ALEAS: a tutoring system for teaching and assessing statistical knowledge[⋆]

Cristina Davino[0000−0003−1154−4209], Rosa Fabbricatore[0000−0002−4056−4375], Daniela Pacella[0000−0003−2343−5069], Domenico Vistocco[0000−0002−8541−6755], and Francesco Palumbo[0000−0002−9027−5053]

University of Naples Federico II, Naples, Italy
{fpalumbo@unina.it}

Abstract. Over the years, several studies have shown the relevance of one-to-one compared to one-to-many tutoring, shedding light on the need for technology-based platforms to assist traditional learning methodologies. Therefore, in recent years, tutoring systems that collect and analyse responses during the user interaction for an automated assessment and profiling were developed as a new standard to improve the learning outcome. In this framework, the tutoring system *Adaptive LEArning system for Statistics* (ALEAS) is aimed at providing an adaptive assessment of undergraduate students' statistical abilities enrolled in social and human sciences courses. ALEAS is developed in the contest of the ERASMUS+ Project (KA+ 2018-1-IT02-KA203-048519). The article describes the ALEAS workflow; in particular, it focuses on the students' categorisation according to their abilities. The student follows a learning process defined according to the Knowledge Space Theory, and she/he is classified at the end of each learning unit. The proposed classification method is based on the multidimensional latent class item response theory, where the dimensions are defined according to the Dublin learning dimensions. In this work, results from a simulation study support our approach's effectiveness and encourage its future use with students.

Keywords: Tutoring system · Multidimensional Latent Class IRT model · Knowledge Space Theory.

1 Introduction

There has been an increasing interest in using technology in education to assist traditional learning methodologies [13]. Intelligent tutoring systems guide learners and help them to fill the gaps in their knowledge [10]. To achieve this, the tutor should correctly diagnose the current state of the student's knowledge so to personalise the learning activities according to individual characteristics [1]. This integrated design significantly improves the effectiveness of the learning process, obtaining an accurate user model tailored to the learner, primarily.

Although there are several technologies developed for teaching and assessing statistics knowledge at the high school and university level [16], there are not many applications specifically aimed at supporting undergraduate students enrolled in social and human sciences courses in the learning of Statistics.

In this framework, we proposed the development of a system to teach and assess knowledge of Statistics with a focus on university students enrolled in social and human degree programs. This system is part of the intellectual outputs of the ALEAS (Adaptive LEArning in Statistics) ERASMUS+ Project[1]. The involved students are less prone to the study of quantitative subjects and therefore are less motivated to master the topic [11]. The system integrates the Knowledge Space Theory (KST; [9]) with the psychometric Item Response Theory paradigm (IRT; [18]) to provide a classification of the learners. The KST organises the full knowledge required to master into a directed acyclic graph structure. The IRT allows the system to assess the ability level of learners and track their progress. In this way, the students' experience can be personalised by selecting the most appropriate set of topics according to her/his status of knowledge [6]. Moreover, since animated graphics can have a considerable effect on the aptitude of learning [15], the ALEAS system is designed to include animated cut-scenes and a tutoring agent to facilitate the learning of some essential statistical topics. The tutoring agent, named "Ronny McStat", reminiscent of the famous statistician Ronald Fisher, welcomes the students and follows them during the learning process.

This article aims to describe the ALEAS system shortly. ALEAS is based on a client-server architecture, where clients are limited to the last generation mobile devices (smartphones and tablets based on the Android operative systems). Here most of the attention is devoted to the algorithm designed to evaluate the learners' ability level and partition them into homogeneous classes. As ALEAS is still in the development stage, we carried out simulations on artificial populations to assess the system's ability to classify the users according to their skills properly. The contribution is organised as follows: Section 2 describes the system architecture, Section 3 introduces the methodological framework ALEAS system is grounded on, Section 4 reports findings from the simulation study and provides an example of feedback for two hypothetical students, Section 5 consists in conclusions and research perspectives.

2 ALEAS: Organisation of the knowledge structure

When designing ALEAS, a preliminary and critical step required was to organise the domain knowledge for the system. Indeed, the different subsets of the domain may not be independent, and the mastery of a specific subset might depend on the mastery of (an)other subset(s). We built a knowledge structure for the basic statistical knowledge exploiting KST [9] and consulting several experts. The resulting knowledge structure consists in ten main nodes, named Topics, and in

[1] https://aleas-project.eu/

the set of the possible relationships among them. It is depicted in Figure 1. The rectangles refer to the Topics and the arrows define the the relationships among them. Moreover, one or more Topics constitute an Area (dotted rectangles in the figure), a more general classification of statistical subjects. On the other hand, each Topic contains several Units that represent the most specific matter of knowledge distinction. For example, the node 'Basic concepts' consists of the following Units: Sample versus population, Taxonomy of variables and levels of measurement, Type of study (observational, correlational, experimental), and Random versus non-random sampling. The user can progress from one node to another one once she/he mastered all the required Topics in an Area following the paths on the knowledge structure. In each Unit, the ability level is evaluated through a multidimensional IRT model, as described below. It is worth to stress that the assessment of students' ability in a multidimensional way represents one of the biggest challenges in the field of education [8], being even more crucial in intelligent tutoring systems.

For ALEAS, we assumed that the student's ability is a multidimensional quantity that grounds on the knowledge structure defined by the Dublin descriptors [12]. The Dublin descriptors qualify the expected outcome of any learning process and serve as bases for the framework for qualifications of the European higher education area. They are typically used as reference dimensions to assess the knowledge a student has achieved within a specific knowledge state. In particular, in ALEAS, the multidimensionality is defined according to the following three of five Dublin descriptors:

- *Knowledge and understanding*: the ability to demonstrate knowledge and understanding including a theoretical, practical and critical perspective on the Topic;
- *Applying knowledge and understanding*: the ability to apply the knowledge identifying, analysing and solving problems sustaining an argument;
- *Making judgements*: the ability to gather, evaluate, and present information exercising appropriate judgement.

All Units are organised including specific learning materials (slides and readings) and fifteen test items, five for each of the three considered descriptors namely knowledge (K), Application (A), and Judgement (J).

3 Methodology: Multidimensional Latent Class IRT model

The IRT is a model-based approach aiming to estimate the probability of correct response to each question for each student, such a probability depends on both her/his ability (typically described by a continuous normal distribution), and on some item characteristics (discriminating power, item difficulty, guessing, and ceiling parameters). In ALEAS, the Dublin descriptors refer to the dimensions that contribute to the definition of the students' ability. Therefore, we assumed the ability as a multidimensional latent trait, then as a statistical tool for our

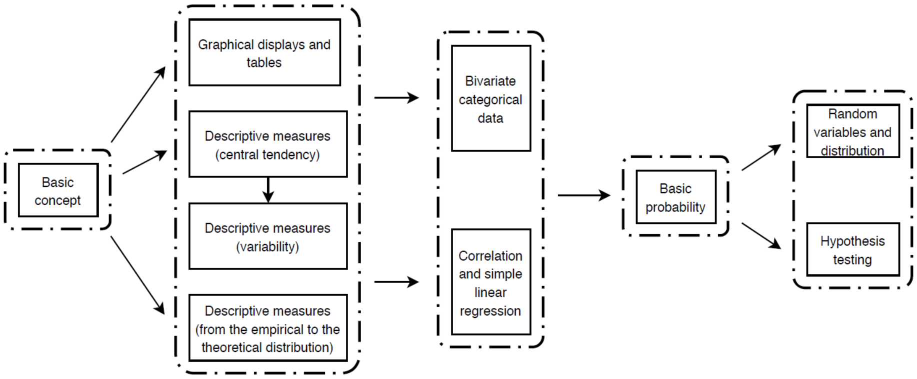
Basic concept
Graphical displays and tables
Descriptive measures (central tendency)
Descriptive measures (variability)
Descriptive measures (from the empirical to the theoretical distribution)
Bivariate categorical data
Correlation and simple linear regression
Basic probability
Random variables and distribution
Hypothesis testing

Fig. 1. Nodes of the Knowledge Space Theory and structure of the relationship among Topics/Areas. Dotted rectangles indicate the Areas in the knowledge structure, whereas solid rectangles indicate the Topics. Arrows represent the required Topics/Areas to masted in order to progress.

purpose, we adopted the class of multidimensional IRT models, proposed by Bartolucci [2].

More in detail, the model we considered is based on the following assumptions:

- *Between-item multidimensionality of the latent traits.* Each item is related only to one latent trait, so that items are divided into different subsets I_d (with $d = 1, \dots, D$) based on D different dimensions. In our model, items were put together according to the three considered Dublin descriptors: K (knowledge and understanding), A (Applying knowledge and understanding), and J (making judgements).
- *Discreteness of the latent traits.* Each latent trait is represented through a discrete distribution with $\xi_1, \dots, \xi_k$ support points defining k latent classes with weights $\pi_1, \dots, \pi_k$. The model assumes that subjects in the same class have the same ability level defined by the corresponding element of the support point vector. Hence, let Θ_s (with $s = 1, \dots, S$) be the discrete random variable of the latent trait of the s^{th} subject, the class weight π_c (with $c = 1, \dots, k$) can be expressed as:

$$\pi_c = P(\Theta_s = \xi_c), \tag{1}$$

with $\sum_{c=1}^{k} \pi_c = 1$ and $\pi_c \geq 0$. It represents the probability that a subject belongs to class c.

Two viable and alternative options allow choosing the number of latent classes k (i.e., the number of support points): a priori based on theoretical knowledge; by comparing the fit of models using different values of k. In the case in point, we exploited a priori theoretical knowledge that statisticians experienced in teaching introductory courses; they suggested the use of $k = 4$ classes, consistently with the four different learning scenarios that will be described in the next section.

- *Two-parameter logistic (2PL) parametrisation.* The 2PL IRT model [4] represents a reduced model of the most general 4PL IRT model, forcing to 0 both the parameters of guessing and ceiling. This setting derives from considering that each item has four possible answers, lowering the impact of guessed answers. Hence, the probability that the subject s correctly answers the dichotomously-scored item i (with $i = 1, \ldots, I$) can be formalised as follows:

$$P(X_{si} = 1|\theta_s, a_i, b_i) = \frac{1}{1 + e^{a_i(\theta_s - b_i)}}. \tag{2}$$

  Where X_{si} is the response of the s^{th} subject at the i^{th} item with realization $x_{si} \in [0, 1]$; $\theta_s \in R$ is the ability of the s^{th} subject; $a_i \in R$ is the item discrimination parameter; and $b_i \in R$ represents the item difficulty.

The estimation of the model parameters is obtained using the *Maximum Marginal Likelihood* (MML) approach [19], and in particular the Expectation-Maximization (EM) algorithm [7]. This algorithm alternates two steps, named E-step and M-step, until convergence. In the E-step, the model estimates each individual's conditional probability belonging to one of the latent classes given her/his response configuration. The M-step consists in maximising the expected value of the complete data log-likelihood based on the posterior probabilities computed in the E-step. The estimation procedure is performed using the R package `MultiLCIRT` [3]. The model is separately applied for every Unit; students are assigned to the latent class that describes their ability upon each Unit completing. The process takes into account the average ability levels reached in each Unit according to the three considered Dublin descriptors, and it provides the learners' categorisation according to their overall performance at the end of each Topic. To this aim, the k-means clustering algorithm [17] is used to classify the learners. At the end of this step, students are provided with a report about their learning outcomes: if they achieve a suitable ability level, they will be allowed to progress to other knowledge nodes, proceeding to a Topic according to the knowledge structure. It is worth noting that the entire Topic is considered complete if the student reports average support greater than zero for all the dimensions. Whenever the student reports average support lower than zero, she/he is encouraged to repeat the related questions: the system identifies the Units to be reiterated.

4 Simulation study

This section describes the simulation study used to test the ability of the model to detect the groups of students with different proficiency levels properly. The study provided us with some evidence about the effectiveness of the model before its use with real-world students. In this phase, we considered a knowledge structure consisting of four Units corresponding to 60 items.

4.1 Design of the Study

The design of the simulation study included the following factors:

- *Item bank.* Firstly, we generated a database of item parameters according to the two-parameter logistic (2PL) parameterization. It included 15 items for each Unit: 5 in Knowledge (K), 5 in Application (A), and 5 in Judgment (J). The difficulty parameter associated with each item was randomly drawn from a standard Gaussian distribution, whereas the discrimination parameters were generated according to a standard log-normal distribution.
- *Item responses.* Item responses were generated, taking into account different ability levels. In particular, concerning the considered Dublin descriptors, we considered as a realistic outcome, four different learning ability levels. In fact, several experts in the subject of Statistics, involved in the ALEAS projects, suggested that the most realistic learning outcome combinations are generally the following:
  1. Poor performance in all the three dimensions;
  2. Good performance in Knowledge and poor performance in both Application and Judgement;
  3. Good performance in Application and average performance in both Knowledge and Judgement;
  4. Good performance in all three dimensions.

  For each learning Unit, $N = 200$ patterns of item responses for each scenario were generated using the `R` package `MAT` [5]. To get the four above specified sub-populations of users, we set the ability level parameters using the `simM3PL` function. In particular, since the latent trait was assumed to follow a normal distribution, assuming $\sigma = 1$, we set $\mu = 2$ for good performers, $\mu = 0$ for average performers, and $\mu = -2$ for poor performers. Moreover, in all the scenarios the correlation between dimensions was set equal to 0.5.
- *Multidimensional Latent Class IRT model.* For each Unit, all the $N = 800$ (200×4) patterns of item responses generated at the previous step were the input for the multidimensional latent class IRT model estimation. As described in Section 3, the model provided us with the following output: matrix of ability levels for each dimension; latent class, and weights of the latent classes; item parameters; posterior probabilities of belonging to the latent classes for each individual. In our model, the number of latent classes was assumed equal to 4, according to the number of simulated learning scenarios. Each student was assigned to the class that corresponds to the highest probability of belonging.
- *Topic-level classification.* The Multidimensional Latent Class IRT model assigns each user to one of the four classes. Then the average ability levels are computed for each of the three Dublin descriptors for all participants. Finally, the k-means clustering algorithm allows obtaining the Topic-level classification for each user.
- *Check of the classification accuracy.* The Adjusted Rand Index (ARI; [14]) allows comparing the Topic-level classification provided by the procedure and

Table 1. Mean and standard deviation of the ARI for all the simulation conditions. In the central column the details about the corresponding population ability mean were also provided.

Simulation condition	**Population ability mean**	**Adjusted Rand Index Mean (SD)**
CASE 1 (n.simules = 1000)	Poor performance: $\mu = -2$ Average performance: $\mu = 0$ Good performance: $\mu = 2$	0.84 (0.11)
CASE 2 (n.simules = 1000)	Poor performance: $\mu \in [-2.2, -1.8]$ Average performance: $\mu \in [-0.2, 0.2]$ Good performance: $\mu \in [1.8, 2.2]$	0.81 (0.13)
CASE 3 (n.simules = 1000)	Poor performance: $\mu \in [-2.5, -1.5]$ Average performance: $\mu \in [-0.5, 0.5]$ Good performance: $\mu \in [1.5, 2.5]$	0.86 (0.04)

the true classification, referred to the one generated learning scenarios. The ARI measures the agreement between two partitions and varies in $[0, 1]$ (random partitioning, partitions perfect agreement); it is widely used to evaluate the overall performance in supervised and unsupervised classification.

The above-described design was replicated 1000 times (CASE 1). ARI means, and standard deviations were used to study the stability of the results.

To assess the model's ability to properly recognise the students according to their ability level, two more simulation studies were run. In CASE 2 and CASE 3 (see Table 1) the ability parameters were generated from Gaussian distribution whose mean parameters were randomly generated from the uniform distribution. The range was ± 0.2 and ± 0.5 in the CASE 2 and CASE 3, respectively. Again, each of these design was replicated for 1000 times.

4.2 Main results

Table 1 shows the mean and the standard deviation of the ARI for all the simulation conditions. Since the ARI lies between 0 and 1, we can conclude that our model was able to recover the starting generated class of the individuals. This result encourages the future use of ALEAS with real-world students.

4.3 ALEAS in action: Results from CASE 1 setting

This section illustrates the results from simulation scenario 1, showing the ALEAS functioning and the type of output report supplied to the students. According to the simulation output, we collected for each student the classification both at the Unit and Topic level, the ability level in each Unit for all the three Dublin descriptor dimensions. At the end of each Topic, students receive preliminary feedback regarding their general performance in that Topic. Figure 2 shows an example of feedback for two hypothetical students. The aim is to provide each student with the assessment on each considered Dublin descriptor, with respect to the overall whole class performance. Each boxplot in Figure 2 refers to the distribution of the mean support of Knowledge, Application, and Judgement. The broken lines join the barycenter of each class (namely the classes obtained from the k-means clustering procedure). Fixing the threshold equal to 0 as the minimum average support to get to pass, the support gained by the two students for each descriptor (represented by rhombuses) indicates that student A (left-hand side) reported low-performance levels on Application and Judgement. In contrast, student B (right-hand side) is a good performer student, especially in Application and Judgement where she/he shows a very high ability (higher than the students belonging to the same class).

To further stress the student's reached level, the mascot Ronny McStat appears in an animated GIF file with an expression according to the level of eval-

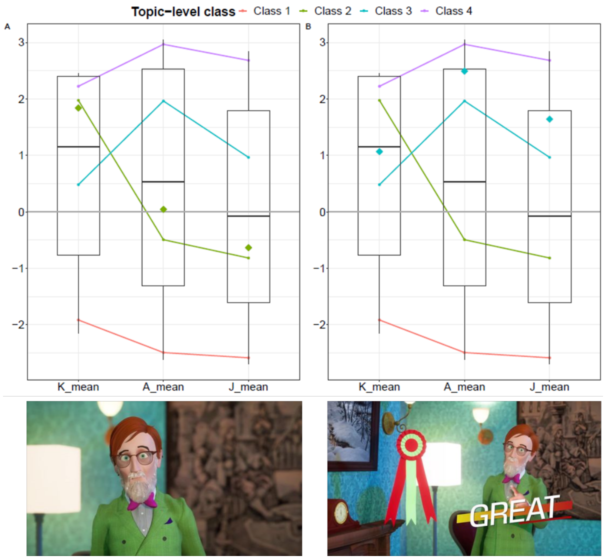

Fig. 2. Topic-level feedback for students A (left-hand side) and B (right-hand side). The boxplots depict the distribution of the ability level means in the sample. Colours indicate the k-means (Topic-level) clusters. Circles represent the class centroids, whereas rhombuses specify the ability level reached by the student. The horizontal gray line defines the ability level required in each dimension to progress to the next Topic. At the bottom of the figure, the animated GIF of Ronny McStat corresponding to the student achievement was reported.

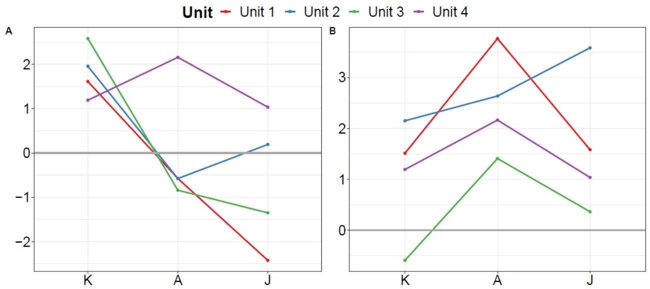

Fig. 3. Unit-level feedback for students A (left side) and B (right side). Colors indicate the different Units. Circles depict the ability level reached by the student. K=Knowledge, A=Application, J=Judgment.

uation (congrats or disappointment expression). The student that properly accomplishes a learning Topic receives a medal from the mascot (student B).

After the Topic-level feedback, users are also provided with a second and more specific report on each Unit (see Figure 3). The students can identify the arguments where they need deepening their knowledge. Therefore, this second report allows us to identify the Units each student needs to repeat. For example, since the student A reached a negative level of ability in judgement (see Topic-level report in Figure 2), according to the Unit-level report in Figure 3 she/he has to repeat the judgement questions in Unit 1 and Unit 3 again.

5 Concluding remarks

We illustrated the ALEAS methodology that is the core of an intelligent tutoring system prototype for teaching and assessing knowledge of statistics in the undergraduate courses in Statistics for students enrolled in human and social sciences courses (in the framework of the homonyms project). It integrates the IRT paradigm with the Knowledge Space Theory, and performs the students multidimensional assessment referring to the learning dimensions Knowledge, Application, and Judgement, which are three of the five learning dimensions that are also known as Dublin descriptors. The system is still in the developing phase. Nevertheless, preliminary results based on the simulation studies indicated that the designed model is adequate in detecting groups of (hypothetical at the current stage) participants, which were simulated according to a different level of abilities. The example in the paper illustrated the assessment feedback that will be provided to a real-world student using the ALEAS. The shown results and several others, not discussed here for the sake of space, portend the ALEAS system effectiveness among the real-world classes students.

References

1. Anderson, J. R., Boyle, C. F., Reiser, B. J.: Intelligent tutoring systems. Science **228**(4698), 456–462 (1985)
2. Bartolucci, F.: A class of multidimensional IRT models for testing unidimensionality and clustering items. Psychometrika **72**(2), 141–157 (2007)
3. Bartolucci, F., Bacci, S., Gnaldi, M.: MultiLCIRT: An R package for multidimensional latent class item response models. Computational Statistics & Data Analysis **71**, 971–985 (2014)
4. Birnbaum, A.: 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. 397–479. Addison-Wesley, Reading (2016)
5. Choi, S. W., King, D. R.: MAT: Multidimensional Adaptive Testing. R package version 2.2 (2014). https://CRAN.R-project.org/package=MAT
6. Davino, C., Fabbricatore, R., Galluccio, C., Pacella, D., Vistocco, D., Palumbo, F.: Teaching statistics: an assessment framework based on Multidimensional IRT and Knowledge Space Theory. In: BoSP SIS2020. Pearson, Milano (*In press*)
7. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society-Series B (Methodological) **39**(1), 1–38 (1977)
8. Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., Maris, G.: Learning meets assessment. Behaviormetrika **45**(2), 457–474 (2018)
9. Doignon, J. P., Falmagne, J. C.: Spaces for the assessment of knowledge. International journal of man-machine studies **23**(2), 175–196 (1985)
10. Elsom-Cook, M.: Student modelling in intelligent tutoring systems. Artificial Intelligence Review **7**(3–4), 227–240 (1993)
11. Fabbricatore, R., Galluccio, C., Davino, C., Pacella, D., Vistocco, D., Palumbo, F.: The effects of attitude towards Statistics and Math knowledge on Statistical anxiety: a path model approach. In: Carpita, M., Fabbris, L. (eds.) ASA Conference 2019 Statistics for Health and Well-being BoSP, pp. 97–100. CLEUP sc, Padova (2019)
12. Gudeva, L. K., Dimova, V., Daskalovska, N., Trajkova, F.: Designing descriptors of learning outcomes for Higher Education qualification. Procedia-Social and Behavioral Sciences **46**, 1306–1311 (2012)
13. Heift, T., Schulze, M.: Errors and intelligence in computer-assisted language learning: Parsers and pedagogues. Routledge, London (2007)
14. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1), 193–218 (1985)
15. Klein, G., Dabney, A.: The cartoon introduction to statistics. Hill and Wang, New York (2013)
16. López Lamezón, S., Rodríguez López, R., Amador Aguilar, L. M., Azcuy Lorenz, L. M.: Social significance of a virtual environment for the teaching and learning of descriptive Statistics in Medicine degree course. Humanidades Médicas **18**(1), 50–63 (2018)
17. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1, pp. 281–297. Cambridge University Press, Oakland (1967)
18. Rasch, G.: Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche (1960)
19. Thissen, D.: Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika **47**(2), 17–110 (2016)