Measuring logical competences and soft skills when enrolling in a university degree course

Bruno Bertaccini, Riccardo Bruni, Federico Crescenzi, Beatrice Donati

1. Introduction

Logical abilities are a ubiquitous ingredient in all those contexts that take into account soft skills, argumentative skills or critical thinking. However, the relationship between logical models and the enhancement of these abilities is rarely explicitly considered. Two aspects of the issue are particularly critical in our opinion, namely: (i) the lack of statistically relevant data concerning these competences; (ii) the absence of reliable indices that might be used to measure and detect the possession of abilities underlying the above-mentioned soft skills. This paper addresses both aspects of this topic by presenting the results of a research that we conducted in between October and December 2020 on students enrolled in various degree courses at the University of Florence. To the best of our knowledge, this is the largest available database on the subject in the Italian University System to date[1] . It has been obtained by a three-stage initiative. We started from an "entrance" examination for assessing the students' initial abilities. This test comprised ten questions, each of which was centered on a specific reasoning construct. The results we have collected show that there is a widespread lack of understanding of basic patterns that are common in the everyday way of arguing. Students then underwent a short training course, using formal logic techniques in order to strengthen their abilities, and afterwards took an "exit" examination, replicating the structure and the questions difficulty of the entrance one in order to evaluate the effectiveness of the course. Results show that the training was beneficial.

2. Data and methods

The "entrance" test was administered to 272 students in October 2020. The short training course was scheduled in November 2020 and was not compulsory. This characteristic and the students' overall difficulties in self-organizing their study time during the health emergency due to the COVID pandemic have led to fewer "exit" exams (67). The data collected through the two exams were used to: a) estimate initial logical abilities of students engaged in a university experience; b) obtain an evaluation of the effectiveness of the short training course by comparing the abilities measured before and after attending the course itself. Both the "entrance" and "exit" exams we scheduled have the same structure in terms of type (logical constructs), number (10, one per construct), and questions difficulty. The considered logical constructs are: *Double negation* (item code N); *Disjunction negation* (item code D); *Conjunction negation* (item code C); *Hypothetical reasoning* (item code IMPL); *Sufficient and necessary conditions* (item code NEC); *Negation of the universal quantifier* (item code NU); *Negation of the existential quantifier* (item code NE); *Modus tollens* (item code MT); *Syllogism* (item code S); *Multiple steps*

[1] We were unable to find traces in the literature of other datasets on the topic available among other Italian universities

Bruno Bertaccini, University of Florence, Italy, bruno.bertaccini@unifi.it, 0000-0002-5816-2964
Riccardo Bruni, University of Florence, Italy, riccardo.bruni@unifi.it, 0000-0003-2695-0058
Federico Crescenzi, University of Florence, Italy, federico.crescenzi@unifi.it, 0000-0002-0701-4398
Beatrice Donati, University of Florence, Italy, beatrice.donati@unifi.it, 0000-0002-4707-8476

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Bruno Bertaccini, Riccardo Bruni, Federico Crescenzi, Beatrice Donati, *Measuring logical competences and soft skills when enrolling in a university degree course*, pp. 41-46, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-304-8.09, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci, *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the opening conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-304-8 (PDF), DOI 10.36253/978-88-5518-304-8

deduction (item code DED). These constructs correspond to what are in our experience ten of the most recurring errors made by undergraduate students. These errors have been identified in many years of teaching experience but also on the basis of the logical tradition that identifies some constructs underlying our way of reasoning. Each close-ended question (item) presents 4 answers, only one of which is true. 1 point was awarded for a correct answer, no points were assigned to missing or wrong answers. We are confident that this framework could be a good method for measuring logical abilities of students. This hypothesis is at the basis of the Item Response Theory (IRT).

Item Response Theory (IRT) is a methodology to investigate the relationship between an individuals' response to an item of a test on an overall measure of the ability that the item was intended to measure (Demars, 2010; Bartolucci et al., 2016). Knowing the item difficulty is useful when building tests to match the trait levels of a target population. For these reasons, IRT has been used proficiently either to score tests or surveys and in test development/assessment (Chen et al., 2005; Lee et al., 2008).

In presence of binary data - as those just described, that typically correspond to a set of n individuals that give wrong or correct responses to a set of items of a test/questionnaire, the main assumptions of IRT models are: unidimensionality (for each individual i who underwent the test, the responses given to the whole set of items depend on the individual ability θ_i), local independence (for each individual, the given responses are independent given the individual ability θ_i) and monotonicity (the conditional probability of responding correctly to a certain item j, known as Item Characteristic Curve, is a monotonic non-decreasing function of θ_i).

At the core of all the IRT models is the item response function (IRF). The IRF expresses the probability of getting the item j "correct" (i.e. $Y_{ij} = 1$) as a function of item characteristics and the individual's latent (i.e. unobserved) trait/ability level θ_i. In IRT literature, we distinguish between one-parameter (known also as the Rasch model), two-parameters and three-parameters logistic IRT models. Intuitively, each model extends the previous one with an additional parameter. The IRF for the three-parameters (3PL) model is:

$$P(Y_{ij} = 1|a_j, b_j, c_j, \theta_i) = c_j + (1 - c_j)\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \tag{1}$$

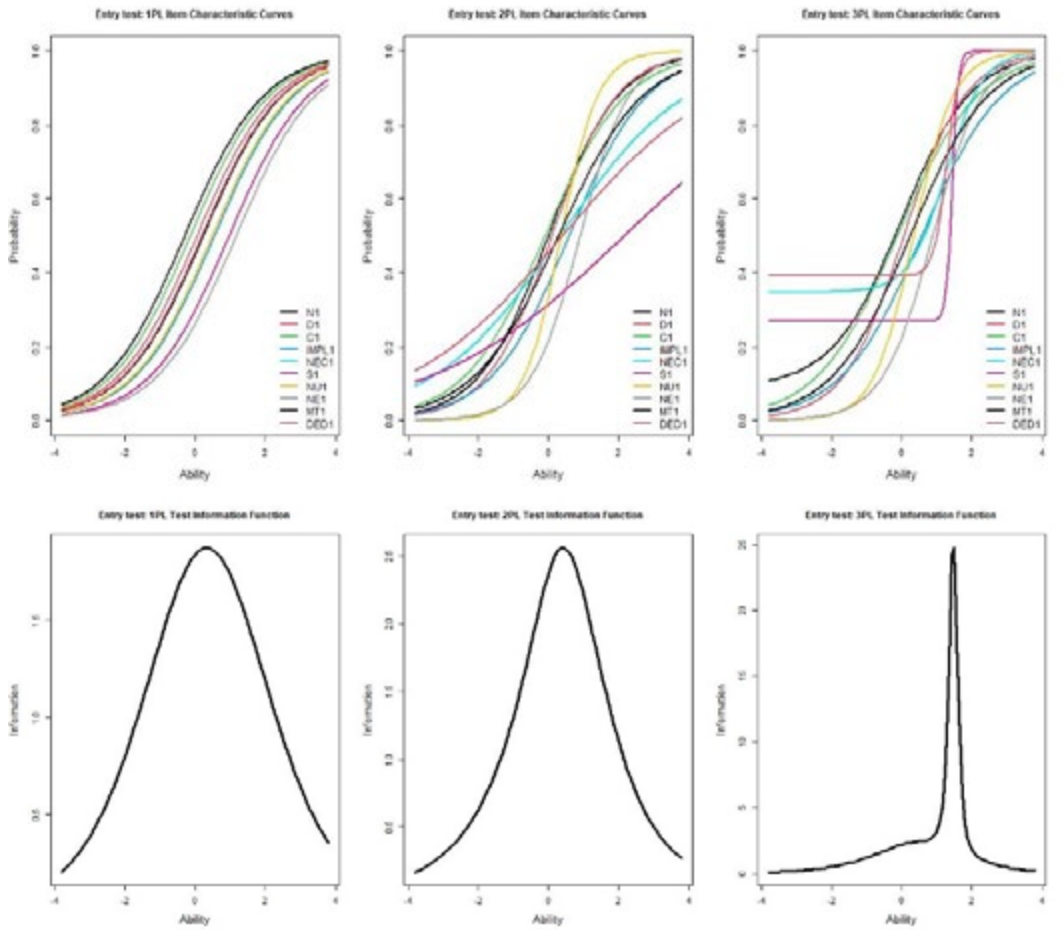
This function describes the probability for an individual with latent ability θ_i to endorse an item j where b denotes the item difficulty, a denotes the item discrimination and c is a parameter for guessing). Under this general configuration, higher difficulty estimates indicate that the item is harder (i.e., higher latent ability to answer correctly), and higher discriminability estimates indicate that the item has better ability to tell the difference between different levels of ability θ. Moreover, individuals with zero ability have a nonzero chance of endorsing any item, just by guessing randomly. For the sake of completeness, the guessing parameter c is not involved in the two parameters logistic (2PL) IRF function, while both the guessing parameter c and the discrimination parameter a are not involved in the one-parameter logistic (1PL, also known as Rasch) model. As usual in IRT modelling, if a parametric model for the ability distribution is not assumed, then the usual two-parameters and three-parameters logistic models present identifiability problems not encountered with the 1PL model (Haberman, 2005). These problems could be solved by imposing substantial constraints such as assuming that the ability latent trait follows a standard normal distribution. Otherwise it is possible to constrain the discriminating parameter of a reference item (usually the first one) to 1 and its threshold difficulty parameter to 0, leaving free the mean and the variance of the ability distribution (still expected normally shaped) (see Bartolucci et al., 2016). Software to estimate such class of models is available for R in the library Ltm (Rizopoulos, 2006).

All logistic IRT models were applied to our data looking for the best parameter estimations, i.e. the most reliable fitting. Results will be presented in the next section.

3. Results

The estimation of the logical abilities of students who undertook the entry test was the first objective of this work. Starting from the simplest one, we have applied all three the IRT logistic models presented above to the data collected administering the "entrance" test. Figure1 shows the Item Characteristic Curves (ICCs) and Test Information Functions respectively from the Rasch, the 2PL and the 3PL models.

Figure 1: Entrance test: Item Characteristic Curves and Test Information Functions obtained after estimating the whole class of logistic IRT models



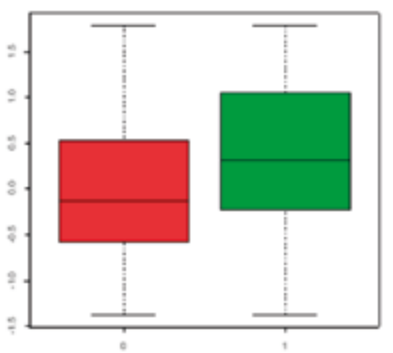

```
Likelihood ratio test
Model 1: 1PL
Model 2: constrained 2PL
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -1755.9
2  18 -1743.7  7 24.364  0.0009831 ***
```

```
Likelihood ratio test
Model 1: constrained 2PL
Model 2: constrained 3PL
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  18 -1743.7
2  28 -1735.4 10 16.623    0.08312 .
```

In particular, the test information functions reported in the bottom panel of Figure 1 are simply the sum of the first derivatives of the ICCs (also called Item Information Curves) in the top panel. Ideally, a good test/questionnaire should provide a good coverage of a rather wide range of latent ability levels. In this case, the information curve should be normally shaped and centred around zero. Otherwise, the test may identify a limited range of ability levels. The 1PL information curve, although centred on a value slightly greater than zero, showed a satisfactory coverage of the range of the possible abilities. Nevertheless, from the analysis of the 1PL model item-fit statistics (here not reported due to lack of space) we observed that 3 items might not fit the 1PL model so well. Also the Likelihood Ratio Test statistic (LRT, presented below Figure 1) suggests an upgrade to the 1PL model.

The 2PL model is more suitable than the Rasch one for describing our data (from the item-fit statistics only one item might still not be in line with the model). The item 2PL ICCs shows that some items provide more information about latent ability for different ability levels. In general, the higher the estimate of the item discriminability the higher the item's capability to provide information about ability levels around the point where there is a 50% chance of getting the item right (i.e. the steepest point in each ICC slope). Instead, the LRT statistic did not provide us with sufficient evidence in favour of the 3PL model (its information curve is quite far from normality), although this model is able to show how the students have tried to guess the answers of the 3 more difficult items (corresponding to NEC, S and DED logical constructs). Individual abilities were then estimated through the 2PL model.

Figure 2: Entrance test ability distribution: students taking just the "entrance" test (red) and student having taken also the "exit" test (green).

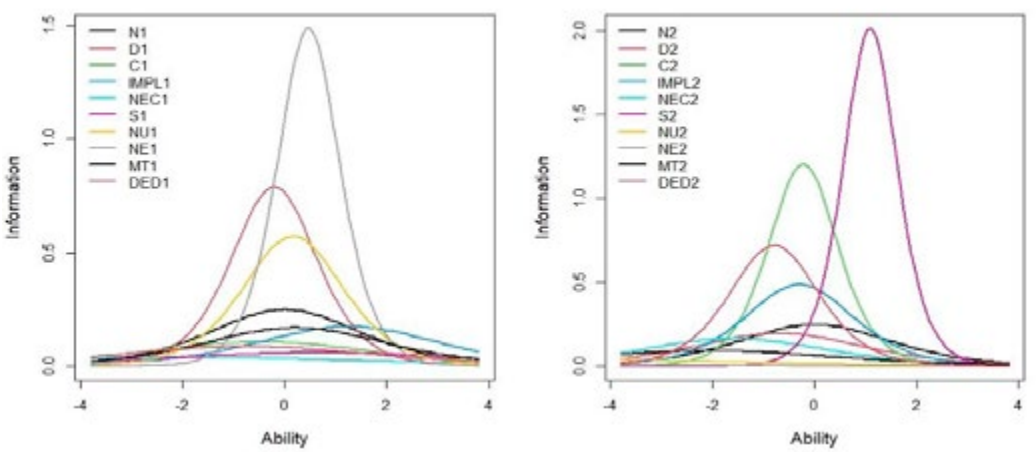


More stable results were obtained limiting the analysis of "entrance" test responses to those students who underwent the short training course and took also the "exit" test. Figure 2 shows the differences in the distribution of abilities of the entrance test respectively for those students who took only that test (red) and those students who also took the exit one (green). The application of the 2PL model to this reduced dataset (see Figure 3a for the estimated Item Information Curves) produced item-fit statistics whose p-values gave no evidence of incoherent or misfitting items. Moreover, there was no evidence against the hypothesis that we were measuring only a

single latent trait (hypothesis of unidimensionality). Interestingly, some ICCs show a different level of information in ability before and after the training course. As these are estimated by the response patterns given by all students who attended the two tests, a plausible reason of this change could lie in the fact that taking the course may have changed the attitude towards the understanding of some constructs. Of course, there may still be a source of randomness in responses because no penalty was assigned in case of incorrect response. The abilities estimated for this subset of students followed an almost perfect standard normal distribution (see Figure 4a).

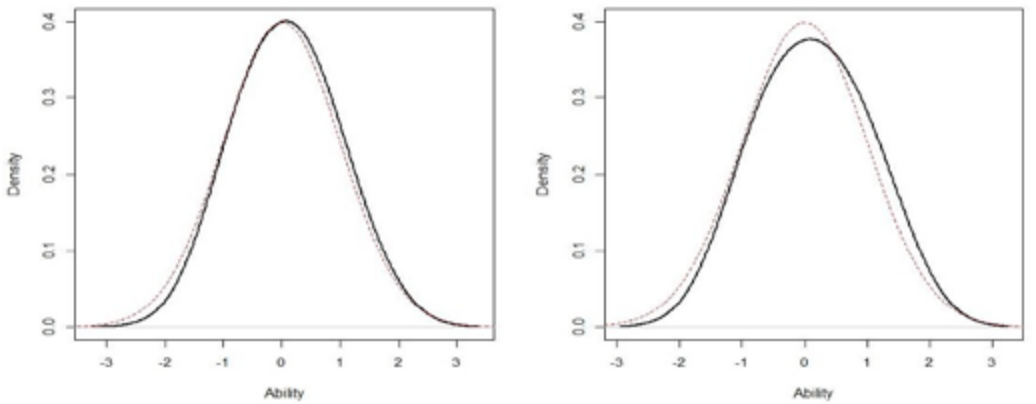
To obtain an evaluation of the effectiveness of the short training course by comparing the abilities measured before and after attending the course itself, we estimated the 2PL model also on responses related to the "exit" test (see Figure 3b and Figure 4b respectively for the estimated Item Information Curves and the distribution of the estimated individual's latent ability). The comparison should be done at individual level to obtain an estimate of the course effect on students' logical abilities. Unfortunately, abilities estimated by the two models are standardized and, consequently, incomparable. The only way to solve this issue is to resort to some test equating techniques. Test equating is a statistical procedure to ensure that scores from different test forms can be compared and used interchangeably. There are several methodologies available to perform equating, some of which are based on the Classical Test Theory (CTT) framework and others are based on the Item Response Theory (IRT) framework (Gonzalez and Wiberg, 2017). Within the IRT framework, if each test form is performed independently or separately in time, their respective parameters will be on different scales and thus incomparable. Equating coefficients solves this problem by transforming the item parameters so that they are all on the same scale. In particular, in this work the abilities estimated with the "entrance" test were transformed to the scale of the "exit" form with the direct equating mean-mean method. Other popular IRT methods for equating pairs of test forms are the mean-sigma, Stocking-Lord and Haebara (Kolen and Brennan, 2014).

Figure 3: Tests underwent by students who completed the short training course: Item Information Curves of the "entrance" test (panel a) and "exit" test (panel b)



We performed the comparison using the "equateIRT" library developed for the R environment for statistical computing (Battauz, 2015). The course effect was thus estimated with a paired sample t-test for differences in abilities. The average difference of 2.07 in the ability estimated before and after taking the course confirms the validity and effectiveness of the programmed training course in incrementing logical abilities of academic students.

Figure 4: Tests underwent students who completed the short training course: distribution of the estimated individualâs latent ability for the “entrance” test (panel a) and “exit” test (panel b)



4. Conclusions

In this paper we presented the results of a research concerning the logical abilities of students enrolled in various degree courses at the University of Florence. This is the first study of this kind and this preliminary data analysis is already very promising and will help us phrasing the test items and refine the entire process. Looking at the data we can already confirm that the “entrance” test results are significant. This convinced us to strongly advise our University to design an internal policy that may become standard, testing all students and providing a mandatory logic course if their ability is below a certain threshold.

We wish to thanks Prof. Sandra Furlanetto of the University of Florence for giving life to this interesting project and providing us with the related data.

References

Bartolucci, F., Bacci, S., Gnaldi, M. (2016). *Statistical analysis of questionnaires: a unified approach based on R and Stata.* Chapman & Hall, CRC Press, Boca Raton (FL).

Battauz, M. (2015). EquateIRT: An R Package for IRT Test Equating. *Journal of Statistical Software*, **68** (7), pp 1–22.

Chen, C.M., Lee, H.M, Chen, Y.H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, **44** (3), pp. 237–255.

DeMars, C. (2010). *Item response theory.* Oxford University Press, Oxford (GB).

Gonzales, J., Wiberg, M. (2017). *Applying Test Equating Methods - Using R.* Springer, Cham (CH).

Haberman, S. J. (2005). *Identifiability of Parameters in Item Response Models With Unconstrained Ability Distributions*. Educational Testing Service, Princeton, NJ. Research Report.

Kolen, M. J., Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd ed.).* Springer-Verlag, New York.

Lee, Y., Palazzo, D. J., Warnakulaso oriya, R., Pritchard, D. E. (2008). Measuring student learning with item response theory. *Physical Review Special Topics - Physics Education Research*, **4**(1)

Rizopoulos D. (2006). Ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, **17** (5), pp 1–25.