Some Statistical analyses of an Exam of a first course in Mathematics for Architects

Jorge Luis Bazán[†] , Sergio Camiz[‡]
[†]*Departamento de Ciencias, Pontificia Universidad Católica del Perú, Lima*
[‡]*Dipartimento di Matematica Guido Castelnuovo, Sapienza Università di Roma, Italia*
email: [†]*jlbazan@pucp.edu.pe,* [‡]*sergio.camiz@uniroma1.it*

Recibido el 31 de agosto del 2011; aceptado el 15 de setiembre del 2011

We present some statistical analyses to evaluate a data set, obtained from exams based on multiple response tests, considering two methods, based on different rationale. Tandem Analysis, an exploratory technique consisting in a Correspondence Analysis followed by a Hierarchical Classification, and the Psychometric Analysis that is based on both Classical and Item Response Theory Analysis were considered. As a case study, we used a data set of a final examination of Basic Mathematics, a test of 46 items, submitted to 180 students in Architecture. As results, the Tandem Analysis showed a relatively independent behaviour of small groups of items, correlated with at least three distinct factors, and partitions in 4 and 8 classes of the students, according to their performance. The Psychometric analysis showed that both the raw and the Rasch scores of the tests were normal, presented high reliability, and confirmed that the test structure was not unidimensional. In addition, the Item analysis indicated that the test could be improved by eliminating some items, whose behaviour was not in agreement with the others. Eventually, the exploratory analysis provides an interesting framework in which the psychometric analysis gives more details that may be taken as a guide to improve the elaboration of exams.

Keywords: Item Response Model, Tandem Analysis, Exams, Math for Architecture, Assessment.

Se presentan dos análisis estadísticos para evaluar datos obtenidos de exámenes basados en pruebas de respuesta múltiple teniendo en cuenta dos métodos de análisis, basados en fundamentos diferentes. Se consideran Análisis Tandem, una técnica exploratoria que consiste en un análisis de correspondencias seguido de una clasificación jerárquica, y el Análisis psicométrico que se basa en el análisis de ítems clásico y usando la Teoría de la Respuesta al Ítem. Como estudio de caso, se utilizó datos de un examen final de Matemática Básica, una prueba de 46 ítems respondida por 180 estudiantes de Arquitectura. Como resultado, el análisis de Tandem mostró un comportamiento relativamente independiente de pequeños grupos de ítems, en asociación con al menos tres factores distintos, y particiones de 4 y 8 clases de los estudiantes, de acuerdo con su desempeño. El análisis psicométrico demostró que tanto el puntaje fila como las puntuaciones Rasch de las pruebas fueron normales, presentan una alta fiabilidad y confirmó que la estructura de la prueba no era unidimensional. Además, el análisis de ítems indica que la prueba se podría mejorar mediante la eliminación de algunos ítems, cuyo comportamiento no estaba en acuerdo con las demás. En conclusión, el Análisis Exploratorio proporciona un marco interesante en el cual el Análisis psicométrico da mayores detalles, que pueden ser empleadas como guía para mejorar la elaboración de exámenes.

Palabras Claves: Model de respuesta al Item, Análisis de Tandem, Exámenes, Matemática para Arquitectos, Evaluación.

1 Introduction

The raise of needs of basic mathematics courses for an increasing number of university students is the result of new needs of this knowledge for careers that did not traditionally took advantage of mathematics in their original structuring. As a consequence, courses are now offered to students whose interest and ability in mathematics are normally very low, due to both their specific interests and their different frame of mind [14]. For this reason, all the teaching of mathematics might be reconsidered, at least in what concerns the service courses, that is the courses for non-mathematicians, in the programs choices, the way of lecturing, and the examination techniques.

The second author is involved since 30 years and more in such courses, so that he tried a new way of teaching to cope with the problems encountered so far: in particular, to adapt to the frame of mind of his students the topics carried out in his lecturing. An important cause of his reflection has been the broadcast of personal computers with the corresponding reduced need of manual calculus abilities. Nowadays nobody does computations by hand any more and had better rely on computer programs specific of his field of activity to perform what he/she needs. Thus, a general knowledge of mathematics might involve more the structures that constitute the different mathematical theories, a capacity of understanding its main features and their use and, in case, the ability to use effectively some computer programs. In this framework, a program of introductory courses may concern the present mathematical language and notation, based mostly on logic and set theory, algebraic and topological fundamentals, including filters for the limits theory, just to quote the most innovative topics.

In parallel with the definition of the topics, a reflection was carried out concerning the evaluation. It was

evident since many years that the classical Italian examination of mathematics, essentially based on both written exercises and theoretical oral exam, was not acceptable any longer. In particular, in the career for architects, in which the second author is involved, an average of only $15 - 20\%$ of successes (independent from the instructor) was indeed a problem for the faculty, that considered the courses of mathematics a true bottleneck for the undergraduate students.

Also, the particular situation of the Italian universities must be taken into account when considering the way the courses are carried out: specific features are the high number of the officially attending students (not always present, since the attendance is not everywhere compulsory) and the total freedom that they have in choosing their examination dates (among up to ten possible per year) and, once failed, repeat them as many times as they like, until the end of their studies. This freedom forces the instructor to perform a highest number of exams, in which the selection may be very important. For these reasons, the availability of a reliable examination tool that speeds up the process may be really helpful. In particular, attention was drawn to the multiple response tests, that may be automatically evaluated thanks to the forms scanning. If effective, they can dramatically speed up and improve the quality of the evaluation itself.

Such a test, to be technically good, should be able to:

1. prevent that a student passes the exam just by answering at random;
2. identify the least level to pass the exam;
3. scale the students according to their performance.

In the ideal case, in which all items are alike and all possible answers are either fully correct (only one per item) or fully wrong (all others, in an equal way), the simple number of correct answers would fit. In this case, the admittance level must be larger than the number of correct answers that can occur by chance at a given probability level, say 5% (e.g., for 46 items with five given answers, a minimum of 14-15 correct answers should be required). As an alternative, a penalty given to the wrong answers could force the student to think carefully which answer to choose or to leave it blank.

To remain in the framework of the closed answer tests, other possibilities may concern:

a) a weight to be given to the items, according to their relative difficulty; by no means, this is necessary when different students are submitted to different items; this may be given: a priori, according to the examiner's evaluation and a posteriori, according to the share of correct answers given to the item;

b) a scaled note given to each answer to a given item, according to their relative correctness or completeness.

It is evident, that the organization of a good testing procedure requires an important data base of items from which to extract the tests, each of them thoroughly weighed if needed. For this purpose, not only a good identification of items and answers is necessary, but a check should be performed to ensure that the weights are fair.

Classical methods of exploratory data analysis may be taken into account to analyze the tests results. In particular, referring to the classical Tandem Analysis, they are able: i) to identify possible outliers: they may be either students or answers that show a behavior totally different from the others; ii) to identify groups of items and group of students with analogous behavior: this could help in understanding which items may concern similar issues or which students found problems in some specific items; iii) to identify a suitable dimension of a sufficient representation space: this could help in understanding whether a single factor or several independent ones contribute to the final evaluation. As an alternative, several model-based methods have been developed specifically for the study of test data, as discussed in [11].

In this work, we want to compare both classes of methods. In particular we wanted to ascertain

i) to what extent a classical exploratory technique is able to analyze this kind of data and which results may be obtained;

ii) to study the quality of the test itself in its ability to evaluate in a reasonable way the student's attitude;

iii) to ascertain to what extent the model-based methods are able to give more information and more useful for an evaluation of the quality of the testing procedure.

We considered convenient to start from the teaching experience of the second author, who introduced multiple response tests at least as a part of both partial and final examinations. Thus, we examined a session of his automatic tests, submitted to the students in order to understand the effectiveness of his teaching.

For this purpose, we used two approaches that seemed to us suitable to be taken into account. As exploratory tool the Tandem Analysis [2], in this case consisting of Correspondence Analysis followed by Hierarchical Classification of the candidates: a typical exploratory tool, whose ability in synthesizing the information contained into a data table, by revealing both factors and structures, is well known. As model-based method, the Psychometric Analysis (McDonald, 1999) was chosen, which consists in the use of Classical Item Analysis and Item Response Models, specifically Rasch Model, for the evaluation of the fit to for the items in the exam.

2 The data

The data set under study concerns 180 automatic tests carried out at the end of a course on mathematics taken by freshmen in Architecture in the University of Roma, performed in February 2007, right after the end of the course (that lasted October 2006-January 2007). The course consisted of 32 lectures of two hours each on several topics, as outlined here, with the number of lectures in parenthesis: Introduction (2), Logics (3), Set theory

(6), Topology (4), Algebra (8), Analysis (7), Probability (1), and Geometry (1).

The test was focused on the six central subjects, since Geometry was taught at the very end of the course, one-two days prior the test itself. The examination consisted of 46 items of choice selection randomly picked from a data base of 60 items. The following distributions of items resulted: Logics (2), Set theory (18), Topology (8), Algebra (9), Analysis (5) and Probability (4). To each item, 5 possible answers were proposed, but in the evaluation no difference was considered between wrong and missing answer, albeit the students were communicated that the weight of a missing answer was a little smaller than that of a wrong one. In general the items are of basic information. Each candidate received the same set of 46 items, after randomization of both the items and the proposed answers. The time of the proof was an hour.

The following study is focused on the multiple response test, limited to the alternative correct/wrong, without considering the possible alternative wrong answers. Thus, the tests' data were recoded as either 1 = exact or 0 = wrong or missing. In addition, the total number of correct answers, and a tentative dichotomy pass/fail (fixed at 10 correct answers) were added as supplemental variables.

3 The analysis methods

3.1 The Tandem analysis

The exploratory technique we used is the so-called Tandem Analysis [2], that is a factor analysis followed by a classification. The method belongs to the so-called exploratory data analysis techniques [9] as a cognitive model able to suggest a possible structure of the data, based on the search of ordination gradients as factors that influence the variation of the data and classifications that allow the partition of the units according to possible sub-populations in respect to the chosen factors. Albeit severely criticized by [2], since such a procedure may not detect the true classes of a "naturally" existing partition, the method is useful in revealing how a selected number of factors may contribute to partition the units in homogeneous classes, in this way synthesizing the data structure [20]. Indeed, despite the criticism, the Tandem Analysis is broadly used, among others through SPAD package [19], in many investigation fields.

Given the presence/absence nature of our data, we applied first Simple Correspondence Analysis (SCA) [5, 17, 20] and then the hierarchical ascendant classification (HAC) [5, 20, 16] based on the Euclidean distance on the selected factor space and on the [25] criterion to aggregate the classes.

As the variance of the units' coordinates on the factors equal the corresponding eigenvalue, it must be remarked that in the computation of the distances the different factors do not play an equal role, so that the classification is more influenced by the larger factors than by the others. Each class structure may be further analyzed by selecting either the characters or the characters' levels whose value in the class is significantly extreme, in respect to an appropriate statistics, in our case the hyper-geometrical law [1]. All these computations were performed through SPAD package [19].

In the use of such a procedure, two choices are left to the user: the number of factors to take into account for the classification and the level at which to cut the hierarchy to obtain a partition. For these tasks we referenced to [22] test and [8] method respectively albeit, given the exploratory nature of our study, we did not pay a strict attention to their applicability and their results. For the same reason, in the interpretation of the classes, we used the probabilities associated to the characters only to sort them in order of importance and not as true probabilities, given the critics that this method sometimes received.

3.2 The Psychometric Analysis

The Psychometric Analysis is routine for evaluation of exams to ensure that scores are as reliable and valid as possible. It can be applied to improve or validate the tests of achievement for educational purposes. For this purpose, a large set of methods was developed based on different models, methods, and techniques. For this paper, we took an eclectic perspective, considering the exam under three different points of view: i) the definition of the scale of abilities, in order to evaluate the students' performance; ii) the item analysis, to test the quality of the questions that formed the test; iii) the test analysis [4] for an overall evaluation of the exam structure. To define the scale of abilities and the transformed scores to evaluate the student's achievement in Mathematics, that resulted from the examination under study, we compared three different models of Item Response Theory (IRT) [3] based on 1, 2, and 3 parameters. To evaluate the items we used the classic item analysis with raw scores, based both on the 1-parameter (referred as the analysis for raw scores) and the Rasch model (analysis for Rasch scores) [7, 26]. To evaluate the tests, we used unidimensionality, reliability and normality analysis.

3.2.1 Definition of the scale of abilities

For the exam, two sets of scores have been calculated: raw and transformed scores. The raw total score is the number of correct answers, that is one point for each item answered correctly. Since the items were 46, the raw scores range from 0 to 46. The raw scale is the traditional form in which the results of an exam are reported: thus, high values of the raw score correspond to a high achievement. Since on the opposite, the proportion of correct answers is inversely tied to the difficulty of the items, a high proportion of correct answers to an item corresponds to its easiness.

The transformed scores are frequently used to analyze and report the results of a test of academic performance; they are also used in international student performance assessments, such as [23]. A popular transformation of the scores is obtained through the Rasch model but other dichotomous IRT models can be considered.

The dichotomous IRT models are described according to the number of parameters upon which they depend. The 3PL is so called because of three parameters. In this

model, if in an examination students $i = 1, \ldots, n$ are submitted to $j = 1, \ldots, k$ questions, the likelihood function, assuming conditional independence, is given by

$$L(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

where $y_{ij} = 1$ if the i-th student answers correctly the j-th question and $y_{ij} = 0$ otherwise, and the probability of is given by the following logistic model:

$$p_{ij} = c_j + (1 - c_j) \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

where θ_i is the person's ability parameter and (a, b, c), are the item parameters, corresponding to discrimination, location, and guessing respectively. Here, a_j is the logistic's slope, b_j is the average (along the x axis), and c_j is the probability to guess the answer by chance.

The 2-parameter and 1-parameter models (2PL and 1PL) differ from this one in that both do not consider random answers, thus $c_j = 0$ and 1PL assumes also that the items have equivalent discrimination ability, thus $a_j = 1$. The Rasch model for dichotomous data is often regarded as an 1PL IRT model. However, rather than being a particular IRT model, proponents of the model regard it as a model that possesses particular properties and derivations.

Additional details including estimation techniques can be revised in [3]. In this paper we adopted a Bayesian perspective with the Markov Chain Monte Carlo method (MCMC), [18, 13]. The three models were fit through the use of WinBUGS software. This estimation method estimates the abilities of all subjects, including the extreme cases (those who answered either all items correctly or none). In order to compare this three models, the Deviance Information Criterion (DIC), [15], the Expected Bayesian Information Criterion (EBIC), and the Expected Akaike Information Criterion (EAIC) as showed by [6] were considered. For these criteria, the smaller is the value, the better is the model. The codes for this models were obtained using *BayesianModeling* software (Bazan, 2011).

3.2.2 Item analysis

Once identified the scale obtained assuming the Rasch model, the quality of individual items of the test was evaluated by using both the statistics of classical item analysis as those obtained assuming the Rasch model. These statistics were calculated through the programs STATA and Winsteps [21] respectively. In this late case, Conditional Maximum Likelihood Estimation was considered as described in [3].

Statistics of the classical item analysis considered are: i) the proportion of students answering an item correctly; if this is highest or lowest the item is considered to have markedly reduced discrimination power; ii) the item-total correlation test, that is, the Pearson correlation of an item to the sum of the others; this is performed to check whether any item is consistent with the rest of the items; if it is not correlated, it may be discarded; iii) the item discrimination coefficient, a point-biserial correlation between items and total scores; it should be positive, so that the students that answered correctly tend to have higher scores. iv) item non-response rate, the percentage of students that omitted or did not answer the item; this may be a sign of problems in the formulation of the item.

Statistics under Rasch models are the Infit and Outfit indicators and their z values [24].

The item-analysis report includes a series of "flags" signalling the presence of one or more conditions that might indicate a problem with an item. The following conditions, based on [23], [4], and [7], determine a flag: i) the percentage of right answers either under 20% or over 80%; ii) the non-response rate higher than 15%; iii) the item-total correlation significant at 0.01 level for a two-tailed hypotheses (that in our case of $N = 180$ corresponds to be lower than 0.192): a validity indicator; iv) the point-biserial correlation significant at 0.01 level for a two-tailed hypotheses (that in our case of $N = 180$ corresponds to be lower than 0.192): a discrimination indicator; v) Rasch goodness-of-fit indexes, Infit and Outfit, either lower than 0.9 or higher than 1.10, or InZ and OuZ values either lower than -2 or higher than $+2$.

From a psychometric point of view, an item is considered to behave poorly if it receives at least two of these flags. It should be withdrawn if it behaves poorly according to both models of analysis.

3.2.3 Test analysis

Once the item analysis is concluded and the good items are selected, the test analysis should be performed. It may evaluate the test's normality, reliability, and unidimensionality, but it is not restricted to only these aspects. The normality is not a necessary condition to consider the test's quality, but only a check to define the kind of statistical analysis that may be performed for inference purposes. The reliability is an important condition, as it checks for consistency of a set of measurements as those that compose an exam. Indeed, reliability does not imply validity: a reliable measure may measure something consistently, but not what one wants to measure. Finally, the unidimensionality is the most important premise for using the Rasch model. If a test is not unidimensional, the Rasch model should not be applied.

To evaluate the normality we used the one sample Kolmogorov-Smirnov test. The evaluation of the reliability include the Cronbach's Alpha and the Pearson Reliability Index (PRI). The Cronbach's Alpha is computed by correlating the scores of all individual items with the overall score of the test. Tests with high reliability, i.e. those with high internal consistency, will achieve an alpha equal or larger than 0.75 on the scale $[0, 1]$, where high score indicates high reliability. The minimum alpha for an acceptable reliability is 0.7.

The PRI "indicates the replicability of a person ordering we could expect if the sample of persons was given another parallel set of items measuring the same construct" [7, pag. 40]. The PRI expresses the ability of the test to discriminate sufficiently among the levels of the sample. Low values of PRI indicate either a narrow

range of person measures, or a small number of items. To increase PRI one can test persons with more extreme abilities (both high and low) or lengthen the test. Improving the test targeting may help slightly: an index of 0.9 discriminates into 3 or 4 levels; an index of 0.8 discriminates into 2 or 3 levels; and an index of 0.5 discriminates into 1 or 2 levels. On the other hand, low item reliability means that the sample is not big enough to precisely locate the items on the latent variable. The literature concerning this index is very limited (see WINSTEPS).

To evaluate unidimensionality, we used the Martin-Löf test, implemented in the SAS package [12], the factor analysis for dichotomic variables, using the tetrachoric correlation matrix, and Winstep's principal components of the correlation matrix of standardized residuals. Indeed, Orlando et al. (2000) suggest that a set of items even with several eigenvalues larger than 1, and therefore with more than one factor, be still "enough" unidimensional to be analyzed through a Rasch model. These authors argue that a test can be considered unidimensional provided that enough items -say 80%- have loadings larger than 0.35 on the first factor. According to [10], even a test with several factors can be considered unidimensional, on condition that the first extracted component explains at least 40% of the total inertia.

4 Results

4.1 Tandem Analysis

The SCA shows a pattern of eigenvalues rather regularly descending, so that the identification of a suitable cut-point is not so easy. As the data table is composed by 0s and 1s it is very sparse and its chi-square has no sense. In any case it results non-significant, so that neither the goodness of fit nor the [22] tests should be applied. Nevertheless, the partial chi-squares associated to the first three factors are all significant at 5% level, so that we decided to limit the study to these first factors, that summarize 16.45% of the total inertia: a value rather low, a sign of the limited power of this analysis in this kind of study.

The examination of the high contributions of the items to these factors (as well as the further ones) is usually limited to very few items: for the first axis, D15 (ambiguous, concerning the inverse of an injective function), D05 (ambiguous, concerning the properties of the equality among integers), D02 (concerning the application of a logical rule), and D16 (asking whether the square is injective) have the highest contributions, from 21 to 10%, summarizing over 63% of the total axis inertia. In practice, they appear to be opposed to all others, that are in a rather compact cloud around and on the other side of the origin. Indeed, items D02 and D05, that received the least exact answers are the farthest from the origin, but the supplemental variables, all of them increasing with the number of good answers, are oriented in the opposite directions, since the number of answers played a more important role in the evaluation than the answers difficulty. Indeed, the axis represents certainly an axis of success, considering the significant correlation (0.33) of the total number of correct answers with this axis. The second axis marks the opposition between the items D15 and D05, whose contribution summarize 42%, with the items D01 (logic), D36 (limit as accumulation point), D23 (on the dimension of an affine manifold), and D02 on the side of D15 and D19 (when a function has its inverse) on the side of D05. This may only be interpreted as some co-occurrence of exact answers to these items. On the third axis, the highest contributions are given by D27 (are generators of a vector space independent?) and D21 (affine varieties as translations of subspaces), with a lower contribution of D05, summarizing 38% of the total inertia. Once again, this may be interpreted as a co-occurrence of exact answers to these items.

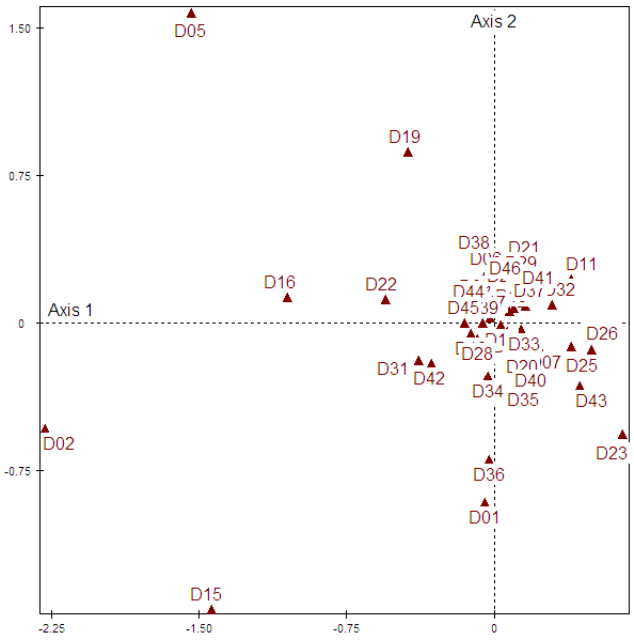

Figure 1. *Correspondence analysis of the good-bad answers to the tests: representation of the items on the plane spanned by the factors 1 and 2.*

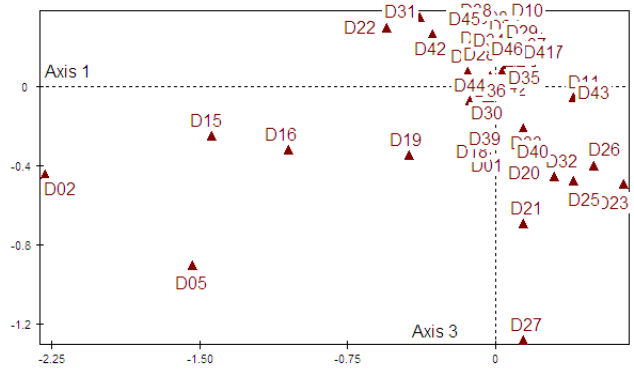

Figure 2. *Correspondence analysis of the good-bad answers to the tests: representation of the items on the plane spanned by the factors 1 and 3*

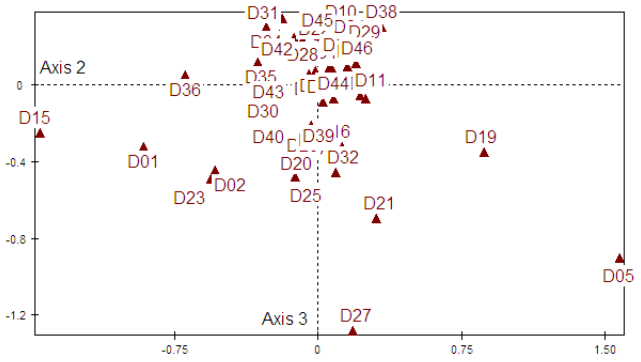

Figure 3. *Correspondence analysis of the good-bad answers to the tests: representation of the items on the plane spanned by the factors 2 and 3.*

In Figures 1, 2 and 3, the scatter diagrams of the items on the factor planes spanned by the factors 1-2, 1-3, and 2-3 respectively are reported: it is evident that no principal pattern exists, shared by a large set of items, but rather polarization of few items along each axis. This explains the low variation of the eigenvalues' size. As a consequence, only very few items are well represented in this 3-dimensional space, say D15, D05, D27, D21 with a percentage ranging from 68 to 36%. An interpretation of this situation could be that these items resulted the more difficult, with very few right answers; their scattering on the plane should mean that not the same students answered these items correctly.

To study the distribution of the students in this 3-dimensional space, we used a hierarchical classification based on the Euclidean distance in this space and the [25] minimum inertia criterion; as cutpoint we applied the [8] criterion. Two encapsulated partitions in 4 and 8 classes appeared to be of interest, the latter resulting by cutting in two each class of the previous one. For their interpretation, we selected the items to which the students in the class answered either correctly or wrongly with a frequency in the class significantly different from the overall frequency at a 1% significance level, all the non-quoted being non-significantly different from the overall means. It must be emphasized that the outlined values represent tendencies and not a common behaviour in the class. The classes may be described as follows:

Class 1 of 4 (55 students) The average of good answers of these students is 25.2 against 20.7 of the total. Their preferred good answers were D27, D26, D25, D21, D32 (all concerning linear manifolds), and D23 (connection), whereas they failed very often the answers D05 and D15. The class may be subdivided into: Class 1 of 8 (35 students) These students answered correctly to items D23 (manifolds), D40 (limit of a filter base), and D26 (Grassman's rule) and badly to D05 and D15, with totals of correct answers approaching 27. They are the best students, so that we may suspect that they answered wrongly to the said items because of their ambiguity. Class 2 of 8 (20 students) These students answered correctly to D21 and D27 and badly to D31 (convergence of a filter base), with a total of 22, thus lower than the previous class. Class 2 of 4 (65 students) The totals of these students are in average lower than in general, only 18 against 20.7. They answered more correctly to items D10 (symmetric difference of sets) and D38 (definition of derivative), but worst than the mean to D15, D16, D27, and D05. They may be subdivided into: Class 3 of 8 (39 students) These students answered more badly to items D15, D27, D16, and D05. Class 4 of 8 (26 students) These students totalled only 13 good answers, much lower than the average. Nevertheless, those who failed the exam are here less than in the total. It is interesting that they distinguished for their correct answers to items D38, D10, D08 (easy: definition of a set), and D29 (inverse of a matrix). Class 3 of 4 (27 students) The students answered correctly to nearly 18 items, lower than the general mean. Despite of this, they answered correctly to items D05, D16, D19 (when a function has an inverse), and D22 (definition of linear manifold). They may be subdivided into: Class 5 of 8 (15 students) All students passed the exam, answering correctly to items D16 and D5. Class 6 of 8 (12 students) These students have low totals (13). As a consequence, the percentage of rejected is higher than the average. Nevertheless, they answered well to items D05, D19, and D02. Class 4 of 4 (33 students) These students too answered correctly to less than 18 items in average, but answered well to items D15 and D02 and failed the item D21. They may be subdivided into: Class 7 of 8 (25 students) All these students passed the exam, with a good answer to D15. Class 8 of 8 (8 students) Here the worst students are found, totalling 14 good answers in average, thus with many non-received. Nevertheless they answered correctly to items D15 and D02.

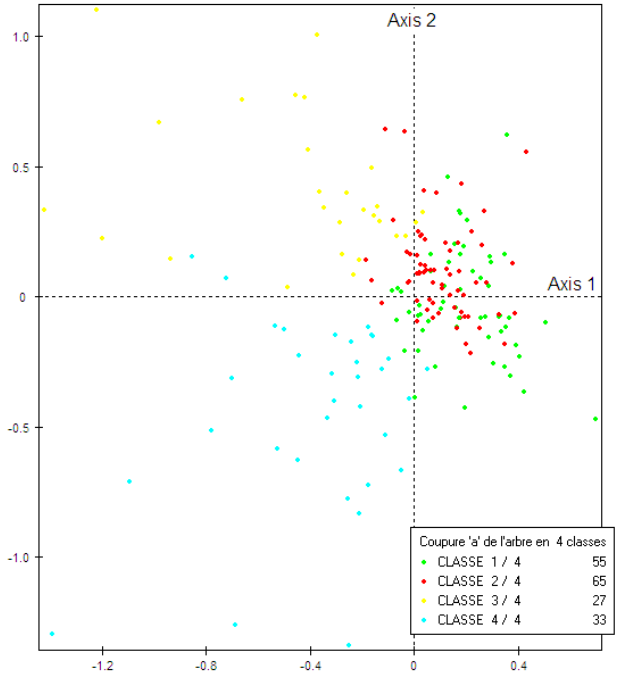

Figure 4. *Correspondence analysis of the good-bad answers to the tests: representation of the students on the plane spanned by the factors 1 and 2.*

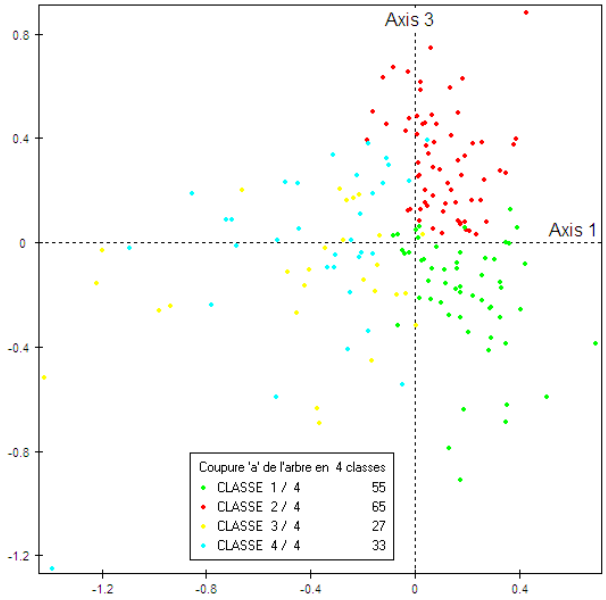

Figure 5. *Correspondence analysis of the good-bad answers to the tests: representation of the students on the plane spanned by the factors 1 and 3*

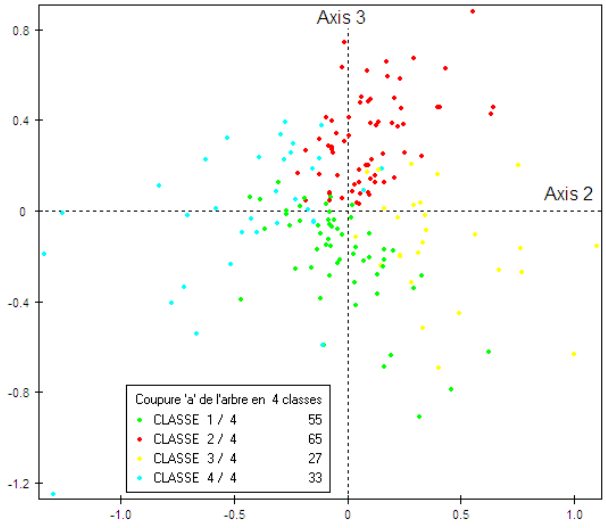

Figure 6. *Correspondence analysis of the good-bad answers to the tests: representation of the students on the plane spanned by the factors 2 and 3.*

In Figures 4, 5 and 6, the scattering of the candidates on the three considered factor planes is represented: the digits indicate to which class of 4 the candidate is attributed.

The pattern of the units on the factor planes shows a rather uniform distribution, with some evidence of the two classes 3 and 4 of 4 that are distinguished from the other on the first axis and between themselves on the second, and of the 1 and 2 of 4 that are separated along the third axis, so that the four classes appear well distinguished on the factor plane spanned by the factors 2 and 3. On the opposite, the structure of the eight classes partition seems much more complicated and is not reported here.

Summarizing, at a first glance, the results of SCA appear far from exhaustive for the description of such a data structure, if any. Of course, this may depend on the structure of the exams itself, without a common strong structure that may be clearly identified. Thus, one may be doubtful of a common behavior of the students, as well as a clear way to identify an order tied to their true quality.

4.2 Psychometric Analysis results

4.2.1 Definition of the scale of abilities

In Table 1 the results are shown of the application of the three IRT models to the exam. According to DIC criterion, the better model results 3PL, according EAIC 2PL is the best, and 1PL is the best according to EBIC. Thus, there is no consensus about the best model to choose. In addition, the estimation of abilities under the three models is similar, as is showed by their highest correlation structure, reported in columns 5 and 6: the least correlation is between 3PL and 1PL (Rasch model) and is really very high ($r = 0.974, p < 0.001$).

Based on these results, we decided to work with the first scale using the Rasch model, because when the students are evaluated to monitor their progress, it is more important to have scores for all subjects than only for those who can provide information for the model (namely, the non-extreme scores). Also, the first scale was preferred over the others by parsimony since a similar information is provided by a simpler model. Setting the scale mean at 100, a scale ranging from 50 to 150 was created, with a standard deviation of 15: the values for the items will be shown in Table 2. However, guessing parameter in 3PL is considered for Item analysis.

Table 1. *Performance of Bayesian model comparison criteria to select the best IRT model*

	Bayesian Models Comparison Criteria				Correlation structure to abilities	
Models	DBAR	DIC	EAIC	EBIC	2 PL	3 PL
1 PL	8883	9085	9335	1880163	.985	.974
2 PL	8736	8968	9280	2260896		.990
3 PL	8728	8942	9364	2641768		

r pearson correlation between estimated abilities under this models.

Table 2. *Item Analysis Index for the Math Exam (N=180)l*

	Rasch model *					Classical Item Analysis				3 PL Model
Item	Diff **	IN. MSQ	IN. ZSTD	OUT. MSQ	OUT ZSTD	% correct answer	Non response	Item rest corr.	Point Biserial correlation	Guessing probability ***
D01	118.16	1.01	0.05	1.12	0.54	0.16	0.11	0.31	0.26	0.10
D02	139.04	1.12	0.41	1.47	0.88	0.05	0.12	0.05	0.02	0.06
D03	73.87	0.94	-0.74	0.88	-0.89	0.72	0.11	0.38	0.33	0.24
D04	85.84	0.92	-1.49	0.85	-1.89	0.56	0.17	0.47	0.41	0.21
D05	125.23	1.27	1.41	1.84	2.41	0.11	0.26	-0.04	-0.08	0.10
D06	98.21	0.90	-1.52	0.91	-1.09	0.38	0.27	0.49	0.44	0.18
D07	102.38	1.03	0.34	1.03	0.24	0.33	0.32	0.35	0.30	0.18
D08	79.50	0.96	-0.60	0.95	-0.50	0.65	0.08	0.39	0.33	0.22
D09	78.26	1.00	-0.03	0.94	-0.52	0.67	0.09	0.36	0.31	0.25
D10	88.13	0.92	-1.56	0.94	-0.71	0.53	0.13	0.47	0.42	0.23
D11	115.56	0.99	-0.07	0.99	-0.04	0.18	0.14	0.33	0.29	0.11
D12	81.53	0.88	-2.06	0.81	-2.13	0.62	0.12	0.49	0.44	0.20
D13	76.55	0.94	-0.81	0.98	-0.16	0.69	0.07	0.38	0.32	0.21
D14	78.26	0.93	-1.11	0.92	-0.75	0.67	0.10	0.42	0.36	0.21
D15	116.83	1.37	2.57	1.75	3.07	0.17	0.31	-0.09	-0.14	0.14
D16	118.16	1.13	0.92	1.38	1.59	0.16	0.31	0.17	0.12	0.12
D17	91.95	0.98	-0.40	0.95	-0.61	0.47	0.24	0.41	0.35	0.21
D18	99.02	0.96	-0.60	0.92	-0.84	0.37	0.23	0.43	0.37	0.16
D19	120.30	1.14	0.92	1.17	0.70	0.14	0.31	0.16	0.11	0.11
D20	95.04	0.95	-0.86	0.94	-0.70	0.43	0.24	0.44	0.39	0.19
D21	99.85	0.94	-0.85	0.94	-0.59	0.36	0.23	0.44	0.39	0.18
D22	120.30	1.19	1.21	1.40	1.53	0.14	0.26	0.07	0.02	0.12
D23	118.16	1.01	0.10	1.25	1.10	0.16	0.41	0.28	0.23	0.11
D24	87.75	1.05	0.83	1.03	0.37	0.53	0.17	0.33	0.26	0.23
D25	101.95	0.85	-2.11	0.80	-2.09	0.33	0.38	0.55	0.50	0.13
D26	99.85	0.88	-1.80	0.85	-1.67	0.36	0.36	0.51	0.46	0.16
D27	116.83	0.96	-0.32	1.05	0.26	0.17	0.24	0.37	0.32	0.09
D28	96.21	1.16	2.51	1.26	2.87	0.41	0.24	0.20	0.14	0.23
D29	78.26	1.01	0.22	0.98	-0.17	0.67	0.07	0.34	0.28	0.26
D30	85.45	1.04	0.65	1.00	0.01	0.57	0.10	0.35	0.29	0.28
D31	112.05	1.08	0.72	1.02	0.15	0.22	0.31	0.28	0.23	0.14
D32	109.39	0.88	-1.24	0.84	-1.19	0.24	0.28	0.49	0.45	0.09
D33	98.21	0.87	-2.13	0.81	-2.30	0.38	0.34	0.53	0.48	0.14
D34	99.02	1.02	0.29	1.06	0.66	0.37	0.38	0.36	0.3	0.18
D35	100.26	1.02	0.30	1.00	0.00	0.36	0.43	0.36	0.31	0.19
D36	108.88	1.01	0.10	0.99	-0.07	0.25	0.32	0.35	0.30	0.15
D37	96.61	1.00	-0.03	1.02	0.21	0.41	0.27	0.39	0.33	0.18
D38	88.90	0.97	-0.49	0.97	-0.42	0.52	0.21	0.41	0.35	0.21
D39	104.60	0.97	-0.34	0.89	-0.95	0.30	0.37	0.42	0.36	0.14
D40	107.40	0.96	-0.47	1.08	0.57	0.27	0.37	0.40	0.35	0.15
D41	101.10	0.93	-1.02	0.95	-0.49	0.34	0.36	0.46	0.40	0.17
D42	113.17	1.19	1.62	1.19	1.04	0.21	0.18	0.16	0.11	0.15
D43	106.92	0.88	-1.38	0.87	-1.03	0.27	0.17	0.50	0.45	0.12
D44	94.26	0.97	-0.58	0.94	-0.80	0.44	0.18	0.42	0.36	0.18
D45	82.72	0.95	-0.91	0.88	-1.39	0.61	0.17	0.43	0.38	0.26
D46	90.04	1.05	0.91	1.01	0.14	0.50	0.09	0.34	0.28	0.27

Statistics obtained in Winsteps.** Mean of scale was set at 100. *** Under the 3PL model (used to estimate c_j), an item is susceptible to guessing when its "guessing probability" is equal to or higher than the inverse of the number of alternatives. In this case, since there are five alternatives, an item should not have a guessing probability equal to or higher than 0.20.

4.2.2 Item analysis

In Table 2 the different index obtained by both the classical item analysis and the Rasch model are shown. The first five columns report the index of fit under the Rasch model, that is the estimated item's difficulty, the Infit and Outfit indexes with their associated z test-values. The following four columns report the results of classical item analysis: the percentage of correct answers, the non-response rate, the item-total correlation and the point-biserial correlation. In addition, the last column shows the guessing parameter c_j under the 3PL model, that is used to estimate the guessing probability.

From the inspection of both Infit and Outfit indices, it results that, for a fixed a $5 - 95\%$ probability interval in which to consider the item acceptable under the Rasch model, the items D05 (equality relation), D12 (product of sets), D15 (injective function), D25 (vector subspace dimension), D28 (vector space basis dimension), and D33 (homeomorphism) lack the fit to the Rasch model. In addition, the items D12, D25, and D33 resulted too easy, whereas the items D5, D15, and D28 resulted too difficult.

By considering that the percentage of correct answer for each item should be within the interval $(0.2, 0.8)$ to accept the item, we found that the items D01, D02 (both on logics), D05, D11 (product of cats and dogs), D15, D16 (square function), D19 (existence of inverse function), D22, D23 (dimension of a manifold), and D27 (generators of a vector space) may not be accepted according to this criterion, as the number of correct answers is too low: for this reason they are not able to sufficiently discriminate the students' ability. The selection would be higher considering that only 13 items have non-response rate lower than 15%. This is an alert that may not be ignored, as it depends either on the formulation of the question or on the inadequate preparation of the students, or on both.

In the following two columns, the item-total and the Point-Biserial correlations are reported. 0.2 is the lower limit for both statistics to consider acceptable the item: thus, D02, D05, D15, D16, D19, D22, D28, and D42 (continuous function) have too low correlations to be considered consistent with the other items.

In the last column the guessing probability under the

3PL is reported. In our case, in which five alternatives are proposed, an item should not have a guessing probability equal to or higher than 0.20, the inverse of the number of alternatives. Thus, the items D03, D04, D08, D09, D10, D12, D13, D14, D17, D24, D28, D29 (inverse of a matrix), D30 (filters), D38 (derivative), D45, and D46 (both on probability) there is a chance that the candidates answered at random, this way hoping to catch the correct answer by chance.

Summarizing, the item analysis shows that most items in the test do not meet all criteria for a good psychometric behavior. Indeed, only 18 items are acceptable according to all tests performed.

4.2.3 Test analysis

The test analysis include the evaluation of the test's normality, reliability, and unidimensionality. The results show that the distributions of both raw and Rasch scores are normal, that there are no significant differences in reliability among them, and that the test is not unidimensional.

Normality For both the raw and Rasch scores we found the Kolmogorov-Smirnov test values of $Z = 1.298(p = 0.07 > 0.05)$ and $Z = 0.991(p = 0.28 > 0.05)$ respectively. In both cases the distributions are close to the normal. Also, both the Histograms and the Normal Q-Q plots of both distributions confirm that the normality is maintained despite the outliers, especially those scoring at the bottom of the distribution.

Reliability To test the reliability we used, for the raw scores, the Cronbach's Alpha coefficient and, for the Rasch scores, the Winsteps' PRI, which is equivalent to the traditional Cronbach's Alpha (once the extreme cases have been excluded). The results indicate that the test reliability is equivalent for both scores, that is moderate: $\alpha = 0.851$ and $PRI = 0.84$, respectively. Both coefficients are above the lowest recommended threshold (0.7), so that we may say that the reliability of test is sufficiently good.

One-dimensionality In Table 3 are reported some the results of some analyzes run to test the underlying unidimensionality of the 46 items under examination.

Table 3. *Results of the Uni-dimensionality Analysis.*

Indicator	Methods	
	AF*	ACP**
Number of factors with eigenvalues over 1	12	5
% variance explained by the first factor	29.26	5.63
% cumulative variance	85.24	23.78
% items with loading over 0.35 in the first factor	84.78	15.22

* Factor Analysis of dichotomic items using tetrachoric correlation matrix

** Principal Component analysis of standardized residual correlations for items under Rasch model

All the used criteria point out the existence of several latent factors. Indeed, they do not agree on the suggested dimension: the Scree plot, based on factor analysis of the tetrachoric correlation matrix, indicates twelve factors with eigenvalues larger than 1; the Winstep's principal components of the correlation matrix of standardized residuals reveals five factors, the first of which explains only less than 6 percent of the total variance. Both are much more than the three factors suggested by the Malinvaud test for correspondence analysis, but nonetheless they confirm more dramatically the non-unidimensionality of the exam.

In summary, the results show that the distributions of both raw and Rasch scores are normal, there are no significant differences in reliability using raw or Rasch scores, but that the test is not unidimensional.

5 Discussion and conclusion

The particular situation of Italian university studies, with its total freedom to choose the examination dates and to repeat freely the failed exams, raises the problem of a highest number of exams to carry out along the academic year. Thus, the availability of an effective examination tool to speed up the process is most appreciated, on condition that it is coherent with the lecturing and able to correctly estimate the students' level.

The aim of this study was to evaluate the quality of the multiresponse test proposed to the freshmen students in Architecture at the end of a non-traditional course. The reason was to understand to what extent the proposed concepts, that belong to the fundamentals of well established theories, could be retained by the students and, at the same time, if the proposed items are consistent for the purpose.

The results showed in section 4, in particular the highest rate of non-responses, indicate that the level of preparation of the students was not adequate. One may wonder if it depends on the quality of the course itself or on the scarce interest that introductory courses in mathematics arouse, even the least traditional ones. Indeed, the results of this kind of examination are in line with the traditional ones, but the same a general improvement of the lecturing, a selection of the questions and/or a better formulation could be of help. Concerning the latter point, a possible improvement of this test would be to remove the items D02, D05, D15, D16, D19, D22, D28, and D42 to obtain a shorter but more consistent version. It is interesting to observe that nearly all of these items are separated from the others on the negative side of the first factor of correspondence analysis (Figure 1). This might indicate some coherence between the two techniques, albeit the exploratory analysis results are not so detailed and rich in information as those of the psychometric anal-

ysis. On the opposite, the position of this "bad" items on the correspondence analysis axes, joint with their evaluation through the psychometric indexes, could contribute to the axes interpretation.

In general, improvements are needed in the examination under consideration from the standpoint of design and topics of mathematics that seeks to assess the type and format of items used. In particular, a larger item data base, composed of subsets homogenous on the point of view of the topics would be a better background from which to extract the items in each occasion in a balanced way. The control of difficulty of the items could be another interesting feature.

Eventually, the adoption of analysis tools to assess the quality of the tests incorporating the methodology presented can be an important help in this regard. Whereas the exploratory analysis could be useful to remove the questions further from the others on the factor space, but without a deep knowledge on the reasons underlying this choice, the adoption of psychometric analysis methods is helpful to understand in detail the problems tied to some questions and fix them. In this sense, even if it is more difficult to handle, the psychometric analysis is an important complement to the exploratory one, due to its better evaluation of both items' and students' particular features.

Acknowledgements

This work was granted by both institutions to which the authors are affiliated. In addition, the second author was granted by the bilateral agreement between Sapienza Università di Roma and Universidad Nacional de Ingenieria de Lima, and by the Istituto Italiano di Cultura in Lima. All institutions are gratefully acknowledged.

1. Agresti A (1992)."A survey of exact inference for contingency tables". *Statistical Science*, 7(1), 131-153.
2. Arabie P, Hubert L (1994). "Cluster analysis in marketing research". In R Bagozzi (ed.), *Advanced methods of marketing research*, pp. 160-189. Blackwell, London.
3. Baker F, Kim S (2004). Item Response Theory. 2 edition.
4. Bazán JL, Millones O (1998). "Evaluación psicométrica de las pruebas CRECER 98." *Análisis de los Resultados y Metodología de las Pruebas Crecer*, pp. 171-195.
5. *Benzécri J (1973). L'Analyse des données.* Dunod, Paris.
6. Bolfarine H, Bazán JL (2010). "Bayesian Estimation of the Logistic Positive Exponent IRT model." *Journal of Educational Behavioral Statistics*, **35-6**, 693-713.
7. Bond T, Fox C (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Lawrence Erlbaum, Philadelphia, PA.
8. Calinski T, Harabász J (1974). "A dendrite method for cluster analysis". Communications in Statistics Theory and Methods, **3**(1), 1-27.
9. Camiz S (2001). "Exploratory 2- and 3-way Data Analysis and Applications". *Lecture Notes of TICMI*, **2**. Http://www.emis.de/journals/TICMI/lnt/vol2/lecture.htm.
10. Carmines E, Zeller R (1979). *Reliability and validity assessment*, volume 17. Sage Publications, Inc, London.
11. Chatterji M (2003). *Designing and using tools for educational assessment.* Allyn and Bacon, Boston, MA.
12. Christensen K, Bjørner J (2003). "SAS macros for Rasch based latent variable modelling". *Technical report 13*, Dept. of Biostatistics, University of Copenhagen.
13. Fox J (2010). *Bayesian item response modeling: Theory and applications.* Springer Verlag, New York.
14. Gardner H (1985). *Frames of Mind: The Theory of Multiple Intelligences.* Basic books, New York.
15. Gelman A, Carlin J, Sterman H, Rubin D (2004). *Bayesian Data Analysis.* Chapman and Hall/CRC, Boca Raton.
16. Gordon A (1999). *Classification.* Chapman & Hall/CRC, Boca Raton, FL.
17. Greenacre M (1984). *Theory and Application of Correspondence Analysis.* Academic Press, London.
18. Kim S (2001). "An evaluation of a Markov chain Monte Carlo method for the Rasch model". Applied Psychological Measurement, **25**(2), 163-176.
19. Lebart L, Morineau A, Lambert T, Pleuvret P (1999). *SPAD – Systéme Pour L'Analyse des données.* Cisia-Ceresta, Paris.
20. Lebart L, Morineau A, Piron M (1995). *Statistique exploratoire multidimensionnelle.* Dunod, Paris.
21. Linacre J (2009). *Winsteps (version 3.68) [Computer Software].* Winstep.com, Beaverton, OR.
22. Malinvaud E (1987). "Data analysis in applied socioeconomic statistics with special consideration of correspondence analysis". In Marketing Science Conference. HEC-ISA, Joy en Josas.
23. OECD (2005). "PISA 2003 technical report". http://www.oecd.org/dataoecd/49/60/35188570.pdf. (downloaded May 20th, 2010).
24. Stone BWM (1979). "Best Test Design. Rasch Measurement." MESA Press, Chicago, IL.
25. Ward J (1963). "Hierarchical Grouping to optimize an objective function". Journal of American Statistical Association, **58**(301), 236-244.
26. Wilson M (2004). *Constructing measures: An item response modeling approach.* Lawrence Erlbaum, Philadelphia, PA.