universiteit hasselt

Peabody Picture Vocabulary Test - Revised data: a Bayesian approach to Item Response Theory

Serena Arima

External Supervisor :

Prof. Luca Tardella

Dipartimento di Statistica, Probabilitá e Statistiche Applicate.

University of Rome "La Sapienza"

Internal Supervisor :

Dr. Annouschka Laenen

Universiteit Hasselt

Thesis submitted in partial **fulfilment of the requirements for** the degree of

Master of **Science in Bio**statistics

2006-2006

Abstract

Background

Item Response Theory is the area of psychometry that deals with the problem of constructing and analyzing psychological and sociological tests. By applying a fully Bayesian approach to this methodology, we analyze a data set obtained administering the Italian translation of the well-known Peabody Picture Vocabulary Test - Revised (PPVT-R) to a sample of Italian children. In the original English version the items are believed to be in increasing difficulty order. One main aim of this thesis is to evaluate if and how much the translation leads to violations of the increasing difficulty ordering. This aspect is important since, in the original version, *basal* and *ceiling* level of the test are determined assuming items in increasing difficulty order.

Methods

Classical item response models, as **1PL** and **2PL** have been applied to PPVT-R data. These models have been extended by including covariates.
Parameters estimation has been performed using a complete Bayesian approach that in this case resulted more flexible than the classical approach. In particular, the flexibility of the Bayesian approach has been underlined with respect to the analysis of an incomplete data matrix, due to the nature of the stopping rule, and to the item difficulty comparisons. Several decision rules for the comparison of item difficulties have been analyzed. We propose a further more general *alternative* method that allows to compare item characteristic curves taking into account also the ability distribution.

Results and Conclusions

1PL, **2PL** models and model with covariates have been estimated using MCMC methodology: the goodness of fit of the models has been analyzed using posterior predictive p-values and the performance of the three models has been compared using AIC, BIC and DIC indices. The model with covariates resulted to be the best in terms of information criteria. Therefore the comparisons of the item difficulties were performed using this

model.
The different decision rules for the item difficulty comparisons have been compared and an ordering of the items for each criterion has been drawn. The criteria agree on concluding violation of the increasing difficulty order: from the analysis of the results, we can conclude that the test can be improved by modifying the ordering and by translating the English terms in Italian words of more common use.

Key Words: item response models, Bayesian approach, MCMC, latent variables, posterior predictive p-values.

Contents

1 Introduction

In recent years evaluating people using tests is very common: we have been tested since the elementary school, in order to evaluate our abilities in mathematics, in reading, in writing, till job interviews, in which the grade of aptitude for a particular job position is measured.

Since tests are so widely used, there has been a considerable interest among psychologists and statisticians in developing a theory that allows to improve educational and psychological tests. Item response theory (IRT), introduced by the mathematician George Rasch (1901-1945), is the area of psychometry that deals with the problem of tests construction, item calibration and with the evaluation of latent ability the test aims at measuring. Item response models are latent trait models in which the probability of correct responses are modeled as function of examinees' ability and as function of items characteristics, such as their difficulty levels and their discriminant powers.

The literature regarding item response models is developing but several problems are still unsolved. In this thesis, using a real data set, the Peabody Picture Vocabulary Test - Revised (PPVT-R) data, we will mainly focus on three aspects: the parameters estimation, the comparison of the difficulty levels of two or more items and the problems concerning the literal translation of a test.

We will explain the most widely used item response models and we will analyze the issue of the parameters estimation: since classical methods, as the maximum likelihood method, do not produce estimators that present desirable properties, the first point of this thesis is to show how the estimation problem can be overcome applying a complete Bayesian approach. We apply classical item response models, as 1PL and 2PL, to the PPVT-R data, estimating all the parameters using a Bayesian approach; the same estimation procedure is used to fit a more complex model involving covariates. We underline the simplicity and flexibility of this approach in terms of estimation and interpretation of the parameters.

After comparing the fit of these models and choosing the "best" of them, we will discuss the difficulties in translating a test by using the data obtained administering the Italian translation of the Peabody Picture Vocabulary Test - Revised (PPVT-R).

The comparison of item difficulties is another crucial point of this thesis: taking advan-

tages from the Bayesian approach, we will develop an *alternative* measure to compare items that allows us to analyze how severe is violation of the increasing difficulty order. This measure is based on the difference of the item characteristic curves averaged on the ability distribution. Therefore it takes into account not only the difficulty parameters but also the distribution of the ability. We will also analyze the problem of the presence of missing data due to stopping rules, underlying the advantages and flexibility of the Bayesian approach.

2 Methodology : Item Response Theory

Many people have been tested at least once in their lives. For example, in school, their abilities in mathematics, reading and writing have been tested. The use of test is of great importance when people are selected for a job on the basis of test results: these tests allow to select the more qualified person not only in terms of knowledge but also in terms of aptitude for a particular job.
Given such importance of tests, their construction must be done very carefully since a lack of quality of tests can invalidate their usefulness.
Therefore the construction of a test involves mainly psychologists and sociologists, for the definition of well formulated questions and answers, but it also involves statisticians for different motivations. First of all, the analysis of the answers to the questions must be done with appropriate statistical models; secondly, an accurate statistical analysis is necessary to check whether the assumptions made in the use of the test are respected in the data. The area of psychometry that deals with such problems is called Item Response Theory (IRT): IRT investigates which assumptions are necessary when using a test in a specific way, and develops statistical methods to check whether the assumptions are plausible for the group of people which take the test.

2.1 Background

Item response theory is a modern area of psychometry; it started to be developed in the last twenty years and it is nowadays in continuous evolution (see for example [2],

[12],[27],[13]) . Most statisticians are turning towards IRT because classical testing methods and measurements procedures have a number of shortcomings.

The most important shortcoming is that examinee characteristics and test characteristics cannot be separated. In fact, the examinee characteristics we are interested in are their abilities measured by the test: in the classical test theory framework, the ability is expressed by the *true score*, which is defined as " the expected value of observed performance on the test of interest". It implies that the *true score* is test-dependent: it means that if a test is *hard*, the examinee's ability will be evaluated as *low*, and if a test is *easy*, the examinee's ability will be evaluated as *high*.

The difficulty of a test item is defined as "the proportion of examinees who answer the item correctly".

These two definitions imply that whether an item is hard or easy depends on the ability of the examinees, and the ability of the examinees depends on whether the items are hard or easy. Hence in this approach it is very difficult to have objective measures of the ability and of the difficulty that allow comparisons between examinees who take different tests and comparisons between items whose difficulties are measured using different groups of examinees.

Two other shortcomings of the classical approach are the definition of reliability and the standard error of measurements. The reliability, defined as the extent to which a test is repeatable and yields consistent scores, can be calculated as "the correlation between test scores on parallel forms of a test" ([5]): since from a practical point of view the definition of parallel test is impossible, the available reliability coefficients provide either a lower bound estimates of reliability or reliability estimates with unknown biases.

The problem concerning the standard error is that it is a function of test score reliability and it is assumed to be the same for all examinees.

Furthermore the classical theory is test-oriented rather than item-oriented: it means that classical test theory does not enable us to make predictions about how an individual will perform on a given item ([13]).

Psychometricians turn towards IRT since one of its cornerstone is the **invariance** property and it is the major distinction from the classical test theory. This property implies that the parameters that characterize the items do **not** depend on the ability distributions

of the examinees and the parameters that characterize an examinee do **not** depend on the set of test items. Moreover IRT aims at having the following characteristics:

- item characteristics are not group-dependent
- scores describing ability are not test-dependent
- the model is expressed at the item level rather than at the test level
- the model does not require parallel tests for assessing reliability
- the model provides measures of precision for the ability scores

.

2.2 IRT assumptions

Item response theory rests on two basic postulates:

1. The performance of examinees on a test item can be explained by a set of factors called abilities or *latent* traits (*latent* because they are not directly observable)
2. The relationships between examinees' item performance and their abilities can be described by a monotonic increasing function called *item characteristic function* or *item characteristic curve* (ICC).

The ICC is a mathematical function that relates the probability of success on an item to the ability measured by the test. Many possible item response models exist, differing in the mathematical form of the ICC or in the number of parameters in the model.
One of the main properties of the item response model is that item and ability parameters are postulated as *invariant*: this property is obtained by incorporating information about the items into the ability-estimation process and by incorporating information about the examinees' abilities into the item-estimation process.
Like all statistical models, also item response models are based on particular assumptions. An assumption common to IRT models is the ***Unidimensionality***: it states that only one ability is measured by a set of items in a test. This assumption is very strict and cannot be always respected since several cognitive, personality and test-taking factors can affect the test performance. For example, tests designed to evaluate mathematical ability are,

of course, influenced by reading ability: in fact, apart from an examinee's mathematical skills, if he/she is not able to read and understand the questions of the problems, he/she will probability answer wrongly to most of questions.

What is required for the ***Unidimensionality*** assumption to be met is the presence of a "dominant" factor that influences the test performance: this "dominant" factor should be the ability measured by the test. In the example reported above, this assumption can be met, for example, asking questions orally or writing very schematic questions.

The second assumption is the ***Local Independence***: we assume that when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent, that is we are assuming *exchangeability*. It means that the abilities specified in the model are the only factors influencing examinees' response to test items. Formally, the ***Local Independence*** assumption can be written as follows: let U_i be the response to item i of a randomly chosen examinee. Let $P(U_i|\theta)$ denote the probability of the response of a randomly chosen examinee with ability θ; $P(U_i = 1|\theta)$ is the probability of a correct response and $P(U_i = 0|\theta) = 1 - P(U_i = 1|\theta)$ is the probability of an incorrect response.

The property of ***Local Independence*** states that:

$$P(U_1, U_2, ..., U_k|\theta) = P(U_1|\theta)P(U_2|\theta)...P(U_k|\theta) = \prod_{i=1}^{k} P(U_i|\theta) \tag{1}$$

Intuitively this property states that the relationships among examinee's responses to several test items are due only to the abilities influencing performance on the items: after conditioning on ability, the ability is "partialled out" and the examinees' response can be considered as independent.

The third assumption related to the **ICC**: it states that the relationships between the examinees' performance and latent traits are described by a monotonic increasing function, called ICC. It is a monotonic increasing function such that the probability of a correct response to an item increases as the ability increases.

A primary distinction among different latent trait models is in the mathematical form of the corresponding item characteristic curves.

In the following section, we will analyze a particular class of these models in which the ICC

is a logistic function. Furthermore other models, as the normal ogive item response model, are available ([17] and [3]). The logistic function has the advantage to be more mathematically tractable than other functions and it has also important statistical properties (see for details [23]).

2.3 One, Two and Three parameter logistic model

2.3.1 One-parameter logistic model

The one-parameter logistic model (**1PL**) is one of the most widely used IRT models.The ICC for this model is given by

$$logit(P_j(\theta_i)) = \theta_i - b_j \quad i = 1, 2, ..., n \quad j = 1, ..., k \tag{2}$$

where $P_j(\theta_i)$ is the probability that the examinee i with ability θ_i answers to the item j correctly, b_j is the item j difficulty parameter, k the number of items in the test and n the number of examinees.

In this logistic model, $P_j(\theta_i)$ is an S-shaped curve with values between 0 and 1 over the ability scale. For different values of the difficulty parameter, the curves vary only in their location on the ability scale and they never cross each other as shown in the figure 1. This is due to the fact that in **1PL** model it is assumed that item difficulty is the only item characteristic that influences examinees' performances. Note also that the lower asymptote of the ICC is zero: this specifies that examinees of very low ability have almost zero probability to answer correctly.

The b_j parameter is the value on the ability scale where the probability of a correct response is 0.5. This parameter can be interpreted as a location parameter, indicating the ICC position in relation to the ability. The greater the value of b_j, the greater the ability required for an examinee to have 50% probability to answer correctly.

The **1PL** model is also called Rasch model, in honor of his developer, the Danish mathematician George Rasch. The model was introduced in 1952 during his work in the Military Psychology Group in Copenhagen: as reported in [19], the model was developed in order to revise the intelligence test used by the military, the so-called IGP test (more details in [1], pages 2-24).

The main motivation for which the Rasch model is so famous is the following: it follows directly from the assumption that the unweighted sum of right answers given by a person will contain all the information needed to measure that person's ability and to calibrate items ([4]). It is the *only* latent trait model that is consistent with "number right" scoring: it means that in the Rasch model the number of correct score is the minimal sufficient statistics for the ability parameter θ. All the other latent trait models lead to more complex scoring rules that, in the estimation procedure, involve unknown parameters for which such satisfactory estimators do not exist ([13],).
Anyway, the Rasch model may not be adequate for the solution of certain measurements problems. The alternative, using more elaborate and more appropriate models, as two or three-parameter models, introduces many problems with respect to parameters estimation.

2.3.2 Two-parameter logistic model

Item characteristic curves for the two-parameter logistic model (**2PL**) are given by the equation

$$logit(P_j(\theta_i)) = a_j(\theta_i - b_j) \quad i = 1, 2, ..., n \quad j = 1, ..., k \tag{3}$$

where $P_j(\theta_i)$ and b_j are defined as 2. The parameter a_j is called discrimination parameter. It is proportional to the slope of the ICC at the point b_j on the ability scale. Items with steeper slopes are more discriminant and then they are more useful for separating examinees into different ability levels than items with less steep slopes. It must be underlined that the discriminant parameter not only fixes the slope of the curve but it determines the shape of the entire item characteristic curve. The item discrimination parameter can assume values from $-\infty$ to ∞: however negatively discriminating items are discarded from ability tests because something is wrong with an item if the probability of answering correctly decreases as examinee ability increases. As for the Rasch model, the lower asymptote of ICC is zero: it means that also **2PL** makes no allowance for guessing behavior, that is it does not take into account the fact that examinees can guess. ICC for different values of the difficulty and discriminant parameters are shown in the figure 1.

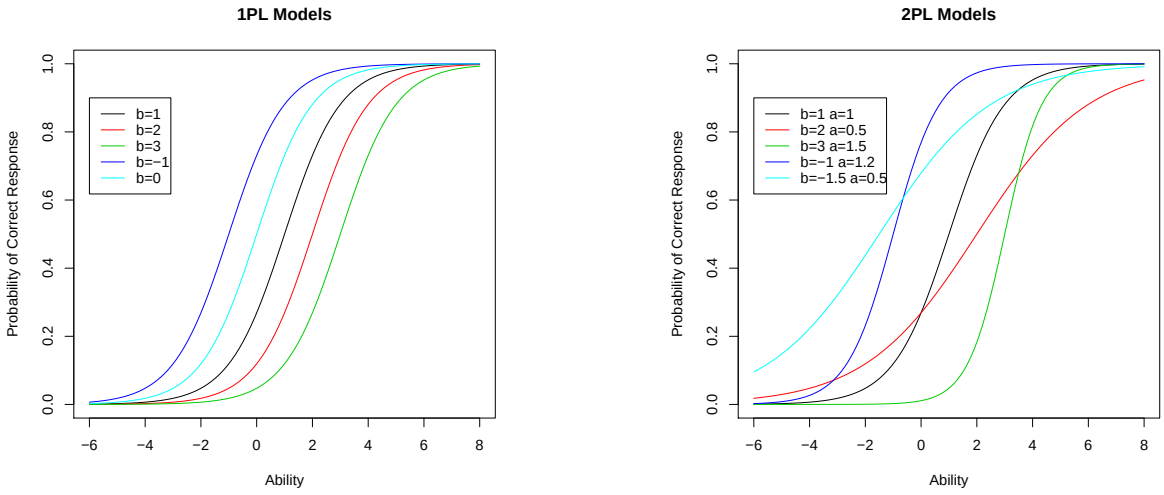

Figure 1: ICC for **1PL** and **2PL** models

2.3.3 Three-parameter logistic model

The mathematical expression for the three-parameter logistic model is

$$P_j(\theta_i) = c_j + (1 + c_j)\frac{exp(a_j(\theta_i - b_j))}{1 + exp(a_j(\theta_i - b_j))} \quad i = 1, 2, ..., n \quad j = 1, 2, .., k \tag{4}$$

where $P_j(\theta_i)$, b_j and a_j are defined as in 3. The additional parameter in the model c_i is called *guessing-parameter*: it provides a nonzero lower asymptote for the ICC and represents the probability of an examinee with low ability answering the item correctly.

2.4 Parameter Estimation

2.4.1 Classical estimation methods

One of the most important step in applying IRT is the estimation of the parameters. In IRT, the probability of a correct response depends on the examinee's ability, θ, and on the parameters that characterize the items. A complication for the estimation procedure is that θs are latent variables, that is unobservable variables: if θ were observable, the problem of the estimation would simplify and reduce to a problem similar to the regression models. Similarly, if the item parameters were known, the estimation of ability would be

straightforward.

The classical estimation procedure is based on the maximization of the likelihood, defined as:

$$f(U_1, U_2, ..., U_k) = \prod_{i=1}^{N} \prod_{j=1}^{k} P_{ij}^{U_{ij}} (1 - P_{ij})^{1-U_{ij}} \tag{5}$$

where U_i is the response pattern of examinee i to n items. The values of the item and ability parameters that maximize this function are called *joint* maximum likelihood estimates (JML). The determination of these estimates can be done in two steps: in the first step, initial values for the ability parameter are chosen. These values are then standardized and, treating the ability values as known, the item parameters are estimated. In the second step, we treat the item parameters as known and we estimate the ability parameters. This procedure is repeated until convergence.

The joint maximum likelihood procedure is quite appealing, but it has several drawbacks. First, the ability estimates with perfect and zero scores do not exist. It means that item parameter estimates for items that are answered correctly (or incorrectly) by all examinees do not exist (see [13]). Furthermore, this procedure gives consistent results for **1PL** model but, since we want to estimate item and ability parameters simultaneously, it is proved that it does not yield consistent estimates of item and ability parameters in the **2PL** and **3PL** ([21],[13]).

An alternative approach to overcome this problem is to apply the *marginal* maximum likelihood method (MML). In this method, we consider the examinees as having been selected randomly from a population; then, by specifying a distribution for the ability parameters, we can integrate them out of the likelihood function. The resulting marginal maximum likelihood estimates have asymptotic properties: the item parameter estimates are consistent as the number of examinees increases. Once the item parameters have been estimated, they are treated as known and the ability parameters can be estimated. Numerical procedure, as EM algorithm, are used to marginalize, by integration, the likelihood (see details in [5]).

However in some situations, these numerical approximation procedures may fail. It is proved that failures of the methods are quite common in **1PL** and **3PL** models. The main motivation for this problem is that the number of unknown parameters under these

models increases with the number of examinees. As underlined in [7], EM algorithm has several drawbacks: it may fail to converge, it sometimes converges to local maximum and it does not produce estimates of the standard error of the maximum likelihood estimators. Furthermore another important drawback is the fact that MML treats the ability parameter as a nuisance parameter, while in most of the case the ability parameters are the ability of interest.
An alternative approach to MML is the *conditional* maximum likelihood method in which the likelihood is conditioned on a sufficient statistics for the parameter θ: this method is applicable only to the **1PL** in which the sum of the correct score is the sufficient statistics for θ ([21]). Since such sufficient statistics do not exist for more complex models, this method is not applicable to **2PL** and **3PL** models.

2.5 Bayesian approach to IRT

As written in the previous section, maximum likelihood estimators, in general, in the latent trait models do not present desirable properties. This is mainly due to the presence in such models of "structural" and "incidental" parameters. The "structural" parameters are the item parameters and the "incidental" parameters are the ability parameters. As suggested in [31], [32] and [33], when several parameters have to be estimated simultaneously, and when both structural and incidental parameters are involved, a Bayesian solution to the estimation problem may be appropriate. This is particularly true when prior information about the parameters is available.

2.5.1 Hierarchical IRT models

In the Bayesian approach the basic idea is to consider θ and the item parameters $\xi = (a, b)$ as random variables with their own prior distributions that summarize the prior information.
Hierarchical models help in understanding such multiparameter problems since observable outcomes are modeled conditionally on certain parameters, which themselves are given by a probabilistic specification in terms of further parameters, known as hyperparameters.

1PL, **2PL** and **3PL** models can be written as hierarchical models: the first level of the hierarchy is specified by the relationships between the probability of correct response with item and ability parameters; the second level is specified by the prior distributions of the ability, difficulty and guessing parameters. The third level is defined by the prior distributions of the hyperparameters specified in the second level.
It is possible to represent such hierarchical models graphically, by using the so-called *directed acyclic graphs* (DAG): for example, under **2PL** model, the responses U_{ij} are independent, conditional on the parameters P_{ij}. For examinee i and item j, each P_{ij} is function of θ_i, of the location parameter b_j and of the slope parameter a_j. θ_i are independently drawn from a normal distribution, with mean μ and variance σ^2.

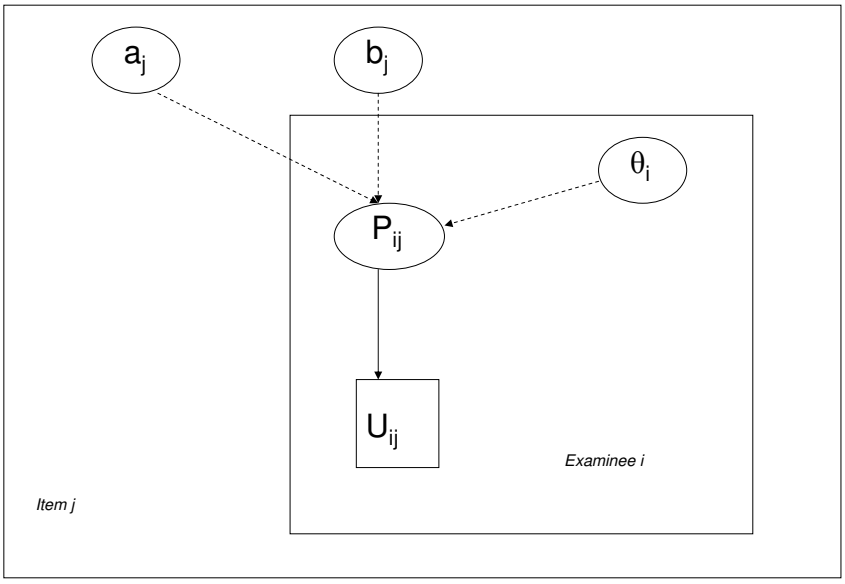

Figure 2: *A directed acyclic graph (DAG) for the **2PL** model*

Figure 2 shows a directed acyclic graph based on the following assumptions: each vari-

able in the model appears as a node in the graph. Directed links correspond to direct dependencies. The solid arrow indicates probabilistic dependency; dashed arrows indicate functional relationships. The model is directed because links between nodes are represented by arrows. It is defined acyclic for the following motivations: let v be a node of the graph and V the set of all nodes. A "parent" of v is defined as any node with an arrow extending from it an pointing toward v. A "child" of v is defined as any node on a direct path beginning from v. The directed acyclic graph is equivalent to assume that the joint distribution of all random quantities is fully specifies in terms of each node's conditional distribution, given its parents, that is:

$$P(V) = \Pi_{v \in V} P[v|parents(v)] \tag{6}$$

where $P(\cdot)$ denotes a probability distribution. This factorization not only allows extremely complex models to be built from local components, but it also provides an efficient basis for implementation of MCMC method ([11]).

2.5.2 The problem of the identification

An problem concerning latent trait models is their identification: consider **2PL** model

$$logit(P_j(\theta_i)) = a_j(\theta_i - b_j) \quad i = 1, 2, ..., n \quad j = 1, ..., k \tag{7}$$

We assume a normal prior distribution for the ability parameters, $\theta \sim N(\mu, \sigma^2)$; for the difficulty parameter b_i we assume a normal prior, $b_i \sim N(0, \delta)$ and for the discriminant parameter a_i we assume a truncated normal prior, $a_i \sim N(0, \nu)I(a_i > 0)$ where $I(\cdot)$ is the indicator function for a_i.
For the ability parameters we assume θ normally distributed with mean μ and variance σ^2: it means that we believe that the distribution of the examinees ability population is bell shaped and we also assume that the n students taking the test are a random sample from this population. The model has $n + 2k + 2$ unknown parameters. From the equation 7 we note that this model is overparametrized: in fact, one can multiply the ability parameter by a constant and divide the discrimination parameter by the same

constant and preserve the model. The usual way of removing this identifiability problem is to impose some restrictions on the item parameters, say $\prod_{j=1}^{k} a_j = 1$ and $\sum_{j=1}^{k} b_j = 0$. An alternative solution, proposed in [3], is to choose specific values for the parameter of the ability distribution. Following [3] and [32] , we set $\mu = 0$ and $\sigma^2 = 1$.

2.6 Bayesian estimation in the IRT models

Consider **2PL** model. In the Bayesian approach the basic idea is to consider θ and the item parameters $\xi = (a, b)$ as random variables with their own prior distributions that summarize the prior information. The joint posterior distribution $f(\theta, \xi|\mathbf{U})$ for these parameters is obtained by combining, through the Bayes' theorem the information from the data, represented by the likelihood $f(\mathbf{U}|\theta, \xi)$, and the prior information as follows:

$$f(\theta, \xi|\mathbf{U}) \propto f(u|\theta, \xi) f(\theta, \xi) \tag{8}$$

where $f(\theta, \xi)$ is the joint prior distribution for the parameters θ and ξ. The marginal posterior distributions of θ and ξ can be obtained by integration of the joint posterior distribution: in fact, the posterior distribution for the parameter θ,$f(\theta|\xi, \mathbf{U})$, is defined as

$$f(\theta|\xi, \mathbf{U}) = \int f(\theta, \xi|\mathbf{U}) d\xi \tag{9}$$

and the posterior distribution for the parameter ξ,$f(\xi|\theta, \mathbf{U})$, is defined as

$$f(\xi|\theta, \mathbf{U}) = \int f(\theta, \xi|\mathbf{U}) d\theta \tag{10}$$

The integration procedure, as usual, is not straightforward and several numerical methods have been developed.

In this thesis MCMC method will be applied. MCMC methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. The state of the chain after a large number of steps is then used as a sample from the desired distribution. The quality of the sample improves, of course, as a function of the number of steps. The samples can then be used to estimate functional of the distribution, typically mean, variance, standard error, HPD intervals and so on.

In particular, we will apply a powerful MCMC method, the Gibbs sampler (see [9]). In this thesis we implement the Gibbs sampler using the WinBUGS software (Medical Research Council Biostatistics Unit, Cambridge, www.mrc-bsu.cam.ac.uk/bugs). WinBUGS applies Gibbs sampling iteratively, drawing samples from the full conditional distributions of the model parameters through the adaptive rejection sampling algorithm ([10],[11]).
To check the convergence of the chains several diagnostic methods have been proposed (see [11]). In this thesis, we check the chains convergence looking at their trace plots and at the autocorrelation plots; we also calculate Geweke and Gelman-Rubin statistics and plot the results.

3 Peabody Picture Vocabulary Test - Revised Data

In this section we analyze a data set obtained administering the Italian translation of the well-known Peabody Picture Vocabulary Test - Revised (PPVT-R) to a sample of Italian children. Our main aim is to apply IRT model to the PPVT-R data focusing, in particular, on the analysis of the item difficulties. In particular we are interested in controlling if the items are in order of increasing difficulty because, as we can see, this aspect is very important for the reliability of the test results.

3.1 Introduction

Peabody Picture Vocabulary Test - Revised (PPVT-R) is an individually administered, norm-referenced test of hearing vocabulary: it is the leading measure of receptive vocabulary for standard English and it is a screening test of verbal ability.
In the original version, each form contains 175 test items arranged in order of increasing difficulty. Each item has four simple, black-and-white illustrations arranged in a multiple-choice format. The examiner provides a stimulus word orally: the subject's task is to indicate the picture which best illustrates the meaning of the stimulus word. The words included in the tests are relative to several categories: animals, man-made objects, human actions (gerunds), nature scenes, plants, inanimate objects, adverbs ...
The English version of the PPVT-R test is designed for persons of an age between 2.5

and 40 years who can see and hear reasonably well, and understand English to some degree. When used with native speakers of English, it can be used as a scholastic aptitude test, since vocabulary is a strong predictor of school success. It can be used as an initial screening device for pre-school children who may have high ability, low ability, or a language disorder. Furthermore it is also helpful in screening foreign-speaking students who planned to attend English-speaking universities.

Since the items are arranged in ascending order of difficulty, it is proved that in order to evaluate in a reliable way the testers' abilities, it is not necessary to ask all the questions, but only some of them: in PPVT-R test, subjects are tested from a *basal* [1] of eight consecutive correct responses to a *ceiling* [2] of six errors in eight consecutive questions. This stopping rule assumes that the subjects would not answer additional items correctly if the test were to be continued beyond the ceiling item. Furthermore, the presence of a well-calibrated stopping rule is quite important especially in test with a large number of items: in fact, the same budget can be used to administer the test to a larger amount of people since we suppose that most of them will not answer to all questions.

However, it must be noted that the actuality of the stopping rule is extremely relevant in determining the degree of confidence that can be placed in the results of the test: in fact, a wrong calibrated stopping rule can compromise the results of the entire test. Furthermore, as we will analyze in the next section, although the stopping rule is well calibrated, the results of the test will be anyway compromised if the items are not in order of increasing difficulty.

[1] Basal Level: for individually administered test, it is defined as the point on test, associated with a given level of skill, for which the examiner is confident that all the items prior to that item would be answered correctly (considered too easy).

[2] The upper limit of ability that a test can effectively measure or for which reliable discriminations can be made. For individually administered tests, the ceiling refers to the point during the administration, after which, all other items will no longer be answered correctly (considered too difficult)

3.2 The Italian translation of the PPVT-R

In this Chapter, we analyze the results obtained by a pilot study of a group of researchers of the university of Rome "La Sapienza". The test was administered to a sample of 2857 children keeping the items' ordering of the original version and using the same illustrations. The main aim of this study is to apply IRT models, focusing on the evaluation and the comparison item difficulties; in particular, we are interested in analyzing if the translated items are in order of increasing difficulty.
The problem of the translation of the PPVT-R test has been largely discussed in [26], in which it is translated in Mexican language: the authors underlined a series of factors that must be considered in translating a test. First of all the fact that a translated word and an original word, although expressing identical concepts, may be of different degrees of difficulty in the new and in the original languages. Secondly, a concept may be not present in the new culture; thirdly, a word may possess a single meaning in one culture but possess multiple meanings in the other. Therefore, according to [20], the literal translation must be done very carefully in the sense that the experts should translate by replacing the intended concept with one which judged to be similar. They also suggest that a re-ordering of the items would be attempted based on an analysis of the item difficulty.
In our case, we will focus on the problem of the order of the items in terms of increasing difficulty and we will try evaluate the degree of violation of this order. We focus our attention on this problem because the increasing order of the items is very influent on the stopping rule.

3.2.1 Data Description

The data set we analyze is a 2696×180 matrix: the first 175 columns are relative to the answers to the 175 items (0 for a wrong answer and 1 for a correct answer) and the last 5 columns contain personal information as age, school class, gender, the zone of Italy in which they live (North, Center, South and Isles). The age is registered in months and it ranges between 1 year and 13 years.
Classical models as **1PL** and **2PL** models will be applied to this data set; the heterogeneity of the data will be taken into account appropriately extending **2PL** model including age

and the geographic zone as covariates in the model.
As written in the previous section, the children do not answer all the 175 questions: examiners must respect a stopping rule in the sense that they stop to convey questions when the examinee totals 6 wrong answers in 8 consecutive questions. It implies that the matrix of data we consider is an incomplete matrix, because we do not have the same number of answers for each examinee. In particular, the figure 3 shows, on the left side, the proportion of children that answer respectively to 1,2,...,175 questions and in the right panel the proportion of correct responses for each item.

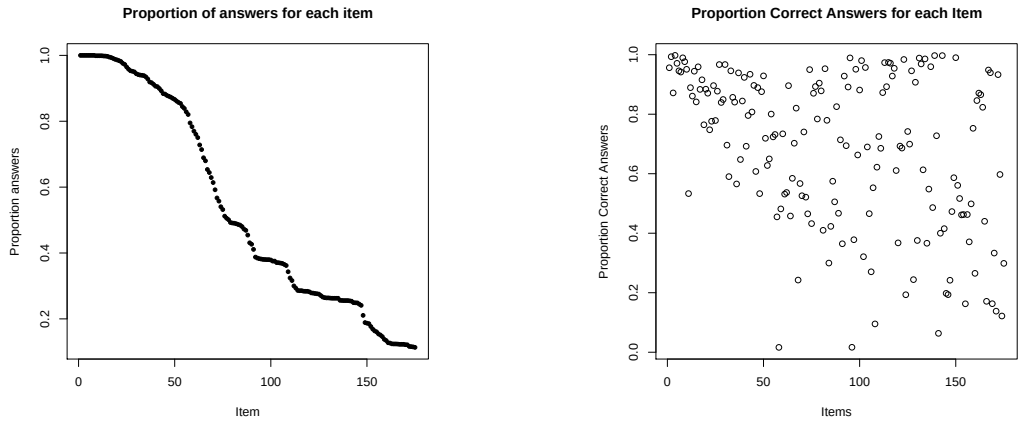

Figure 3: *Left panel: proportion of answer for each item.Right panel: proportion of correct answers for each item.*

As we can see, only the 11% of the children answer to all the 175 questions and about 50% of them answer to at least half of the questions. Focusing on the plot on the right, we can notice that if the items were in increasing order, we should see a decreasing trend in the proportion of the correct answers. At first glance, as we can see from the plots in figure 3, it seems that violations of the order of difficulty of the items are clear: in particular, items as 50, 58 and 96 have a proportion of correct responses smaller than 1%. Item 11 also seems not to be in the right position: in fact with respect to its position, it should be one of the easiest items, but the proportion of correct answers to this item is only 50%. On the contrary, for some items, for example 168, 167, 172 we have a proportion of correct answers larger than 90%.

In the next sections, we apply item response models and investigate their performance in the PPVT-R data in order to gain some insights on the ordering of items' difficulty. In the following section we focus on this problem highlighting the simplicity and the flexibility of the Bayesian solution.

3.3 Incomplete data matrix problem

The matrix of data we have to analyze is an incomplete matrix: as shown in the figure 4, we do not have the same number of observations for each subject, but we have missing responses.

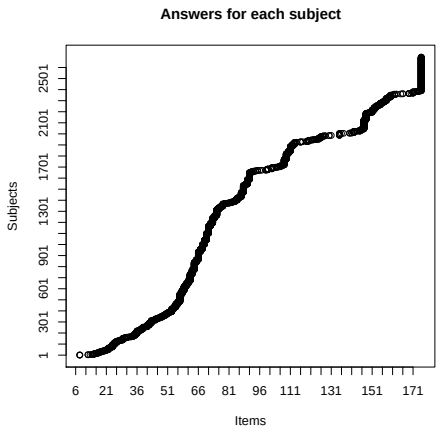

Figure 4: *Proportion of answers for each subject*

The problem of missing data has been widely treated in literature: Little and Rubin in [15] and [22] established the well-known taxonomy for the three types of the missing data processes:

MCAR (Missing Completely At Random) The missingness is independent of both observed and unobserved data

MAR (Missing At Random) Conditional on the observed data, the missingness is independent of the unobserved measurements

MNAR (Missing Not At Random) The missingness depends on observed and unobserved data

In the context of IRT, the analysis of the missingness process is quite difficult to catalogue in one of these three categories. In fact, in this framework, the missingness mechanism depends on the latent traits: it means that the probability of non response depends on the latent ability of the responder.
An interesting traditional approach to missing data problem in IRT is reported in [12]. They consider three possible answers: a correct answer, a wrong answer and a missing answer. The probabilities of correct or wrong answers are modeled using classical item response models (as **1PL** or **2PL**) while the probability of non-responses, depending only on the ability, is modeled using a multinomial regression model. Furthermore this approach is not completely adequate to our data set: in fact in PPVT-R data, the non-response can be ascribed mainly to the stopping rule which is, of course, strictly connected to the ability of the responders.
Therefore the missingness mechanism involved in our data set cannot be treated as MAR because the missingness mechanism depends on the ability by means the stopping rule; it should be treated as MNAR where the not random mechanism incorporates the stopping rule.
Bayesian and classical approaches propose different methods to solve this problem: the difference is mainly based on the fact that the two approaches consider the stopping rules in different way.

3.3.1 Stopping Rule Influence

The problem of the relevance of the stopping rules in a Bayesian and not-Bayesian approach is one of the main issues between researchers involved in these two main streams. In a Bayesian approach, inferences should depend only on the observed data, not on the reason why these data are collected. It implies that any difference in the data that makes no difference in the posterior probability distribution is irrelevant for the inference. It directly follows from the likelihood principle.
On the other hand, in the classical approach stopping rules influence inferences in a sig-

nificant way.
The following example could help us to understand the problem: consider an experiment consisting of a sequence of independent and identically distributed binary observations, as tossing a coin, in which the outcomes are labeled as 1 for a success or 0 for a failure. One stopping rule specifies that the researcher stops after one hundred observations are made, while another stopping rule specifies that the researcher stops after observing 50 successes. In a Bayesian prospective the two stopping rules do not influence the inference: in fact, since the two likelihood functions that we obtain from the two experiments differs only for a constant, then according with the likelihood principle, the posterior distributions of the involved parameters we obtain are the same.
In a classical approach the consequences of these two different designs are relevant: the main problem is that in the first design, the number of trials is fixed and the number of successes treated as a random variable. On the contrary in the second design, the number of successes is fixed and the number of trials is treated as random variable. These differences might have consequences in terms of inference and ignoring stopping rules can mislead the inferential conclusions.
Involving the stopping rule in this framework is not an easy task: in fact, in our case we can model the probability of non-response to an item as function of the stopping rule which is function of all possible responses to the previous eight questions and function of the ability. Modeling this non-response functioning seems to be quite difficult and the implementation would not be straightforward.
On the other hand, the Bayesian approach allows to solve easily the problem considering only the observed data, that is using the only available proportion of the incomplete data matrix.

3.3.2 1PL Model, 2PL Model and Model with Covariates

We apply the models described in the previous chapters to this data set, in particular **1PL** and **2PL**.
We also consider an extension of IRT models that consists of the inclusion of covariates that seems to be related to the ability of the children. The most straightforward extension

of IRT models to covariates inclusion is discussed in [2]: they argue that IRT models can be seen as multilevel models, where the first level of the model describes the relationships between the observed item scores and the ability parameters and the second level describes the relationships between the latent traits and several covariates.
In this section, we apply a hierarchical model whose first level consists of **2PL** model and the second level consists of a regression model between the latent traits and some predictors. Taking into account the fact that the design is unbalanced with respect the distribution of the age in the geographic zone, we choose to include only the variables age and gender and not to include the variable geographic zone because of a possible confounding effects with the variable age.
Therefore the first and the second level of the model are specified by the following equations:

$$logit(P_j(\theta_i)) = a_j(\theta_i - b_j) \quad j = 1, 2, ..., k \quad i = 1, ..., N \tag{11}$$

$$\theta_i = \beta_{Age} * Age_i + \beta_{Gender} * Gender_i \quad j = 1, 2, ..., N \tag{12}$$

that is equivalent to write

$$\theta_i \sim N(\beta_{Age} * Age_i + \beta_{Gender} * Gender_i, 1) \tag{13}$$

In the third level, we specify all the prior distributions. For the **2PL** model specified in the first level, we consider the same prior distributions reported in the previous chapters; for the regression parameters β_{Age} and β_{Gender} we assume flat normal distributions.

3.4 Results

In the figure 5, we report some convergence diagnostics for the parameter b_1 in the **2PL** model. Since similar plots are obtained for the other models and for all parameters, we can conclude that the chains converge and we can use the estimated parameters to make inference.

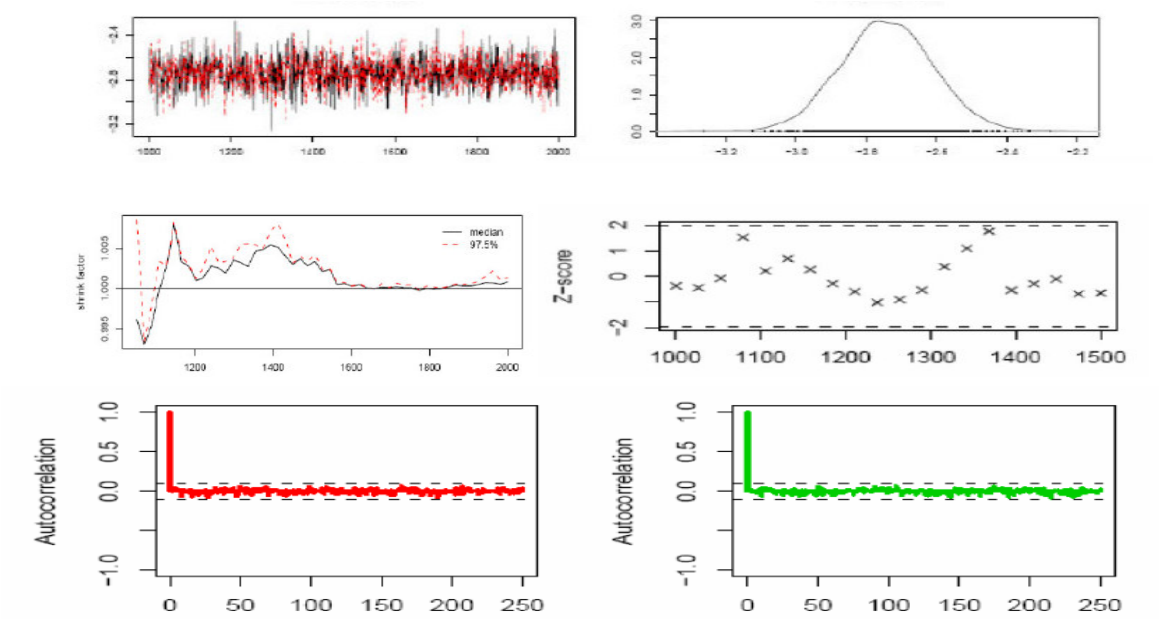

Figure 5: *Convergence diagnostics for the parameter* b_1 *for the* ***2PL*** *model: trace plots, density plot, BGR (Brooks,Gelman and Rubin) plots, Geweke and autocorrelation plot*

In the figure 6, we report the plot of the estimated posterior means for the difficulty parameter of the **2PL** model for each item: points with suspicious positions in terms of difficulty order are highlighted in red.

In the table 1, we report posterior summary statistics for the parameters *beta.age* and *beta.sex*: as we can see, the ability does not seem to be influenced by sex since the credibility interval is almost centered in 0. Furthermore, the variable age, that was centered in order to improve convergence of the chain, influences the ability: this is a quite expected conclusion since we can suppose that, for a normally intelligent person, the language skills increase with the age.

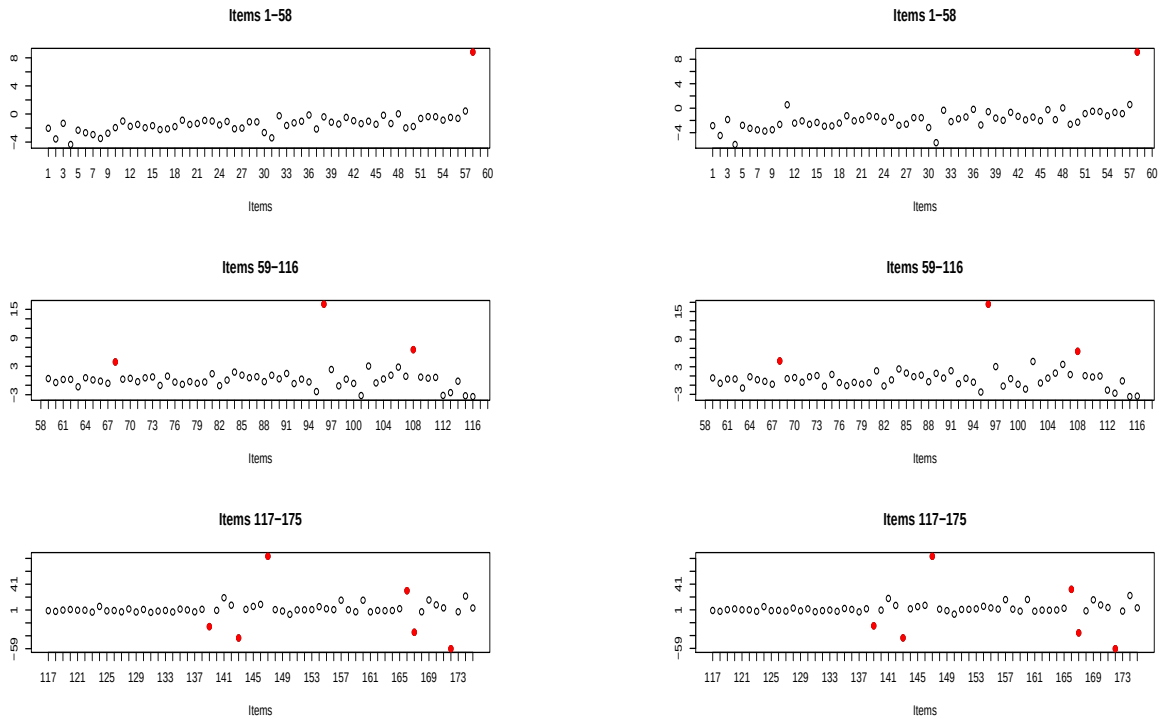

Figure 6: **2PL** *model (left panel) and model with covariates (right panel): posterior means of the difficulty parameters for each items*

Parameter	**Mean**	**Sd**	**MC Error**	**25%**	**Median**	**97.5%**
beta.age	0.04518	0.001072	0.000234	0.03813	0.04411	0.05259
beta.sex	0.009189	0.0306	0.004109	-0.05239	0.007566	0.0757

Table 1: *Model with covariates: posterior means, standard deviations, MC errors, first quartiles, medians and third quartiles for the posterior distributions of the parameters beta.age and beta.sex*

3.5 Goodness of fit and model comparison

3.5.1 Goodness of fit measures

Assessing the fit of the item response models is not a straightforward task. The main difficulty is that the possible number of response patterns (2^I for a test with I binary items) is large even for moderately long assessments, leading to sparse contingency tables, so the standard chi-square tests do not apply directly ([27]).
The situation is hardly better for IRT modeling under the Bayesian framework.
In this thesis we will apply the posterior predictive model-checking (PPMC), a popular Bayesian model diagnostic tool ([9]): it has an intuitive appeal since it is simple and can provide graphical and numerical evaluation of the model misfit. The method compares the observed data with the data predicted by the model with a number of diagnostic measures that are sensitive to model misfit. Any systematic differences between observed and predicted data indicate a possible failure of the model.
A posterior predictive distribution is defined as

$$p(y^{rep}|y) = \int p(y^{rep}|\omega)p(\omega|y)d\omega \tag{14}$$

In practice, test quantities or discrepancy measures $D(y, \omega)$ are defined and the posterior distribution of $D(y, \omega)$ is compared to the posterior predictive distribution of $D(y^{rep}, \omega)$; substantial differences between them indicate model misfit.
A quantitative measure of lack of fit is the tail-area probability also known as the PPP-value:

$$P(D(y^{rep}, \omega) \geq D(y, \omega)|y) = \int_{D(y^{rep},\omega)\geq D(y,\omega)} p(y^{rep}|\omega)p(\omega|y)dy^{rep}d\omega \tag{15}$$

Because of the difficulty in dealing with the equation 14 and 15 analytically, [9] suggested simulating replicated (or *posterior predictive*) data sets from the PPD.
Several discrepancy measures have been proposed in literature ([9] and [10]); [27] and [28] analyzed the performance of different discrepancy measures in their application to item response models.
In this thesis, we choose to evaluate the fit of the models using two of the diagnostic measures proposed in [27]: we will compare the observed and fitted scores and the observed

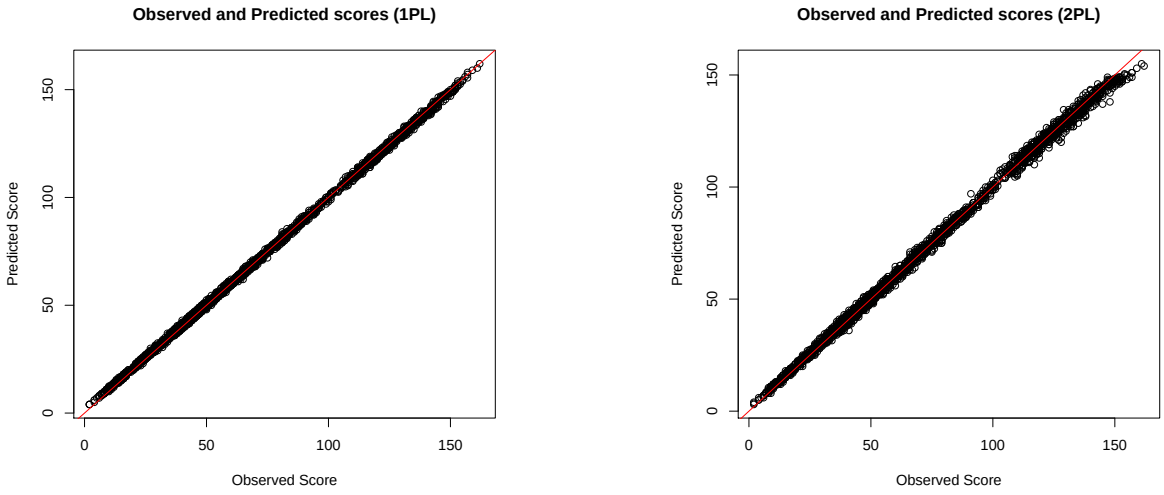

Figure 7: Observed and Predicted Scores for **1PL** and **2PL** model

and fitted odds ratio.

In figure 7, we report the plot of observed and predicted scores: as we can see, there is a good correspondence between them since the points are well aligned on the red straight line for both **1PL** and **2PL** models.

Furthermore, as observed by [27], the total scores for the items are sufficient statistics for the difficulty parameters in the simplest possible IRT model, i.e. the Rasch model. Because almost all parametric IRT models build on the Rasch model, the model-predicted scores are expected to match the observed scores, especially because each item is usually given to a large number of individuals. Therefore this measure will probably not be very powerful since it is possible that any IRT model, even one not adequate for the data, will predict these scores well. Practically, examining this measure can help to check the computations: if the computations show that the model cannot predict the scores for the item well, than will be a strong indication of some problems in the computer program used to implement the model.

Another largely used measure of fit for IRT models is the biserial correlation (see [27]): since it is proved in [27] that the biserial correlation is a powerful measure for detecting misfit of the Rasch model but it may not provide much insight about any misfit for more

complex models, we decide not to apply this measure to our data.

A widely used measure for the evaluation of the fit are the sample odds ratios: since there are no parameters in any IRT models that directly address how items interact/associate with each other, odds ratios are discrepancy measures that capture the associations among items and will probabily be effective in detecting possible model misfit.

They are defined as follows: let $n_{kk'}$ denote the number of the individuals scoring k on the first item and $k^{'}$ on the second item, $k, k^{'} = 0, 1$. The sample odds ratio is defined as

$$OR_{ij} = \frac{n_{11}n_{00}}{n_{10}n_{01}} \tag{16}$$

The quantity on the right side in the above definition is the sample odds ratio corresponding to the population odds ratio defined as

$$\frac{P(item\ \ i\ \ correct|item\ \ j\ \ correct)/P(item\ \ i\ \ wrong|item\ \ j\ \ correct)}{P(item\ \ i\ \ correct|item\ \ j\ \ correct)/P(item\ \ i\ \ wrong|item\ \ j\ \ correct)} \tag{17}$$

Since odds ratios are measures of association, their examination should help researchers to detect if the model can adequately explain the association among test items.

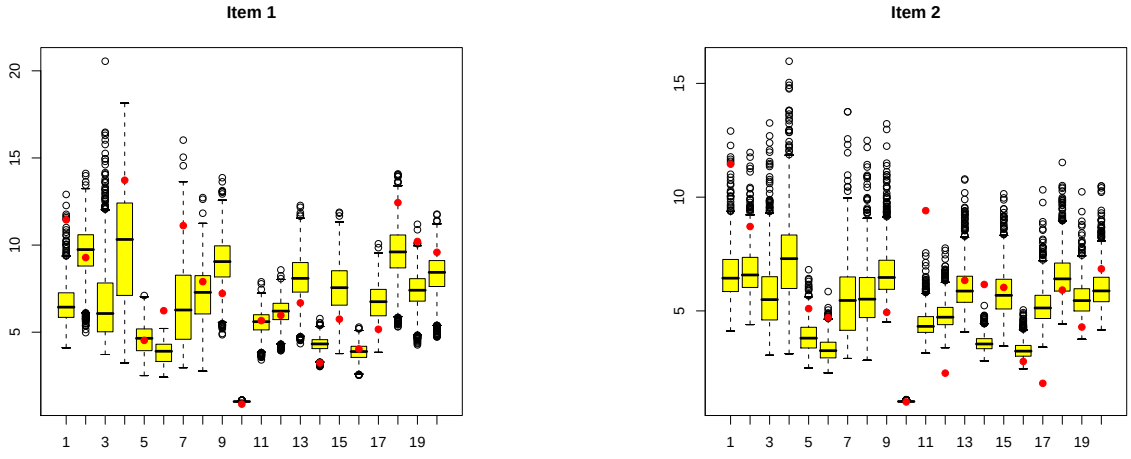

Figure 8: **2PL** *model : observed and predicted odds ratios between items 1, 2 and the first 20 items*

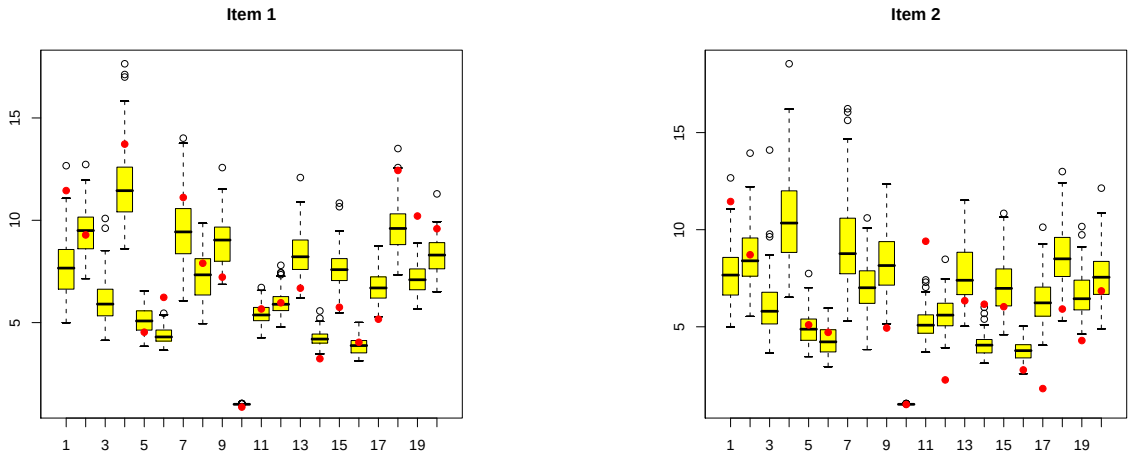

Figure 9: *Model with covariates: observed and predicted odds ratios between items 1, 2 and the first 20 items*

Figures 8 and 9 show box plots of the predicted odds ratios between items 1, 2 and the first 20 items. The red dots denote the observed odds ratios. As we can see, most of the dots lie within the 95% credible intervals for all items: the model with covariates seems to fit better than **2PL** model since more than 90% of the observed values fall in the credibility interval. The same happens for the other items.

3.5.2 Model selection criteria

In this thesis we will mainly focus on three model selection criteria, AIC, BIC and DIC; they are alternatives to the Bayes Factors that in the Bayesian framework represents the dominant method for model testing. Bayes Factors are defined as follows: suppose that we observe data X and we want to compare two model M_1 and M_2 defined by $f(x|\theta_1$ and $f(x|\theta_2)$. The Bayes factor is defined as:

$$B(x) = \frac{\pi(M_1|x)/p(M_1)}{\pi(M_2|x)/p(M_2)} \tag{18}$$

Unfortunately, while the Bayes Factors are rather intuitive, as a practical matter they are often quite difficult to calculate ([9]).
AIC, Akaike Information Criterion, is defined as

$$AIC = -2logL(\theta^*|y) + 2p \tag{19}$$

where θ^* the estimates of the parameters means (or medians or modes) in the model and p the number of parameters in the model.
Smaller AIC values are better. It does not need models to be nested; its main drawback is that it tends to be biased in favor of more complicated models, because the log-likelihood tends to increase faster than the number of parameters.
BIC, Bayesian Information Criterion, is defined as:

$$BIC = -2logL(\theta^*|y) + 2p * log(n) \tag{20}$$

where p is the number of parameters and n the sample size.
This statistic can also be used for non-nested models. Given any two estimated models, the model with the lower value of BIC is the one to be preferred. The main drawback is that BIC tends to choose models that are too simple due to heavy penalty on complexity.
DIC, Deviance Information Criterion, is a new statistics introduced by the developers of WinBUGS ([29]); it is defined as

$$DIC = Mean(-2logL(\theta_t|y)) - Mean(-2logL(\theta_t|y) - 2logL(\theta^*|y)) \tag{21}$$

$-2logL(\theta_t|y)$ is the deviance calculated as the average of the log-likelihoods calculated at the end of an iteration of the Gibbs Sampler and $2logL(\theta^*|y))$ is the log-likelihood calculated using the posterior means of θ. The second expression, $(-2logL(\theta_t|y) - 2logL(\theta^*|y))$ is the penalty for the over-parameterizing the model. It is important to note that DIC assumes the posterior mean to be a good estimate of the stochastic parameters. If this is not so, because of extreme skewness or even bimodality, then DIC may not be appropriate (see [11]).
In the table 2, we report AIC, BIC and DIC values for the three models.

Model	DIC	AIC	BIC
1PL	5816980	75007179	15938552
2PL	180983	178016.9	188941.4
Covariates	179865	177205	186561.6

Table 2: *DIC, AIC and BIC for* ***1PL****,* ***2PL*** *and* ***Covariates*** *model*

As we can see all criteria agree on the fact that the model with covariates presents a better fit of the data. Therefore in the following sections, item difficulty comparisons will be done using this model.

3.6 Item difficulty comparison

As written in the previous sections, the aim of this thesis is to analyze the difficulty levels of the items and verify whether possible violations of the increasing difficulty ordering occur.

Several methodologies have been proposed in order to compare item difficulties: in the classical approach (see e.g [13]) the comparison of the difficulty levels of two or more items is based on the comparison of the difficulty parameter estimations. An important shortcoming of this procedure is that the comparison of the difficulty of two items is based only on the comparison of the difficulty parameters estimates: it implies that the decision rule that compares item difficulties does not take into account the uncertainty of the estimates.

This problem is overcome in the Bayesian approach: in fact, the comparison of the difficulty of two items is based on the comparison of the entire posterior distributions of these two items. It means that we can obtain a comparison of the difficulty levels in terms of probability: in fact, the probability the item j is more difficult than item $j+1$ is defined as:

$$Pr(\pi(a_j|U) >= \pi(a_{j+1}|U)) = \mathbf{E}(I_{j,j+1}) \tag{22}$$

where $\pi(a_j|y)$ and $\pi(a_{j+1}|y)$ are the posterior distributions of the parameters a_j and

a_{j+1}, y the data and $I_{j,j+1}$ an indicator function that is equal to 1 if $a_{j|U}$ is larger than $a_{j+1|U}$ and 0 viceversa for each value of the posterior distributions.
It implies that the expression 22 compares item difficulties taking into account the uncertainty of the estimates.
The main drawback of this method is that it involves only the difficulty parameters, that is it takes into account only the location of the item characteristic curves but not their entire shapes. To better understand the problem, consider the following example: suppose to have item characteristic curves as those represented in the figure 10: for the curves in the left panel, the difficulty parameters for the item 1 and 2 are respectively $a = 1$ and $a = 2$, while for the curves in the right panel the parameters of the two curves are respectively $a = 1$ and $b = 1$ for item 3 and $a = 0.3$ and $b = 0.3$ for item 4.

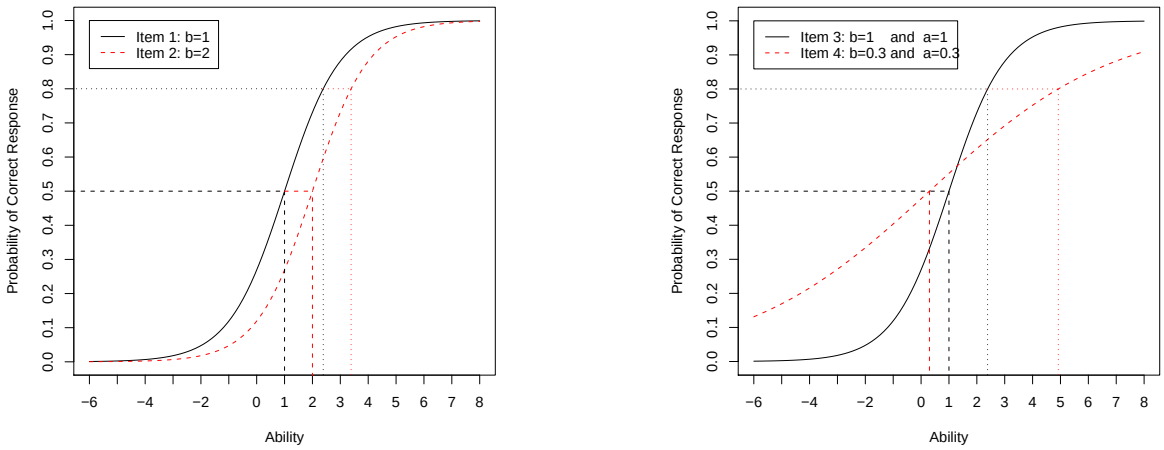

Figure 10: *Example of two item characteristic curves*

As we can see, for the item 1 and item 2 the criterion 22 let us to conclude that item 1 is easier than the item 2: in fact, as highlighted in figure 10, in which we plot the ability necessary to answer correctly to the item 1 and the item 2 with probabilities 80% and 50%, since the two curves do not cross each other, the ability one needs to answer correctly to the item 1 is always smaller than the ability one needs to answer correctly to the item 2.

When the curves cross each other, as shown in the right panel of the figure 10, applying the criterion 22 that compares the values of the difficulty parameters, we can conclude that item 2 is easier than item 1: in fact, looking at the plot in the figure 10, we can note that the ability that we need to have a 50% probability to answer correctly to item 2 is smaller than that we need for item 1. Furthermore, this conclusion does not take into account the fact, that, for example, the ability that we need to have a 70% probability to answer correctly to item 1 is larger than those that we need for item 2. We can conclude that this criterion allows us to have a partial evaluation of the item difficulty, since it is based only on the comparison of the difficulty parameters.

A possible solution of this problem has been proposed by [25], in which they compare the areas under the item characteristic curves of the two items we want to compare: from the definition of item characteristic curve, it follows that the larger it is, the easier the item is. In fact, for a very easy item, the probability of correct response is high for all the ability levels and therefore the area under the ICC is large.

Following [25], the area under the curve is calculated by integrating the expression of the ICC over the ability scale: assuming that the ability ranges in an interval (A, B), the difference between the area under the ICC of the item j and k is calculated as follows:

$$\int_A^B \frac{exp(a_j(\theta - b_j)}{1 + exp(a_j(\theta - b_j)} - \frac{exp(a_k(\theta - b_k)}{1 + exp(a_k(\theta - b_k))} d\theta \quad = \tag{23}$$

$$[\log(1 + exp(a_j(\theta - b_j))) - \log(1 + exp(a_k(\theta - b_k)))]_A^B \tag{24}$$

As noticed in [18], the problem concerning this approach is the choice of an appropriate interval for the ability, that is an opportune choice of the values A and B; in [25], according to their experience, they integrate on the interval $(-3, +3)$, but this choice is, of course, arbitrary. In fact, consider the right panel of the plot in the figure 10: if we choose to integrate on the interval $[-6, 0]$, then the item 2 is easier than the item 1. On the other hand, if we choose to integrate on the interval $[2, 6]$, we then conclude that the item 1 is easier than the item 2.

To avoid this problem, [18] suggested to integrate on the entire theta scale range, that is

on $(-\infty, +\infty)$: they proved that, under the **1PL** and **2PL** models the expression 23 is reduced to the difference between the difficulty parameters,that is

$$\int_{-\infty}^{+\infty} \frac{exp(a_j(\theta - b_j)}{1 + exp(a_j(\theta - b_j)} - \frac{exp(a_{j+1}(\theta - b_{j+1})}{1 + exp(a_{j+1}(\theta - b_{j+1}))} d\theta = b_j - b_k \tag{25}$$

It must be underlined that this equality is not true in the case of the **3PL** model, for which more complex expressions need [18].

We note that the problems discussed make sense when we have item characteristic curves that cross each other; in fact, when two curves do not cross each other, there is stochastic dominance of one item with respect to the other, that is

$$P_j(\theta) < P_{j'}(\theta) \tag{26}$$

then

$$E_{f_\theta}[P_j(\theta)] < E_{f_\theta}[P_{j'}(\theta)] \tag{27}$$

for all possible $f\theta$.

Therefore when the ICC curves do not cross each other, the inequality in 27 is true for all possible distribution of $f(\theta)$ and the criteria 25 and 23 lead to the same conclusions. On the other hand, when the curves cross each other, the inequality in 27 depends on the probability distribution of θ.

For the criterion 25 we observe that it assumes the ability uniformly distributed on the entire interval, that is the density function of θ is uniform. This assumption is in contrast with the assumptions of the item response models: in fact, item response models assume that the ability are distributed as a standard normal distribution. Therefore, the criterion 25 assumes distributions for the ability parameters which are not the same assumed by the item response model used to estimate the difficulty parameters that 25 wants to compare. In order to take into account the ability distribution in comparing item complexities, we propose the following alternative: we calculate the differences written in 25 weighted with the ability distribution, that is we calculate the following expression:

$$\int_{-\infty}^{+\infty} \left[\frac{exp(a_j(\theta - b_j)}{1 + exp(a_j(\theta - b_j)} - \frac{exp(a_{j+1}(\theta - b_{j+1})}{1 + exp(a_{j+1}(\theta - b_{j+1}))} f(\theta) \right] d\theta = E_{N(0,1)}[P_j(\theta)] - E_{N(0,1)}[P_{j'}(\theta)]$$

where $f(\theta)$ is the density function of the ability distribution. In our model, we suppose that the ability is distributed as a standard normal: therefore we assume $f(\theta) \sim N(0, 1)$.

Easy Items	Word	Difficult Items	Word
Item 172	*Presuntuoso*	Item 147	*Cassare*
Item 143	*Pescatore*	Item 166	*Ingegnoso*
Item 167	*Fratellanza*	Item 174	*Costernazione*
Item 139	*Illuminazione*	Item 141	*Laminato*
Item 150	*Svuotato*	Item 160	*Compulsare*
Item 31	*Ligneo*	Item 96	*Sassofono*
Item 4	*Scopa*	Item 157	*Timpano*
Item 2	*Palla*	Item 169	*Deciduo*
Item 115	*Medico*	Item 170	*Telaio*
Item 30	*Balena*	Item 58	*Frullare*

Table 3: *List of the first ten easier and more difficult items and their correspondent words for the criterion 25*

The expression in 28 solves the problem of the arbitrariness of integration range choice, that is the choice of the interval (A, B), and it also takes into account a distribution for the ability parameter coherent with the model assumptions.
As we can see, a closed form solution for the expression in 28 does not exist; that's why we approximate this expression using the MC method.

3.6.1 Results item comparison

Tables 3 and 4 show the results in terms of item difficulties obtained by applying the criteria reported respectively in 25 and 28.
We compare the posterior probabilities that the item i is more difficult than the item j, for $i, j = 1, 2, ..., 175$ using both criteria.

As we can see, the two criteria give similar results. Furthermore some conclusions of the criterion in 25 seem to be not plausible with respect the objective difficulty of the correspondent word. For example, the easiest item for this criterion is 172, that corresponds to the word "presuntuoso" (in English "conceited"): looking at the correspondent

Easy Items	Word	Difficult Items	Word
Item 4	*Scopa*	Item 58	*Frullare*
Item 139	*Illuminazione*	Item 96	*Sassofono*
Item 143	*Pescatore*	Item 108	*Balaustra*
Item 2	*Palla*	Item 141	*Laminato*
Item 8	*Candela*	Item 155	*Ellisse*
Item 5	*Ape*	Item 124	*Cingere*
Item 9	*Pianta*	Item 128	*Lubrificare*
Item 1	*Automobile*	Item 145	*Allettare*
Item 10	*Leggere*	Item 174	*Costernazione*
Item 16	*Collo*	Item 146	*Stame*

Table 4: *List of the first ten easier and more difficult items and their correspondent words for the criterion 28*

picture, this term cannot be considered the easiest one since only the 10% of the individuals answered correctly. The same can be said for the terms "fratellanza" (in English "brotherhood") or "svuotato" (in English "empty").
Therefore we can say that some conclusions drawn with this criterion may be misleading. For the criterion 28 one of the most difficult item is the item 58 that corresponds to the word "frullare", in English "to whip": this term does not seem to be so difficult and it is widely used in the common slang, but its picture represents a whisker, which is used not to whip but to whisk. The same for the item 96, that corresponds to the word "sassofono", in English "saxophone": the picture shows four wind instruments, that is a trumpet, transverse flute, saxophone and an horn. Therefore it is difficult for the children to distinguish the different instruments.
Particular items are also item 11, "scala a pioli" (in English "ladder"), and item 48, "mobile" (in English "furniture"). For the item 11, it is considered quite easy in the English scale, but for Italian children it seems to be quite difficult: in fact, in order of difficulty it is in the 80th position. This is mainly due to the fact that in English they have different

words to indicate the different type of stairs: they have ladder, perron, staircase ... In Italian we have only one word, "scala", and several adjectives to distinguish the different types of stairs. Furthermore, the term "scala a pioli" is not largely used.
The item 48 corresponds to the word "mobile": this item resulted quite difficult for Italian children because the word "mobile" can have several meanings: it can mean furniture but also, an more commonly, movable. The correspondent picture shows a stair, a balcony, a candelabrum and a sofa: one of the possible meaning of the term "mobile" is referred to sofa, but it is not the more common meaning.
Another difficult question is the item 160: the correspondent word is "compulsare", an Italian translation of the English term "perusing": the Italian translation of the term is, of course, right but the term "compulsare" is not of the usual language.
On the other hand, items 139 and 143 resulted too easy to stay in those positions: in fact, the word correspondent to item 139 is "illuminazione" (in English "lighting") and for the item 143 it is "pescatore" (in English "fisherman"). Both words are of common use and the correspondent pictures quite clear.
Therefore, we can conclude that, with respect our decision rule 28, there are some violations of the increasing difficulty order of the items. This violations can be due to several motivations: first of all we notice that some words are not completely opportunely translated, in the sense that they translated literally without taking into account the use of the word in the common slang. Furthermore some pictures do not represent appropriately the correspondent word.
As stated in [26], a translated word and an original word, although expressing identical concepts, may be of different degrees of difficulty in the new and in the original languages. Therefore, according to [20], the literal translation must be done very carefully in the sense that the experts should translate by replacing the intended concept with one which judged to be similar. In this case, this study can be used as a pilot study: these results can help psychologists to modify the test in terms of term translations and item ordering.

4 Conclusion

In this thesis we mainly focused on the application of the Bayesian approach to the item response theory: we highlight the advantages of this approach in terms of flexibility and applicability. In particular, we applied the classical item response models, as **1PL** and **2PL** to the PPVT-R data; the classical models have been extended to take into account and evaluate the influence of important covariates on the ability parameters.

The goodness of fit of the three models, **1PL**, **2PL** and model with covariates, have been checked using the posterior predictive p-value approach, using as discrepancy measures the observed and predicted scores and the observed and predicted odds ratios. It results that the model with covariates fits the data better than the other models, as also confirmed by the AIC, BIC and DIC indices.

Once we selected a model that fits the data in a plausible way, we focused on the problem of the item difficulty comparison: we reported several criterion proposed in the literature and we highlighted their limitations. Then we proposed an alternative method expressed by the equation 28, in which the difficulty levels of two items are compared using the entire item characteristic curves and taking into account the ability distribution.

The different criteria have been compared using the PPVT-R data and the advantages of the proposed criterion have been underlined. The difficulty of the translation of the test have been highlighted: in particular, we noticed that some words resulted more difficult for the Italian children than for the English children because of the translation was not completely adequate. Furthermore we also noticed that the ordering of the items should be adapted taking into account the degree of use of the term in the common slang.

In this thesis we chose to apply a Bayesian approach to classical item response theory: in particular, we chose to use a standard normal distribution for the ability parameters. Further extension of this work may include the analysis of possible alternative for the ability distributions and alternative decision rules to evaluate the difficulty of the items.

References

[1] Boomsma A., van Duijn M.A.J, and Snijders T.A.B. *Essays on Item Response Theory*. Springer, 2000.

[2] R.J. Adams, M Wilson, and M. Wu. Multilevel item response models: an approach to errors in variables regression. *Journal of educational and behavioural statistics*, 22:47–76, 1997.

[3] J.H. Albert. Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling. *Journal of Educational Statistics*, 17(3):251–269, 1992.

[4] L. Bertoli-Bersotti. An order-preserving property of the maximum likelihood estimates for the Rasch model. *Statistics and Probability Letters*, 61:91–96, 2003.

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

[6] B.P. Carlin and S. Chib. Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of Royal Statistical Society. Series B*, 57(3):473–484, 1995.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via EM algorithm. *Journal of Royal Statistical Society*, 39:1–38, 1977.

[8] G.H. Fischer. On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, 46:59–77, 1981.

[9] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman and Hall, London, 1995.

[10] W.R. Gilks, N.G. Best, and K.K.C. Tan. Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *J. Stat. Plan. Inf.*, 88:99–115, 1995.

[11] W.R. Gilks, S. Richardson, and D.G. Spiegelhalter. *Markov Chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

[12] Y. Goegebeur, P. De Boeck, G. Molenberghs, and G. del Pino. A local-influence-based diagnostic approach to a speeded item response theory model. *Applied Statistics*, 55:647–676, 2006.

[13] R.K. Hambleton, H.. Swaminathan, and Rogers H.J. *Fundamentals of Item Response Theory*. Sage Publications, 1991.

[14] G.H. Ironson and M.J. Subkoviak. A Comparison of Several Methods of Assessing

Item Bias. *Journal of Educational Measurements*, 16(4):209–225, 1979.

[15] R.J.A. Little and D.B Rubin. *Statistical Analysis with missing data (2nd edn)*. Wiley, New York, 2002.

[16] F. Lord. Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23:157–162, 1986.

[17] F.M. Lord. *A theory of test scores*. Psychometric Society, New York, 1953.

[18] N.S. Raju. The Area Between Two Item Characteristic Curves. *Psychometrika*, 4:495–502, 1988.

[19] G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute of Educational Research, Copenhagen, 1960.

[20] J.S. Renzulli and D.H. Paulus. A Cross-Validation study of the item ordering of the Peabody Picture Vocabulary Test. *Journal of Education Measurement*, 6(1):15–20, 1969.

[21] Item response theory: Parameter estimation techniques. *Essays on Item Response Theory*. Marcek Dekker, 1992.

[22] D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[23] L. Sanathanan. Some Properties of the Logistic Model for Dichotomous Response. *Journal of American Statistical Association*, 69(347):744–749, 1974.

[24] K. Seock-Ho. An evaluation of a Markov Chain Monte Carlo Method for the Rasch Model. *Applied Psychological Measurement*, 25(2), 2001.

[25] L.A. Shepard, G. Camilli, and M. Averill. Accounting for statistical artifacts in item bias research. *Journal of Education Statistics*, 9:93–128, 1984.

[26] A.J. Simon and L.M. Joiner. A Mexican Version of the Peabody Picture Vocabulary Test. *Journal of Educational Measurement*, 13(2):137–143, 1976.

[27] S. Sinharay, M.S. Johnson, and H.S. Stern. Posterior predictive assessment of item response theory models. *Applied psychological measurements*, 30(4):298–321, 2006.

[28] Sandip Sinharay. Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement*, 42(4):375–394, 2005.

[29] D.J. Spiegelhalter, N.G. Best, B.P Carlin, and A. van Der Linde. Bayesian measures of model complexity and fit. *Journal Of The Royal Statistical Society Series B*, 64(4), 2002.

[30] K.S. Sujit. Bayesian Estimation and Model Choice in Item Response Models. *Journal of Statistical Computation and Simulation*, 72:217–232, 2002.

[31] H. Swaminathan and J.A Gifford. Bayesian Estimation in the Rasch model. *Journal of educational statistics*, 7(3):175–191, 1982.

[32] H. Swaminathan and J.A Gifford. Bayesian Estimation in the two-parameter logistic model. *Psycometrika*, 50(3):349–364, 1985.

[33] H. Swaminathan and J.A Gifford. Bayesian Estimation in the three-parameter logistic model. *Psycometrika*, 51(4):589–601, 1986.