Analysis of questionnaires for Virtual Learning Environments based on Item Response Theory

Oscar Yair Ortegon Romero[1], **Guilherme Medeiros Machado**[1], **Leandro Krug Wives**[1]

[1]PPGC – Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{oyromero,g.medeiros,wives}@inf.ufrgs.br

Resumo. *Este artigo apresenta um modelo para o planejamento e a criação de questionários adaptativos utilizados em ambientes virtuais de aprendizagem. O modelo apresentado combina o uso da Teoria de Resposta ao Item (TRI) com a análise histórica de questionários. Com base nisso, propõe-se uma metodologia para a categorização e ranqueamento das questões pertencentes aos questionários. Tal ranking provê um feedback valioso para o professor ou tutor, que pode então refinar e adaptar o questionário. Os resultados dos experimentos mostram que, levando-se em consideração os parâmetros da TRI, é possível extrair um ranking das questões mais aptas ao ensino de determinado tópico. Espera-se que o uso da metodologia auxilie o docente na elaboração de questionários mais concisos e eficazes.*

Abstract. *This paper presents a model for the design and creation of virtual adaptive evaluations for e-learning environments, combining Item Response Theory (IRT) along with log analysis of previous questionnaires. The proposed model allows the definition of a methodology for the ranking and categorization of questions. Such ranking provides valuable feedback to the teacher or tutor who can refine and adapt the questionnaire. Experiment results reveal that IRT parameters are sufficient for ranking and selecting questions that are more appropriate to teach specific topics. We believe that this approach should become an essential tool for the creation of questionnaires that are more concise and effective in the context of virtual courses.*

1. Introduction

Distance Learning (DL) has brought a new paradigm to teaching strategies and relies on computer-aided tools for the exchange of learning objects, enabling interclass communication, and assessing learners' knowledge through online tests. The most common tools used to support DL are Virtual Learning Environments (VLE), such as Moodle.

One category of VLE that has been proven very effective is adaptive learning systems. Adaptive systems, according to [Brusilovsky 2001], are those that treat the problem of "one-size-fits-all", i.e., when users with different preferences and backgrounds receive the same standardized content. Learning environments are the most successful applications of adaptive strategies [Brusilovsky 2001]. One of the reasons is because educational profiles, such as the previous knowledge of users, should be taken into consideration when presenting new content.

DOI: 10.5753/cbie.sbie.2019.509

Computer Adaptive Tests (CAT) are a critical kind of examination used inside adaptive learning environments, and they are used to deal with the multitude of learners with different backgrounds. In CAT, an algorithm manages the presentation and the selection of questions in addition to deciding dynamically when the test should be finished. In the end, the test verifies students answers and estimate each student level of knowledge [Chalhoub–Deville and Deville 1999].

The relevance given by teachers to the tests used for the evaluation is still very high; a factor that can affect the quality of the evaluation is evident in the tendency they have to use the tests with the bank of questions that they prepared much more than any other type of tests [Darling-Hammond 2000]. That leads us to think about the additional knowledge that teachers must have to create a test that manages to adequately measure one or several levels of knowledge acquired by the student.

Item Response Theory (IRT) is one of the methodologies that can be used to analyze banks of questions. IRT is becoming popular in the educational field because it has been successfully used in qualitative processes of psychological and educational evaluation. It is used to measure and evaluate students acquired knowledge and the development of necessary skills in some subject [Vendramini 2002]. IRT is a framework for modeling student responses on a set of assessments. It is used to describe the relationship between the proficiency of a student and the likelihood of correctly answering a test item. IRT seeks to find a theoretical description to explain the behavior of empirical data generated from the application of a psychometric instrument over the questionnaires. Such theoretical description helps in evaluating the technical quality of each question and also estimates the level of knowledge each student has on a specific topic.

In this context, this work presents a methodology to apply IRT over a set of non-adaptive questionnaires. The main goal of such methodology is, by using previous answers given to one questionnaire, to perform a selection of the essential questions of such questionnaires. In this sense, we believe that delivering such ranking of questions to a teacher can help in the improvement of the other questions. Such refinement is a way of adapting the questionnaires to the learners' knowledge and decreasing the mean error rate.

This work is structured as follows. The next section presents the general concepts that support this work. Section 3 presents investigations proposed in the literature with a different approach that combines IRT for evaluating the learning process. Then, we present and test the methodology for the realization of this approach, presenting experiments. Finally, the last section concludes our work and presents and future work.

2. Background

As the learning process evolved and as a consequence of the success achieved by tests in the evaluation area, a need has evolved to develop a theoretical framework to allow the validation of the interpretations and inferences made from tests and allow estimation of measurement errors inherent in any process of this type. This general framework, called Classical Test Theory (CTT) allowed establishing a functional relationship between observable variables based on empirical scores obtained by subjects in tests or in the elements that compose them and the unobservable variables. In this context, IRT was born as an alternative solution to the problems generated by the relationship between the results obtained by a given subject and the error resulting from the measurement process [Fernández and Hambleton 1992].

2.1. Item Response Theory

According to [Fernández and Hambleton 1992], IRT is a methodology that estimates the ability(s) of an individual in an area of knowledge and the characteristics of the items considered relevant for evaluation, i.e., that may interfere with the response given by a particular examinee to an item.

In this context, skill is a latent variable, i.e., a variable that cannot be measured directly, differently from variables such as weight, height, and temperature. Therefore, variables such as anxiety, satisfaction, intelligence, knowledge, which are not directly measured, are classified as latent; this type of variable is measured from observable secondary related variables. In the case of competence, for instance, the second variable is given by the respondent to an item. IRT proposes models for latent variables and is currently applied in several areas.

IRT has as its basic unit the item, i.e., each question of a test, the test being then a set of items and estimates the parameters that are their characteristics, such as difficulty (b), discrimination (a), and random hit probability (c). According to [de Andrade et al. 2000], three main models use these parameters, and they are described as follows:

Logistic model of one parameter (aka the Rasch model): analyzes that the probability of hitting an item depends only on the level of difficulty of that item and the level (of ability) of the subject in the measured variable.

Logistic model of two parameters: considers the same as the previous model plus item discrimination.

Logistic model of three parameters: besides the parameters previously described, it also considers the casual hit of the item by the examinee of low ability, denoted by Birnbaum (1968), who introduced this parameter to the model because students with low ability sometimes give correct answers.

The item parameters are invariant in a population. It means that, no matter what the average skill of the group, the parameters will be the same, i.e., they are independent of ability.

2.2. Information Function

According to [Baker and Kim 2017], the term information and its statistical meaning were defined as the reciprocal of the variance with which a parameter could be estimated. Statistically, the magnitude of precision with which a parameter is estimated is inversely related to the size of the variability of the estimates around the value of the parameter. The variance of the estimator is denoted by δ^2, and the amount of information, denoted by *I*, then is given by the following equation: $I = 1/\delta^2$ (A)

IRT estimates the value of the ability parameter for an examinee. From Eq. (A), the amount of information at a given ability level is the reciprocal of this variance. If the amount of information is large, it means that an examinee whose actual ability is at that level can be estimated with precision; that is, all the estimates will be reasonably close to the real value. If the amount of information is small, it means that the ability cannot be estimated with precision and the estimates will be widely scattered about the actual ability.

The information function has great importance in the use of the tests since it allows us to choose the one that contributes with more information on the range of ability that we are interested in measuring. It is also very useful in building the test. From a bank of calibrated items (i.e., from which we have estimated its parameters) we can select those that allow an Information function to fit particular objectives. For more details about IRT See [Baker and Kim 2017].

Next, we present a general overview of works that were taken into account as a conceptual basis for the formation of the methodology proposed.

3. Related Work

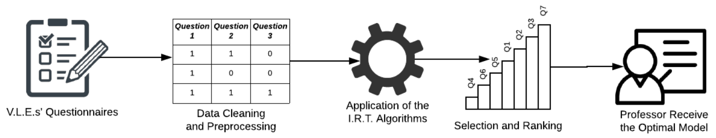
In literature, we can find studies combining item analysis methodologies with IRT. One of such works is the one of [Santos and de Rezende Guedes 2005], which presents a computational tool for the elaboration of adaptive evaluation using as a conceptual base IRT and CAT. In that work, a methodology is proposed for the calculation of the level of difficulty and the ability of the student, using means and medians for the scores obtained from the answers to the questions, the expected values are calculated, and experiments are elaborated with two evaluation models.

A study of the reading development level of Chinese students is presented in [Tian et al. 2017]. In this case, IRT parameters are applied to find a relationship between their values and a method to modify item's options that do not have a reasonable behavior to the data.

A multidimensional IRT temporal model named T-BMIRT is proposed by [Huang and Wu 2017], and it is compared with traditional IRT in online learning studies. The study raises the importance that students, during different moments, may have different levels of knowledge.

The application of IRT in online environments still needs to be studied and disseminated in the research environment, as indicated by [Jatobá et al. 2017], and this field generates a large study space to use techniques that implement IRT. In this case, results showed that the use of online environments based on CAT and IRT is still quite limited, which motivates us to go deeper into the subject and try to contribute to the research process since the methodology proposed generates input for the creation of a CAT with VLE's question banks.

The works mentioned above leave open the need to know if it is possible to identify which questions are more important within an evaluation process and how we could link the evaluated contents with the questions of the questionnaires. The proposed methodology aims to initially address this classification process so that in the next stage, the resulting information can be used as input into the creation of more accurate adaptive tests.


V.L.E.s' Questionnaires
Data Cleaning and Preprocessing
Application of the I.R.T. Algorithms
Q4
Q6
Q5
Q1
Q2
Q3
Q7
Selection and Ranking
Professor Receive the Optimal Model

Question 1	*Question 2*	*Question 3*
1	1	0
1	0	0
1	1	1

Figure 1. Overview of the Process

4. Questions selection methodology

The methodology presented in this section has the goal of providing a helpful tool for teachers, which can pick and rank the most representative questions in a questionnaire. Such selection is made by using IRT Logistic models and a set of strategies, defined in this work, to select and rank the questions. By providing the teacher with this set of questions, it is expected to help in the improvement of their questionnaires, enhancing students' performance in the tests. Since question selection is made by following IRT parameters and not only by text analysis along with answers statistics, this method can guarantee the selection of questions that really cover essential parts of the knowledge, and also can contribute for decreasing the errors.

In Figure 1, it is provided an overview of the proposed methodology for the analysis, selection, and ranking of questions.

4.1. Data Cleaning and Preprocessing

The first step of the methodology is data acquisition and preparation. Even though it is assumed the dataset will come from questionnaires of existing VLE, it is necessary to analyze the answers provided by students as well as the content of the questions before IRT Logistic Models could be applied.

The first constraint for the collected data is that it should be aggregated into questionnaires. To such aggregation, it is necessary to guarantee that the contents and the order of questions do not change when applied to different sets of students (of different classes, for instance). Each questionnaire now can be organized into a matrix where each line represents a student, and each column represents a question. Matrix cells are filled with the students' grades to each question. If the constraint is guaranteed, the data can be analyzed and cleaned [Bong Na et al. 1999].

4.2. Application of Item Response Theory

The questionnaires are then submitted for analysis by using MIRT, an R library that analyzes dichotomous data [Rizopoulos 2006] and computes the IRT parameters, including difficulty, discrimination, guessing, and the maximum amount of information. The process involves two phases:

- **Computing Parameter Logistic Models (Rasch, 2PL, 3PL):** a binary matrix of each questionnaire is created considering the following criteria: if the student has reached at least 75% of the grade of a question, it is then replaced by 1, otherwise by 0; then with this binarization, three logistic models of IRT are computed.
- **Creating optimal model:** After generating the three IRT logistic models (Rasch, 2PL, 3PL) described in figure 2. We identified the questions with the highest coefficient in the questionnaires for each logistic model(3).

4.3. Selection and Ranking

The final step refers to the selection of the subset of more important questions. According to [de Andrade et al. 2000], an item presents a more significant amount of information when it has a high discriminative index and a low success rate. In the analysis of the Item Characteristic Curve and Item Information Function applied in a test, it is possible to construct a classification related to the amount of information and discrimination, highlighting the following elements:

1. Good information, good discrimination, and a reasonable chance of success.
2. Lots of information, high discrimination, and low probability of chance.
3. Low information and low discrimination.
4. Out of trial standards by IRT.

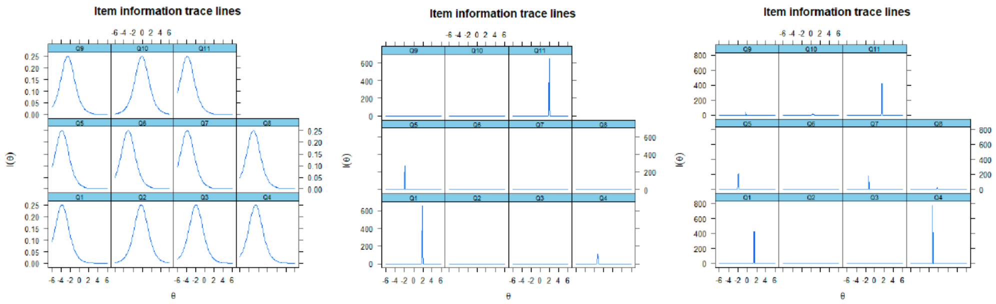
Under these conditions, items that meet conditions 1 and 2 are considered candidates to the optimal model. For the others, they need to be reevaluated to correct problems such as cohesion, clarity of required skill, correction of alternatives, or even the layout of the item in the test. Ultimately, if an item does not meet conditions 1 and 2, it is discarded as it does not meet the required criteria. Finally, the ranking was created giving priority to the questions that had a higher difficulty level.

5. Experiments

To validate the proposed methodology, two experiments were performed; The first one considers a course on “Data Classification and Searching” offered at the Institute of Informatics at UFRGS, considering the periods 2016-I, 2016-II, and 2017-II. From this course were obtained the answers of two questionnaires: Algorithms Complexity and Hashing.The second experiment was performed for the course of Electrical Engineering, also at UFRGS, during the periods 2016-II, 2017-I, 2017-II, and 2018-I, were the answers of 12 questionnaires were obtained. The steps proposed in the last section were followed. Below we detail their application.

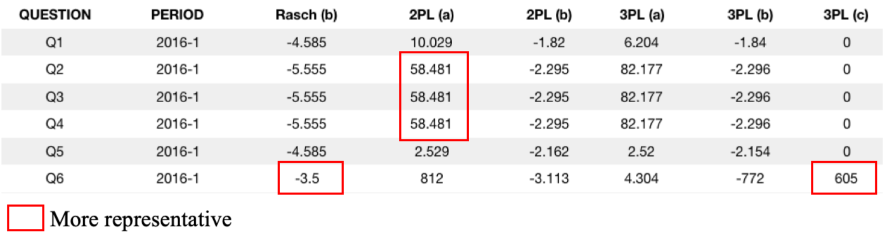
- **Data Cleaning and Preprocessing**. The questionnaires were cleaned to eliminate unfinished and multiple attempts. In *Questionnaire 1: Hashing*, from 46 records originally presented in the 2016-1 period, 15 records were excluded, and from 40 records originally presented in 2016-II period, 14 records were excluded. In *Questionnaire 2: Algorithms Complexity*, there were no changes. During cleaning, only 72% of the 111 records collected were kept, which represents 81 attempts. Once cleaned, it is necessary to binarize CSV files.
- **IRT application**. The next step consists of generating the three IRT logistic models (Rasch, 2PL, 3PL) as described in figure 2. From the visual analysis of the generated models, we have then identified the questions with the highest coefficient in the questionnaires for each logistic model.
- **Selection and Ranking**. Finally, the criterious described previously (item 4.3) for the questions candidate to the optimal model were applied; the results are showed in Tables 1 and 2; the optimal model is computed analysing the matches between questions in distinct periods and ranked for his difficulty level. Then, the final selection and ranking the questions for the optimal model are:
For the *Hashing Questionnaire:* Q3, Q4, Q5, Q6.
And the *Algorithm Complexity Questionnaire:* Q1, Q3.

Figure 2. IRT Analysis for Complexity Questionnaire. Period:2016-I



Rasch Model | 2PL Model | 3PL Model

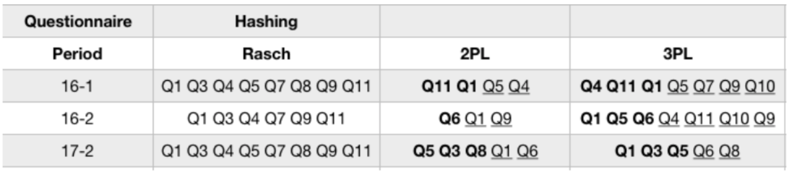
Figure 3. Coefficients - Complexity Questionnaire - Period:2016-I



QUESTION	PERIOD	Rasch (b)	2PL (a)	2PL (b)	3PL (a)	3PL (b)	3PL (c)
Q1	2016-1	-4.585	10.029	-1.82	6.204	-1.84	0
Q2	2016-1	-5.555	58.481	-2.295	82.177	-2.296	0
Q3	2016-1	-5.555	58.481	-2.295	82.177	-2.296	0
Q4	2016-1	-5.555	58.481	-2.295	82.177	-2.296	0
Q5	2016-1	-4.585	2.529	-2.162	2.52	-2.154	0
Q6	2016-1	-3.5	812	-3.113	4.304	-772	605

□ More representative

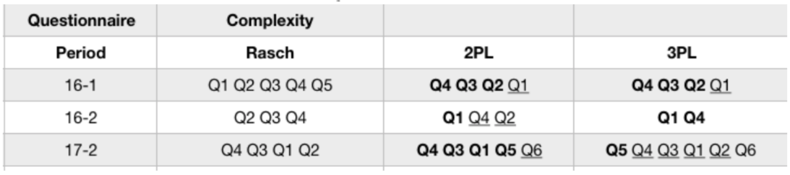
Table 1. Hashing Questionnaire



Questionnaire	Hashing		
Period	Rasch	2PL	3PL
16-1	Q1 Q3 Q4 Q5 Q7 Q8 Q9 Q11	**Q11 Q1** Q5 Q4	**Q4 Q11 Q1** Q5 Q7 Q9 Q10
16-2	Q1 Q3 Q4 Q7 Q9 Q11	**Q6** Q1 Q9	**Q1 Q5 Q6** Q4 Q11 Q10 Q9
17-2	Q1 Q3 Q4 Q5 Q7 Q8 Q9 Q11	**Q5 Q3 Q8** Q1 Q6	**Q1 Q3 Q5** Q6 Q8

Q: Low amount of info | **Q**: big amount of info | Q: Good amount of information

Table 2. Algorithms Complexity Questionnaire



Questionnaire	Complexity		
Period	Rasch	2PL	3PL
16-1	Q1 Q2 Q3 Q4 Q5	**Q4 Q3 Q2** Q1	**Q4 Q3 Q2** Q1
16-2	Q2 Q3 Q4	**Q1** Q4 Q2	**Q1 Q4**
17-2	Q4 Q3 Q1 Q2	**Q4 Q3 Q1 Q5** Q6	**Q5** Q4 Q3 Q1 Q2 Q6

Q: Low amount of info | **Q**: big amount of info | Q: Good amount of information

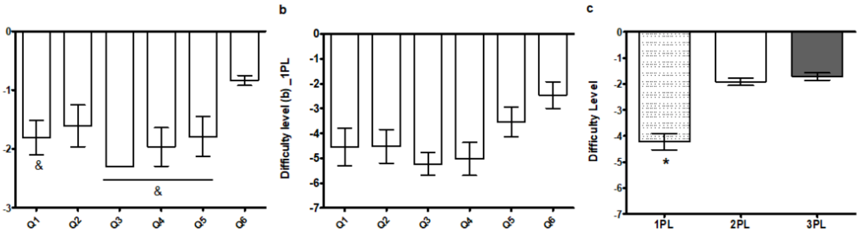
The results of each question were analyzed using one-way ANOVA[Field 2009] followed by Tukey post hoc test in order to determine differences between evaluating periods. Besides, differences among evaluating periods and questions were established for the IRT models considered. More specifically, the differences between models considering the difficulty level were assessed by one-way ANOVA because the parameter (b) are presented in the 3 models analyzed, while student's t-test verified differences using the discriminatory index because the parameter (a) only is presented in two models(2PL and 3PL), in the Rasch model (a =1). P-values$<$0.05 were considered significant, and data are expressed as mean ± standard error (S.E.). Next, a statistical analysis is presented for each questionnaire.[1].

5.1. Statistical analysis - Algorithms Complexity Questionnaire

The analysis of the difficulty level of the ”Algorithms Complexity” questionnaire considering model 3PL showed that question number 6 had a higher difficulty level compared to questions 1, 3, 4 and 5 since $F_{(5,17)} = 3.256, p < 0.05 (seeFigure\ 4a)$. A similar pattern was seen in the logistic model 1PL, evidencing once again that question number 6 was more difficult than the others $F_{(5,17)} = 2.833, p = 0.65, n.s. Figure\ 4b$. In order to analyze the general difficulty levels of the questionnaire, the three IRT logistic models were compared. There was a significant difference between 1PL and the other models $(F_{(5,17)} = 3.256, (p < 0.05) Figure\ 4c)$. Suggesting that this model could be more sensible to determine the difficulty level, which in this case was easy.

Taking each question of this questionnaire into consideration, one way-ANOVA did not show significant differences between them. Moreover, no significant differences (P $>$ 0.05) were found between the evaluating periods when the following variables were evaluated: total qualification and time spent in the questionnaire.

Figure 4. Algorithms Complexity Questionnaire



Difficulty levels by each question using 3PL (a) and 1PL (b). Evaluation of difficulty level by means of IRT models (c). $^{\&}$Significant differences from question 6. *Significant differences from the other evaluating periods. Data analyzed using one-way ANOVA. Significance accepted p$<$0.05.

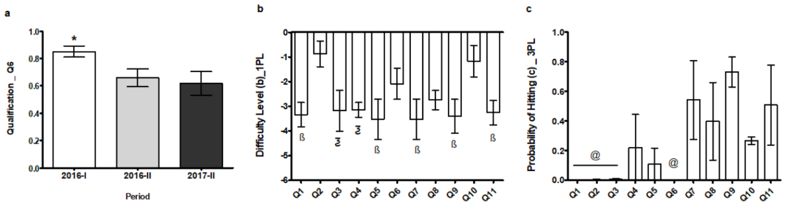
5.2. Statistical analysis - Hashing Questionnaire

As depicted in Figure 5a, there was a significant difference in question 6, indicating that it was perceived as more difficult in 2016-I than in the other periods $(F_{(2,93)} = 3.408, p < 0.05)$. No additional differences were found considering the participants' results along the evaluating periods.

One-way ANOVA revealed a significant difference between questions when the difficulty level of the questionnaire was established employing the 1PL model $(F_{(10,32)} = 2.311, p < 0.05)$; pairwise comparison showed that questions 2 and 10 were more difficult than 1, 5, 7, 9 and 11. Besides, questions 3 and 4 were easier when compared to question 2, as shown in Figure 5b. No significant differences were observed between question number 6 and questions 2 and 10. The 3PL model was carried out in order to identify the “guessing” probability. There was a significant difference between question 9 when compared to questions 1, 2, 3, 5, and 6, as shown in Figure 5c.

[1] Detailed statistical analysis available at: https://github.com/oOrtegon/Experiments

Figure 5. Hashing Questionnaire



Result of a specific question by each period (a). Difficulty levels by each question using 1PL (b). Probability of hitting by each question using 3PL (c). *Significant differences from the other evaluating periods.$^{\beta}$Significant differences from questions 2 and 10.$^{\Im}$Significant differences from question 2. $^{@}$Significant differences from question 9. Data analyzed by one-way ANOVA. Significance accepted p$<$0.05.

6. Analysis and Disscusion

From the previous analysis, it was possible to apply a methodology proposed. To discuss the results, the following questions are proposed:

Q1 : Are all the questions necessary? Are some of them more important than others?
Q2 : Can it be established a ranking of importance of the questions in a questionnaire?
Q3 : Can the position of the question in the ranking indicate poorly formulated questions?
Q4 : Can some questions be easier or more difficult for one group of students?
Q5 : Can the concentration made by IRT analysis be evaluated for identifying if a question is poorly formulated, or if it is a difficult question?

In the case of Q1 and following IRT, all questions are necessary if the assumptions of unidimensionality and local independence are fulfilled, although these questions can be classified according to the previously mentioned criteria. The questions mentioned above were classified by different criteria such as the ranking of amount of information, level of discrimination, and random success, these classifications can be questioned in the case of Q2. The position in the ranking cannot indicate questions directly with errors, as indicated in Q3; otherwise, the questions that did not comply with the criteria proposed cannot be measured with IRT. In this case, there should be a need to reformulate them in order to correct problems such as cohesion, clarity of skill required, alternatives or layout of the item in the test, as proposed by [de Andrade et al. 2000]. Questions Q4 and Q5 can be demonstrated with the results of the statistical analysis. [2]

7. Conclusion

The application of the proposed methodology can be a useful tool for teachers. Knowing which questions are contributing more to the learning of their students can help teachers in the task of refining and adapting their questionnaires. As future work, we propose performing online tests using the output of the methodology at the beginning of a class to have the feedback from the teacher about the helpfulness of the methodology on both the questionnaires refinement and in the decreasing of the mean error rate as a whole. Also, we believe that an ontology could be designed for mapping the knowledge area of

[2] This work is partially funded by FAPERGS-RS.

each questionnaire, helping the teacher to identify the areas their students are performing worse. It would also allow making comparisons between the evaluated groups, as suggested by [Millán et al. 2013]. These efforts would allow a better classification between the evaluated content and the questionnaires.

References

Baker, F. B. and Kim, S.-H. (2017). Test calibration. In *The Basics of Item Response Theory Using R*, pages 105–125. Springer.

Bong Na, W., Marshall, R., and Lane Keller, K. (1999). Measuring brand power: validating a model for optimizing brand equity. *Journal of product & brand management*, 8(3):170–184.

Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11(1-2):87–110.

Chalhoub–Deville, M. and Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, pages 273–299.

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education policy analysis archives*, 8:1.

de Andrade, D. F., Tavares, H. R., and da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*.

Fernández, J. M. and Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de psicología/The UB Journal of psychology*, (52):41–66.

Field, A. (2009). *Descobrindo a estatística usando o SPSS-2*. Bookman Editora.

Huang, J. and Wu, W. (2017). T-bmirt: Estimating representations of student knowledge and educational components in online education. In *Big Data (Big Data), 2017 IEEE International Conference on*, pages 1301–1306. IEEE.

Jatobá, V., Valdivia-Delgado, K., Farias, J., and Freire, V. (2017). Testes adaptativos computadorizados baseados na teoria de resposta ao item em sistemas e-learning: Uma revisão sistemática da literatura. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 273.

Millán, E., Descalço, L., Castillo, G., Oliveira, P., and Diogo, S. (2013). Using Bayesian networks to improve knowledge assessment. *Computers & Education*, 60(1):436–447.

Rizopoulos, D. (2006). ltm: An r package for latent variable modeling and item response theory analyses. *Journal of statistical software*, 17(5):1–25.

Santos, F. D. and de Rezende Guedes, L. G. (2005). Testes adaptativos informatizados baseados em teoria de resposta ao item utilizados em ambientes virtuais de aprendizagem. *RENOTE*, 3(2).

Tian, X., Han, X., Cheng, H. N., Chang, W.-C., Liao, C. C., Sun, J., Zhu, X., and Liu, S. (2017). Applying item response theory to analyzing and improving the item quality of an online chinese reading assessment. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 754–759. IEEE.

Vendramini, C. M. M. (2002). Aplicação da teoria de resposta ao item na avaliação educacional. *Temas em avaliação psicológica*, pages 116–130.