

# Towards a Tailored Hybrid Recommendation-based System for Computerized Adaptive Testing through Clustering and IRT

Wesley Silva<sup>1</sup><sup>a</sup>, Marcos Spalenza<sup>1</sup><sup>b</sup>, Jean-Rémi Bourguet<sup>2</sup><sup>c</sup> and Elias de Oliveira<sup>1</sup><sup>d</sup>

<sup>1</sup>Postgraduate Program of Informatics (PPGI), Federal University of Espírito Santo, Vitória, Brazil

<sup>2</sup>Department of Computer Science, Vila Velha University, Vila Velha, Brazil

**Keywords:** Technology-enhanced Learning, Computerized Adaptive Testing, Intelligent Tutoring Systems, Item Response Theory, Clustering, Recommendation-based System, Collaborative Filtering, Content-based Filtering.


**Abstract:** Creating a student individualized evaluation path composed by a sequence of activities is a hard task and requires efforts and time for teachers. In such cases, the activities have to be well adjusted to the latent knowledge of specific students groups. In this paper, we propose a hybrid system that automatically selects and recommends activities based on a historical evolution of past students during the teaching-learning process. Our system is supported by the hybrid usage of Item Response Theory and techniques of clustering to output different kinds of recommendations as filters to select activities and build the tailored evaluation path.


## 1 INTRODUCTION


Nowadays we can easily encounter a large set of information from various sources such as Facebook, Twitter, LinkedIn, and so forth. But, this huge bunch of data can be confusing for users who want to select particular items and discard others. A recommendation-based system is a well known solution to solve this issue. For example, if users explicitly indicate a preference for a style of music called MPB (*Brazilian Popular Music*), the recommendation system may recommend some *Tom Jobim's* songs. Such an approach is founded on content-based filtering (see Pazzani and Billsus, 2007). On the other hand, it exists a collaborative filtering approach (see Herlocker et al., 2004) in which the system generates a group of similar users in terms of interests producing a recommendation based on the analysis of their characteristics. For example, in the case of a musical platform, the user profiles would be defined in terms of the songs already consumed or liked through the service. With this a priori knowledge, the system can generate a group of similar users in terms of interest and perform a set of recommendations based on the analysis of the singular characteristics in the group.


In the field of educational data, some similar situations exist, especially when teachers have to select assessment items according to some expected performances of their students. Considering the ongoing grades in a discipline, it is possible to group current and past students by the similarity of their performances in different moments of the teaching-learning process. For example in the seminal works of Oliveira et al. (2013), a recommendation-based system can select assessment items that are expected to be compatible with coherent educational objectives.

In this paper, we propose a new implementation of this system by clustering students in accordance with their similar past performances and recommending an evaluation path that should maximize their future performances based on the data collected by similar past students. The characterization of the assessment items is supported by the Item Response Theory (IRT) that generates descriptors based on probabilities of success in function of presupposed student latent traits. IRT allows both qualitative and quantitative items analysis to support the construction of an evaluation path (Baker, 2001). Therefore, our hybrid recommendation-based system deals with both the students characteristics and the probabilities of success. With our data processing, we can select a neat sequence of items to build a tailored evaluation path individualized for each student. If a student has a certain ongoing latent trait, our system can progressively route this student through a steady and coherent

<sup>a</sup> <https://orcid.org/0000-0001-9103-0536>

<sup>b</sup> <https://orcid.org/0000-0002-3826-1500>

<sup>c</sup> <https://orcid.org/0000-0003-3686-1104>

<sup>d</sup> <https://orcid.org/0000-0003-2066-7980>

evaluation path. By identifying their weaknesses and strengths, the system recommends the most suitable activities to improve students performances. Therefore, our system try to soften the exams recommending questions guided by the detections of students learning gaps as advocated by Perrenoud (1998).

The organization of this paper is structured as follows. In Section 2, we briefly detail some works which have similarities with our approach in this paper. In Section 3, we introduce the theories and techniques that support our system. The description of our tailored hybrid recommendation-based system and its particular filtering approaches is presented in Section 4. Finally, we conclude the paper with some considerations and perspectives in Section 5.

## 2 RELATED WORKS

The taxonomy of educational data mining applications proposed by Bakhshinategh et al. (2018) designates student modeling as an essential task to perform behavior predictions. Student models can be separate into two broad categories: expert-centric or data-centric (Mayo and Mitrovic, 2001). While the expert-centric approach relies on experts to identify the skills required to solve a problem by providing the structure of the model, the data-centric approach relies on using the evaluation data to uncover the student abilities (Johns et al., 2006).

Consequently, the trails left by the past students assessments represent strategic data about the effectiveness of the teaching-learning process. By observing the students on their multidimensional characteristics during the teaching-learning process, the lecturer can isolate key performance indicators which could hinder students progress (Lieberman, 1990). Since the learning evaluations aim to measure variables not directly observed, it is relevant to use standardized and certified assessment processes to understand in detail what is measured (Baker, 2001). Following such approaches, it is now recognized that assessments can provide new inputs for possible recommendation-based systems in terms of next learning steps guided by detected learning gaps (Bakhshinategh et al., 2018). Since information retrieval is a pivotal activity in technology-enhanced learning (TEL), Manouselis et al. (2011) emphasize that deployments of recommender systems has attracted a lot of interest.

In the literature, the large part of the recommendation-based systems during the teaching-learning process mainly suggests learning resources sets. For example, in the work of Romero et al.

(2007), the system uses web mining techniques for recommending links to visit to students. Lee et al. (2010) propose an architecture to recommend contents that can reinforce areas in which a particular student needs improvements. In the paper of Shishehchi et al. (2010), a semantic recommender system for e-learning is released by means of which, learners are able to find and choose the right learning materials suitable to their field of interest. The system described by Timms (2007) provides error feedbacks and hints to support the students confronted to a series of problems. The aforementioned system is able to determine the level of hints that students need during a problem solving.

Lastly, the acronym CAT (standing for Computerized Adaptive Testing) has been conceptualized as the set of the methods for administering tests that are adapted to the examinee's ability level (Lee et al., 2010). On the other hand, derived from Psychometry, the Item Response Theory (IRT) represents a possible support for CAT, as it can better explain the results of a given evaluation (van der Linden and Hambleton, 2013). As stated by Roijers et al. (2012), IRT can be used to inform students about their competence and learning, and teachers about students progresses. According to Sinharay et al. (2006), model checking in IRT is an underdeveloped area. This last work examines the performance of a number of discrepancy measures for assessing different aspects of fit of common IRT models and the creation of specific recommendations. Unlike the usual IRT models, MixIRT models do not assume a single homogeneous population. Rather, these models assume that there exist latent subpopulations in the data (Sen and Cohen, 2019). In their work, Johns et al. (2006) propose to train IRT models to predict how the student should fare on the next problem based on past students performances on previous problems. The prediction reaches an accuracy of 72% whether a student would answer a multiple choice problem correctly. In their paper, Wauters et al. (2010) explore the possibility of designing an adaptive item sequencing by matching the difficulty of the items to the learner's knowledge level in intelligent tutoring systems. Lee and Cho (2015) propose a method to select items and create a customized assessment sheet for adaptive testing considering both the learner's ability and characteristics. Farida et al. (2011) propose a method to generate exercises from the learner's progression observed through the information collected. Lee et al. (2010) developed an intelligent tutoring system for English learning that provides content suitable for specific levels of ability supported by an IRT-based approach. Finally, in the approach of Yeung (2019), IRT was

coupled with Knowledge Tracing Modeling (i.e. modeling students' knowledge to determine when a skill has been learned).

As we argued in introduction, establishing assessment criteria suitable to measure learning variations encountered is a well known challenge (Perrenoud, 1998). This challenge increases in proportion to the volume of students in the class. Nevertheless supported by virtual learning environments, the applications of activities and assessments are facilitated, enabling more efficient uses of the information (see Spalenza et al., 2018). In their works, Oliveira et al. (2013) used clustering and classification-based techniques to tackle the individual learning gaps through word processing of the students answers. In this paper, we extend this approach by using IRT to a new tailored hybrid recommendation-based system.

### 3 PRELIMINARIES

Nowadays, it is common to generate artificial datasets to support new prototypes (see Bourguet, 2017). We assigned to the students a set of pseudo activities to simulate assessments. Such datasets were generated for three different types of distributions.

#### Normal Distribution

The composition of our normal law-based distribution is supported by different proportions parameters: 70% of  $\mathcal{N}(5,3)$ , 15% of  $\mathcal{N}(2,1)$  and 15% of  $\mathcal{N}(8,1)$ . Figure 1 presents the histogram of the grades means.

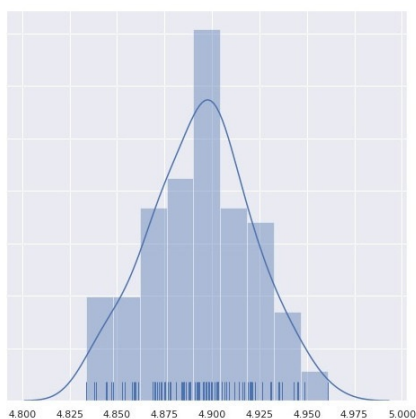


Figure 1: Normal Distribution.

#### Uniform Distributions

The composition of our first uniform law-based distribution is supported by different proportions parameters: 25% of  $\mathcal{U}(1,3)$ , 25% of  $\mathcal{U}(4,6)$ , 25%  $\mathcal{U}(7,9)$  and 25% of  $\mathcal{U}(0,10)$  and 25%. Figure 2 presents the histogram of the grades means.

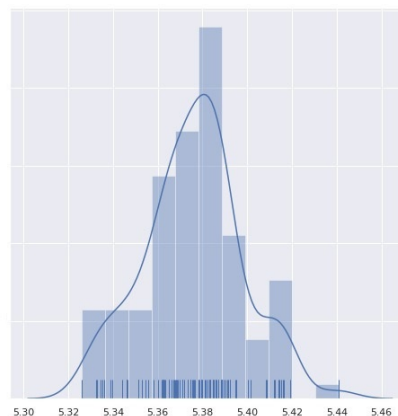


Figure 2: Uniform Distribution with Floats.

The previous distribution deals with float numbers while the second uniform law-based distribution (set with the same parameters as above) generates only integer numbers.

Figure 3 presents the histogram of the grades means for the second uniform law-based distribution.

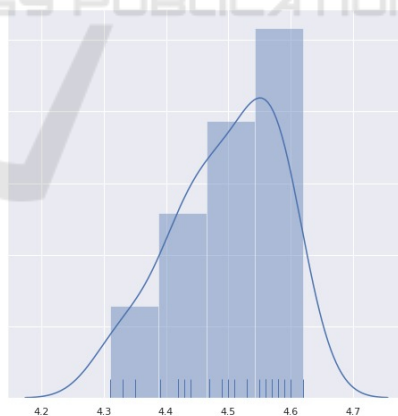


Figure 3: Uniform Distribution with Integers.

Here, our intention is to generate different datasets representing as close as possible real-world cases about students evaluations.

## The Clustering Process

The clustering process aims to identify students who are similar to others according to their performances. Clustering by the *k-means* technique establishes *centroids* according to a specified number *k*. In our case, we assume three groups of performances in a classroom: a high, medium and low performing group. Similarity by cosine distance was used to classify students in the *clusters*. The evaluation of the *clusters* was performed by checking the distribution density of the grades in each cluster.

Figure 4 shows some metrics (intra-cluster grades densities on the left side and intra-cluster density of euclidean distances on the right side) after the clustering of the normal distribution.

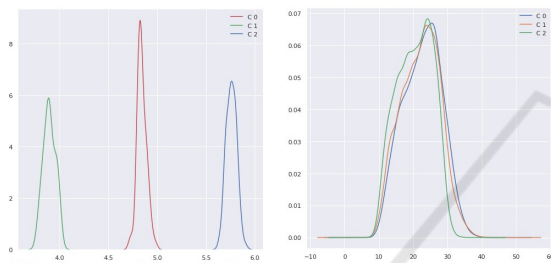


Figure 4: Clustering Metrics (Normal Distribution).

Figure 5 shows the same metrics after the clustering of the first uniform distribution.

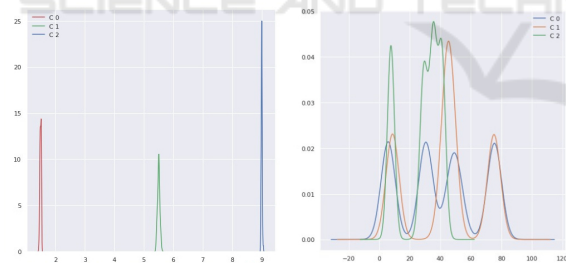


Figure 5: Clustering Metrics (First Uniform Distribution).

Figure 6 shows the same metrics after the clustering of the second uniform distribution.

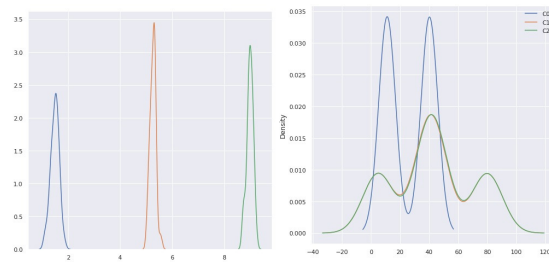


Figure 6: Clustering Metrics (Second Uniform Distribution).

examinee possesses some amount of the underlying ability (also called latent trait) materialized as an ability score (i.e. a numerical value denoted  $\theta$ ) on a rating scale. IRT advocates that depending of a certain ability level, there will naturally be a probability denoted  $P(\theta)$  with which an examinee will answer correctly to the item. The S-shaped curve of this function is called item characteristic curve, and each item have its own. The relation between this probability and the ability score is modeled by a logistic (i.e. sigmoid) function. Such functions relate the natural assumption that the probability will be low for examinees with weak abilities and high for examinees with great abilities, the probability tends to zero at the lowest levels of ability and tends to 1 at the highest. The function depends on parameters describing some properties of each item.

**The Difficulty.** Whatever the number of parameters used to define the characteristic curve of an item, the parameter difficulty is always present. The difficulty of an item  $i$  denoted  $\delta(i)$  is the ability score which corresponds to a probability of success equal to 0.5.

**The Discrimination.** A second important feature of the item is its discriminative power, i.e. its capacity to differentiate examinees (i.e. distinguish those who succeed from those who fail the item) in relation to their underlying ability score. The discrimination of an item  $i$  denoted  $\alpha(i)$  is the maximum slope of the characteristic curve of the item (i.e. the slope of the geometric tangent passing through the inflection point of the curve). Therefore, the slope can be more or less inclined: the more the slope is steep, the more the item is discriminative and inversely.

**The Pseudo-guessing.** Even if, the examinee doesn't have any skill in the scope being evaluated, he or she may have a non-null probability of correctly answering the item. This is particularly the case for evaluations based on multiple-choices. The pseudo-guessing of an item  $i$  denoted  $\gamma(i)$  is the probability of success in the item corresponding to

## The Item Response Theory

IRT (Baker, 2001) has been considered by many experts as a milestone for the modern Psychometrics and an extension of the Classical Test Theory (CTT). While CTT is founded on the proposition that measurement error, a random latent variable, is a component of the evaluation score (Traub, 1997), IRT considers the probability of getting particular items right or wrong given the ability of the examinees. Each

the minimal underlying ability score.

There are different possible models (with one, two or three parameters aforementioned) to build the functions representing the characteristic curves of the items. Equation 1 presents the model with three parameters.

Let an item  $i$  and  $\delta(i)$  (resp.  $\alpha(i)$ ,  $\gamma(i)$ ) its difficulty (resp. discrimination, pseudo-guessing), the probability of success in the item  $i$  for an examinee with an ability score of  $\theta$  is defined as follow:

$$P(\theta) = \gamma(i) + \frac{1 - \gamma(i)}{1 + e^{-1.7\alpha(i)(\theta - \delta(i))}} \quad (1)$$

The notion of precision (or information), defined in Equation 2 assumes an essential role since it indicates in particular on which portion of the underlying ability scale (i.e. for which category(s) of examinees) the precision of the item is the highest. An hard item will give very few information about the examinees with the weakest skills and in the contrary an easy item will not be an accurate test for the examinees with the strongest skills.

Let an item  $i$ ,  $\alpha(i)$  (resp.  $\gamma(i)$ ) its discrimination (resp. pseudo-guessing),  $P(\theta)$  the probability of success for an examinee with an ability score in of  $\theta$ , the precision of such ability score is defined as follow:

$$I(\theta) = 2.89 \alpha(i)^2 \left( \frac{1 - P(\theta)}{P(\theta)} \right) \left( \frac{P(\theta) - \gamma(i)}{1 - \gamma(i)} \right)^2 \quad (2)$$

Figure 7 shows in the top left corner the density distribution; in the top right the parameters estimated by the model IRT, normalized to mean 0 and standard deviation 1; in the bottom left the few discriminative characteristic curves; and finally in the bottom right the information curves.

Figure 8 shows the same kind of plots as the ones shown in Figure 7. The notable differences are that the characteristic curves are more discriminative forming sigmoid curves. The information curves are concentrated on the right of 0 representing the ability of the items to be descriptive in such levels of latent trait.

Figure 9 shows the same kind of plots as the ones shown in Figure 7 and Figure 8. The notable differences are that the characteristic curves are very discriminative and describe easy items. Due to the geometry of the data, it was not possible to calculate the covariance that is an information required to build information curves.

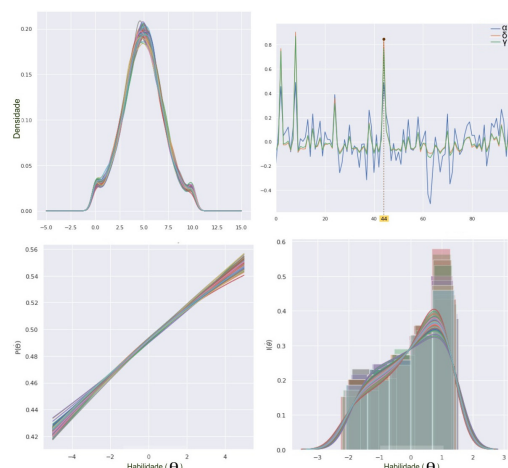


Figure 7: IRT Analysis for Normal Distribution.

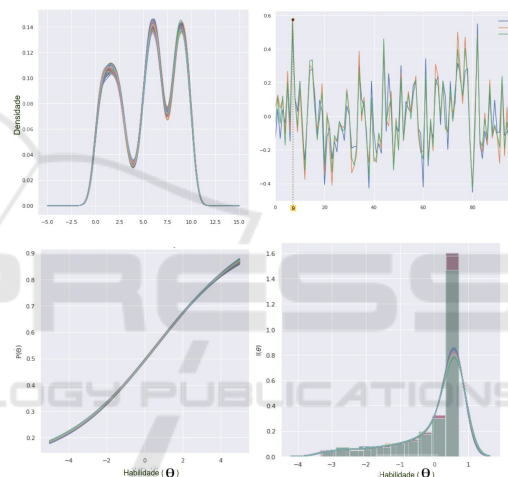


Figure 8: IRT Analysis for Uniform Distribution.

## 4 SYSTEM DESCRIPTION

We argue that successful past actions may be applied to similar students to stimulate their developments. Therefore, we propose to select items with a controlled probability of success that match with the estimated capacity of a given student. Our tailored hybrid recommendation-based system deals with both the students characteristics and the probabilities of performances.

### System Workflow

Our system manages the students in models of performances by gathering together similar students and recommend initiatives to solve the learning gaps. The adoption of data mining and machine learning tech-

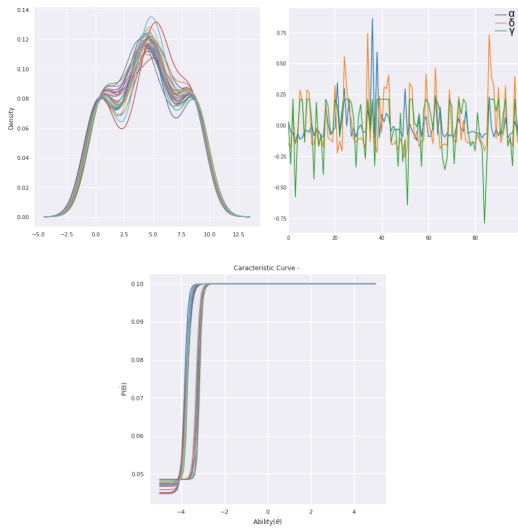


Figure 9: IRT Analysis for Uniform Distribution.

niques helps to better understand the performances behaviors of the students. Figure 10 represents the global work flow of our approach.

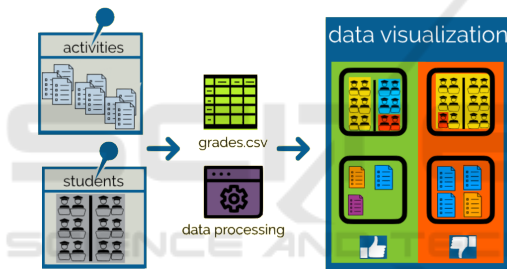


Figure 10: Recommendation-based System Workflow.

In the processing step, student grades represented by a spreadsheet file icon are used as predictor parameters. To implement the recommendation system with a collaborative filtering approach, we used the clustering technique, in order to group similar students grades vectors. Treating each student’s list of grades as a vector, a clustering algorithm is able to group the most similar vectors related to a centroid. On the other hand, to discriminate the evaluation items, supported by IRT, our system can generate for each item the values of the so-called parameter of difficulty, discrimination and pseudo-guessing. Therefore, the system can cluster students and activities. A desirable situation is illustrated on the left side highlighting a fair variation of students and activities. On the other hand, on the right side occurs an undesirable situation in which it was not possible to separate groups of activities and students. By combining the information generated by these two techniques, the system is able to compose personalized assessments path for each student.

## Historical Data

Our Historical Data module represents a foundation for our system in the sense that it represents an historical dataset supporting the process to compute the recommendations. First, our approach is based on a cartesian approach, then we presuppose that our pool of activities  $P$  is stratified and sequentially organized such that  $P = A_0 \cup \dots \cup A^n$  with  $A^i$  a set of activity corresponding to the  $i$ -th scope of a given discipline.  $A_0$  represents a set of initial activities aiming to preliminarily evaluate the students. The evaluation path of a student is a function such that for a student  $s_i$  we have  $E(s_i) = (A_i^0, \dots, A_i^n)$  and  $\forall j \in \llbracket 1, n \rrbracket$  we have  $A_i^j \subseteq A^j$  and  $|A_i^j| = N$ . Note that  $a(k, A_i^j)$  returns the  $k$ -th activity from the set of the activities realized by the student  $s_i$  in the level (or strate)  $t$ . The function  $g : S, P \rightarrow [0, 10]$  will return a grade in function of a given student performing a given activity. When the activities were not assigned to the student, the function can return a value NA (i.e. Not Available).

In the case of our simulation, we artificially generate 1000 students that are confronted to 100 activities, the first activity is always the same after what we randomly pick up one activity among three possible activities and repeat this process consecutively until the end. At the end, all the students will have start their evaluations path in the same way before to proceed on singular roads (sometimes similar) until a certain point.

Our recommendation process is performed using both cluster and IRT analysis as explained thereafter.

## Recommendation Process

When a student will perform a new activity at a certain level  $t$  ( $1 \leq t \leq n$ ), the system considers that his latent trait is actually the pondered (by item difficulty  $\delta$ ) mean of all the grades he obtained until to pass the new activity as described in Equation 3.

$$\theta_t(s_i) = \frac{1}{t} \sum_{j=1}^t \frac{\sum_{k=1}^N \delta(a(k, A_i^j)) \cdot g(s_i, a(k, A_i^j))}{\sum_{k=1}^N \delta(a(k, A_i^j))} \quad (3)$$

The set of past activities of a student at a given level  $t$  is outputted through the function  $A^p : S \rightarrow P$ . Each time a student performed a given set of activities, the vector of his past activities is upgraded by adding a new position at the end of the vector with the aforementioned activities. The function  $g^p : S \rightarrow [0, 10]^{t \times N}$  will associate a student with his current vector of grades at the level  $t$ .

### Recommendation Guided by Difficulty

The first option to guide the recommendation is to use the parameter of difficulty and the clustering. After to have upgraded the vector of past activities, the system selects the set of students  $S(s_i)$  who performed the same past activities as those of the student  $s_i$  such that  $S(s_i) = \{s_j | A^p(s_j) = A^p(s_i)\}$ . Thus, the system proceeds to a clustering task using the vector  $g^p(s_i)$  and the vectors of the set  $\bigcup_{s_j \in S(s_i)} g^p(s_j)$ . Note that a *cluster* is built in relation to the internal similarity  $\rho$  of its members.

Let  $C_p(s_i)$  the set of students currently present in the same cluster as  $s_i$ , the set of the  $\kappa$  activities recommended for the student  $s_i$  is denoted  $R_1^\kappa(s_i)$  and is described in the Equation 4. Note that  $\kappa$ argmax will select the arguments from the  $\kappa$  maximum scores.

$$R_1^\kappa(s_i) = \kappa \operatorname{argmax}_{\substack{a \in A_j^{t+1} \\ \text{s.t. } s_j \in C_p(s_i)}} \sum_{l=t+1}^n \frac{\sum_{k=1}^N \delta(a(k, A_j^l)) g(s_i, a(k, A_j^l))}{(n-t) \sum_{k=1}^N \delta(a(k, A_j^l))} \quad (4)$$

### Recommendation Guided by Discrimination

The second option to guide the recommendation is to use the parameter of discrimination. As explained in Section 3, the discrimination is the capacity to differentiate examinees (i.e. distinguish those who succeed from those who fail the item) in relation to their underlying ability score. The higher the value of the parameter  $\alpha$ , the more the item is considered discriminating. To guide the interpretation of the parameter  $\alpha$ , Baker (2001) offers an evaluation grid: null if  $\alpha = 0$ , very weak if  $\alpha \in [0, 01; 0, 34]$ , weak if  $\alpha \in [0, 35; 0, 64]$ , moderate if  $\alpha \in [0, 65; 1, 34]$ , strong if  $\alpha \in [1, 35; 1, 69]$ , very strong if  $\alpha > 1, 70$  and perfect if  $\alpha$  tends to  $+\infty$ . In the case of our recommendation-based system, the student will be challenged by recommending an evaluation in a certain level of knowledge that corresponds to the student's latent trait. Instead of using the  $\kappa$ argmax operator as previously, the system selects a set of items that correspond to the level of the student by applying a threshold parameter. Once selected, these items are ranked using their own parameters of discrimination. Let a student  $s_i$  at a level  $t$ , and his latent trait  $\theta_t(s_i)$  as described in Equation 3, the system builds a partial preorder on the set activities such that  $\forall (a_j, a_k) \in A^{t+1}$  we have:

$$(a_j, a_k) \in \leq \Leftrightarrow |\theta_t(s_i) - \delta(a_j)| \leq |\theta_t(s_i) - \delta(a_k)| \quad (5)$$

Let  $d$  a threshold s.t.  $d \in \llbracket N, M \rrbracket$  with  $M$  the total number of available activities, the system selects the

$d$ -th closest difficulties in relation to the latent trait of the student by applying a function  $D : S, A^k, N \rightarrow A^k$ . After what, as described in Equation 6, a  $\kappa$ argmax operator is applied to the set in order to select the  $\kappa$  items that will challenge the most the student.

$$R_2^\kappa(s_i) = \kappa \operatorname{argmax}_{a \in D(s_i, A^{t+1}, d)} \alpha(a) \quad (6)$$

### Recommendation Guided by Pseudo-guessing

Note that the two last recommendations can be used as filters, applying one after another. The system gets a last filter considering the pseudo-guessing parameter. With this filter, the chance can be minimized by selecting the activities with minimal pseudo-guessing as described in Equation 7.

$$R_3^\kappa(s_i) = \kappa \operatorname{argmin}_{a \in D(s_i, A^{t+1}, d)} \gamma(a) \quad (7)$$

Applying all the filters together, the system can allocate weights to the filters by setting different values for  $\kappa$ .

Let  $\kappa_1, \kappa_2, \kappa_3$  the weights for the different filters of recommendations  $R_1^{\kappa_1}, R_2^{\kappa_2}$  and  $R_3^{\kappa_3}$ , a recommended set of activities for a given student is described in Equation 8.

$$W(s_i) = \bigcap_{j \in \llbracket 1, 3 \rrbracket} R_j^{\kappa_j}(s_i) \quad (8)$$

### Simulation of an Evaluation Path

An evaluation path is represented in Figure 11 for each of the three distributions. As you can observe on the plots situated on the left side, a given student belongs to different clusters along his evolution through the evaluation path. The best score for the student grades vector in relation to its cluster is informed on the vertical axis. The left side of the figure shows the cosine similarities distribution inside the student cluster at some levels of the path.

The clusters performed on the scaled values of the IRT parameters (difficulty, discrimination and pseudo-guessing) are represented on the right side of Figure 12 through the characteristic curves of the items. In this figure, three different clusters ( $C_0, C_1$  and  $C_2$ ) were represented and the densities of the scaled values of the IRT parameters are represented on the left side.

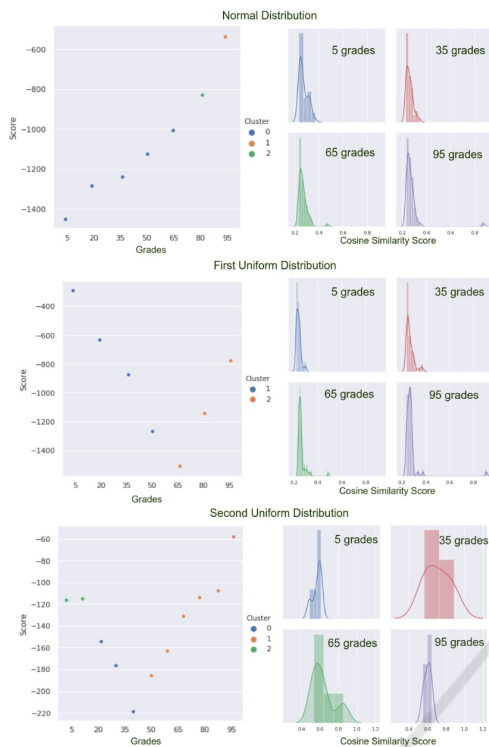


Figure 11: Evaluation Path for a given Student.

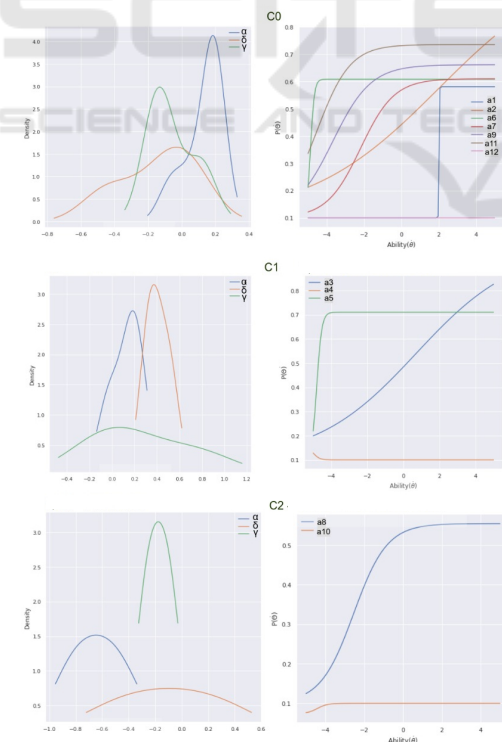


Figure 12: Clustering the Item through the IRT Parameters.

## 5 CONCLUSION

Everyone who teaches has to spend a lot of time creating exams, inspecting and evaluating students activities to discover their learning gaps (Mangaroska et al., 2019). The trails left by the past students assessments represent strategic data about the effectiveness of the teaching-learning process. Some existent systems (see for example Oliveira et al., 2013) can recommend activities indicated for similar profiles that already received recommendations. However, building an appropriate evaluation path in which the levels of knowledge of each student are frequently refreshed and contextualized across the set of available items remained an opened issue.

This article proposes a strategy for selecting the appropriate activities through a tailored evaluation path for each student. We use historical data, i.e. assessments of previous students to build a statistical model in order to predict the future student success. Our system is supported by the common usage of IRT and techniques of clustering to output different kinds of recommendations as filters to select activities. IRT guarantees the employment of a content-based filtering related to the extrinsic qualities of the recommended items while clustering techniques support the collaborative-based filtering related to the intrinsic profiles of the students. As future proposals, we intend to provide an interface for our system to support the selection of the items either by using the clustering techniques or by applying the different filters presented in this work.

## REFERENCES

Baker, F. B. (2001). *The basics of item response theory*. Education Resources Information Center.

Bakhshinategh, B., Zaiane, O. R., Elatia, S., and Ipperciel, D. (2018). Educational Data Mining Applications and Tasks: A Survey of the Last 10 Years. *Education and Information Technologies*, 23(1):537–553.

Bourguet, J. (2017). Purely synthetic and domain independent consistency-guaranteed populations in SHIQ(D). In Lossio-Ventura, J. A. and Alatrasta-Salas, H., editors, *4th Annual International Symposium, SIMBig 2017, Lima, Peru, September 4-6, 2017*, volume 795 of *Communications in Computer and Information Science*, pages 76–89. Springer.

Farida, B., Malik, S., Catherine, C., and Jean, C. P. (2011). Adaptive exercises generation using an automated evaluation and a domain ontology: the odala+ approach. *International Journal of Emerging Technologies in Learning (iJET)*, 6(2):4–10.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recom-



- mender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.
- Johns, J., Mahadevan, S., and Woolf, B. P. (2006). Estimating student proficiency using an item response theory model. In Ikeda, M., Ashley, K. D., and Chan, T., editors, *Intelligent Tutoring Systems, 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006, Proceedings*, volume 4053 of *Lecture Notes in Computer Science*, pages 473–480. Springer.
- Lee, Y. and Cho, J. (2015). Personalized item generation method for adaptive testing systems. *Multimedia Tools Appl.*, 74(19):8571–8591.
- Lee, Y., Cho, J., Han, S., and Choi, B. (2010). A personalized assessment system based on item response theory. In Luo, X., Spaniol, M., Wang, L., Li, Q., Nejdil, W., and Zhang, W., editors, *Advances in Web-Based Learning - ICWL 2010 - 9th International Conference, Shanghai, China, December 8-10, 2010. Proceedings*, volume 6483 of *Lecture Notes in Computer Science*, pages 381–386. Springer.
- Lieberman, D. A. (1990). *Learning: Behavior and Cognition*. International Student Edition. Wadsworth Publishing Company, Belmont, CA.
- Mangaroska, K., Vesin, B., and Giannakos, M. N. (2019). Cross-platform analytics: A step towards personalization and adaptation in education. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge, LAK 2019, Tempe, AZ, USA, March 4-8, 2019*, pages 71–75. ACM.
- Manouselis, N., Drachler, H., Vuorikari, R., Hummel, H., and Koper, R. (2011). Recommender systems in technology enhanced learning. In *Recommender systems handbook*, pages 387–415. Springer.
- Mayo, M. and Mitrovic, A. (2001). Optimising its behaviour with bayesian networks and decision theory. *International Artificial Intelligence Education Society*, 12:124–153.
- Oliveira, M. G., Marques Ciarelli, P., and Oliveira, E. (2013). Recommendation of programming activities by multi-label classification for a formative assessment of students. *Expert Systems with Applications*, 40(16):6641–6651.
- Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In Brusilovsky, P., Kobsa, A., and Nejdil, W., editors, *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 325–341. Springer.
- Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes. towards a wider conceptual field. *Assessment in Education: Principles, Policy & Practice*, 5(1):85–102.
- Roijers, D. M., Jeuring, J., and Feelders, A. (2012). Probability estimation and a competence model for rule based e-tutoring systems. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, LAK '12*, page 255–258, New York, NY, USA. Association for Computing Machinery.
- Romero, C., Ventura, S., Delgado, J. A., and Bra, P. D. (2007). Personalized links recommendation based on data mining in adaptive educational hypermedia systems. In Duval, E., Klamma, R., and Wolpers, M., editors, *Creating New Learning Experiences on a Global Scale, Second European Conference on Technology Enhanced Learning, EC-TEL 2007, Crete, Greece, September 17-20, 2007, Proceedings*, volume 4753 of *Lecture Notes in Computer Science*, pages 292–306. Springer.
- Sen, S. and Cohen, A. S. (2019). Applications of mixture irt models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, 17(4):177–191.
- Shishchchi, S., Banihashem, S. Y., and Zin, N. A. M. (2010). A proposed semantic recommendation system for e-learning: A rule and ontology based e-learning recommendation system. In *2010 International Symposium on Information Technology*, volume 1, pages 1–5.
- Sinharay, S., Johnson, M. S., and Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4):298–321.
- Spalenza, M. A., Nogueira, M. A., de Andrade, L. B., and Oliveira, E. (2018). Uma Ferramenta para Mineração de Dados Educacionais: Extração de Informação em Ambientes Virtuais de Aprendizagem. In *Anais do Computer on the Beach*, pages 741–750, Florianópolis, Brazil. <https://computeronthebeach.com.br/>.
- Timms, M. J. (2007). Using item response theory (IRT) to select hints in an ITS. In Luckin, R., Koedinger, K. R., and Greer, J. E., editors, *Artificial Intelligence in Education, Building Technology Rich Learning Contexts That Work, Proceedings of the 13th International Conference on Artificial Intelligence in Education, AIED 2007, July 9-13, 2007, Los Angeles, California, USA*, volume 158 of *Frontiers in Artificial Intelligence and Applications*, pages 213–221. IOS Press.
- Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, 16:8–13.
- van der Linden, W. J. and Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.
- Wauters, K., Desmet, P., and Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, 26(6):549–562.
- Yeung, C. (2019). Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. In Desmarais, M. C., Lynch, C. F., Merceron, A., and Nkambou, R., editors, *Proceedings of the 12th International Conference on Educational Data Mining, EDM'19, Montréal, Canada, July 2-5, 2019*. International Educational Data Mining Society.