Self-evaluation test for student guidance: the case of the University of Bologna

Mariagiulia Matteucci[1], Stefania Mignani
Statistics Department
University of Bologna
Via Belle Arti 41,
40126 Bologna, Italy
E-mail: m.matteucci@unibo.it ; stefania.mignani@unibo.it

Roberto Ricci
Regional Institute for Educational Research
Via Ugo Bassi, 7
40121 Bologna, Italy
E-mail: ricci@irreer.it

Abstract In this work, the theme of student guidance in the university context is outlined. In order to introduce the competence evaluation for student guidance, a new guidance project based on self-evaluation tests of the University of Bologna (Italy) is presented. In the paper, the issues of item specification and test calibration are explored, with reference to item response theory models. Models are for multiple-choice and binary items and they assume that a single ability is responsible for the student performance in the test. The test calibration is conducted for the Statistics Faculty and the results are discussed, together with the future developments of the project.

Keywords: competence evaluation, item response theory, student guidance, test calibration.

INTRODUCTION

In the educational field, the concept of guidance has become fundamental. Generally, the word “guidance” has a double meaning: it can be defined as an individual process able to develop instruments for decision or it may characterize a set of interventions that act on individuals to support them in the same decision process. With reference to the latter definition, guidance can be viewed as a powerful tool for schools and universities to improve their formative path.

[1] Correspondece should be addressed to Mariagiulia Matteucci, Statistics Department, University of Bologna, via Belle Arti 41, 40126 Bologna (Italy). Phone: +39.0512094628, fax: +39.051232153, e-mail: m.matteucci@unibo.it

With reference to the university system, in the last few years an increasing importance has been given to many guidance initiatives supporting the entire student's career. Based on the student life cycle, three different moments of guidance can be distinguished: before entrance (choice of athenaeum, faculty, and degree course), *in itinere* (organization of teaching and services as tutoring, libraries, and student facilities), and after the degree conferring (job training, placement, and working experiences). Furthermore, the last stage implies the measurement of the agreement between the degree and the job in terms of job and salary satisfaction, progression of the career, and use of the acquired skills.

In this paper, we will focus on the entry guidance, which is associated to the very first phase of the student life cycle: the choice of the faculty. The entry guidance is essential in order to introduce students into an appropriate formative path which should both turn out in a satisfactory study career and prevent difficult learning and dropping out. In particular, attention will be given to the case of the University of Bologna (Italy). Recently, the Italian University has undergone a complex process of reformation that led to the proliferation of degree courses within the single faculties. As a consequence, many universities have developed several proposals to support high school students in the choice of both the faculty and the degree course. It is well known how much a wrong choice of the faculty may compromise the student's career in terms of performance and satisfaction: therefore, the introduction of effective guidance instruments has become essential.

The paper is organized as follows. First of all, the main guidance instruments in the entrance phase concerning university are reviewed, and the concept of competence test as a new possibility for guidance is introduced, with reference to the case of the University of Bologna. Secondly, the topic of competence evaluation is developed both by giving some methodological notes on building appropriate tests in order to evaluate abilities and by illustrating the problem of test calibration. In particular, the item response theory (IRT) approach is considered through two different models: the multiple-choice model (Thissen & Steinberg, 1984) and the three-parameter logistic model (Birnbaum, 1968). Then, the results of the test analysis and calibration are presented for the Faculty of Statistics. Finally, some suggestions for further development of the project are discussed.

STUDENT GUIDANCE AT UNIVERSITY: THE CASE OF BOLOGNA

The comparative choice of the faculty is a crucial phase in the career of all the students. For this reason, guidance instruments are provided by the faculties to integrate students' individual process of decision. In this section, the classical instruments for entry guidance are reviewed. Then, a new guidance project of the University of Bologna is presented.

Student guidance is a fundamental activity for all the universities. Usually, universities provide guidance services in order to help students with a wide range of educational matters and offer specialist support. Of course, the first and most important guidance instrument is information. Students are always informed about different faculties and degree courses, examinations, university regulations, and so on. Even if good information is a valid starting point for entry guidance, it should not be the only instrument to be used. This is the reason why many faculties developed further methods to strengthen the guidance initiatives as psychological

support, meeting with students, and submission of aptitude questionnaires.

As far as the University of Bologna is concerned, the main guidance initiatives can be synthetized as follows:

- Guidance Days;
- Open Days;
- Guidance Service.

The Guidance Days represent the connection point between high school students and the University of Bologna. The meeting is organized in the fair of Bologna and lasts from two to three days. During this period, all the faculties have the possibility to present their didactic programmes to students by using oral presentations, practical activities, and multimedia. The event is usually very successful because students can get much information about all the faculties at the same time. If they are completely undecided, they have the opportunity to learn about different academic programmes while, if they already have a favourite faculty, they can talk with enrolled students, PhD students, and professors. Guidance is also oriented to high school teachers in order to illustrate them the possible university paths and make them aware of new opportunities. On the contrary, the Open Days are managed autonomously by the single faculties. During these days, typically twice in a year, high school students are invited both to visit the faculty and to experiment some practical activities, which pertain the specific subjects. Finally, the Guidance Service includes a group of psychologists who organize different activities for students, from the Guidance Days to individual interviews and aptitude questionnaires. Some of these initiatives try to overcome the role of guidance as simple information in order to involve the student actively and require his/her personal efforts in the decision process. In particular, the role carried out by the Guidance Service is fundamental for a linear and successful decision process. Individual interviews and aptitude tests are valid instruments to let the students understand their passions and interests; furthermore, they are useful occasions of debate and exchanging of views. Despite all the efforts, the problem of student guidance is still crucial in the University of Bologna. This is a direct consequence of the complex academic structure, in terms of degree courses and geographical configuration. In fact, the data referred to the 2007/2008 academic year show that there are 23 different faculties containing 252 degree courses. The number of enrolled students is around 90.000. Furthermore, Bologna has a multi-campus university, with a central structure and four separate campuses in Cesena, Forlì, Ravenna, and Rimini. With reference to these problems, the guidance process turns out to be fundamental especially to prevent students' dropping out caused by a wrong choice of the faculty.

Besides the classical methods for student guidance, the University of Bologna has carried out a new project, coordinated by the Guidance Service, in order to improve guidance in the entry phase. The project focuses on the coherence between the initial abilities of the student and the contents of the degree programmes. The project investigates competence of secondary school students in order to support them in the guidance phase of the faculty choice. The tool used is a faculty-specific test made up of general culture and specific knowledge multiple-choice items. Therefore, the distinguishing feature is that the set of items evaluates competences rather than aptitudes. Competence tests are widely used in educational testing to evaluate single or multiple abilities. Because the competence assessment has been highlighted as a more powerful instrument in predicting the student performance respect to the usual aptitude evaluation, a self-

evaluation test has been introduced also for student guidance. In this sense, we expect that students can draw information from their performance in the test to take a more aware decision. The items contain questions related to the peculiar subjects of each faculty. Therefore, this instrument should be used by the students to understand both if they are really interested to these subjects and if they are able to succeed or need further studies. The test is online and it is computer-based. Furthermore, the test is supported by an automatic evaluation system, which provides the students with the number of correct responses for each section and the possibility of correcting the test.

The online test consists of two sections. The first one includes general culture items and it is common to all the faculties that accepted to take part in the project. In more details, for each respondent, 10 general culture items are randomly selected from an item bank of 30 items by using a block design. The items concern five main topics: actuality, civic culture, geography, and both general humanistic and scientific subjects. The second section of the test is faculty-specific and consists of 20 fixed items referring to the contents taught in each faculty. In both sections the items are multiple-choice with 5 alternatives, with only one correct answer.

The data have been collected in the period from May 2006 to May 2007. The number of respondents for the nine faculties which joined the project is presented in Table 1. These faculties do not represent the entire formative offer of the University of Bologna, but all the others will be included in the next stages of the project.

Table 1 Number of respondents

Faculty	*Respondents*
Agriculture	360
Arts and Humanities	2314
Economics	1263
Education Science	565
Foreign Languages and Literature	1748
Pharmacy	1300
Political Science	3246
Psychology	1498
Statistics	324
Total	12618

The initial phase of the project has been carried out in two different moments: the item formulation according to specific rules and the item calibration on a group of respondents. A calibration analysis has been conducted on the administered tests by using the item response theory approach. The models implemented are described in the following section, after a review about the features of multiple-choice items.

COMPETENCE TESTS: ITEM FORMULATION AND TEST CALIBRATION

The accurate formulation of an item in a test

An appropriate item formulation of a test is a process that requires a careful pondering on several aspects that are equally important. It is unfortunately quite common that the item writing phase is underestimated; therefore, the quality and the amount of information given by a test is poor. A good item formulation is a circular process that foresees several aspects with item changes on the basis of the results of the calibration.

As known, the test theory offers a large variety of items which responds to different research purposes. In this section, the attention will be turned to multiple-choice items because the online test consists of such items. A standard multiple-choice consists of two basic parts: a problem, usually called *stem*, and a list of alternatives which contains one correct answer and a number of incorrect alternatives, the so called *distractors*.

The first important step in the formulation of a multiple-choice item test is to distinguish between objectives which can be appropriately assessed by using this kind of items and objectives which would be better assessed by some other means. Multiple-choice items have advantages and limitations just as any other type of test items, they are appropriate for the application in many different knowledge areas, and they are adaptable to various levels of learning outcomes. Moreover, multiple-choice items are easily amenable to item analysis, which enables the researcher to improve the item by replacing distractors that do not function properly.

In order to write well structured multiple-choice items, some basic rules have to be respected. The first one is the identification of the item objective: it is opportune that every item tends to evaluate only one educational aim, which has to be clearly expressed in the stem. It is definitely better that the stem is stated in positive form. Positive items are more appropriate to measure the attainment of most educational objectives. Another relevant phase in the item construction is the design of the alternatives. Response alternatives that overlap create undesirable situations. Moreover, if the alternatives are too much heterogeneous, the student's task becomes unnecessarily confusing. Alternatives that are parallel in content help the items to present a clear-cut problem which is capable of measuring the attainment of a specific aim. The alternatives should be similar as much as possible. Similar in length, in grammatical formulation, in answer suggestions, and so forth.

At the end of this very short and not exhaustive formulation of the main guidelines for a good item preparation, it may be opportune to reflect on the frequently used alternatives "all of the above" and "none of the above". These two alternatives are usually inserted in the item when the test writer has trouble coming up with a sufficient number of distractors. Such writers emphasize quantity of distractors over quality. Unfortunately, the use of either of these alternatives tends to reduce the effectiveness of the item (Haladyna & Downing, 1989).

Models for test calibration: the item response theory approach

The assessment of student performance is a crucial issue in the educational testing. In a given phase of a learning process, the competence evaluation can be typically carried out by analyzing the results of a questionnaire containing a set of items related to the ability to be assessed. Since ability is not directly observable and measurable, it is referred to as a *latent trait* assumed to underlie the test results. The evaluation of a latent trait can be achieved by using the item response theory (IRT) approach, commonly implemented in educational testing. IRT is a measurement theory whose roots can be traced back in the thirties and forties but it was first formalized in the sixties with the fundamental work of Lord and Novick (1968). Nevertheless, IRT has been intensively applied only recently, especially in the educational field.

Simply, an IRT model describes the trace line or conditional probability of a response given the latent variable, for an item with categorical responses (Thissen & Steinberg, 1986). Therefore, the relationship between the *observable* examinee's performance in the test and the *unobservable* latent ability is synthetized. The predominant role of IRT in testing motivated the decision to perform the calibration of the guidance tests by using IRT models. Within the test development, the calibration phase consists in the analysis of the item properties, in order to select proper items to be included in a test. When the items are correctly calibrated and they are set on the same scale, the estimated item parameters are taken as known and used to characterize the latent ability for examinees who produce a particular response pattern.

In order to perform the item calibration, many IRT models can be implemented. Mainly, the choice of suitable models depends on the data structure. In our context, multiple-choice items suggest the use of a model supporting nominal polytomous data. Nevertheless, data reduction from polytomous to dichotomous may be useful to overcome the complexity of a model for polytomous responses. For these reasons, two models are used simultaneously in the calibration phase: the multiple-choice model (MCM) developed by Thissen and Steinberg (1984) in order to analyse the behavior of multiple-choice items and their specific response alternatives, and the three-parameter logistic (3PL) model (Birnbaum, 1968) to characterize the item properties with binary data. The models are for observed item response data and the latent trait involved is considered a random variable, in the context of marginal maximum likelihood (MML) estimation (Bock and Aitkin, 1981).

Consider a set of k items with m categorical response alternatives. Furthermore, consider a completely latent response category *"0"* to take into account the so called "totally undecided individuals", i.e. the examinees who don't know the correct answer and guess. According to the MCM, the relationship between the response y to item j, with $j=1,\ldots,k$, and the latent ability θ is expressed through the following logistic transformation

$$P(y_j = h \mid \theta) = \frac{\exp(\alpha_h \theta + \delta_h) + \gamma_h \exp(\alpha_0 \theta + \delta_0)}{\sum_{l=0}^{m} \exp(\alpha_l \theta + \delta_l)}. \qquad (1)$$

Therefore, the probability of responding in the category h, with $h=1,\ldots,m$, conditional to the latent ability, depends on a slope parameter α_h and on an intercept term δ_h. Each parameter is referred to a specific item j but here we use a

reduced formulation of the model to keep the specification simple. The response probability of the "don't know" category is

$$P(y_j = 0 \mid \theta) = \frac{\exp(\alpha_0 \theta + \delta_0)}{\sum_{l=0}^{m} \exp(\alpha_l \theta + \delta_l)}. \qquad (2)$$

Consequently, a parameter γ_h is introduced in (1) to represent the unknown proportion of individuals who choose each option randomly. This parameter is allowed to be a function of estimated parameters. The α_h's represent the ordering between the options: for well calibrated items, the correct response has the biggest positive value while the other alternatives are associated to low and intermediate estimates. Particularly, what we expect as the ability increases is that the correct response has an increasing probability of been selected while the distractors have a decreasing probability to be chosen. We also allow a non-monotonic trend for the incorrect response curves, i.e. increasing for low ability levels and decreasing for high ones. The δ_h's reflect the selection relative frequency; in fact, for alternatives with similar values of α_h, those with larger δ_h are chosen more frequently.

Model (1) is principally used in the item calibration to perform a preliminary graphical analysis of the different response alternatives for each item in the test. In fact, it can be adopted usefully to create a response curve for each alternative as a function of ability. When the correct alternative has a monotonic increasing S-shaped response curve and the incorrect options are associated with non-monotonic or decreasing trends, the item is well described by the model. The MCM presents several complexities due to the high number of parameters involved in the estimation process. Besides, in a practical context we are more interested on whether students could identify the correct answer. Therefore, a model for binary data, which is more stable and easy to interpret, is considered.

The 3PL model can be applied to dichotomous items, with $m = 2$ response categories. Usually, the data are coded as "1" for a right answer and "0" for a wrong one. Once more, the existence of a group of individuals who don't know the correct answer to item j is considered. The probability of a correct response to item j is described by the 3PL model as follows

$$P(y_j = 1 \mid \theta) = \gamma_j + (1-\gamma_j)\frac{\exp[D\alpha_j(\theta-\beta_j)]}{1+\exp[D\alpha_j(\theta-\beta_j)]}, \qquad (3)$$

where α_j, β_j, γ_j are the item parameters and $D = 1.702$ is a scaling constant, so that the model is set in the *normal metric*.

The interpretation of the item parameters is quite straightforward. The α_j is the *discrimination parameter*, i.e. it reflects the capability of the item to differentiate between the examinees with different ability levels. The higher the α_j is, the more discriminating the item and the steeper the item characteristic curve (ICC) are. In fact, from a geometrical point of view, α_j is proportional to the slope of the ICC at the point $\theta = \beta_j$.

The β_j represents the *difficulty parameter* for the item and its values are collocated on the same scale of θ. The β_j is a location parameter because it defines the position of the ICC respect to the ability values. Particularly, as it increases the ICC moves to the right, i.e. a higher level of ability is required to have the same

probability of a correct answer. On the other hand, as the β_j decreases the ICC moves to the left side.

Finally, the γ_j is called *guessing parameter* or, more precisely *pseudo-chance level parameter* (Hambleton & Swaminathan, 1985). Geometrically, it is the lower asymptote of the ICC and it represents the probability of examinees with low ability to correctly answer the item *j*. According to the 3PL model, the probability of a correct response is never zero because a guessing factor is introduced. For this reason, the point on the horizontal axis where the β_j is equal to the ability level θ corresponds to $(1+\gamma_j)/2$ on the vertical axis. In this case, the probability of a correct response is exactly the mean value between the highest and lowest probabilities of success. In the IRT applications to educational assessment and knowledge tests, the guessing parameter should never be excluded. In fact, the hypothesis of not guessing examinees is not reliable in this context while it may be likely for psychological tests.

THE CALIBRATION OF THE STATISTICS FACULTY TEST

In this section, data from the 20 specific items of the Statistics test are considered. The items concern different subjects: mean values (1-4), probability (5-9, 19), logic (10-14), and data interpretation (15-18, 20). The number of respondents is 324, collected from May 2006 to May 2007. The 51,85% of the sample consists of males while the 48,15% of females. Furthermore, the 62,65% of the students comes from high school (*Italian liceo*), the 33,33% from polytechnic school, and the 4,02% from vocational school.

The calibration has been conducted by using the software Multilog 7.0 (Thissen, 2003). All the models are estimated with the MML method via the EM algorithm (Dempster et al., 1977). The method of contrasts with the deviation matrices has been chosen to impose constraints on the item parameters. Missing are treated as missing at random (MAR).

First of all, the number of correct, incorrect and omitted responses for all the items is shown in Table 2.

Table 2 Percentage of correct, incorrect and omitted responses

Item	*Correct*	*Incorrect*	*Omitted*	*Item*	*Correct*	*Incorrect*	*Omitted*
1	79.94	15.12	4.94	11	85.80	10.49	3.70
2	78.09	16.05	5.86	12	46.60	48.77	4.63
3	66.98	29.32	3.70	13	58.02	37.35	4.63
4	82.72	10.80	6.48	14	62.35	34.26	3.40
5	57.41	37.96	4.63	15	40.43	54.94	4.63
6	40.12	54.94	4.94	16	47.84	48.46	3.70
7	61.42	33.95	4.63	17	91.67	4.32	4.01
8	55.25	40.74	4.01	18	88.89	5.56	5.56
9	66.98	28.70	4.32	19	62.65	33.33	4.01
10	78.40	16.36	5.25	20	29.01	67.90	3.09

We can notice that the items present different levels of difficulties: in fact, the percentages of correct responses have a wide range of variation, from 29.01 (item

20) to 91.67 (item 17). The results suggest that the total test score, in terms of sum of correct responses, may not be a valid indicator of student's performance.

Secondly, the MCM is implemented to understand the item behavior. The response category curves have been estimated for the 20 items. As an example, Figures 1 and 2 show the curves for item 4 and item 6, respectively. Item 4 has a proper behavior while item 6 is not a good one. The horizontal axis represents the ability values, typically from –3 to 3, and the vertical axis the probability range [0,1]. A curve for each response alternative is presented, expressing the probability of being selected as a function of ability according to the (1). The correct alternative is represented by the scattered line. Item 4 (see Fig. 1) strictly respects the MCM features, because the scattered curve is monotonic increasing and S-shaped while the curves associated with the distractors are decreasing or non-monotonic. Therefore, the correct alternative is not preferable for low ability levels but it is associated with an increasing probability of being selected as ability increases.

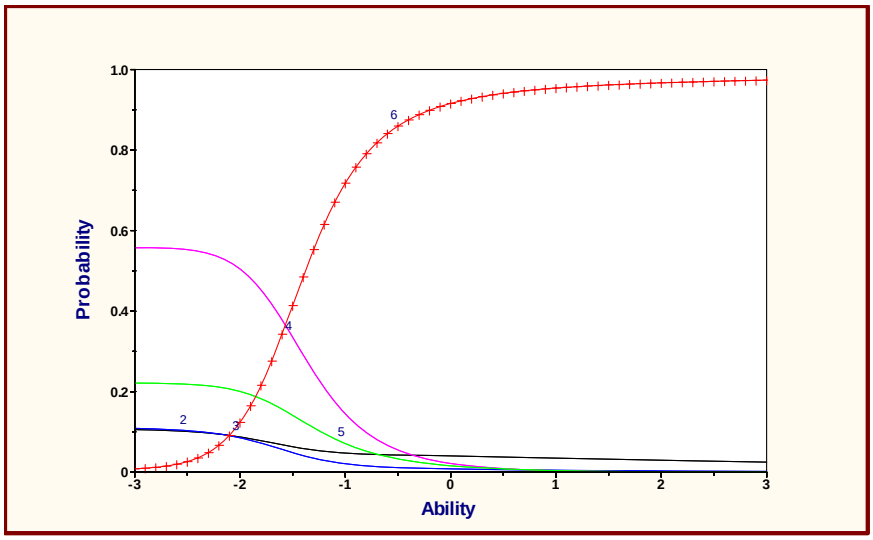

Fig. 1 – Response category curves, item 4

On the other hand, item 6 is a clear example of a not proper item with a non-monotonic trend for the correct response, especially decreasing for high values of the latent trait (see Fig. 2).

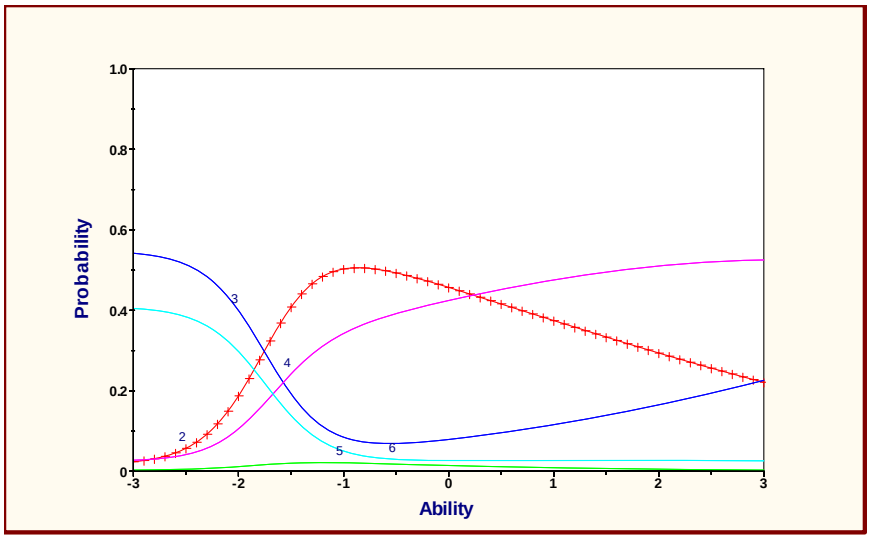

Fig. 2 – Response category curves, item 6

Totally, nearly half of the items are associated with acceptable response curves. In the item analysis, the behavior of incorrect options is very important: the item characteristics not only depend on the stem itself but also on the attractiveness of the different response alternatives. Nevertheless, the MCM has identification problems and involves the estimation of a high number of parameters (for 20 multiple-choice items with 5 alternatives the number of parameters to be estimated is 280).

Restricting the data format to the binary case, the 3PL model can be used to estimate the item parameters. A Bayesian prior has been chosen for the logit of the guessing parameters of all the 20 items. In particular, a Gaussian prior with mean equal to –1.4 (logit of 0.2) and standard deviation equal to 1 has been used, because the number of response categories is 5. Table 3 shows the item parameter estimates in the traditional normal metric.

Table 3 Item parameter estimates, 3PL model

Item	α	β	γ	*Item*	α	β	γ
1	0.71	-1.45	0.20	11	0.65	-2.09	0.16
2	0.95	-1.19	0.16	12	0.82	0.47	0.17
3	0.64	-0.73	0.11	13	1.07	0.35	0.34
4	0.66	-1.99	0.16	14	0.54	-0.50	0.12
5	0.92	0.31	0.31	15	1.08	0.92	0.23
6	-0.07	-7.97	0.19	16	0.55	0.56	0.17
7	0.52	-0.32	0.18	17	0.87	-2.55	0.16
8	1.08	0.21	0.24	18	0.76	-2.46	0.18
9	0.93	-0.56	0.13	19	6.59	-0.02	0.27
10	0.87	-0.93	0.35	20	0.48	2.00	0.14

Theoretically, the α's may take values in $]-\infty,+\infty[$ but in practice estimates from 0.3 to 2 are acceptable. A negative discrimination would reflect a decreasing

probability for the correct response as ability increases, and also very low values would result in the inability of differentiating between the examinees. On the other hand, extremely high values would create a step function, with probability of success equal to 1 or to the guessing parameter without intermediate values. The discrimination estimates in Table 3 are quite low, except few cases. Extreme values are noticed for item 6 (negative estimate) and item 19 (very high estimate). Both items have been excluded from the current version of the test. The predominance of negative estimates for the difficulty parameter β suggests that the item characteristic curves are shifted on the left side of the ability range. With respect to the ability scale, there are only few items with higher difficulties, i.e. items 5, 8, 12, 13, 15, 16, 20. These items require a higher ability level to be successfully answered. In order to calibrate the test, items with different difficulties are needed. The results suggest the introduction of more discriminating and difficult items. Finally, the guessing parameter γ seems to be quite moderate for all the items.

To roughly investigate the model fit, the observed and expected proportions for the correct response can be compared. The results are given in Table 4.

Table 4 Observed and expected correct proportion comparison, 3PL model

Item	*Obs.*	*Exp.*	*Item*	*Obs.*	*Exp.*
1	0.8409	0.8404	11	0.8910	0.8915
2	0.8295	0.8250	12	0.4887	0.4859
3	0.6955	0.6940	13	0.6084	0.6037
4	0.8845	0.8834	14	0.6454	0.6441
5	0.6019	0.5974	15	0.4239	0.4226
6	0.4221	0.4223	16	0.4968	0.4945
7	0.6440	0.6409	17	0.9550	0.9567
8	0.5756	0.5723	18	0.9412	0.9406
9	0.7000	0.6952	19	0.6527	0.6424
10	0.8274	0.8224	20	0.2994	0.3003

No meaningful discrepancies are noticed between the observed and the expected proportions for the correct response, supporting a good model fit. Nevertheless, the problem of goodness of fit is still open in IRT and a deep analysis on the residuals should be conducted. With 2^n possible response patterns and a sample size of $n=324$ respondents, the problem of sparse data is clear: standard goodness-of-fit statistics cannot be used. Research is very active in this sense, for example see Maydeu-Olivares and Joe (2005).

Finally, to investigate the distribution of the examinees respect to the latent ability θ, expected a posteriori (EAP) scores have been calculated (Bock & Mislevy, 1982). This method is based on the mean of the posterior distribution of θ, given the observed response patterns. Each student is assigned to a single score to evaluate his/her performance in the test. The histogram of the relative frequencies related to the 324 examinees is presented in Fig. 3.

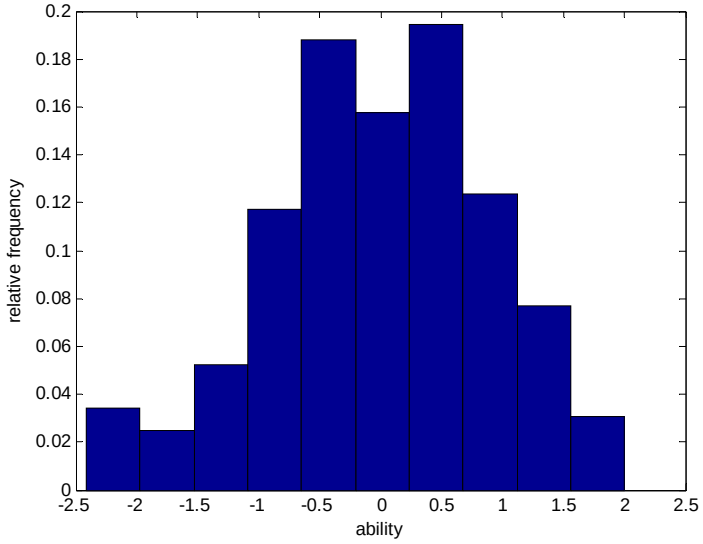

Fig. 3 Histogram of relative frequency, EAP scores

Scores are observed in the range [-2.5; 2] and the ability distribution is rather symmetric. No extreme values are noticed and more than half students are collocated in the range [-0,5; 0,5] of ability.

CONCLUDING REMARKS

The paper illustrates the first results of an experimental project, coordinated by the Guidance Service of Bologna University, for investigating knowledge of secondary school students in order to support them in the faculty choice and provide a useful tool for the entry guidance phase. A faculty-specific test containing both general culture and specific knowledge multiple-choice items is used instead of the common aptitude evaluation. This peculiarity introduces the competence evaluation within student guidance.

An appropriate test can be built following the item response theory approach. In particular, a good test should contain items with proper behavior respect to the ability to be measured. A well developed item bank should contain items with specific properties, i.e. with high discrimination power and different levels of difficulty. Therefore, we have carried out the test calibration to estimate the item parameters in order to select the proper items. In the paper, we have focused on the test developed by the Faculty of Statistics. Similar results in terms of item features have been obtained for the other faculties which joined the project (see Matteucci, 2007).

At the moment, the first phase of the Guidance Project is over because the item bank has been built and the tests are online. All the faculties will be included in future developments of the project. With reference to the student evaluation, the idea is to proceed with finding a simpler method for student classification, respect to the computation of EAP scores in order to provide the students with a qualitative feedback on their performance. Furthermore, we believe that a complete guidance action cannot be conducted only evaluating student competences. The test should be completed with a section regarding the

investigation of both aptitudes and interests, in order to provide the students with an exhaustive educational profile.

To conclude, we believe that the competence evaluation tests can be powerful tools both to guide students in an acquainted choice of the faculty and to prevent the drawbacks due to a barely aware decision process.

Acknowledgements We would like to thank Dr. Bernard P. Veldkamp (University of Twente) for his support and guidance in the phases of test design and calibration and for his persevering and still active collaboration for the Guidance Project.

REFERENCES

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. (In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores* (pp. 397-424). Reading, MA: Addison-Wesley).

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

Bock, R.D. & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.

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

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

Haladyna, T.M. & Downing, S.M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37-50.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. (Boston: Kluwer Nijhoff Publishing).

Maydeu-Olivares, A. & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: a unified framework. *Journal of the American Statistical Association*, 100(471), 1009-1020.

Matteucci, M. (2007). *Item response theory models for the competence evaluation: towards a multidimensional approach in the university guidance*. PhD thesis, Statistics Department, University of Bologna (Italy).

Thissen, D. (2003). *Multilog 7.0. Multiple, categorical item analysis and test scoring using item response theory*. (Lincolnwood, IL: Scientific Software International).

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

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