th=zeros(n,Q);

av1=zeros(m,k);
av2=zeros(m,k);
bv=zeros(m,k);
th_m=zeros(n,Q);
th_s=zeros(n,Q);


for i=1:n,
   for j=1:k,
if y(i,j)==8
  y(i,j)=0;
end;
   end;
end;
```

```
d=zeros(n,k);
for i=1:n,
    for j=1:k,
if y(i,j)==9
   d(i,j)=0;
else
   d(i,j)=1;
end;
    end;
end;

h=waitbar(0,'Simulation in progress');

for kk=1:m

    % SIMULATE THE MEAN VECTOR AND THE VAR-COV MATRIX OF THETA

    theta_mean_row=mean(th);
    theta_mean=theta_mean_row';
    v_n=v_zero+n;

 thi1=th(:,1);
 thi2=th(:,2);
 squarei1=power(thi1,2);
 squarei2=power(thi2,2);
 sommai1=sum(squarei1);
 sommai2=sum(squarei2);
 prod_th=thi1.*thi2;
 sommai12=sum(prod_th);
 th_mean1=theta_mean(1);
 th_mean2=theta_mean(2);
 S(1,1)=sommai1-n*(th_mean1)^2;
 S(1,2)=sommai12-n*(th_mean1*th_mean2);
 S(2,1)=S(1,2);
 S(2,2)=sommai2-n*(th_mean2)^2;

    lambda_n=lambda_zero+S+(k_zero*n/(k_zero+n))*(theta_mean-mu_zero)*(theta_mean-
mu_zero)';
    sigma_theta=iwishrnd(lambda_n,v_n);

    mu_n=(k_zero/(k_zero+n))*mu_zero+(n/(k_zero+n))*theta_mean;
    k_n=k_zero+n;
    sigma_n=sigma_theta/k_n;

    mu=mu_n+sigma_n*randn(Q,1);
```

```
     % SIMULATE ZETA

      lp=th*a-ones(n,1)*b;
      bb=normcdf(-lp,0,1)   ;
      u=rand(n,k);
      tt=(bb.*(1-y)+(1-bb).*y).*u+bb.*y;
      z=norminv(tt,0,1)+lp;

 for i=1:n,
    for j=1:k,
if d(i,j)==0
 z(i,j)=0;
end;
  end;
end;

    % SIMULATE THETA

  z_t=z';
      L=chol(sigma_theta);
  A=a';
  B=A*L;
  beta=b';

  d_t=d';
  theta_zero_hat=inv(B'*B)*(B')*(z_t+(d_t.*((beta-A*mu)*ones(1,n))));
  I=eye(Q);

  th_zero_mean=(I+(B'*B))\(B'*B)*theta_zero_hat;
  th_zero_var=inv(I+(B'*B));
  chol_th_zero_var=chol(th_zero_var);
  theta_zero=th_zero_mean+chol_th_zero_var*randn(Q,n);
  theta=mu*ones(1,n)+L*theta_zero;
  th=theta';

  th_mean=mean(th);
  th(:,1)=th(:,1)-th_mean(1,1);
  th(:,2)=th(:,2)-th_mean(1,2);

 %%%%%%%%%%%  simulation for the single observations
  % for i=1:n
  %  th_i=th(i,:);
  %  th_i=(th_i)';
  %  z_i=z(i,:);
  %  z_i=(z_i)';

  %theta_zero_i_hat=(B'*B)\B'(z_i+beta-A*mu);

  %th_zero_mean=(I+(B'*B))\(B'*B)*theta_zero_i_hat;
  %th_zero_var=inv(I+(B'*B));
  %chol_th_zero_var=chol(th_zero_var);
```

```
   %theta_zero_i=th_zero_mean+chol_th_zero_var*randn(Q);
   %theta_i=mu+L*theta_zero_i;
   %end



    % SIMULATE ITEM PARAMETERS

    x=[th -ones(n,1)];
    mu_xi=inv((inv(sigma_xi_zero)+x'*x))*(((inv(sigma_xi_zero))*mu_xi_zero)*ones(1,k)+x'*z);
    var_xi=inv(inv(sigma_xi_zero)+x'*x);
    C=chol(var_xi);
    xi=mu_xi+C*randn(Q+1,k);
    a=xi(1:Q,:);
    b=xi(Q+1,:);
    a1=xi(1,:);
    a2=xi(2,:);


    a1=(a1-a1(1,31))/(a1(1,1)-a1(1,31));
    a2=(a2-a2(1,1))/(a2(1,31)-a2(1,1));
    xi(1,:)=a1;
    xi(2,:)=a2;


   av1(kk,:)=a1;
   av2(kk,:)=a2;
   bv(kk,:)=b;

    th1(kk,:)=th(:,1)';
    th2(kk,:)=th(:,2)';

var_theta_one=sigma_theta(1,1); % variance of theta1
var_theta_two=sigma_theta(2,2); % variance of theta2
cov_theta=sigma_theta(1,2) % covariance of theta1,2

var1(kk,:)=var_theta_one;
var2(kk,:)=var_theta_two;
cov(kk,:)=cov_theta;

  waitbar(kk/m)

end

close(h)
```

Appendix E

Guidance questionnaires

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

SEQUENCE OF GENERAL CULTURE ITEMS

1) Quando è stato firmato il Trattato di Roma che ha dato vita alla Comunità Economica Europea?

a) 1991
b) 1950
c) 1957*
d) 1960
e) 1992

2) Il Regno Unito (UE=Unione Europea; UME=Unione Monetaria Europea)

a) Appartiene all'UE ma non all'UME*
b) Appartiene all'UME ma non all'UE
c) Appartiene ad entrambe
d) Non appartiene a nessuna delle due
e) Appartiene all'UME con possibilità di veto

3) Quale tra queste regioni italiane è a Statuto Speciale?

a) Veneto
b) Lombardia
c) Liguria
d) Basilicata
e) Valle d'Aosta*

4) In Italia la Festa della Repubblica si celebra il

a) 4 novembre
b) 20 settembre
c) 2 giugno*
d) 24 maggio
e) 9 febbraio

5) Quale tra queste nazioni fu una colonia italiana?

a) Eritrea*
b) Egitto
c) Algeria
d) Marocco
e) Tanzania

6) Di che nazionalità è lo scrittore e Premio Nobel per la Letteratura Gabriel García Márquez?

a) Messicano
b) Venezuelano
c) Cileno
d) Brasiliano
e) Colombiano*

7) A quale delle seguenti cifre si avvicina maggiormente la popolazione italiana attuale?

a) 100 milioni
b) 80 milioni
c) 60 milioni *
d) 30 milioni
e) 10 milioni

8) Un file di word ha estensione

a) doc*
b) xls
c) ord
d) wor
e) mac

9) Galileo Galilei è vissuto prevalentemente nel secolo:

a) ‘400;
b) ‘500;
c) ‘600;*
d) ‘800;
e) ‘900.

10) Quale di questi Paesi fa parte della penisola iberica?

a) Portogallo*
b) Grecia
c) Bretagna
d) Lussemburgo
e) Paesi Bassi

11) Cosa si intende con il termine inglese impeachment?

a) voto di sfiducia
b) messa in stato di accusa*
c) voto di fiducia
d) sostituzione anticipata dell’eletto da parte degli elettori
e) crisi di governo

12) Che cosa si intende per ONG?

a) Organismo Non modificato Geneticamente
b) Organizzazione Non Governativa*
c) Organizzazione Nazionale Genitori
d) Organizzazione Naturalisti e Geografi
e) Ordine Nazionale Geometri

13) Quando entrò in vigore la Costituzione italiana?

a) 1945
b) 1946
c) 1948*
d) 1949
e) 1950

14) La Camera dei Deputati e il Senato della Repubblica sono eletti per:

a) 5 anni*
b) 4 anni
c) 3 anni
d) 2 anni
e) 1 anno

15) Quale personaggio di Shakespeare viveva in Danimarca?

a) Otello
b) Re Lear
c) Romeo
d) Amleto*
e) Macbeth

16) Pandora

a) Nella mitologia greca: la prima donna*
b) Nella mitologia mesopotamica: la prima donna vasaio
c) Nella mitologia romana: una divinità silvestre
d) Nella mitologia indiana: la donna che scoprì l'oro
e) La mitica fondatrice di Verona

17) Il Mali è

a) una moneta
b) un movimento di liberazione internazionale
c) un paese dell'Africa*
d) un arcipelago della Polinesia
e) un ballo folcloristico

18) La scala Mercalli:

a) Misura il rischio sismico nelle varie regioni d'Italia
b) Misura l'intensità dei terremoti
c) Misura le onde sonore
d) Misura gli effetti dei terremoti*
e) Misura l'indice della Borsa Valori di Milano

19) Il PIL è:

a) il valore del debito estero di un paese
b) il valore della produzione internazionale di un paese
c) il valore dell'insieme dei beni e servizi prodotti in un paese*
d) il valore della produzione agricola in un paese
e) il valore del costo della mano d'opera

20) Carlo Rubbia è premio Nobel di:

a) medicina
b) fisica*
c) letteratura
d) chimica
e) economia

21) L’etimologia è:

a) la disciplina che studia la derivazione e la formazione delle parole*
b) la disciplina che studia la derivazione e la formazione degli stili letterari
c) la disciplina che studia la retorica dei testi linguistici
d) la disciplina che studia la retorica dei testi letterari
e) la disciplina che studia le abitudini degli animali

22) Chi è Nelson Mandela?

a) l'ex Segretario dell'ONU
b) un cantante di musica etnica
c) un uomo politico sudafricano liberato da una lunga prigionia nel 1990*
d) un poeta di colore sostenitore dell'apartheid
e) un calciatore brasiliano

23) In quale città si assegna il Premio Nobel?

a) Strasburgo
b) Bruxelles
c) Copenhagen
d) Parigi
e) Stoccolma*

24) Normalmente viene chiamato "quarto potere":

a) l'insieme delle sentenze della magistratura
b) la magistratura militare

c) la polizia
d) la scuola
e) la stampa*

25) Il Parlamento italiano è:

a) Bicamerale*
b) Bipolare
c) Bipartisan
d) Unicamerale
e) Federale

26) Che cosa è un elzeviro?

a) Un tessuto
b) Un copricapo
c) Un articolo culturale*
d) Un componimento poetico
e) Una pianta

27) Gregor Mendel è noto per:

a) il secondo principio della termodinamica
b) la teoria della trasmissione genetica dei caratteri*
c) la teoria dell'evoluzione per selezione naturale
d) la teoria della relatività generale
e) la legge di gravitazione universale

28) Quale delle seguenti città non è sul mare?

a) Sidney
b) La Paz*
c) New York
d) Turku
e) Bergen

29) Il DNA descrive:

a) il processo di duplicazione delle cellule di un essere vivente
b) il sistema nervoso centrale di un organismo vivente
c) il patrimonio genetico di un essere vivente*
d) il sistema nervoso periferico di un organismo vivente
e) il sistema immunitario di un organismo vivente

30) Nel sistema solare quale pianeta è più vicino al Sole?

a) Luna
b) Mercurio*
c) Giove
d) Venere
e) Saturno

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

AGRICULTURE

1) Le colture erbacee sono:

a) colture annuali o poliennali di consistenza non legnosa*
b) specie impiegabili per il miglioramento dei prati spontanei
c) colture (annuali o poliennali) di altezza inferiore a due metri
d) colture per l'alimentazione del bestiame
e) colture annuali di consistenza non legnosa

2) La temperatura dell'aria si misura con:

a) termometri*
b) barometri
c) igrometri
d) densimetri
e) freatimetri

3) Le colture da energia sono utilizzate per:

a) produrre carburanti*
b) migliorare l'ambiente
c) migliorare l'alimentazione del bestiame
d) migliorare l'efficienza della lavorazione
e) la difesa fitopatologica delle piante

4) Un ambiente si dice che è umido se:

a) c'è alta umidità relativa dell'aria*
b) c'è alta umidità assoluta dell'aria
c) è freddo
d) la temperatura è alta
e) la pressione atmosferica è bassa

5) Il paesaggio è…

a) … ritenuto di pregio solo se non è stato modificato dall’uomo
b) … una realtà immutabile
c) … determinato dalle sole componenti naturali del territorio
d) … una determinata parte di territorio, così come è percepita dalle popolazioni, il cui carattere deriva dall’azione di fattori naturali e/o umani e dalle loro interrelazioni*
e) … studiato soprattutto dagli artisti

6) Per realizzare un giardino o un parco occorre conoscere soprattutto…

a) … il tipo di suolo e l’ubicazione dell’intervento
b) … le caratteristiche di alcune delle essenze vegetali del luogo
c) … le condizioni climatiche del luogo ed il tipo di suolo*
d) … se è necessario eseguire interventi di sistemazione del terreno
e) … il costo degli interventi necessari

7) Cos'è un disciplinare di lotta integrata?

a) una persona che coordina l'utilizzo dei principi attivi
b) l'elenco e le modalità di applicazione dei principi attivi permessi*
c) una legge che punisce gli agricoltori
d) un insetto che ristabilizza l'ecosistema agricolo
e) un ispettore fitosanitario

8) Cosa si intende per insetti pronubi?

a) insetti che non si accoppiano
b) insetti che portano le nubi
c) insetti che permettono la fecondazione delle piante*
d) insetti che pungono
e) insetti che si autofecondano

9) Cosa s'intende per marketing?

a) è un sinonimo di mercato
b) è una promozione commerciale fatta in tv
c) è una delle tipologie della grande distribuzione
d) è una funzione organizzata e un insieme di processi volti a creare comunicazione e a conseguire valori*
e) è l'incontro tra operatori di un comparto produttivo

10) Cos'è la filiera produttiva?

a) è la disposizione dei negozi all'interno di un centro commerciale
b) è un tipo di gioco a biliardo
c) è un insieme coordinato di operatori che si occupano di un medesimo comparto*
d) è un filato di seta
e) è una catena di negozi della medesima insegna

11) Cosa si intende per la granulometria del terreno?

a) proporzioni relative delle varie categorie di particelle costituenti il terreno*
b) modo di aggregazione delle particelle costituenti il terreno
c) grado di dispersione delle particelle costituenti il terreno
d) grado di disgregazione di un terreno
e) proprietà chimiche di granuli costituenti il terreno

12) L'acqua utilizzata per l'irrigazione può compromettere la qualità del suolo?

a) assolutamente no
b) sempre
c) sì, in relazione alle sostanze disciolte*
d) sì, in relazione alle tecniche di irrigazione
e) sì, in relazione al periodo dell'anno

13) Il drenaggio dei terreni coltivati ha lo scopo di ….

a) impedire i movimenti franosi
b) apportare acqua alle piante
c) creare il giusto equilibrio fra il contenuto di acqua e di aria nel terreno*
d) risparmiare acqua
e) apportare fertilizzanti

14) L'idrologia agraria è la scienza che studia:

a) il comportamento dell'acqua nel suolo*
b) i fenomeni di assorbimento dell'acqua nelle piante
c) gli aspetti economici dell'irrigazione
d) le reazioni chimiche tra l'acqua e le altre sostanze presenti nel terreno
e) la crescita dell'apparato radicale delle piante

15) L’enologo ha competenze nel settore:

a) viticolo
b) tecnologico
c) sensoriale
d) controllo e gestione dell’azienda vitivinicola*
e) equivale al sommelier

16) Cosa si intende per vino D.O.C.?

a) vino a denominazione di origine certificata
b) è un marchio di qualità europeo
c) vino a denominazione di origine controllata*
d) vino a denominazione di origine controllata e garantita
e) vino di qualità

17) Cos’è un antiossidante?

a) una sostanza che si ossida*
b) un additivo alimentare
c) una sostanza presente nei cibi biologici
d) un radicale libero
e) una vitamina

18) Che cosa si intende per zuccheri?

a) gli additivi usati per dolcificare gli alimenti
b) il glucosio
c) il saccarosio
d) composti mono e disaccaridi*
e) tutti i dolcificanti raffinati

19) Cos’è un prodotto di IV gamma?

a) un prodotto di elevata qualità
b) un prodotto di elevata tecnologia
c) un prodotto di scarto
d) un prodotto contenente un’elevata quantità di antiossidanti
e) un prodotto confezionato e pronto all’uso*

20) Cosa si intende per FDA?

a) Federazione Dietisti Alimentaristi
b) l’ente di controllo europeo sugli alimenti
c) Food and Drug Administration*
d) un marchio di qualità
e) un alimento per diabetici

ALMA MATER STUDIORUM - UNIVERSITA’ DI BOLOGNA

ARTS AND HUMANITIES

1) Quale tra queste città non è un porto del Mediterraneo ?

a) Marsiglia

b) Alessandria d'Egitto

c) Napoli

d) Cadice *

e) Patrasso

2) Quale dei seguenti è uno stato della Repubblica Federale Tedesca?

a) La Franconia

b) L'Alsazia

c) La Neckar

d) Il Kaiserslautern

e) La Turingia *

3) Con quale di questi paesi confina l'Egitto ?

a) Etiopia

b) Israele *

c) Tunisia

d) Giordania

e) Ciad

4) Le prime forme di scrittura sono attestate:

a) In Mesopotamia *

b) In India

c) In Cina

d) A Creta

e) In Messico

5) Quale di queste opere non è di Niccolò Machiavelli?

a) Principe

b) Istorie fiorentine

c) Ricordi *

d) Mandragola

e) Discorsi sopra la prima Deca di Tito Livio

6) Quale di queste opere non è stata scritta nel Trecento?

a) Il Filocolo
b) Orlando innamorato *
c) Africa
d) Decameron
e) I Trionfi

7) Il trattato *Dei delitti e delle pene* fu scritto da:

a) Marsilio Ficino
b) Campanella
c) Hobbes
d) Locke
e) Beccaria *

8) La Critica della ragion pura è un'opera filosofica di:

a) Bacone
b) Cartesio
c) Galilei
d) Kant *
e) Nietzsche

9) Che cos'è un format televisivo?

a) Uno standard relativo alla dimensione delle immagini
b) Uno standard relativo alla definizione delle immagini
c) La struttura di base di un programma televisivo *
d) Un reality show
e) Una tecnica di misurazione dell'ascolto

10) Che cosa è una "natura morta" in pittura?

a) Un dipinto non finito
b) Un genere pittorico dedicato alla resa di fiori, frutta, oggetti, ecc. *
c) La rappresentazione della morte
d) Un paesaggio con rovine
e) Un genere pittorico dedicato alla rappresentazione della morte di Cristo

11) Quale di questi personaggi è un noto regista teatrale?

a) Luca Ronconi *
b) Luca Rigoni
c) Luca Giurato
d) Franco Quadri
e) Franco Franchi

12) Che cos'è il Lied?

a) Il capo di un partito politico
b) Una composizione vocale *
c) Un brano d'opera
d) Un quartetto d'archi
e) Una danza popolare tedesca

13) Quale di questi titoli si riferisce a un'opera di Mozart?

a) Don Carlo
b) Don Pasquale
c) Don Giovanni *
d) Don Camillo
e) Don Chisciotte

14) Cosa sono i mass media?

a) I mezzi di comunicazione di massa come la radio, la televisione, ecc. *
b) I comportamenti tipici dell'uomo medio
c) Le medie d'ascolto di un programma radiofonico, televisivo, ecc.
d) I programmi televisivi di successo
e) Le grandi manifestazioni di massa

15) Che cos'è l'encausto?

a) Un discorso ironico e irridente
b) Una tecnica di incisione a caldo di tavole lignee
c) Una tecnica pittorica che usa colori sciolti nella cera *
d) L'acquisto inconsapevole di opere d'arte rubate
e) La scottatura provocata da colori a base di soda caustica

16) Cosa indica lo *share* di una trasmissione televisiva?

a) Il numero degli spettatori che hanno guardato la trasmissione
b) La percentuale d'ascolto riferita al numero complessivo degli spettatori *
c) Il gradimento della trasmissione
d) Il numero di interruzioni pubblicitarie
e) Il posizionamento nel palinsesto

17) Il verso “Le donne i cavalier l’arme gli amori” è:

a) un decasillabo
b) un novenario
c) un dodecasillabo
d) un endecasillabo *
e) un ottonario

18) Chi è il regista di Roma città aperta?

a) Federico Fellini
b) Vittorio De Sica
c) Martin Scorsese
d) Giuseppe Ferrara
e) Roberto Rossellini *

19) Che cosa è un “piano-sequenza”?

a) Un lungo movimento della macchina da presa cinematografica
b) Un tipo particolare di pianola meccanica
c) Un affresco composto di più elementi organizzati in sequenza
d) Una sequenza cinematografica girata senza stacchi di montaggio *
e) Un movimento della macchina da presa realizzato grazie all'uso del dolly

20) Di cosa si occupa la semiotica?

a) Delle tecniche diagnostiche
b) Dei sistemi di segni *
c) Dei linguaggi non verbali
d) Dell'interpretazione dei simboli
e) Della segnaletica stradale

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

ECONOMICS

1) Se il prezzo di Forlilat passasse da 100 a 200 il rendimento del titolo risulterebbe pari

a) Al 100%*
b) Al 200%
c) Al 50%
d) A -50%
e) A 100

2) Se controllo due società, A e B, ove A è tassata più pesantemente rispetto a B, nel caso in cui ne abbia la possibilità, mi conviene imputare maggiori costi (deducibili fiscalmente)

a) Alla prima piuttosto che alla seconda*
b) Alla seconda piuttosto che alla prima
c) Equiripartirli tra le due società
d) E' impossibile dirlo
e) Dipende dalla moneta

3) Il lancio di un nuovo prodotto richiede un investimento iniziale di 20. Ogni unità prodotta può essere venduta a 2 mentre comporta un costo pari ad 1. Quante unità minime andrebbero vendute per realizzare un profitto?

a) Almeno 21*
b) Almeno 11
c) Almeno 201
d) Almeno 100
e) Almeno 56

4) Per obbligazioni più rischiose gli investitori dovrebbero richiedere rendimenti inferiori

a) Vero
b) Falso*
c) Non si può dire
d) Dipende dall'andamento dell'economia
e) Dipende dai tassi di interesse

5) Nel periodo 2003-2004 il Prodotto Interno Lordo (PIL) nel paese A ha registrato un tasso di crescita annuo del 6%, mentre il PIL del paese B è cresciuto al tasso annuo del 4%. Si consideri che nel periodo in considerazione la popolazione è diminuita nel paese A mentre è aumentata nel paese B. Quale dei due paesi presenta il più elevato livello di PIL pro-capite nel 2004?

a) Il paese A
b) Il paese A se la sua popolazione è aumentata di un tasso superiore al tasso di crescita del PIL
c) Il paese B, se la sua popolazione è diminuita ad un tasso superiore al divario dei tassi di crescita dei due paesi
d) Il paese con il più elevato livello di PIL nel 2003
e) Non è possibile rispondere alla domanda in base ai soli dati forniti*

6) La Banca Centrale Europea è:

a) La principale Banca Commerciale dei paesi Europei
b) La principale Banca Commerciale dei paesi che hanno adottato l’Euro come valuta nazionale
c) L’istituzione sopranazionale che ha il compito di condurre la politica monetaria comune dei paesi che hanno adottato l’Euro come valuta nazionale*
d) L’istituzione sopranazionale che ha il compito di condurre la politica monetaria comune di tutti i paesi Europei
e) L’organo comune europeo che svolge funzioni equivalenti a quelle svolte a livello nazionale dal Ministero del Tesoro

7) Che effetto può avere l’aumento del prezzo del petrolio sul costo della vita in Italia?

a) Nessun effetto, perché l’Italia è un produttore marginale di petrolio
b) Tende a produrre un aumento del costo della vita essendo una risorsa in larga parte importata dall’Italia e ampiamente utilizzata da cittadini e imprese italiane*
c) Tende a produrre una riduzione del costo della vita essendo una risorsa prodotta in Italia e esportata all’estero
d) Nessun effetto, perché i cittadini italiani non consumano direttamente petrolio ma consumano i suoi derivati come ad esempio la benzina
e) Nessun effetto o al limite tende a ridurre il costo della vita, perché se facesse aumentare il costo della vita i cittadini italiani non utilizzerebbero il petrolio

8) La frase “Le esportazioni italiane sono diminuite nel 2004 rispetto al 2003” significa che:

a) Il valore dei beni e servizi italiani venduti all’estero è diminuito rispetto al 2003*
b) Il valore dei beni e servizi esteri venduti in Italia è diminuito rispetto al 2003
c) La differenza tra il valore dei beni e servizi italiani venduti all’estero e quelli esteri venduti in Italia è diminuita rispetto al 2003
d) La differenza tra il valore dei beni e servizi italiani venduti all’estero e quelli esteri venduti in Italia è aumentata rispetto al 2003
e) Il valore dei beni e dei servizi italiani prodotti nel 2004 è diminuito rispetto al 2003

9) Se il tasso di interesse sui prestiti fosse del 10% annuo quanto dovrebbe restituire il signor Caio, dopo un anno, per estinguere un prestito di 1.000 € ottenuto dalla propria banca?

a) 1010 €
b) 10 €
c) 100 €
d) 1100 €*
e) 900 €

10) Se si afferma che un’impresa ha un ruolo di monopolista sul mercato dei cioccolatini si intende che:

a) L’impresa è libera di scegliere il prezzo a cui vendere i cioccolatini
b) L’impresa è libera di decidere quanti cioccolatini immettere sul mercato
c) L’impresa è la più grande impresa tra quelle operanti nel mercato dei cioccolatini in termini di volumi di vendite
d) L’impresa è l’unica impresa operante sul mercato dei cioccolatini*
e) L’impresa è di proprietà dello stato

11) Chi è il creditore?

a) Un vicino antipatico
b) Un estraneo molestatore
c) Uno che deve pagare un debito
d) Uno che deve pagare una oblazione
e) Uno che deve ricevere un pagamento*

12) Chi è mediatore?

a) Chi compra dal fabbricante e rivende al commerciante
b) Il commerciante stesso
c) Chi si interpone nei contrasti familiari per evitare liti
d) Chi compie un affare altrui per averne ricevuto incarico
e) Chi mette in rapporto due persone con la conclusione di un affare*

13) Che cos’è il pegno?

a) Un oggetto materiale nel proprio patrimonio
b) Un pagamento
c) Una sanzione irrogata dalla P.A.
d) Un provvedimento del giudice
e) Una cosa mobile vincolata a garanzia*

14) Che cos’è il contratto?

a) Un documento scritto
b) Un regolamento privato o pubblico
c) Uno scambio di merci
d) Un pagamento
e) Un accordo concreto, con vincoli patrimoniali*

15) Che cos'è la responsabilità civile?

a) Un particolare compito o incarico da svolgere
b) Un dovere che incombe sui genitori in quanto tali
c) Una sanzione irrogata dalla pubblica amministrazione
d) Un giudizio di cattiva condotta
e) Una obbligazione di risarcire il danno conseguente, in generale, alla violazione di una norma*

16) Che somma si ottiene impiegando per un anno, a un tasso di interesse annuo del 4%, un capitale di 250 euro?

a) 320€
b) 835€
c) 400€
d) 260€*
e) 290€

17) Il fatturato annuo di una società è sceso da 450 a 360 milioni di euro. Calcolare il decremento percentuale

a) -25%
b) -20%*
c) -15%
d) -90%
e) -30%

18) Qual è la probabilità che lanciando due monete equilibrate si ottengano due teste?

a) 3/4
b) 1/18
c) 3/36
d) 1/2
e) 1/4*

19) A quanto ammonta lo sconto se una cucina che costa 10.000 € è scontata del 15%?

a) 1.200 €
b) 1.000 €
c) 1.500 € *
d) 150 €
e) 750 €

20) Una scatola contiene 16 penne: 8 rosse e 8 nere. Qual è la probabilità che estraendo due penne dalla scatola queste siano entrambe nere?

a) 9/25
b) 1/4*
c) 1/11
d) 8
e) 1

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

EDUCATION SCIENCE

1) FANTASIA sta a LEGGENDA come REALTA' sta a ...

a) Storia*
b) Quotidianità
c) Illusione
d) Mito
e) Passato

2) AUMENTARE sta a DIMINUIRE come SOMMARE sta a ...

a) Togliere
b) Sottrarre*
c) Ridurre
d) Dividere
e) Accumulare

3) Quale delle seguenti parole NON è un sinonimo di occulto:

a) Accorto*
b) Nascosto
c) Esoterico
d) Invisibile
e) Coperto

4) Quale delle seguenti parole NON è un sinonimo di posticipare:

a) Rinviare
b) Ritardare
c) Rimandare
d) Posporre
e) Muovere*

5) Quale delle seguenti parole NON è il contrario di blando:

a) Deciso
b) Forte
c) Energico
d) Vigoroso
e) Collerico*

6) Che cosa studia la Pedagogia generale

a) La relazione educativa fra adulto e bambino
b) Le finalità generali dell'educazione in relazione ai differenti contesti della formazione*
c) L'ascolto dei bisogni formativi dell'allievo
d) La comunicazione scuola-famiglia
e) Le modalità di organizzazione e funzionamento dei sistemi scolastici

7) La Didattica ha per oggetto:

a) La valutazione scientifica degli apprendimenti
b) I metodi di insegnamento e di istruzione in genere*
c) La conduzione dei gruppi d'apprendimento
d) L'insegnamento individualizzato
e) I rapporti fra scuola ed extrascuola

8) Chi ha scritto "Democrazia ed educazione"?

a) John Dewey*
b) Jerome Bruner
c) Giovanni Gentile
d) Paulo Freire
e) John Locke

9) Chi dei seguenti autori NON è uno dei classici della sociologia?

a) Max Weber
b) Karl Marx
c) Emile Durkheim
d) Charles Darwin*
e) Georg Simmel

10) Dal punto di vista metodologico un limitato numero di persone che rappresenta l'intera popolazione viene denominato:

a) Universo
b) Campione*
c) Gruppo

d) Variabile
e) Unità d'analisi

11) Il positivismo può essere definito come la convinzione che:

a) Le scienze sociali devono enfatizzare gli aspetti positivi della vita sociale
b) La conoscenza si basa su evidenze empiriche che derivano dall’osservazione*
c) I sociologi devono fare un ampio uso di statistiche
d) Gli uomini sviluppano relazioni a fin di bene
e) Lo sviluppo storico-sociale comporta sempre un miglioramento

12) La meritocrazia è una caratteristica della società in cui lo status sociale si basa su:

a) La classe sociale
b) La ricchezza
c) La famiglia
d) Il potere
e) I risultati conseguiti*

13) Qual è il principale modo per aiutare un bambino a sviluppare i primi nuclei di autoconoscenza e la curiosità verso se stesso?

a) La lettura
b) Il gioco*
c) Il divieto
d) Il sonno
e) Il cibo

14) Secondo il cognitivismo, come funziona la mente nel processo di apprendimento?

a) Opera come un sistema capace di integrare ed elaborare informazioni *
b) Agisce come una scatola nera che contiene ogni piccola unità di informazione
c) Si limita a registrare i dati provenienti dall’ambiente esterno
d) Acquisisce passivamente le nozioni attraverso processi di imitazione
e) Agisce per riflessi condizionati generati dallo scambio tra stimolo e risposta

15) Secondo la teoria dell’apprendimento elaborata da Piaget, la comparsa delle capacità di astrazione caratterizza lo stadio definito:

a) Delle “operazioni concrete”
b) Delle “operazioni formali”*
c) Dell’”intelligenza sensomotoria”
d) Dell’”intelligenza pre-operatoria”
e) Delle “azioni interiorizzate”

16) L’apprendimento per “condizionamento operante” elaborato da B.F. Skinner consiste nel conseguire un apprendimento:

a) grazie al rinforzo legato ad una ricompensa*
b) attraverso la ripetizione di una serie di comportamenti nella stessa sequenza
c) grazie all’imitazione del comportamento di altri soggetti
d) attraverso l’assorbimento dei dati provenenti dal mondo esterno nelle strutture innate del soggetto
e) attraverso la memorizzazione di nuove esperienze

17) Il colonialismo è stato una forma di:

a) Gestione dei nuovi territori scoperti e colonizzati nelle Americhe e in Africa a partire dal XVI secolo
b) Di imperialismo basato sulla subordinazione politica, economica e culturale attuato da alcune nazioni europee su territori e culture non occidentali*
c) Di vita umana e animale realizzata in colonie organizzate
d) Di gestione delle relazioni commerciali che hanno legato a partire dal XVI secolo madre-patria e colonie sparse sul pianeta
e) Di organizzazione della gestione del tempo libero

18) Il “nomadismo” è:

a) Una pratica di spostamento comunitario tipica degli zingari o rom
b) Una pratica di mobilità regionale tipica delle culture “primitive”
c) Una forma di mobilità in un territorio dettata dalla ricerca di nuove risorse*
d) Una forma di mobilità legata al fenomeno delle emigrazioni
e) Un modo di pensare se stessi non radicati a specifici territori e società

19) Per “tabù” si intende:

a) Un oggetto/persona che non si può nominare
b) Un oggetto/persona di natura magica con cui un gruppo si identifica
c) Un divieto o proibizione rituale riguardante oggetti/persone investiti di sacralità*
d) Le pratiche magico-sacrali riservate a oggetti/persone al centro di un rituale
e) Una danza rituale

20) La teoria detta della “selezione naturale” indica che:

a) I più forti vengono selzionati e sono i soli a sopravvivere
b) Coloro che si trovano a possedere i caratteri più adatti alla sopravvivenza li trasmettono ai discendenti *
c) L’evoluzione della specie si spiega con la sopravvivenza del più forte sul piano genetico
d) Una specie più forte deriva da un’altra considerata inferiore
e) I processi genetici sopprimono i tratti più deboli della specie

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

FOREIGN LANGUAGES AND LITERATURE

1) Il Viaggio in Italia di Goethe alimenta uno dei grandi miti culturali dell'Europa alle soglie della modernità: quale?

a) L'idealizzazione del Bel Paese
b) L'esaltazione dei valori mediterranei
c) Il tentativo di ricostruire e comprendere le antiche civiltà classiche
d) Il "Grand Tour" come esperienza personale e simbolica fondamentale nella formazione del giovane intellettuale*
e) L'esotismo

2) Quale tipo di rapporto storico viene modernamente drammatizzato nella Tempesta di W. Shakespeare?

a) Il rapporto tra uomo e natura
b) Il rapporto tra diritto naturale e diritto divino
c) Il rapporto tra colonizzatore (Prospero) e colonizzato (Calibano)*
d) Il rapporto tra Dio e l'uomo
e) Nessun rapporto perché Prospero è escluso dalla civiltà

3) Nella tragedia classica (si pensi all'Antigone sofoclea, tanto per indicare un paradigma di riferimento) è più importante il peso specifico che nell'opera esercita il mito o la storia?

a) È fondamentale la vicenda storica per comprendere il senso tragico dell'opera
b) Mito e storia innescano un conflitto nella vicenda di Creonte e Antigone
c) Mito e storia sono assenti: si tratta di faide famigliari
d) Il mito, come si conviene nel tragico antico, è fondamentale per produrre il significato della tragedia*
e) Un nucleo di verità sarà pure presente, ma l'inverosimiglianza della vicenda non permette di credere alla tragedia

4) Nella frase "stanchi, dopo la camminata in montagna, ritornammo ai nostri tetti" quale figura retorica è riconoscibile fra quelle sottoelencate?

a) Antonomasia
b) Sineddoche
c) Metafora
d) Litote
e) Metonimia*

5) Quale dei seguenti cinque versi danteschi è costituito da un “endecasillabo di settima”?

a) E come quei che con lena affannata*
b) Uscito fuor del pelago alla riva
c) Si volge all’acqua perigliosa e guata
d) Così l’animo mio, ch’ancor fuggiva
e) Si volse a retro a rimirar lo passo

6) Nella parola italiana "sciogliere" quanti suoni individuate?

a) Sei*
b) Otto
c) Nove
d) Dieci
e) Sette

7) Quale, fra le lingue sottoelencate, appartiene alla famiglia camito-semitica?

a) Armeno
b) Ebraico*
c) Turco
d) Persiano
e) Curdo

8) Analizzando contrastivamente il British English e l’American English, in quale settore della lingua collochereste le più vistose differenze?

a) Fonetica
b) Morfologia
c) Sintassi
d) Lessico*
e) Semantica

9) In quale delle sottoelencate relazioni semantiche collochereste la coppia di parole italiane “reale1 e reale2?”

a) Sinonimia
b) Antonimia
c) Polisemia
d) Iponimia
e) Omonimia*

10) Quale, fra le lingue dell’Europa settentrionale sottoelencate, non appartiene alla famiglia indoeuropea?
a) Islandese
b) Norvegese
c) Estone*
d) Svedese
e) Danese

11) Traduzione tecnico-scientifica: quale di queste affermazioni è giusta?

a) Più tecnico e specifico un testo, più è facile tradurlo
b) Più tecnico e specifico un testo, più è opportuno che un traduttore sia coadiuvato da un esperto del settore*
c) Un buon dizionario tecnico risolve quasi tutti i problemi di traduzione
d) Più tecnico e specifico un testo, più è difficile tradurlo
e) Più tecnico e specifico un testo, più è importante avvalersi di un esperto che sappia anche tradurre

12) Un approccio adeguato alla traduzione poetica deve identificare prevalentemente

a) Il tessuto ritmico, fonetico e retorico del verso poetico*
b) La recitabilità del verso poetico
c) La struttura sintattica del verso poetico
d) La poetica dell'autore
e) Nessun approccio è adeguato perché tradurre la poesia è pressoché impossibile

13) Un approccio adeguato alla traduzione letteraria deve basarsi prevalentemente su

a) L'analisi testuale e contestuale*
b) L'autore e la sua biografia
c) La sintassi della prosa
d) Le rime e la musicalità della prosa
e) Nessun elemento particolare perché tradurre letteratura significa riscrivere il testo

14) L'addetto al turismo culturale è una figura professionale

a) Sovrapponibile alla guida turistica tradizionale
b) Sovrapponibile all'operatore di agenzia di viaggio
c) È un "tour operator" esclusivamente dedicato alla promozione culturale
d) È un operatore degli Urp (Uffici relazioni con il pubblico)
e) È una figura professionale con competenze linguistiche, artistiche e culturali specifiche e competenze organizzative che opera per mettere in luce e promuovere gli aspetti culturali legati al turismo*

15) L'esperto linguistico d'area per l'economia è una figura professionale di

a) Esperto dei processi di internazionalizzazione dell'economia
b) Esperto del linguaggio specifico utilizzato in ambito economico e aziendale e del suo linguaggio corrispondente in lingua straniera*
c) Traduttore economico in ambito aziendale
d) Interprete di comunità
e) Mediatore linguistico

16) Perchè quando si parla di rappresentazione in campo culturale è inevitabile alludere alla sua ambiguità?

a) La rappresentazione non è la cosa rappresentata, ma è facile confonderla con essa*
b) La rappresentazione non è ambigua, al contrario è chiara ed intelligibile
c) L'ambiguità deriva dal fatto che la rappresentazione è sempre opaca, non è mai chiara

d) La rappresentazione è un modo per avere accesso alla realtà
e) Rappresentare è ripresentare l’oggetto, quindi vi è una sostanziale identità tra di essi

17) È corretto affermare che la cultura dell’Oriente ha contribuito in profondità a formare e ad influenzare la cultura dell’Occidente o tra i due ambiti va vista una frattura non componibile?

a) Non è vero. Occidente e Oriente hanno avuto storie distinte e pienamente autonome a cui ha contribuito la mancanza di veri e propri canali di comunicazione e di scambio
b) È vero, la cultura Occidentale è stata fortemente influenzata, in diverse epoche storiche, dalle culture orientali tanto da averne assorbito valori e forme*
c) Occidente e Oriente sono portatori di modelli di civiltà irriducibili tra loro
d) L’Oriente ha sempre avuto nell’Occidente un nemico con il quale ha ingaggiato una secolare tradizione di guerre e scontri
e) L’Occidente, sviluppatosi prima dell’Oriente, ha contributo in profondità allo sviluppo delle culture orientali

18) Quale definizione di colonialismo ed imperialismo appare più precisa alla luce delle diverse rese storiche?

a) Il colonialismo e l’imperialismo sono coincidenti e rappresentano una occupazione spaziale di un territorio
b) Il colonialismo e l’imperialismo sono fenomeni del passato, ormai totalmente assenti nel mondo contemporaneo
c) Pur essendo categorie diverse anche sul piano storico, colonialismo e imperialismo possono essere, in generale, distinti sul piano spaziale: il primo rappresenta un potere che si fonda sulla espansione e sulla occupazione, quindi sul dominio diretto di colonie, il secondo è legato al controllo e al potere che viene esercitato dalla metropoli senza necessariamente contare su vere e proprie colonie*
d) Il colonialismo e l’imperialismo posso essere intesi come progetti di espansione dell’Occidente
e) L’imperialismo è la forma storica contemporanea del colonialismo

19) Cosa definisce le fondamenta dello stato nazione moderno e su quali presupposti ideologici si fondano le narrative di nazione dei nazionalismi?

a) Il vincolo differenziale è quello propugnato, in campo culturale, dal romanticismo: della identità di lingua, sangue e territorio*
b) La costituzione è il documento formale che sancisce l’esistenza di una nazione
c) È la nascita a stabilire la nazionalità, anche modernamente
d) Riconoscersi nella stessa storia di fondazione nazionale
e) La continuità delle istituzioni attestano l’esistenza della nazione

20) Modernità e modernizzazione sono categorie coincidenti?

a) Modernità e modernizzazione sono sinonimi
b) La modernità è il risultato della modernizzazione
c) La modernizzazione è possibile solo quando c’è la modernità
d) Nell’età moderna sono coincidenti
e) No, la modernizzazione può essere un processo implicito alla modernità ma si può avere anche una modernizzazione senza modernità, quindi si tratta di concetti al contempo interdipendenti e autonomi*

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

PHARMACY

1) Da che cosa è costituita la membrana cellulare?

a) da glucidi e DNA
b) da proteine e lipidi*
c) da proteine e acqua
d) da lipidi e vitamine

2) Gli enzimi sono:

a) catalizzatori biologici*
b) catalizzatori inorganici
c) catalizzatori industriali
d) acidi nucleici

3) Il principale polisaccaride di riserva degli animali è:

a) glucosio
b) saccarosio
c) glicogeno*
d) amido

4) Il ciclo di Krebs si svolge:

a) nei mitocondri*
b) nei ribosomi
c) nei cloroplasti
d) nei cromosomi

5) Qual è la funzione principale dell'adenosintrifosfato (o ATP)?

a) è la moneta di scambio energetico*
b) è il mediatore di alcuni ormoni
c) entra nella composizione del DNA
d) è una vitamina

6) Le gonadi sono:

a) ormoni
b) organi degli apparati genitali maschile e femminile*
c) organi dell'apparato genitale femminile
d) ghiandole dell'apparato urinario

7) Con il termine composto si intende:

a) una sostanza omogenea separabile in sostanze più semplici per mezzo di trasformazioni chimiche*
b) un miscuglio di più elementi
c) un miscuglio omogeneo capace di variare gradualmente la sua composizione
d) una sostanza omogenea solida

8) Il numero Avogadro esprime il numero di:

a) atomi contenuti in una molecola
b) protoni contenuti in un atomo
c) molecole contenute in una mole di molecole*
d) elettroni delocalizzati in un metallo in condizioni standard

9) Quale tra i seguenti elementi è un gas nobile?

a) Si
b) Mo
c) Ge
d) He*

10) Nella grammomolecola di una specie chimica è presente un numero di molecole:

a) variabile
b) diverso se la molecola è mono-, bi- o poli-atomica
c) pari a $6{,}02*10^{23}$ *
d) che dipende dalla temperatura

11) Gli isotopi di un elemento:

a) sono separabili*
b) non sono separabili
c) sono separabili solo se differiscono per il numero di protoni
d) sono separabili solo se gassosi

12) L'ossigeno liquido ha un punto di ebollizione di -183° C, cioè di:

a) -90° K
b) 183° K
c) -456° K
d) 90° K *

13) 3000 calorie equivalgono a:

a) 30 kcal
b) $3*10^{6}$kcal
c) $3*10^{-6}$kcal
d) 3 kcal *

14) Un'auto si muove di moto vario. Essa impiega 40 minuti per percorrere 20 km. Quale è la sua velocità media espressa in m/s? Quale è la sua velocità istantanea al tempo t = 20 min?

a) vm = 8,33 m/s; vi = indeterminata *
b) vm = 4,67 m/s; vi = 60 km/h
c) vm = 8,33 m/s; vi = 30 km/h
d) vm = 4,67 m/s; vi = 30 km/h

15) Una barca si muove sulla corrente di un fiume con una velocità media v1=2,5 m/s nel verso opposto della corrente. La corrente ha una velocità media v2=6 km/h. Quanto tempo impiega la barca a percorrere 1 km?

a) 30 min
b) 15 min
c) 600 s
d) 1200 s *

16) $(a^{m})^{n}$ è uguale a:

a) a^{m+n}
b) a^{m-n}
c) a^{m*n} *
d) $a^{m/n}$

17) Quale delle disuguaglianze è corretta?

a) $4,38*10^{2} > 4,38*10^{3}$
b) $6,37*10^{-4} > 3,52*10^{-6}$ *
c) $9,45*10^{3} > 3,25*10^{5}$
d) $54,3*10^{4} > 5,43*10^{5}$

18) Di quanti gradi è l'angolo giro?

a) 90°
b) 45°
c) 360°*
d) 180°

19) L’espressione $k^{(a-b)}$ è uguale a:

a) $k^a - k^b$
b) $k^a * k^b$
c) k^a / k^b *
d) $k^b – k^a$

20) 0/30 è uguale a:

a) 0 *
b) 30
c) ∞
d) i

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

POLITICAL SCIENCE

1 L’affermazione “non tutti gli abitanti di Forlì sono nativi di Forlì” a quale delle seguenti affermazioni equivale?

a) Alcuni abitanti di Forlì non sono nati a Forlì *
b) Tutti gli abitanti di Forlì non sono nati a Forlì
c) Non tutti i nati a Forlì sono abitanti a Forlì
d) Non tutti i nati a Forlì non sono abitanti a Forlì
e) Alcuni nati a Forlì non sono abitanti a Forlì

2 Se ho letto 32 pagine in più della metà di un libro e me ne mancano 123, a quale pagina sono arrivato?

a) 91
b) 94
c) 155
d) 187 *
e) 310

3 Luca è più alto di Lia che è più alta di Carla. Maria è più alta di Lia; quindi sicuramente:
a) Luca è più alto di Maria
b) Maria è più bassa di Carla
c) Carla è più bassa di Maria *
d) Lia è più alta di Luca
e) Luca e Maria hanno la stessa statura

4 L’insieme lato-angolo-cateto-ipotenusa si associa a:
a) trapezio
b) parallelogramma
c) pentagono
d) esagono
e) triangolo *

5 La frase “Tutte le volte che gioco a briscola perdo” è falsa. Allora è sicuramente vero che:

a) tutte le volte che gioco a briscola vinco
b) almeno una volta ho vinto giocando a briscola *
c) sono fortunato al gioco
d) tutte le volte che gioco a tresette perdo
e) se perdo a un gioco di carte, sicuramente è briscola

6 Che cosa vuol dire “imbelle”?

a) armato, dotato di risorse
b) uomo rozzo
c) ragazze vistosamente truccate
d) che si oppone all’autorità
e) vile e fiacco *

7 Quale delle seguenti parole è scritta in forma corretta?

a) scenza
b) proficuo *
c) comincierà
d) migliardario
e) umigliare

8 Quale dei seguenti periodi è corretto?

a) credevo che tu eri uscito
b) la perizia ha dimostrato che l’incendio fosse doloso
c) ho paura che Lia si è offesa
d) sebbene fosse ricco, viveva come un barbone *
e) lo conobbi quando andai alle elementari

9 Che cosa significa la parola oblio?

a) finestrino di un'imbarcazione
b) dimenticanza totale e persistente *
c) arma tipica delle legioni romane
d) voce del verbo oblire
e) pozzo inaccessibile, oscuro e profondo

10 Qual è il significato del termine “interstizio”?

a) piccolo spazio che separa due corpi o due strati *
b) posizione assunta dal sole nell’arco dell’anno
c) forma di difficoltà respiratoria
d) territorio fra due frontiere
e) la fine delle ostilità dopo un conflitto

11 Due avvenimenti, uno importante per il mondo intero - la scoperta dell’America - l’altro importante per l’Italia e per l’Europa - la morte di Lorenzo de’ Medici detto il Magnifico -sono avvenuti nello stesso anno. Quale?

a) 1453
b) 1492 *
c) 1509
d) 1606
e) 1648

12 "Dei delitti e delle pene" è stato scritto da:

a) Edgar Allan Poe
b) Il Marchese De Sade
c) Silvio Pellico
d) Cesare Beccaria *
e) Cesare Lombroso

13 Il Cubismo è:
a) una branca della geometria superiore
b) una corrente della pittura del '900 *
c) un tipo di gioco coi dadi
d) un movimento matematico del '900
e) una sindrome depressiva

14 L' Illuminismo è una corrente di pensiero del secolo:
a) XIV
b) XV
c) XVI
d) XVII
e) XVIII *

15 Chi è l'autore dell'opera "Il Principe"?
a) Niccolò Machiavelli *
b) Luigi XIV
c) Benito Mussolini
d) Madre Teresa di Calcutta
e) Dante Alighieri

16 ……………..is yours, this one or that one?
a) that
b) whose
c) where
d) which *
e) what

17 When do you usually ……………….. to the radio?

a) watch
b) see
c) listen *
d) hear
e) log in

18 About 100 people …………..outside the theatre when I arrived.

a) have queued
b) queued
c) were queuing *
d) queue
e) are queuing

19 By the time the teacher arrived, the classroom was empty as the students ………..

a) left
b) will have left
c) were leaving
d) have left
e) had left *

20 Tomorrow is Sunday, so you …………. go school.

a) couldn't
b) shouldn't
c) mustn't
d) don't have to *
e) needn't

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

PSYCHOLOGY

1 Quale parola è un sinonimo di *Contingente*?

A Inevitabile
B Sistematico
C Possibile*
D Vicino
E Lontano

2 Quale parola è un sinonimo di *Irridere*?

A Encomiare
B Divertire
C Ridere
D Dileggiare*
E Lodare

3 Quale termine corrisponde alla seguente definizione?
Esemplare tipico

A Esempio
B Stereotipo
C Caso
D Schema
E Prototipo*

4 Quale parola non è simile alle altre quattro?

A Gotico
B Barocco
C Liberty
D Illuministico*
E Dorico

5 Indicare le due coppie di parole che hanno lo stesso tipo di rapporto logico.

1 Governo/ Nazione
2 Esercito/ Marcia
3 Musica/ Inno
4 Regola/ Disciplina
5 Abito/ Divisa

4-5	**A**
1-3	**B**
3-5	**C***
1-4	**D**
2-5	**E**

6 Preside sta a scuola come colonnello sta a …… : Indicare la parola giusta

A Esercito
B fanteria
C Caserma*
D squadrone
E armata

7 Rispettando la regola alla base della sequenza, trovare il numero che completa la serie.

128 64 32 16 8 4

12	**A**
0	**B**
8	**C**
2	**D***
6	**E**

8 Se K viene prima di C
Y viene prima di C
C viene prima di D
K viene prima di Y
Quale tra le seguenti affermazioni non è corretta?

K è la prima della serie	**A**
Y viene dopo D	**B***
Y non è l’ultima della serie	**C**
Y viene prima di D	**D**
L’ordine non è alfabetico	**E**

9 Se lanciando una moneta è uscito per 20 volte di seguito testa, che probabilità ci sono che al 21° lancio esca croce?

100 %	**A**
80 %	**B**
75%	**C**
60%	**D**
50%	**E***

10 Se lanciando un dato esce il numero 4, che probabilità ci sono che ricompaia il numero 4 lanciando lo stesso dado subito dopo?

Una probabilità su due	**A**
Una probabilità su quattro	**B**
Una probabilità su sei	**C***
Due probabilità su cinque	**D**
Due probabilità su sei	**E**

---

11 Il Sig. Rossi è andato dalla città A alla città B in x ore. Nel viaggio di ritorno, per la stessa strada, la sua velocità media è raddoppiata. Quale è il numero totale di ore impiegate dal Sig. Rossi per il viaggio di andata e ritorno?

2/3 x	**A**
3/2 x	**B***
4/2 x	**C**
2/4 x	**D**
5/3 x	**E**

12

Che cosa si intende per fagocitosi?

l’introduzione all’interno della cellula di grosse molecole solide	**A**
l’introduzione all’interno della cellula di batteri	**B**
l’introduzione all’interno della cellula di detriti cellulari	**C**
un particolare tipo di endocitosi	**D**
tutte e quattro le precedenti	**E***

13 *Due gemelli identici originano sempre da:*

una stessa cellula uovo (ovocita) fecondata con due spermatozoi	**A**
due ovociti fecondati con due spermatozoi	**B**
due ovociti fusi tra loro, fecondati con uno spermatozoo	**C**
un ovocita fecondato con uno spermatozoo	**D***
una madre a sua volta gemella identica	**E**

14 *Quale è in percentuale il maggiore costituente del corpo umano?*

Acqua	**A***
Grasso	**B**
Proteine	**C**
Minerali	**D**
Carboidrati	**E**

15 *In condizioni normali i reni:*

pompano il sangue nella circolazione sistemica	**A**
trasportano il sangue nelle urine	**B**
depurano il sangue dalle sostanze di rifiuto	**C***
ossigenano il sangue	**D**
digeriscono le proteine introdotte con la dieta	**E**

16 *Quale è la velocità della luce?*

3 chilometri al secondo	**A**
30 chilometri al secondo	**B**
300 chilometri al secondo	**C**
300000 chilometri al secondo	**D***
Nessuna di queste	**E**

Sostituire ai puntini il termine corretto

17 *Listen! The bell ………*

rings	**A**
is ringing	**B***
rang	**C**
Has rung	**D**
ringed	**E**

18 *Times were especially hard for people ……… the 1920's*

at	**A**
within	**B**
inside	**C**
to	**D**
during	**E***

19 *We got to the station rather late but nevertheless ……… catch the train*

we were able to	**A***
We can	**B**
We could	**C**
we will be able	**D**
we have to	**E**

20 *……… do you think of my new sweeter?*

why	**A**
whose	**B**
which	**C**
when	**D**
what	**E***

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

STATISTICS

Domanda 1:

Nella classe di Francesco l’insegnante di matematica fa 5 verifiche valutate ciascuna con un punteggio massimo di 50. Nelle prime 3 verifiche Francesco ha un punteggio medio di 30, nella quarta ha preso 40. Quale voto deve prendere per avere alla fine delle 5 prove una media di 30?

a. 30;
b. 20; *
c. 60;
d. 15;
e. 25.

Domanda 2:

L’elevata densità di popolazione di un paese (rapporto tra numero di abitanti e superficie) indica che:

a. l'ammontare della popolazione è molto elevato
b. la superficie del paese è molto vasta
c. l'ammontare della popolazione è elevato rispetto alla superficie *
d. la superficie del paese è molto ridotta
e. la popolazione è uniformemente distribuita sul territorio

Domanda 3:

Durante il 2003 in Emilia Romagna e in Lazio si sono registrati, rispettivamente, 48.215 e 51.147 decessi. Per effettuare un primo confronto sulla sopravvivenza nelle due regioni:

a. è necessario conoscere anche la densità (rapporto tra numero di abitanti e superficie) delle due popolazioni
b. è necessario conoscere anche l'ammontare delle due popolazioni *
c. è sufficiente confrontare i decessi
d. è necessario conoscere anche l'ammontare degli anziani nelle due regioni
e. è necessario conoscere anche le cause di morte nelle due regioni

Domanda 4:

I 25 studenti della classe VB hanno una media in matematica di 7,5 mentre i 21 della VD hanno una media di 7. Si può quindi dire che:

a. tutti gli studenti della VB sono più bravi in matematica di tutti gli studenti della VD
b. tutti gli studenti della VB sono meno bravi di quelli della VD

c. gli studenti della VB sono più bravi di quelli della VD perché sono di più
d. gli studenti della VB sono meno bravi di quelli della VD perché sono di meno
e. è possibile che alcuni studenti della VB siano più bravi in matematica di alcuni della VD. *

Domanda 5:

Se si lanciano due dadi, è più probabile ottenere la somma:

a. 4;
b. 5;
c. 6;
d. 7; *
e. 2.

Domanda 6:

Se si lancia due volte una moneta con testa e croce, qual è la probabilità di ottenere una testa ed una croce?

a. 1/2; *
b. 1/3;
c. 1/4;
d. 1/5;
e. 1/8.

Domanda 7:

Di due diverse lotterie sono stati venduti, rispettivamente, 200 e 350 biglietti. Avendo acquistato 15 biglietti della prima e 22 della seconda, in quale delle due lotterie si ha maggior probabilità di vincere il premio più alto?

a. la prima lotteria; *
b. la seconda lotteria;
c. i dati non sono confrontabili;
d. le due lotterie sono uguali;
e. dipende dal numero complessivo dei biglietti venduti.

Domanda 8:

Una partita di 6000 caramelle alla frutta, tutte confezionate alla stesso modo, è formata di 2000 caramelle all’arancia, 1000 alla fragola e 3000 al lampone. Prendendo una caramella a caso, qual è l’evento più probabile?

a. prendere una caramella alla fragola;
b. prendere una caramella all’arancia;
c. prendere una caramella al lampone;
d. prendere una caramella che non sia al gusto di fragola; *
e. prendere una caramella al lampone o alla fragola.

Domanda 9:

Un bambino che non sa scrivere gioca con quattro lettere di legno che rappresentano le lettere A, M, O, R. Giocando a mettere in fila le lettere, qual è la probabilità che componga la parola ROMA?

a. 1/2;
b. 1/4;
c. 5/24;
d. 1/24; *
e. 23/24.

Domanda 10:

Luigi è arrivato al traguardo prima di Luca, ma non di Franco. Giacomo è arrivato dopo la sua ragazza Martina, ma prima di Franco. Chi è arrivato ultimo?

a. Luigi;
b. Luca; *
c. Franco;
d. Giacomo;
e. Martina.

Domanda 11:

Nella famiglia Rossi, nel 2004, il padre è andato al cinema una volta ogni due mesi, la madre una volta al mese ed i figli una volta alla settimana.

a. Tutti sono andati al cinema almeno una volta mese;
b. Solo i figli sono andati al cinema almeno una volta al mese;
c. Solo la madre è andata al cinema almeno una volta al mese;
d. Tutti sono andati al cinema almeno una volta alla settimana;
e. Tutti, tranne il padre, sono andati al cinema almeno una volta al mese. *

Domanda 12:

Da un’indagine effettuata presso quattro classi prime di una scuola superiore sul numero di studenti che posseggono una *play station* è emerso quanto segue:

Classe	**Numero studenti**	**Numero di *play station***
1A	20	9
1B	28	12
1C	20	10
1D	30	11

Qual è la graduatoria decrescente delle classi rispetto alla presenza relativa di *play station*?

a. 1C, 1A, 1B, 1D; *
b. 1D, 1B, 1A, 1C;
c. 1B, 1D, 1C, 1A;
d. 1A, 1B, 1C, 1D;
e. 1C, 1B, 1A, 1D.

Domanda 13:

"Se non ci fosse corrente elettrica, il treno non partirebbe". Posto che questa affermazione sia vera, quale delle seguenti proposizioni è vera?

a. Se c'è corrente elettrica, il treno partirà;
b. Se il treno non parte, significa che non c'è corrente elettrica;
c. Il fatto che il treno non parta, implica che non c'è corrente elettrica;
d. Se il treno parte, vuol dire che c'è corrente elettrica; *
e. Se c'è corrente elettrica, il treno non partirà.

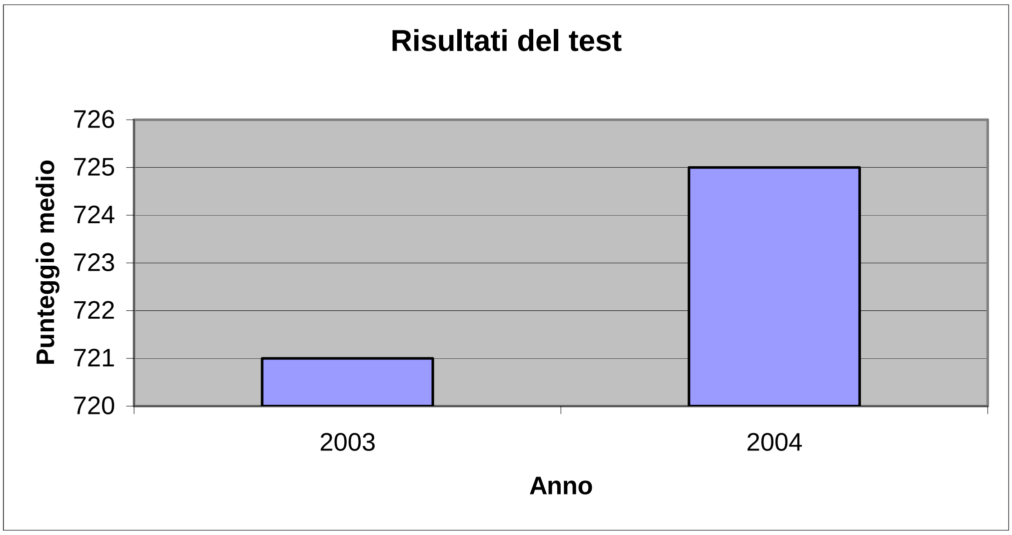
Domanda 14:

Vera l'affermazione "se piove la strada è bagnata", osservando la strada bagnata si può dire che:

a. nulla si può dire;
b. è piovuto;
c. è probabile che sia piovuto;*
d. hanno lavato la strada;
e. si è rotto un tubo dell'acquedotto.

Domanda 15:

Il risultato medio di un test (punteggio massimo 1000) somministrato in due anni consecutivi in una stessa classe è riportato nel seguente grafico:

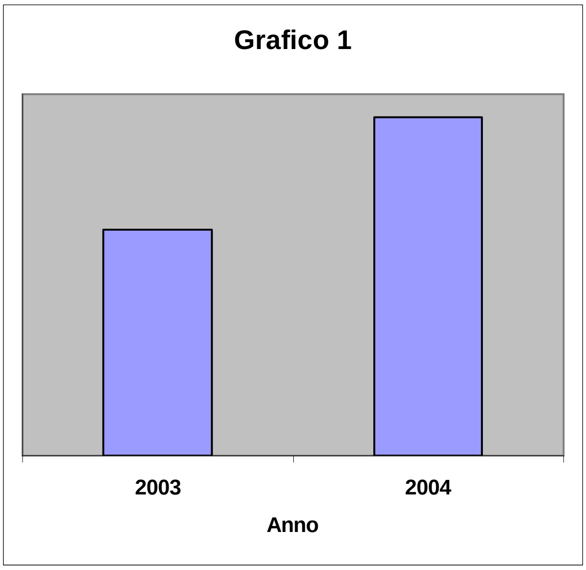


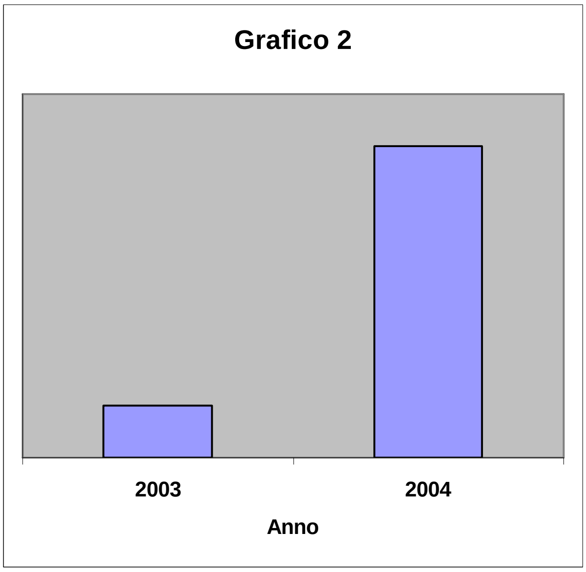
Dal grafico si può concludere:

a. nel 2004 c'è stato un modesto miglioramento dei risultati; *
b. nel 2004 c'è stato un notevole miglioramento dei risultati;
c. per poter confrontare i risultati dei due anni bisognerebbe conoscere il numero degli studenti che hanno risposto al test in ciascuna delle due prove;
d. il risultato del 2004 è migliore di quello dell'anno precedente di circa il 4%;
e. il risultato del 2004 è peggiore di quello dell'anno precedente di circa il 4%.

Domanda 16:

Quale dei due grafici rappresenta la crescita maggiore?





a. Il grafico 1;
b. il grafico 2;
c. non è possibile rispondere perché non è nota l’unità di misura dell’asse verticale di entrambi i grafici; *
d. non è possibile rispondere perché sarebbero necessari anche i dati relativi agli anni precedenti;
e. non è possibile rispondere perché sarebbero necessari anche i dati relativi agli anni successivi.

Domanda 17:

Agli studenti della VB è stato chiesto di indicare se possiedono un cellulare. E’ risultato che il 70% degli studenti possiede un cellulare. Si ha pertanto che:

a. 1l 50% degli studenti della VB non possiede il cellulare;
b. Solo la metà degli studenti della VB possiede il cellulare;
c. Sono di più quelli che non lo posseggono rispetto a quelli che ne hanno uno;
d. Sono di meno quelli che lo posseggono rispetto a quelli che non lo posseggono;
e. Il 30% degli studenti della VB non possiede il cellulare. *

Domanda 18:

Le persone iscritte ad una polisportiva sono classificate secondo l’attività svolta:

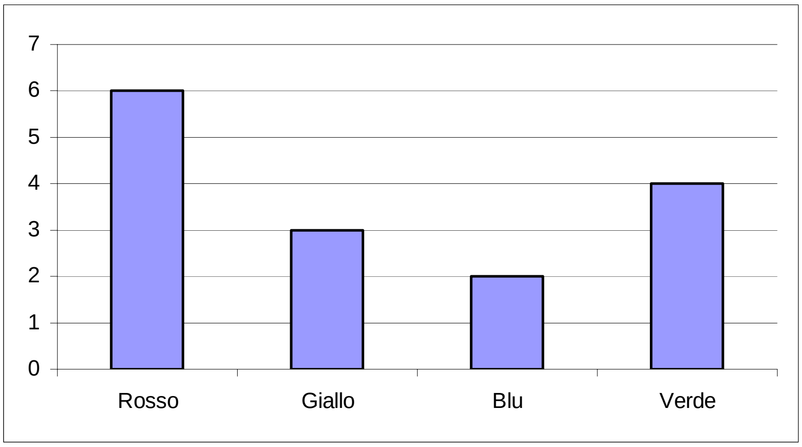
Attività	Iscritti
Calcio	52
Tennis	25
Volley	28
Basket	45

Quanti sono complessivamente gli iscritti?

a. 150; *
b. 4;
c. 100:
d. 77;
e. 73.

Domanda 19:

In un sacchetto ci sono biglie di 4 colori diversi. Il grafico seguente mostra il numero di biglie di ciascun colore.



Qual è la probabilità di estrarre una biglia gialla?

a. 20% *
b. 50%
c. 60%
d. 30%
e. 40%

Domanda 20:

Se in un dato anno a Imola si sono verificati meno incidenti stradali che a Bologna, si può affermare che:

a. è più pericoloso guidare a Imola che a Bologna;
b. non è possibile stabilire in quale città il rischio è più alto; *
c. è più pericoloso guidare a Bologna che a Imola;
d. il rischio di incidente nei due comuni è uguale;
e. il rischio di incidente dipende dal numero delle auto in circolazione a Imola .