


Content Modeling in Smart Learning Environments: A systematic literature review


Alberto Jiménez-Macías

(Universidad Carlos III de Madrid, Leganés, Spain)

 <https://orcid.org/0000-0002-1148-742X>, albjimen@it.uc3m.es)


Pedro J. Muñoz-Merino

(Universidad Carlos III de Madrid, Leganés, Spain)

 <https://orcid.org/0000-0002-2552-4674>, pedmume@it.uc3m.es)


Margarita Ortiz-Rojas

(Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador)

 <https://orcid.org/0000-0003-2193-8316>, margarita.ortiz@cti.espol.edu.ec)


Mario Muñoz-Organero

(Universidad Carlos III de Madrid, Leganés, Spain)

 <https://orcid.org/0000-0003-4199-2002>, munozm@it.uc3m.es)

Carlos Delgado Kloos

(Universidad Carlos III de Madrid, Leganés, Spain)

 <https://orcid.org/0000-0003-4093-3705>, cdk@it.uc3m.es)

Abstract: Educational content has become a key element for improving the quality and effectiveness of teaching. Many studies have been conducted on user and knowledge modeling using machine-learning algorithms in smart-learning environments. However, few studies have focused on content modeling to estimate content indicators based on student interaction. This study presents a systematic literature review of content modeling using machine learning algorithms in smart learning environments. Two databases were used: Scopus and Web of Science (WoS), with studies conducted until August 2023. In addition, a manual search was performed at conferences and in relevant journals in the area. The results showed that assessment was the most used content in the papers, with difficulty and discrimination as the most common indicators. Item Response Theory (IRT) is the most commonly used technique; however, some studies have used different traditional learning algorithms such as Random Forest, Neural Networks, and Regression. Other indicators, such as time, grade, and number of attempts, were also estimated. Owing to the few studies on content modeling using machine learning algorithms based on interactions, this study presents new lines of research based on the results obtained in the literature review.

Keywords: Content modeling, Smart content, Learning analytics, Smart learning environments, Literature review

Categories: F.1.2, K.3.1, K.3.2

DOI: 10.3897/jucs.106023

1 Introduction

Education has been significantly transformed by the integration of technology. One of these emerging technologies is Smart Learning Environments (SLEs), defined as environments that use technology to make learning more personalized. SLEs can adapt to each student's needs, making learning more engaging and efficient, ultimately helping learners achieve better results in their education [Spector, 2016]. Central to the design and efficacy of such environments are three interrelated components: user/student, knowledge, and content. While both user and knowledge modeling contribute to the enhancement of SLEs, this study's primary focus lies on content modeling, which can be defined as the process of estimating or predicting content indicators using a dataset based on students' interactions with the content. Within the Smart Learning Contents (SLC) domain, content modeling plays a crucial role in creating personalized learning experiences. It helps educators identify knowledge gaps and develop more effective learning strategies from sources such as assessments, exercises, videos, lectures, files, and discussion forums [Brusilovsky et al., 2014].

There have been different approaches to implementing content modeling. Some models have been developed using machine learning algorithms that utilize student interactions, such as Item Response Theory (IRT), random forest, regression, and neural networks [Huang and Wu, 2017]; [Benedetto et al., 2020a]; [Xue et al., 2020]. For example, the IRT can be used in questionnaires and surveys. The most common uses in the educational field are test calibration, inferences from items, student modeling, content modeling, and content adaptation. Studies have been carried out by varying the IRT base model to estimate content indicators, such as difficulty, discrimination, and guessing [Martínez-Plumed et al., 2019]; [Martínez-Plumed et al., 2016]; [Kadengye, 2014].

Other models do not employ machine learning algorithms, such as ontologies, predicate theory, or Unified Modeling Language (UML) diagrams [Madhusudhana, 2017]; [Akhras, 2005]; [Brajnik, 2007]. For example, Brajnik [Brajnik, 2007] proposed a framework for analyzing the content and expression of learning objects using concept maps and UML diagrams. Akhras [Akhras, 2005] analyzed students' interactions with content in learning assessment situations using predicate theory. Finally, some studies used machine learning algorithms but did not incorporate student interactions into their proposed models [Alrajhi et al., 2020], [Capuano et al., 2021], [Atapattu et al., 2020]. Among the educational content that can be utilized in content modeling are exercises, videos, and discussion forums.

Various state-of-the-art studies have been conducted on other types of modeling in SLE, such as user modeling and knowledge modeling. For example, different systematic reviews have identified learning parameters associated with students, such as their level of knowledge, behaviors, learning preferences, and emotions [Abyaa et al., 2019]; [Chrysafiadi and Virvou, 2013]; [Desmarais and Baker, 2012]. However, although user and knowledge modeling have gained considerable attention, a notable gap exists in the existing literature.

This systematic literature review explores various aspects of educational content and its indicators, going beyond the specific domain of question-difficulty assessment. While existing systematic reviews, such as those by AlKhuzayy et al. [AlKhuzayy et al., 2021] and Jia et al. [Jia et al., 2020], have focused on the particular scope of item and question difficulty prediction methods, this review takes a broader perspective. This study aimed to comprehensively understand the multifaceted landscape of educational content by encompassing the totality of educational content and its associated indicators. Although previous reviews have provided valuable insights, the present review aims to

bridge this gap by encompassing a broader range of aspects of educational content, investigating various indicators, and clarifying their implications for education and pedagogical practices. Thus, the purpose of this study was to understand the current state of the art in content modeling. To achieve this goal, the following research questions were posed:

- RQ1: For what purpose has content modeling been used?
- RQ2: Which areas or platforms have used educational content modeling?
- RQ3: Which metrics have been used to evaluate the models in virtual systems?
- RQ4: Which techniques or models have been used to model educational content in virtual systems?
- RQ5: What cognitive and meta-cognitive skills are involved in the models for educational content in virtual systems?
- RQ6: Which indicators have been used for content modeling?

The following section presents a detailed description of the research methodology, including a rigorous explanation of the data collection and analysis procedures.

2 Methods

The principles of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [Moher, et al., 2009] were used in this review and divided into the following subsections: data sources, identification of inclusion criteria and exclusion, and study selection and data synthesis.

2.1 Data sources

This systematic review was performed using the following two databases: Scopus and ISI Web of Knowledge (WoK), including available studies until August 2023. Both databases were used because they are multidisciplinary databases in the academic community. The combination of search words in the databases has the following format:

("content model*" OR "exercise model*" OR "assessment model*" OR (smart W/1 content) OR "intelligent content" OR "content metadata" OR "personalized content" OR "content adaptation" OR ((irt OR ("item response theory")) AND parameter) AND ("tutoring system" OR (learning W/1 system) OR "learning environment" OR mooc OR "Massive Open Online Course" OR "Massive Open Online Courses" OR lms).

The wildcard asterisk (*) was included in the keywords to encompass all the words containing the indicated suffix. The operator (W/) indicates the distance between the specified words, regardless of their order. The combination of words in the search included the largest number of papers conducted in content modeling in smart learning environments using machine learning algorithms based on interactions, using wildcards, and approximation operators to include different ways of writing various word classes. We decided to add papers related to the IRT parameter as keywords because IRT is a form of content modeling for estimating item characteristics such as difficulty, guessing, and discrimination. IRT is a term commonly used for user/knowledge/domain modeling and

originated long ago when the term 'ML' was not yet in use. For these reasons, papers related to IRT did not use the same terminology that is currently used. In this literature review, we only searched for papers conducted in smart learning environments. Therefore, we included different ways of describing these systems at the end of the text.

The search was restricted to the following fields: title, abstract, and keywords, as these fields have the most representative terms in this paper. Studies related to content modeling could have been excluded from the present study because they did not include search keywords in the fields described above.

2.2 Inclusion / exclusion criteria

Table 1 shows an overview of the inclusion and exclusion criteria used in this systematic review. Initially, the inclusion criteria encompassed academic research focused on content modeling using machine-learning algorithms derived from student interactions. This refined approach excluded studies focused solely on user modeling, a category that emerged from preliminary search results. In addition, the inclusion criteria required articles to be composed solely in English because of their prevalence in the international academic community, thus excluding articles in Spanish, Chinese, and other languages that were part of the initial results. In addition, articles that exclusively adhered to the fundamental IRT model without any variation were excluded, contributing to a higher level of rigor in the literature review.

| Criterion | Inclusion | Exclusion |
|-----------------|---|---|
| Scopus focus | Content model using machine learning algorithms | Studies that did not focus on the content model |
| Type of article | Conferences papers, journals | Technical reports, only theory |
| Language | English | Non-English studies |
| Others | | Use only the IRT base model, duplicate on both database |

Table 1: Inclusion and exclusion criteria

2.3 Study selection

Initially, 1097 studies were obtained from the search: 926 in Scopus and 171 in WoK. First, 116 duplicate studies were removed from the databases (SCOPUS and WoK).

Next, a review of the article titles was conducted to identify those that aligned with the focus of the search; in this step, 693 studies were excluded. Following a similar procedure, abstracts of the remaining papers were examined, resulting in the exclusion of 150 studies. To identify those incorporating content modeling, the remaining 138 studies underwent a thorough review across sections, including methodology, results, discussion, and conclusions. Ultimately, 21 studies were included in this analysis. Given the limited volume of findings, a literature review was synthesized by encompassing four high-impact conferences and four influential journals within the realm of interest.

The search process encompassed the following conference proceedings: Intelligent Tutoring System (ITS), Artificial Intelligence in Education (AEID), Learning Analytics Knowledge (LAK), and Educational Data Mining (EDM), along with the following

journals: Computers and Education, IEEE Transactions on Learning Technologies, Journal of Learning Analytics, and Journal of Education Data Mining (EDM) for the years 2023, 2022, 2021, 2020, 2019, and 2018. These conferences and journals were chosen because of their significant influence on this research.

Articles cited as references in selected studies were also included. All papers published in these eight venues during the six years were scrutinized without employing search terms. This approach facilitated the identification of new papers, even if they were indexed in SCOPUS or WoK, because search clauses did not constrain this step.

The selected conferences and journals published 1826 studies over the last six years. The initial step was to review the titles of the studies and exclude those that were not relevant to the research topic, leading to the exclusion of 1043 studies. Subsequently, of the 783 studies selected, the same abstracts were examined, separating the 674 studies that did not pertain to content modeling. Finally, 109 articles were subjected to a thorough reading of all sections, yielding 7 studies that pertained to content modeling.

From these 7 studies, a database of 45 studies referenced in the selected studies was compiled. Following the same steps, including a review of the title, abstract, and full article, three other studies relevant to the research topic were obtained.

Finally, in addition to these 10 studies, the remaining 21 identified a total of 31 studies in the present review, as illustrated in the PRISMA flow diagram in Figure 1. Section 4 delineates the findings in terms of the research questions.

3 Results and Discussion

3.1 General results

To understand the evolution of content modeling using machine-learning algorithms, it is essential to understand the publication dates of the selected studies. Figure 2 shows the temporal progression of scientific research in the field studied. The x-axis represents the years of publication of scientific articles, whereas the y-axis indicates the number of articles published each year. The graph presents two distinct lines: a blue line representing the annual publications, and an orange line representing the cumulative publications up to each corresponding year.

The graph provides valuable insights into the dynamic nature of research activities over time. Initially, research was observed in 2005 [Chen et al., 2005], as indicated by the blue line. Subsequently, there was a noticeable gap in the number of research publications until 2012. One hypothesis about the lack of research before 2012 could be that many researchers focused on student modeling, as shown in the literature review conducted in [Chrysafiadi and Virvou, 2013] during that time. Therefore, it was not common to perform content modeling, or it was performed by researchers but was not defined as such.

A discernible increasing trend emerged from 2012 to 2017, signifying an increase in the annual research production. This phase is represented by an upward trajectory in the blue line, illustrating progressive growth in annual publications. For example, Wauters et al. [Wauters et al., 2012], Jarušek and Pelánek [Jarušek and Pelánek, 2012] and Abbakumov [Abbakumov, 2014] proposed a variant of the IRT model to estimate content difficulty. Martínez-Plumed et al. [Martínez-Plumed et al., 2016] estimated IRT indicators, such as guessing, discrimination, and difficulty. In 2017, Huang and Wu [Huang and Wu, 2017] proposed the T-BMIRT model, a temporal multidimensional using IRT to infer student responses to questions.

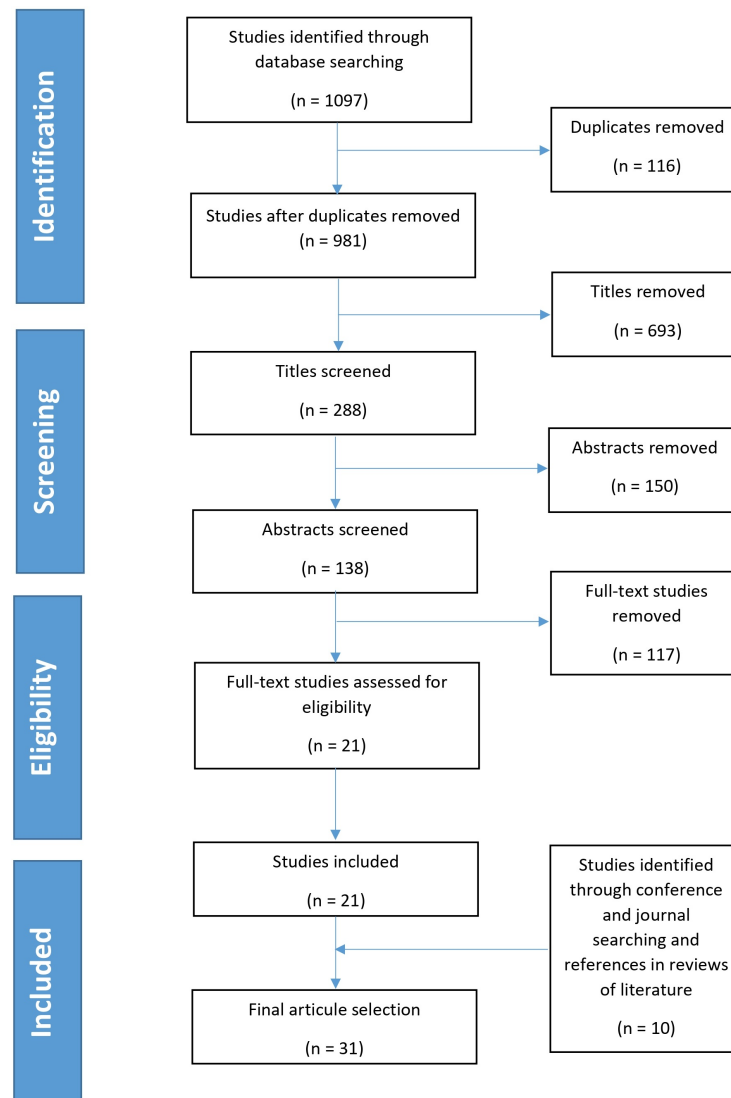


Figure 1: Prisma Flow.

From 2018 to 2023, the graph shows a significant increase in research activity, marked by a substantial increase in the number of articles published annually, which is clearly visible in the ascending blue line. Simultaneously, the orange line rises steadily, indicating the accumulation of articles over time. This cumulative line highlights the aggregate impact of research efforts in each corresponding year, revealing the overall trajectory of steady growth. Table 9 in Appendix A summarizes the results of the selected studies. Rushkin et al.[Rushkin et al., 2018] used log-normal to estimate students' response times to questions. In addition, Converse et al.[Converse et al., 2019]

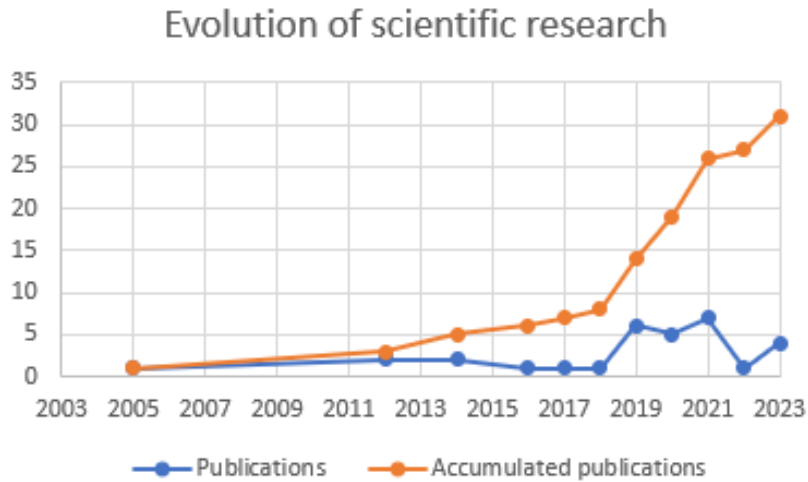


Figure 2: Evolution research.

and Qiu et al.[Qiu et al., 2019] used neural networks to estimate content features such as difficulty and discrimination, respectively. Yaneva et al.[Yaneva et al., 2019] identified Random Forest as the best algorithm for estimating the difficulty of multiple-choice questions. Some studies have also been conducted using IRT: Lalor et al.[Lalor and Wu and Yu, 2019] estimated content difficulty, Martínez-Plumed et al.[Martínez-Plumed et al., 2019] estimated discrimination, and Saxena et al.[Saxena et al., 2021] proposed an IRT++ model to estimate two IRT indicators difficulty and discrimination of two IRT indicators. Jiménez-Macías et al. [Jiménez-Macías et al., 2021] proposed an exercise model using grade, number of attempts, and time spent. Gershon et al. [Gershon et al., 2023] evaluated the stability of item parameters in different massive open online courses (MOOCs) using the item response model (IRT). Dias et al. [Dias et al., 2021] used IRT in Machine Learning to estimate the difficulty, discrimination, and guessing of items in the Fashion MNIST database.

The sharp contrast between the blue and orange lines highlights the transition in annual contributions to the cumulative body of knowledge. The interaction between these lines highlights the evolution of the scientific research landscape, with periods of increased activity, leading to a progressively expanding research base. Table 8 in Appendix A presents a summary of the results of the selected studies, which may be input for further studies by other researchers. In conclusion, this figure visually summarizes the changing research output over the years, highlighting the importance of understanding the trends and contributions of research in the field under review.

We explored our research questions in the following subsections and conduct an in-depth analysis of each question. The results will guide the understanding of the various dimensions of content modeling in virtual educational systems and contribute to the broader discourse surrounding the impact of content modeling. We aimed to identify the indicators, cognitive and metacognitive skills, techniques, models, metrics, areas, platforms, and purposes of educational content modeling through systematic analysis.

3.2 RQ1: For what purpose has content modeling been used?

Content modeling was used for the following purposes. Saxena et al. [Saxena et al., 2021] predicted student responses using dichotomous data by extending the basic IRT model called IRT++, combining 1-parameter and 2-parameter using student responses to questions about mathematical concepts. Furthermore, Converse et al. [Converse et al., 2021] estimated the item indicator difficulty, discrimination, and correlated latent abilities. Deonovic et al. [Deonovic et al., 2020] proposed a new method for analyzing the data generated by massive online learning systems such as Duolingo. In addition, Benedetto et al. [Benedetto et al., 2020a] estimated the difficulty and discrimination of newly created multiple-choice questions using parameters obtained from question text and answer options.

In addition, Chopra et al. [Chopra et al., 2023] analyzed messages posted by university students in online discussion forums to capture the themes and temporal progression of their discourse. Baral et al. [Baral et al., 2021] predicted scores from students' responses to open-ended questions in mathematics using a deep-learning model that uses sentence-level semantic representations. Marinho et al. [Marinho et al., 2023] estimated the difficulty parameters of multiple choice questions in the Brazilian National Secondary Education Exam (ENEM). Jensen et al. [Jensen et al., 2021] explored the possibility of prospectively predicting student success in a short formative assessment as an initial but critical step in implementing intelligent and well-timed suggestions.

Next, Converse et al. [Converse et al., 2019] compared two different model types of neural networks using autoencoders and variational autoencoders. Benedetto et al. [Benedetto et al., 2020b] proposed R2DE (Regressor for Difficulty and Discrimination Estimation), a model that estimates the difficulty and discrimination of each question using question text and answer options. Xue et al. [Xue et al., 2022] converted observed response patterns to continuous latent traits and approximated some continuous functions.

Lehman and Zapata-Rivera [Lehman and Zapata-Rivera, 2018] detected the emotions students experienced while completing non-traditional assessments. Furthermore, Rushkin et al. [Rushkin et al., 2018] estimated the response time for assessment items using log-normal for online courses. Qiu et al. [Qiu et al., 2019] proposed a document-enhanced attention-based Neural Network (DAN) to estimate the difficulty of multiple choice problems in medical exams. Xue et al. [Xue et al., 2020] estimated the difficulty and response time of multiple-choice questions using transfer learning. Yaneva et al. [Yaneva et al., 2019] estimated the difficulty of Multiple-Choice Questions (MCQs) from a high-stakes medical exam.

Next, Gershon et al. [Gershon et al., 2023] evaluated the stability of item parameters in different administrations of a massive open online course (MOOC) using the item response model (IRT). Jiménez et al. [Jiménez-Macias et al., 2023] analyzed the contributions of the proposed model in each scenario and the possible decisions of the teacher based on the results obtained, for example, redesigning the content of the exercise to improve student learning. Dias et al. [Dias et al., 2021] used Item Response Theory to evaluate items' difficulty, discrimination, and guessing, investigating the effect of increasing the number of training samples in the Fashion MNIST database.

Lalor et al. [Lalor and Wu and Yu, 2019] estimated the difficulty of IRT models using response patterns (RPs) generated using artificial crowds of DNN models. Moreover, Martínez-Plumed et al. [Martínez-Plumed et al., 2019] estimated IRT parameters, such as discrimination, difficulty, and guessing, using different algorithms. Jarušek and Pelánek [Jarušek and Pelánek, 2012] estimated problem-solving times using a linear re-

lationship between the student's latent problem-solving ability and the logarithm of the time it takes the student to solve the problem. In addition, Chen et al. [Chen et al., 2005] proposed a personalized learning system based on IRT (PEL-IRT), using the difficulty of the course content and the student's ability to provide individual learning paths for learners. Abbakumov [Abbakumov, 2014] estimated the difficulty level of items and the solution of the "cold start problem." Wauters et al. [Wauters et al., 2012] detected the starting difficulty of content using methods other than the traditional IRT.

Finally, Uto [Uto, 2019] proposed a variant of the IRT model using ratings and semantic features of students' responses to written essay questions using topic models or deep neural networks. In addition, Huang and Wu [Huang and Wu, 2017] proposed a model based on IRT called T-BMIRT, a temporal and multidimensional model capable of predicting student response to assessments. Martínez-Plumed et al. [Martínez-Plumed et al., 2016] analyzed experiments with different datasets and classifiers to identify problems in interpreting IRT parameters: discrimination, guessing, and student ability. Additionally, Jia and Le [Jia et al., 2020] estimated indicators of quiz questions used in an intelligent online tutoring system called "Lexue100" and designed adaptive tests on mathematics.

3.3 RQ2: Which areas or platforms have used educational content modeling?

Table 2 shows the results obtained, where 22 out of 31 studies did not specify the educational level at which content modeling was conducted. Two studies were carried out at the university level: one at the University Higher School of Economics [Abbakumov, 2014], the other in various university programs [Wauters et al., 2012], different periods between fall 2019 and spring 2020 at a public university in the United States [Chopra et al., 2023], different periods such as spring 2015 (2015S), fall 2015 (2015F) and spring 2016 (2016S) in a MOOC [Gershon et al., 2023]. Furthermore, Jia and Le [Jia et al., 2020] implemented the model in an intelligent tutoring system for high school students in China. Also, Marinho et al. [Marinho et al., 2023] used the model in the National Examination of Secondary Education (ENEM), which is composed of four knowledge areas: Languages and Codes (LC), Human Sciences (CH), Natural Sciences (CN), and Mathematics (MT). Jarušek and Pelánek [Jarušek and Pelánek, 2012] employed a tutoring system that combined high school and university students.

Table 3 shows the distributions across the different platforms. Content modeling is used most frequently in Learning Management Systems (LMS), MOOCs, and online courses. Additionally, conversational agents and engineering laboratories were also included in the modeling studies.

3.4 RQ3: Which metrics have been used to evaluate the models in virtual systems?

Table 4 shows the obtained results. Among the metrics found in the selected articles, we have: Accuracy (ACC), Area under the ROC Curve (AUC), Root mean squared error (RMSE), Mean Absolute Error (MAE), Root mean squared difference (RMSD), Spearman Rank Correlation Coefficient (SCC), Kendall Rank Correlation Coefficient (KCC), Absolute value relative bias (AVRB), Word Mover's Distance (WMD), Pearson correlation coefficient, Smirnov test, ANOVA test and correlation (CORR), with RMSE being the most commonly used metric. Most studies have evaluated their proposed content models by comparing them with existing models or by varying algorithms.

| Areas | Studies | Total |
|-------------------------------------|---|-------|
| Combined (High school – university) | [Jarušek and Pelánek, 2012] | 1 |
| High school | [Jia et al., 2020] [Marinho et al., 2023] | 2 |
| University | [Abbakumov, 2014];[Yang et al., 2021] [Wauters et al., 2012];[Chopra et al., 2023] [Jensen et al., 2021];[Gershon et al., 2023] | 6 |
| Not specified | [Xue et al., 2022];[Saxena et al., 2021]; [Converse et al., 2021]; [Benedetto et al., 2020b]; [Xue et al., 2020];[Deonovic et al., 2020]; [Benedetto et al., 2020b];[Uto, 2019]; [Martínez-Plumed et al., 2019]; [Converse et al., 2019];[Qiu et al., 2019]; [Lalor and Wu and Yu, 2019]; [Yaneva et al., 2019];[Rushkin et al., 2018]; [Huang and Wu, 2017]; [Martínez-Plumed et al., 2016]; [Kadengye, 2014];[Chen et al., 2005] ; [Jiménez-Macías et al., 2021] [Jiménez-Macías et al., 2023] [Baral et al., 2021];[Dias et al., 2021] | 22 |

Table 2: Distribution of areas

Saxena et al.[Saxena et al., 2021] evaluated the accuracy of their algorithm in validation and testing against four other models and four baseline IRT models with 1-parameter and 2-parameter IRT variants and obtained a value of 0.7090 using a dataset with 1774 answered questions. In addition, Converse et al.[Converse et al., 2021] evaluated three different parameter estimation techniques and datasets to determine the lowest possible error using the RMSE metric. The datasets consisted of 50 items with 20,000 students, 28 items with 2922 students, 200 items with 50,000 students, and 27 items with 3,000 students. Uto[Uto, 2019] evaluated his model with the RMSE metric calculated between the expected grades and observed mean grades using different datasets with different topics and raters, with a dataset of four essays for 34 students.

Next, Jiménez-Macías et al [Jiménez-Macías et al., 2021] evaluated the accuracy of different classifier algorithms using 200 simulated data points representing student interactions and a model trained with unbalanced data for each class in the grade indicator. Huang and Wu[Huang and Wu, 2017] compared their proposed model with other existing models using the ACC and AUC metrics. The dataset consisted of interactions in 1427 evaluations by 860 students, with the model being trained on 90% of the dataset and 10% for testing, and obtained a result of 0.743 for ACC and 0.815 for AUC.

Additionally, Martínez-Plumed et al.[Martínez-Plumed et al., 2016] used different datasets as inputs in different models to test the effectiveness of their model using the ACC metric. The different datasets consisted of 1745 separate instances in nine different

| Platforms | Studies | Total |
|---|---|-------|
| Intelligent tutoring system | [Jia et al., 2020] | 1 |
| Engineering lab | [Lalor and Wu and Yu, 2019] | 1 |
| Simulation | [Jiménez-Macías et al., 2021] [Jiménez-Macías et al., 2023] | 2 |
| Learning environment | [Xue et al., 2022];[Kadengye, 2014] | 2 |
| Not specified | [Martínez-Plumed et al., 2019]; [Martínez-Plumed et al., 2016]; [Jarušek and Pelánek, 2012]; [Wauters et al., 2012] [Marinho et al., 2023];[Dias et al., 2021] | 6 |
| e-learning platform (LMS,MOOCs, online platform / course) | [Saxena et al., 2021];[Converse et al., 2021]; [Benedetto et al., 2020a];[Xue et al., 2020]; [Deonovic et al., 2020]; [Benedetto et al., 2020b]; [Uto, 2019];[Converse et al., 2019]; [Yaneva et al., 2019]; [Qiu et al., 2019]; [Rushkin et al., 2018]; [Huang and Wu, 2017]; [Abbakumov, 2014];[Chen et al., 2005]; [Yang et al., 2021]; [Jensen et al., 2021]; [Chopra et al., 2023]; [Baral et al., 2021];[Gershon et al., 2023]; | 19 |

Table 3: Distribution of platforms

datasets, and 128 classifiers were used to modify the parameters between 15 different algorithms. In their proposed model, Converse et al.[Converse et al., 2019] used different inputs, such as autoencoders (AE) and variational autoencoders (VAE), to estimate content indicators for discrimination and difficulty, using the metrics AVR, RMSE, and CORR. The model was tested using simulated data from 10,000 students with an assessment of 28 items per student.

Benedetto et al.[Benedetto et al., 2020a][Benedetto et al., 2020b], Yaneva et al. [Yaneva et al., 2019], Jarušek and Pelánek [Jarušek and Pelánek, 2012], and Xue et al. [Xue et al., 2022][Xue et al., 2020] compared these models with other models proposed by different authors using the RMSE metric. Yaneva et al.[Yaneva et al., 2019] estimated the difficulty of multiple-choice questions using a 12038 medical licensing questions. Furthermore, Jarušek and Pelánek[Jarušek and Pelánek, 2012] used their model in 20 schools with more than 5000 users within a tutoring system with 20 types of computer science problems.

Moreover, Lehman and Zapata-Rivera[Lehman and Zapata-Rivera, 2018] and Rushkin et al. [Rushkin et al., 2018] used standard deviations and p-values as metrics for comparing models. Qiu et al. [Qiu et al., 2019] compared different models using the RMSE, MAE, SCC, and KCC metrics to estimate the difficulty of multiple-choice questions with a dataset containing 16,342 questions for 394 students. Martínez-Plumed et al. [Martínez-Plumed et al., 2019] and Wauters et al. [Wauters et al., 2012] used an accuracy

metric to evaluate their models against existing models.

Dias et al. [Dias et al., 2021] evaluated the accuracy of the proposed model using the original data set and the augmented data set. The dataset consisted of a training set of 60,000 examples and a test set of 10,000 examples. Baral et al. [Baral et al., 2021] compared this model with six other Rasch models using machine learning techniques, and the metrics evaluated were the Pearson's correlation coefficient, Spearman's correlation coefficient, MAE, and MSE. The dataset consisted of 150,477 student responses to open-ended questions from 27,199 unique students who responded to 2,076 unique questions. Gershon et al. [Gershon et al., 2023] used statistical tests such as Smirnov's test and ANOVA to determine whether the distributions of the parameters differed significantly in three different courses. A dataset of 1,278 items with 12,338 students was enrolled in three terms.

Next, Jiménez-Macias et al [Jiménez-Macias et al., 2023] evaluated the proposed exercise model using seven different scenarios as inputs, with 300 simulated students. The model was trained using 80% of the dataset and tested using 20% of the data. The metrics evaluated in each scenario were the precision, recall, f1-score, default unnormalized root mean square error (RMSE), and area Under the Curve (AUC). Chopra et al. [Chopra et al., 2023] evaluated the quality of the topics using the inverted Rank-Biased Overlap (IRBO), normalized pointwise mutual information (NPMI), and word embeddings-based similarity. In addition, the word mover's distance (WMD) was used to measure the semantic similarity between topics in adjacent months and to construct topic chains to evaluate the evolution of topics. The dataset contained 32,409 posts created by 449 students from 636 courses.

Finally, Lalor et al. [Lalor and Wu and Yu, 2019] calculated the RMSD metric to evaluate their difficulty model using two parameter estimates: marginal maximum likelihood (MML) and variational inference (VI). Chen et al. [Chen et al., 2005] used a survey with a 5-point Likert scale to evaluate satisfaction with the proposed model for the difficulty of course materials. The dataset included 35 course materials with different difficulty levels and 210 users in the system. Finally, in Abbakumov [Abbakumov, 2014], Jia and Le [Jia et al., 2020], and Deonovic et al. [Deonovic et al., 2020], an evaluation model was not indicated.

3.5 RQ4: Which techniques or models have been used to model educational content in virtual systems?

Table 5 the different options for implementing the content-modeling algorithms. Several models were used in the same study to evaluate the results obtained. Thus, we explain how each algorithm is used below:

- K-Nearest Neighbor: Jiménez-Macias et al. [Jiménez-Macias et al., 2023] used the Nearest Neighbor algorithm with a value of k equal to 10 to predict the grade obtained based on the number of attempts and time spent during the exercise. The authors performed exercise simulations using multiple-choice questions (MCQs) with a probability of correct answers of 7%. The dataset consisted of 300 simulated students, each interacting with the exercise at least once.
- Log-normal: Rushkin et al. [Rushkin et al., 2018] used a log-normal model to estimate time intensity, discrimination, and the influence of the correctness of responses. Time-intensity was defined as a measure of difficulty for each question. The model was tested on 47 HarvardX STEM and non-STEM courses with over 34,000 students and 4,000 multiple-choice assessment questions with multiple attempts. The

| Metrics | Studies | Total |
|--|--|-------|
| Likert scale | [Chen et al., 2005] | 1 |
| None | [Deonovic et al., 2020];[Jia et al., 2020]; [Abbakumov, 2014] | 3 |
| Comparative model with different metrics | [Chopra et al., 2023]; [Marinho et al., 2023] [Jensen et al., 2021]; [Jiménez-Macías et al., 2023]; | 4 |
| Comparative with different inputs | [Xue et al., 2020];[Converse et al., 2019] [Uto, 2019];[Yang et al., 2021] [Gershon et al., 2023];[Dias et al., 2021] | 6 |
| Comparative with different models | [Xue et al., 2022]; [Saxena et al., 2021]; [Benedetto et al., 2020b]; [Benedetto et al., 2020a]; [Converse et al., 2021]; [Qiu et al., 2019]; [Yaneva et al., 2019]; [Lalor and Wu and Yu, 2019]; [Martínez-Plumed et al., 2019]; [Rushkin et al., 2018]; [Huang and Wu, 2017]; [Wauters et al., 2012]; [Martínez-Plumed et al., 2016]; [Kadengye, 2014]; [Jiménez-Macías et al., 2021] [Jarušek and Pelánek, 2012]; [Baral et al., 2021] | 17 |

Table 4: Distribution of metrics

model found that the time spent on correct or incorrect answers did not affect its outcome. Latent Dirichlet Allocation: Chopra et al. [Chopra et al., 2023] used the Latent Dirichlet Allocation (LDA) model to extract latent topics, augmenting it with the bag-of-words (BoW) model that integrates the contextualized representation of words. The Word Mover's distance (WMD) algorithm was employed to quantify the semantic similarity between topics in consecutive months, facilitating the construction of topic chains. The corpus employed in the study encompasses messages authored by university students on online discussion forums from fall 2019 to spring 2020 within a public university in the United States. The dataset comprised 32,409 posts generated by 449 students across 636 distinct courses.

- Regression: Xue et al. [Xue et al., 2020] used linear regression (LR) to predict item indicators: difficulty and response time using input features extracted from a dataset of 18,000 multiple-choice questions from a medical licensing exam. The authors preprocessed the text of the questions using the ELMo model [Baldwin et al., 2021]. Moreover, Benedetto et al. [Benedetto et al., 2020a] used a regression module to estimate content features, including difficulty and discrimination, from questions and item responses using linguistic features. The dataset comprised 11,000 multiple-choice questions with four possible answers from CloudAcademy E-learning. Jiménez-Macías et al. [Jiménez-Macías et al., 2021] used different ma-

chine learning algorithms to estimate the indicators of the exercises, such as the number of attempts, grades, and time spent. They used simulations with 200 interactions across three different difficulty levels of the exercises to evaluate the proposed model. Using a logistic regression algorithm, they obtained the best results based on the proposed metrics.

- Random forest: Benedetto et al. [Benedetto et al., 2020b] tested random forest (RF), decision trees (DT), support vector machines (SVM), and linear regression (LR). They obtained the best performance with a random forest (RF) regressor using the encoded text of questions and answers as the input. The RF model consisted of 250 estimators for estimating the difficulty, each with a maximum depth of 50, and 100 estimators for estimating the discrimination, each with a maximum depth of 25. These indicators were estimated using multiple choice questions. Second, Yaneva et al. [Yaneva et al., 2019] tested random forests, support vector machines, linear regression, Gaussian processes, and dense neural networks (three layers). They obtained the best performance using a Random Forest to predict the difficulty in 12,038 multiple-choice questions, with each item answered by 328 users. The model includes linguistic parameters as features. Finally, Baral et al. [Baral et al., 2021] used the Random Forest and XGBoost algorithms combined with natural language processing techniques to assess responses that merge mathematical expressions and non-mathematical text. The initial dataset encompasses 150,477 student responses to open-ended questions on the ASSISTments online learning platform. The second dataset was used for secondary analysis and encompassed 30,371 student responses evaluated by 12 secondary school mathematics instructors. The data were collected during the spring and fall of 2020.
- Neural Network: Xue et al. [Xue et al., 2022] used artificial neural networks (ANNs) to obtain unbiased estimates of item difficulty and discrimination when data are missing, not at random, or have nonignorable missing values (MNAR) in a Virtual Learning Environment (VLE). They used simulated data from 63,625 students and the same number of items. In addition, Converse et al. [Converse et al., 2021] used a neural network to estimate item indicators, such as difficulty and discrimination, using variational autoencoders (VAE) with three parameter estimation techniques. In another study [Converse et al., 2019], the same authors used neural networks, specifically autoencoders (AE) and variational autoencoders (VAE), to estimate item discrimination and difficulty indicators using the same dataset. Also, Uto [Uto, 2019] proposed a model for automating grade estimation in essay writing using a deep neural network with semantic features and 34 students as users who completed four tasks each. Qiu et al. [Qiu et al., 2019] proposed a Document enhanced Attention-based neural Network (DAN) framework to estimate the difficulty of Multiple-Choice Problems in medical exams using more than 800,000 test logs with information about the questions and answers made by each student. Finally, Yang et al. [Yang et al., 2021] proposed a model for sentiment analysis using voice and text from videos in a MOOC environment. The study did not specify the number of videos analyzed.
- IRT is the most common machine-learning algorithm in content modeling. We only included studies that modified the baseline IRT model in the present study. The following are some of the papers that we found. Saxena et al. [Saxena et al., 2021] combined the 2-parameter IRT model with the 1-parameter IRT model and optimized parameters using adaptive gradient descent and random-normal parameter

initialization to estimate item difficulty and discrimination using 542 students and 1774 questions answered. Huang and Wu [Huang and Wu, 2017] proposed the T-BMIRT (a temporal combined multimodal IRT) model using video learning indicators to predict evaluation indicators for a dataset with 860 students and 1427 assessments. Gershon et al. [Gershon et al., 2023] used the IRT model to estimate the item difficulty and discrimination parameters across different course instances. MITx provided data for three periods of the 'Advanced Introductory Classical Mechanics' course: spring 2015 (2015S), fall 2015 (2015F), and spring 2016 (2016S), involving a total of 12,338 enrolled students. The dataset encompassed 1,713 items in 2015S, 1,780 in 2015F, and 1,760 in 2016S. A subset of 1,278 items was tested across the three periods and referred to as items from the initial response matrices.

In conclusion, the most popular content-modeling algorithm is IRT, and numerous studies have modified the original IRT model to increase its precision. However, other algorithms, such as random forest, neural networks, and regression, have been used to estimate item indicators. Overall, this research question provides a useful overview of the different models and techniques used to model educational content in virtual systems.

3.6 RQ5: What cognitive and meta-cognitive skills are involved in the models for educational content in virtual systems?

Cognitive skills are the mental operations necessary for learning, including memory, attention, perception, and logic. Meta-cognitive skills are abilities to monitor and control one's own thinking processes. In this research question, we identify which skills have been involved in the different models proposed by the authors. Table 6 shows three skills: ability, efficiency, and slowness.

The most implied skill in the content models is student ability. Depending on the particular ability being discussed, the student's ability may be either a cognitive or a metacognitive. In IRT, student ability refers to the latent trait or construct that is being measured based on the student's performance on test or assessment items. Studies found that student ability is related to difficulty [Xue et al., 2020], [Jarušek and Pelánek, 2012] [Chen et al., 2005], [Xue et al., 2022] [Converse et al., 2019], [Jensen et al., 2021], [Gershon et al., 2023], [Dias et al., 2021], discrimination [Martínez-Plumed et al., 2019], [Jia et al., 2020], [Saxena et al., 2021],[Gershon et al., 2023],[Dias et al., 2021] and guessing [Martínez-Plumed et al., 2016], [Jia et al., 2020], [Dias et al., 2021]. In IRT, student ability is not considered a cognitive or metacognitive skill, but rather a latent construct or trait as a measure of a student's proficiency by a test or assessment [Rasch, 1993]. However, cognitive and metacognitive skills may contribute to a student's level of ability as measured by IRT.

In [Jiménez-Macias et al., 2021], student efficiency was proposed as the relationship between the exercise indicators: grade, number of attempts and time spent based on student simulations. Student efficiency encompasses both cognitive and metacognitive skills, as well as other factors such as motivation and engagement, which are also important in assessing a student's learning efficiency [Yu et al., 2021].

Furthermore, in [Rushkin et al., 2018], student slowness was defined as a measure of the time it takes a user to answer a question during an assessment. Student slowness is not a cognitive or metacognitive skill, but rather a description of the rate at which a student completes a task or learns a new concept. Slowness can be influenced by several factors, such as the student's cognitive abilities, metacognitive skills, and external factors. However, slowness is not a skill itself, but rather a characteristic of the student.

| Models | Studies | Total |
|-----------------------------|--|--------------|
| K-Nearest Neighbor | [Jiménez-Macías et al., 2023] | 1 |
| Latent Dirichlet Allocation | [Chopra et al., 2023] | 1 |
| Log-normal | [Rushkin et al., 2018] | 1 |
| Random Forest | [Benedetto et al., 2020b]; [Yaneva et al., 2019] [Baral et al., 2021] | 3 |
| Regresión | [Benedetto et al., 2020a]; [Xue et al., 2020]; [Jiménez-Macías et al., 2021]; | 3 |
| Neural Network | [Xue et al., 2022];[Converse et al., 2021] [Uto, 2019];[Qiu et al., 2019]; [Converse et al., 2019];[Yang et al., 2021] | 6 |
| IRT | [Saxena et al., 2021]; [Deonovic et al., 2020]; [Jia et al., 2020]; [Lalor and Wu and Yu, 2019]; [Martínez-Plumed et al., 2019]; [Huang and Wu, 2017]; [Martínez-Plumed et al., 2016]; [Abbakumov, 2014]; [Kadengye, 2014]; [Jarušek and Pelánek, 2012]; [Wauters et al., 2012]; [Chen et al., 2005] [Marinho et al., 2023];[Jensen et al., 2021] [Gershon et al., 2023];[Dias et al., 2021] | 16 |

Table 5: Distribution of models

In summary, student ability is one of the most commonly used skills in educational models. However, content modeling allows researchers to infer other cognitive and metacognitive skills based on estimated indicators. For example, through content modeling, researchers can determine which content would or would not facilitate a student's help-seeking behavior [Wilson et al., 2005].

3.7 RQ6: Which indicators have been used for content modeling?

Our literature review also found studies that specified the type of data content used, for example, assessment question data [Benedetto et al., 2020a], [Benedetto et al., 2020b], [Qiu et al., 2019] [Yaneva et al., 2019], [Xue et al., 2020] [Marinho et al., 2023], assessment responses [Qiu et al., 2019], [Martínez-Plumed et al., 2019], [Baral et al., 2021]. Table 7 shows the different indicators used in assessment modeling, being difficulty, discrimination, and time spent the most used. Few studies obtained other content indicators. For example, Jiménez-Macías et al. in [Jiménez-Macías et al., 2021] and [Jiménez-Macías et al., 2023] proposed an exercise model with indicators: grade, time spent and number of attempts using simulations-data. Yang et. al. [Yang et al., 2021] analyzed

| Skills | Studies | Total |
|------------|--|-------|
| Efficiency | [Jiménez-Macías et al., 2021] | 1 |
| Slowness | [Rushkin et al., 2018] | 1 |
| None | [Yang et al., 2021] [Chopra et al., 2023] [Baral et al., 2021] [Jiménez-Macías et al., 2023] [Marinho et al., 2023] | 5 |
| Ability | [Xue et al., 2022];[Saxena et al., 2021]; [Converse et al., 2021];[Xue et al., 2020]; [Deonovic et al., 2020];[Jia et al., 2020]; [Benedetto et al., 2020b];[Xue et al., 2020]; [Qiu et al., 2019];[Chen et al., 2005] [Uto, 2019];[Lalor and Wu and Yu, 2019]; [Martínez-Plumed et al., 2019] [Yaneva et al., 2019];[Converse et al., 2019] [Huang and Wu, 2017];[Martínez-Plumed et al., 2016]; [Abbakumov, 2014];[Wauters et al., 2012]; [Kadengye, 2014];[Jarušek and Pelánek, 2012]; [Jensen et al., 2021];[Dias et al., 2021] [Gershon et al., 2023] | 24 |

Table 6: Distribution of skills

the feelings produced by videos using text and voice in a MOOC. Uto[Uto, 2019] estimates the grade on a written essay. Rushkin et al.[Rushkin et al., 2018] proposed a log-normal statistical model to estimate the response time in an assessment question in an online course. Huang and Wu[Huang and Wu, 2017] proposed a model capable of predicting the following student response in assessment. Guessing has been estimated using IRT alone [Martínez-Plumed et al., 2016], [Jia et al., 2020], citedias2021 use. Furthermore, for discrimination, different algorithms such as Regression [Benedetto et al., 2020a], Random Forest [Benedetto et al., 2020b], Neural network [Xue et al., 2022], [Converse et al., 2021], [Converse et al., 2019] and IRT [Martínez-Plumed et al., 2019], [Martínez-Plumed et al., 2016], [Jia et al., 2020], [Saxena et al., 2021], [Gershon et al., 2023],[Dias et al., 2021]. Finally, the most estimated content indicator is the difficulty using the following algorithms: Regression [Benedetto et al., 2020a], [Xue et al., 2020], Random Forest [Benedetto et al., 2020b], [Yaneva et al., 2019], Neural network [Qiu et al., 2019], [Xue et al., 2022], [Converse et al., 2021], [Converse et al., 2019] and IRT [Abbakumov, 2014], [Lalor and Wu and Yu, 2019], [Martínez-Plumed et al., 2019], [Martínez-Plumed et al., 2016], [Jarušek and Pelánek, 2012], [Chen et al., 2005], [Wauters et al., 2012], [Kadengye, 2014], [Jia et al., 2020], [Saxena et al., 2021], [Deonovic et al., 2020], [Marinho et al., 2023], [Jensen et al., 2021]. Based on the articles found, the main conclusion is that difficulty and discrimination were the indicators used so far. Other indicators such as grade, time spent have been studied in the last years in only 7 papers out of 31.

4 Future lines of research

The results of this systematic literature review provide an overview of content modeling using machine learning algorithms. This section outlines the insights gleaned and points to possible avenues for future research that address the identified research questions and gaps. These directions may serve as valuable guidance for other researchers interested in content modeling.

- Exploring new content types: Most authors used multiple-choice questions in their models, except for Uto [Uto, 2019], who proposed a model using written essays. Future research could analyze other types of questions, such as true/false, fill-in-the-blank, multiple-choice fill-in-the-blank, and multiple-choice questions with only one correct answer. For example, the grade distribution for a true/false question could differ from that of a writing question, and the number of attempts for a writing question could not be the same as that for a multiple-choice question. In addition, other types of content were found in the results, such as videos, lectures, files, and discussion forums. For instance, sentiment analysis can be performed on messages made by students in discussion forums.
- Incorporating unexplored indicators: The results demonstrated that difficulty, discrimination, and guessing indicators are the most commonly used. These three indicators are typically estimated using IRT and various machine-learning algorithms. The authors often compared their results with those obtained using IRT to measure the effectiveness of their proposed models. Furthermore, we propose a robust approach for future research. Building on our analyses, we suggest the exploration of previously unexamined indicators. For example, authors rarely use indicators such as grade, number of attempts, and time spent simultaneously in their models. Future research could incorporate the estimation of these indicators and their inclusion in the predictive model, as presented by [Jiménez-Macias et al., 2021].
- Evaluation of content modeling approaches: IRT is still the most widely used algorithm in content modeling, but alternative machine-learning techniques have been adopted in recent years. In particular, researchers have explored using other machine learning algorithms, such as neural networks [Qiu et al., 2019, Xue et al., 2022, Uto, 2019, Converse et al., 2021], random forests [Benedetto et al., 2020b, Yaneva et al., 2019], and regression [Benedetto et al., 2020a, Xue et al., 2020], to model content features. Moreover, model evaluation is a critical factor in the research process to select the best algorithm that can efficiently solve the problems posed, as stated by Raschka [Raschka, 2018]. In the selected studies, the proposed model was compared with other models (from other authors, or the same model with a different algorithm). Additionally, comparisons were made using metrics such as accuracy, precision, F1-SCORE, AUC, RMSE, and MAE, as indicated by Pelánek [Pelánek, 2015]. Further research could delve into contextual variations in the use of metrics in different data sets, identifying optimal choices for specific scenarios.
- Enhancing generalizability across environments: Based on our analyses, we recommend investigating the generalizability of content models in various intelligent environments, including LMS, MOOCs, and conversational agents at different educational levels, including primary, secondary, and university. However, generalizing models is limited by the number of required content interactions. A line of research could define and understand the behavior of different intelligent environments. For

example, the number of students in a MOOC is likely to be much larger than that in an LMS, and understanding the implications of these differences could inform the development of more effective models.

5 Conclusions, limitations and future work

This systematic literature review critically examines the landscape of studies on content modeling using machine learning algorithms based on learner interactions. The results indicate that this is an emerging area that is ripe for future research. One way to contribute is to perform innovative content modeling by leveraging various machine-learning algorithms that go beyond the conventional IRT model. Although the reviewed studies focused mainly on educational exercises, there is scope for developing new models using various types of content, such as discussion forums.

In addition, content models have the potential to serve as an avenue for inferring hitherto unexplored cognitive and metacognitive skills, thereby deepening our understanding of students' teaching and learning processes. In our study, we also detected a boom in content modeling research. These recent efforts are characterized by the integration of new indicators and alternative algorithms, illustrating a trajectory that has developed over the last two years.

Despite the rigor of our methodology, it is crucial to acknowledge its inherent limitations. The selection of studies and their representation may have been influenced by the specific keywords and the search criteria employed. Consequently, there is a possibility that some pertinent studies may not have been included in our analysis, potentially affecting the comprehensiveness of our conclusions. To mitigate this concern, we conducted manual searches across distinguished conferences and journals to enhance the scope and coverage within the domain. Another limitation was the selection of articles published in English to exclude other languages such as Spanish and Chinese. This criterion was used because most of the literature in this area is in this language.

Furthermore, it is important to note that our study was bounded by a restricted time-frame and selection of conferences/journals due to the constraints posed by the availability of pertinent papers for manual review. To overcome these limitations, we recommend that future researchers consider incorporating the term "content modeling" among their chosen keywords. Such an approach would facilitate the identification of relevant studies in future reviews, thereby broadening the perspective and depth of insights within the field.

Regarding implementation practice, our findings suggest several guidelines for optimizing the application of content modeling in smart learning environments. These recommendations encompass diversifying content types to extend beyond educational exercises, incorporating features beyond conventional indicators such as difficulty and discrimination, and embracing machine-learning algorithms tailored to specific objectives while ensuring rigorous evaluation through comprehensive metrics. These practices aim to enrich the understanding of content dynamics, student engagement, and the overall efficacy of content models.

In future work, we intend to develop new content models for intelligent learning environments based on learner interactions that allow us to make inferences about different skills—exploring various types of content, such as true/false and fill-in-the-blank questions, and investigating under-explored indicators, such as sentiment analysis and emotional responses, to obtain further insight into learner engagement. We also plan to incorporate other types of content, such as discussion forums. In addition, we intend to

provide teachers with a tool for visualizing the results obtained in the content model, allowing them to identify potential problems and redesign the content or intervene in the orchestration of the course; for example, adapting the content to better fit the needs of the learners. In addition, there is the potential to assess the generalizability of content models in different intelligent environments, considering factors such as user volume and engagement patterns, to encourage the development of adaptive models.

Acknowledgements

This work was supported in part by the FEDER/Ministerio de Ciencia, Innovación y Universidades–Agencia Estatal de Investigación, through the Smartlet Project under Grant TIN2017-85179-C3-1-R and the H2O Learn Project under Grant PID2020-112584 RB-C31, in part by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307 a project which is co-funded by the European Structural Funds (FSE and FEDER).

References

- [Abbakumov, 2014] Abbakumov, D. “The solution of the “cold start problem” in e-Learning”; *Procedia-Social and Behavioral Sciences*. pp.1225-1231. (2014)
- [Abyaa et al., 2019] Abyaa, A., Khalidi Idrissi, M., and Bennani, S “Learner modelling: systematic review of the literature from the last 5 years”; *Educational Technology Research and Development*, 67, pp.1105-1143. (2019)
- [Akhras, 2005] Akhras, F. N “Modelling the context of learning interactions in intelligent learning environments”; In *Modeling and Using Context: 5th International and Interdisciplinary Conference CONTEXT 2005, Proceedings 5* , pp. 1-14. (2005)
- [AlKhuzaey et al., 2021] AlKhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V “A systematic review of data-driven approaches to item difficulty prediction.”; In *International Conference on Artificial Intelligence in Education*, pp. 29-41. (2021)
- [Alrajhi et al., 2020] Alrajhi, L., Alharbi, K., and Cristea, A. I. “A multidimensional deep learner model of urgent instructor intervention need in MOOC forum posts”; In *Intelligent Tutoring Systems: 16th International Conference, Proceedings 16*, pp. 226-236. (2020)
- [Atapattu et al., 2020] Atapattu, T., Falkner, K., Thilakaratne, M., Sivaneasharajah, L., and Jayashanka, R “What do linguistic expressions tell us about learners’ confusion? A domain-independent analysis in MOOCs”; *IEEE Transactions on Learning Technologies*, 13(4), pp.878-888. (2020)
- [Baldwin et al., 2021] Baldwin, P., Yaneva, V., Mee, J., Clauser, B. E., and Ha, L. A “Using natural language processing to predict item response times and improve test construction”; *Journal of Educational Measurement*, 58(1), pp.4-30. (2021)
- [Baral et al., 2021] Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., and Heffernan, N. T “Improving Automated Scoring of Student Open Responses in Mathematics.”; *International Educational Data Mining Society* (2021)
- [Benedetto et al., 2020a] Benedetto, L., Cappelli, A., Turrin, R., and Cremonesi, P. “Introducing a framework to assess newly created questions with Natural Language Processing”; In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Proceedings Part I 21* , pp. 43-54. (2020a)
- [Benedetto et al., 2020b] Benedetto, L., Cappelli, A., Turrin, R., and Cremonesi, P. “R2DE: a NLP approach to estimating IRT parameters of newly generated questions”; In *Proceedings of the Tenth International Conference on Learning Analytics Knowledge*, pp. 412-421. (2020b)

- [Brajnik, 2007] Brajnik, G. "Modeling content and expression of learning objects in multimodal learning management systems"; *Lecture Notes in Computer Science*, 4556, p.501. (2007)
- [Brusilovsky et al., 2014] Brusilovsky, P., Edwards, S., Kumar, A., Malmi, L., Benotti, L., Buck, D., Ihantola, P., Prince, R., Sirkiä, T., Sosnovsky, S. and Urquiza, J. "Increasing adoption of smart learning content for computer science education"; In *Proceedings of the Working Group Reports of the 2014 on Innovation Technology in Computer Science Education Conference*, pp. 31-57. (2014)
- [Capuano et al., 2021] Capuano, N., Caballé, S., Conesa, J. and Greco, A. "Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis"; *Journal of Ambient Intelligence and Humanized Computing*, 12, pp.9977-9989 (2021)
- [Chen et al., 2005] Chen, C.M., Lee, H.M. and Chen, Y.H. "Personalized e-learning system using item response theory"; *Computers Education*, 44(3), pp.237-255. (2005)
- [Chopra et al., 2023] Chopra, H., Lin, Y., Amin, M., Cavazos, J.G., Yu, R., Jaquay, S. and Nixon, N. "Semantic Topic Chains for Modeling Temporality of Themes in Online Student Discussion Forums."; (2023)
- [Chrysafiadi and Virvou, 2013] Chrysafiadi, K. and Virvou, M. "Student modeling approaches: A literature review for the last decades"; *Expert Systems with Applications*, 40(11), pp.4715-4729. (2013)
- [Converse et al., 2019] Converse, G., Curi, M. and Oliveira, S. "Autoencoders for educational assessment"; In *Artificial Intelligence in Education: 20th International Conference, Proceedings Part II 20*, pp. 41-45. (2019)
- [Converse et al., 2021] Converse, G., Curi, M., Oliveira, S. and Templin, J. "Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders"; *Machine learning*, 110(6), pp.1463-1480. (2021)
- [Deonovic et al., 2020] Deonovic, B., Bolsinova, M., Bechger, T. and Maris, G. "A Rasch model and rating system for continuous responses collected in large-scale learning systems"; *Frontiers in psychology*, 11, p.500039. (2020)
- [Desmarais and Baker, 2012] Desmarais, M.C. and Baker, R.S.D. "A review of recent advances in learner and skill modeling in intelligent learning environments"; *User Modeling and User-Adapted Interaction*, 22, pp.9-38. (2012)
- [Dias et al., 2021] Dias, J., Rodrigues, C.M. and Rodrigues, A.C. "Use and Interpretation of Item Response Theory Applied to Machine Learning." In *Latin American Workshop on Computational Neuroscience*, pp. 15-24 (2021)
- [Gershon et al., 2023] Gershon, S.A.K., Anghel, E. and Alexandron, G. "An evaluation of assessment stability in a massive open online course using item response theory"; *Education and Information Technologies*, pp.1-19 (2023)
- [Ghosh et al., 2020] Ghosh, A., Heffernan, N. and Lan, A.S. "Context-aware attentive knowledge tracing"; In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery data mining*, pp. 2330-2339. (2020)
- [Hammad et al., 2017] Hammad, R., Odeh, M. and Khan, Z.A. "eLEM: A Novel e-Learner Experience Model"; *International Arab Journal of Information Technology*, 14(4A), pp.586-597. (2017)
- [Huang and Wu, 2017] Huang, J. and Wu, W. "T-BMIRT: Estimating representations of student knowledge and educational components in online education"; In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 1301-1306, (2017)
- [Jarušek and Pelánek, 2012] Jarušek, P. and Pelánek, R. "Analysis of a simple model of problem solving times"; In *Intelligent Tutoring Systems: 11th International Conference, Proceedings 11*, pp. 379-388. (2012)

- [Jia et al., 2020] Jia, J. and Le, H. “The design and implementation of a computerized adaptive testing system for school mathematics based on item response theory”; In *Technology in Education. Innovations for Online Teaching and Learning: 5th International Conference, Revised Selected Papers 5*, pp. 100-111. (2020)
- [Jiménez-Macias et al., 2021] Jiménez-Macias, Alberto, Pedro J. Muñoz-Merino, and Carlos Delgado Kloos. “A model to characterize exercises using probabilistic methods”; In *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*, pp. 594-599. (2021)
- [Jiménez-Macias et al., 2023] Jiménez-Macias, Alberto, Pedro J. Muñoz-Merino, and Carlos Delgado Kloos. “Recreation of different educational exercise scenarios for exercise modeling.” In *2023 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1-9. (2023)
- [Jensen et al., 2021] Jensen, E., Umada, T., Hunkins, N.C., Hutt, S., Huggins-Manley, A.C. and D’Mello, S.K. “What you do predicts how you do: Prospectively modeling student quiz performance using activity features in an online learning environment.”; In *LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 121-131 (2021)
- [Jia et al., 2022] Jia, X.U., Tingting, W.E.I., Ge, Y.U., Xinyue, H.U.A.N.G. and Pin, L.Y.U. “Review of Question Difficulty Evaluation Approaches.”; *Journal of Frontiers of Computer Science Technology*, 16(4), pp.734 (2022)
- [Kadengye, 2014] Kadengye, D.T., Ceulemans, E. and Van den Noortgate, W. “A generalized longitudinal mixture IRT model for measuring differential growth in learning environments”; *Behavior research methods*, 46, pp.823-840. (2014)
- [Kaplan and Haenlein, 2016] Kaplan, A.M. and Haenlein, M. “Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster”; *Business horizons*, 59(4), pp.441-450. (2016)
- [Lalor and Wu and Yu, 2019] Lalor, J.P., Wu, H. and Yu, H. “Learning latent parameters without human response patterns: Item response theory with artificial crowds”; In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Vol. 2019, p. 4240. (2019)
- [Lehman and Zapata-Rivera, 2018] Lehman, B.A. and Zapata-Rivera, D. “Student emotions in conversation-based assessments”; *IEEE Transactions on Learning Technologies*, 11(1), pp.41-53. (2018)
- [Madhusudhana, 2017] Madhusudhana, K. “The Cognitive Dimension and Course Content Modeling: An Ontological Approach”; *International Journal of Emerging Technologies in Learning*, 12(5). (2017)
- [Marinho et al., 2023] Marinho, W., Clua, E.W., Martí, L. and Marinho, K. “Predicting Item Response Theory Parameters Using Question Statements Texts.”; In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pp. 1-10 (2023)
- [Martínez-Plumed et al., 2019] Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A. and Hernández-Orallo, J. “Item response theory in AI: Analysing machine learning classifiers at the instance level”; *Artificial intelligence*, 271, pp.18-42. (2019)
- [Martínez-Plumed et al., 2016] Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A. and Hernández-Orallo, J. “Making sense of item response theory in machine learning”; In *ECAI 2016*, pp. 1140-1148. (2016)
- [Moher, et al., 2009] Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. and PRISMA Group “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement”; *Annals of internal medicine*, 151(4), pp.264-269. (2009)
- [Pardos and Dadu, 2018] Pardos, Z.A. and Dadu, A. “dAFM: Fusing psychometric and connectionist modeling for Q-matrix refinement”; *Journal of Educational Data Mining*, 10(2), pp.1-27. (2018)

- [Pardos et al., 2017] Pardos, Z.A., Tang, S., Davis, D. and Le, C.V. “Enabling real-time adaptivity in MOOCs with a personalized next-step recommendation framework”; In Proceedings of the fourth (2017) ACM conference on learning@ scale , pp. 23-32. (2017)
- [Pecheanu et al., 2003] Pecheanu, E., Segal, C. and Stefanescu, D. “Content Modeling in Intelligent Instructional Environments”. In Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, Proceedings, Part II 7 , pp. 1229-1234. (2003)
- [Pekrun et al., 2002] Pekrun, R., Goetz, T., Titz, W. and Perry, R.P. “Academic emotions in students’ self-regulated learning and achievement: A program of qualitative and quantitative research”; Educational psychologist, 37(2), pp.91-105. (2002)
- [Pelánek, 2015] Pelánek, R. “Metrics for Evaluation of Student Models”; Journal of Educational Data Mining, 7(2), pp.1-19. (2015)
- [Qiu et al., 2019] Qiu, Z., Wu, X. and Fan, W. “Question difficulty prediction for multiple choice problems in medical exams”; In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 139-148. (2019)
- [Rasch, 1993] Rasch, G “Probabilistic models for some intelligence and attainment tests”; MESA Press, 5835 S. Kimbark Ave., Chicago, IL 60637; e-mail: MESA@uchicago.edu; web address: www.rasch.org; tele (1993)
- [Raschka, 2018] Raschka, S. “Model evaluation, model selection, and algorithm selection in machine learning”; rXiv preprint arXiv:1811.12808. (2018)
- [Rushkin et al., 2018] Rushkin, I., Chuang, I. and Tingley, D. “Modelling and using response times in online courses”; arXiv preprint arXiv:1801.07618. (2018)
- [Saxena et al., 2021] Saxena, N., Lodaya, V. and Thakur, T. “IRT++: Improving Student Response Prediction With Gaussian Initialisation and Other Modifications”; In 2021 International Conference on Advanced Learning Technologies (ICALT), pp. 166-167. (2021)
- [Spector, 2016] Spector, J.M. “Smart learning environments: Concepts and issues”; In Society for Information Technology teacher education international conference ,pp. 2728-2737. (2016)
- [Suraweera et al., 2005] Suraweera, P., Mitrovic, A. and Martin, B “A knowledge acquisition system for constraint-based intelligent tutoring systems”; (2005)
- [Uto, 2019] Uto, M. “Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability”; In Artificial Intelligence in Education: 20th International Conference, Proceedings, Part I 20 ,pp. 494-506. (2019)
- [Xue et al., 2022] Xue, K., Huggins-Manley, A.C. and Leite, W. “Semisupervised learning method to adjust biased item difficulty estimates caused by nonignorable missingness in a virtual learning environment”; Educational and Psychological Measurement, 82(3), pp.539-567. (2022)
- [Wauters et al., 2012] Wauters, K., Desmet, P. and Van Den Noortgate, W. “Item difficulty estimation: An auspicious collaboration between data and judgment”; Computers Education 58, no. 4 , 1183-1193. (2012)
- [Wilson et al., 2005] Wilson, C.J., Deane, F.P., Ciarrochi, J.V. and Rickwood, D. “Measuring help seeking intentions: properties of the general help seeking questionnaire”; (2005)
- [Xue et al., 2020] Xue, K., Yaneva, V., Runyon, C. and Baldwin, P. “Predicting the difficulty and response time of multiple choice questions using transfer learning.”; In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications , pp. 193-197. (2020)
- [Xue et al., 2022] Xue, K., Huggins-Manley, A.C. and Leite, W. “Semisupervised learning method to adjust biased item difficulty estimates caused by nonignorable missingness in a virtual learning environment”; Educational and Psychological Measurement, 82(3), pp.539-567. (2022)

[Yaneva et al., 2019] Yaneva, V., Baldwin, P. and Mee, J. “Predicting the difficulty of multiple choice questions in a high-stakes medical exam”; In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications , pp. 11-20. (2019)

[Yang et al., 2021] Yang, S., Dai, Y., Li, S. and Zhao, K. “An Automatic Analysis and Evaluation System Used for Teaching Quality in MOOC Environment”; In 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI) , pp. 38-41. (2021)

[Yu et al., 2021] Yu, Z., Gao, M. and Wang, L. “The effect of educational games on learning outcomes, student motivation, engagement and satisfaction”; Journal of Educational Computing Research, 59(3), pp.522-546. (2021)

[Zhao et al., 2020] Zhao, S., Wang, C. and Sahebi, S. “Modeling knowledge acquisition from multiple learning resource types”; arXiv preprint arXiv:2006.13390. (2020)

A Synthesis of Accepted Articles

| Indicators | Studies | Total |
|------------------------|--|-------|
| Voice sentiment | [Yang et al., 2021] | 1 |
| Text sentiment | [Yang et al., 2021] | 1 |
| Reply messages | [Chopra et al., 2023] | 1 |
| Response of assessment | [Huang and Wu, 2017];[Baral et al., 2021] | 2 |
| Number of attempts | [Jiménez-Macías et al., 2021] [Jiménez-Macías et al., 2023] | 2 |
| Grade | [Uto, 2019];[Jiménez-Macías et al., 2021] [Jiménez-Macías et al., 2023] | 3 |
| Guessing | [Jia et al., 2020];[Dias et al., 2021] [Martínez-Plumed et al., 2016] | 3 |
| Time spent | [Xue et al., 2020];[Rushkin et al., 2018] [Jiménez-Macías et al., 2021] [Jiménez-Macías et al., 2023] | 4 |
| Discrimination | [Xue et al., 2022]; [Saxena et al., 2021]; [Converse et al., 2021]; [Dias et al., 2021] [Benedetto et al., 2020a];[Jia et al., 2020] [Converse et al., 2019]; [Gershon et al., 2023] [Martínez-Plumed et al., 2019]; [Martínez-Plumed et al., 2016] [Benedetto et al., 2020b]; | 11 |
| Difficulty | [Xue et al., 2022];[Saxena et al., 2021]; [Converse et al., 2021]; [Jia et al., 2020];[Gershon et al., 2023] [Xue et al., 2020]; [Deonovic et al., 2020]; [Benedetto et al., 2020b]; [Benedetto et al., 2020a]; [Qiu et al., 2019];[Yaneva et al., 2019]; [Lalor and Wu and Yu, 2019]; [Converse et al., 2019];[Chen et al., 2005] [Martínez-Plumed et al., 2016]; [Wauters et al., 2012]; [Abbakumov, 2014];[Jensen et al., 2021] [Jarušek and Pelánek, 2012] [Kadengye, 2014];[Dias et al., 2021] [Marinho et al., 2023] | 22 |

Table 7: Distribution of indicators

| Title (Content) | Approach (Learning environment level of education) | Reference |
|---|--|---------------------------|
| Introducing a Framework to Assess Newly Created Questions with Natural Language Processing (Assessment) | Estimate the difficulty and discrimination of newly created Multiple-Choice Questions (e-learning platform) | [Benedetto et al., 2020a] |
| IRT++: Improving Student Response Prediction with Gaussian Initialisation and Other Modifications (Assessment) | Predict student responses in assessment (online education platform) | [Saxena et al., 2021] |
| The Design and Implementation of a Computerized Adaptive Testing System for School Mathematics Based on Item Response Theory (Assessment) | Estimate the student's response to each question on the assessment (Intelligent Tutoring system "Lexue 100") | [Jia et al., 2020] |
| Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders (Assessment) | Estimate item parameters such as difficulty and discrimination and the student's latent capacity (Examination for the Certificate of Proficiency in English) | [Converse et al., 2021] |
| A Rasch model and rating system for continuous responses collected in large-scale learning systems (Assessment) | Estimate item difficulty and analyze data generated by Duolingo (Duolingo multiplatform) | [Deonovic et al., 2020] |
| Autoencoders for educational assessment (Assessment) | Estimate item parameters by comparing different data in a neural network (Certificate of Proficiency in English) | [Converse et al., 2019] |

| | | |
|---|--|----------------------------------|
| T-BMIRT: Estimating representations of student knowledge and educational components in online education (Assessment) | Propose a multidimensional temporal model to estimate item parameters (Online education system) | [Huang and Wu, 2017] |
| A generalized longitudinal mixture IRT model for measuring differential growth in learning environments (Assessment) | Propose a model that combines a longitudinal Rasch model, a mixture Rasch model and a random item IRT model (Web-based e-learning environment) | [Kadengye, 2014] |
| Making sense of item response theory in machine learning (Assessment) | Compare different models and datasets to understand the parameters used in IRT (Not specific) | [Martinez-Plumed et. al. , 2016] |
| Semisupervised Learning Method to Adjust Biased Item Difficulty Estimates Caused by Nonignorable Missingness in a Virtual Learning Environment (Assessment) | Propose a semi-supervised learning model for converting response to latent features and approximating them to a function (Virtual learning environments) | [Xue et al., 2022] |
| R2DE: A NLP Approach to Estimating IRT Parameters of Newly Generated Questions (Assessment) | Estimate the difficulty and the discrimination of question (E-learning platform) | [Benedetto et al., 2020b] |
| Modelling and Using Response Times in Online Courses (Assessment) | Estimate response time in assessment item (Online course) | [Rushkin et al., 2018] |

| | | |
|--|--|-------------------------------|
| Predicting the Difficulty of Multiple-Choice Questions in a High-stakes Medical Exam (Assessment) | Estimate the difficulty of Multiple-Choice Questions (MCQs) a high-stakes medical exam (High-stakes Medical Exam) | [Yaneva et al., 2019] |
| Predicting the Difficulty and Response Time of Multiple-Choice Questions Using Transfer Learning (Assessment) | Predict the difficulty and response time (High-stakes Medical Exam) | [Xue et al., 2020] |
| Learning latent parameters without human response patterns: Item response theory with artificial crowds (Assessment) | Estimate difficulty with IRT models using response pattern (RP)s generated from artificial crowds of DNN models (Not specific) | [Lalor and Wu and Yu, 2019] |
| Item response theory in AI: Analysing machine learning classifiers at the instance level (Assessment) | Estimate IRT parameters such as discrimination, difficulty and guessing using different algorithms (Not specific) | [Martinez-Plumed et al. 2019] |
| Analysis of a simple model of problem-solving times (Assessment) | Estimate problem solving times (University and high school students) | [Jarušek and Pelánek 2012] |
| Personalized e-learning system using Item Response Theory (Assessment) | Provide individual learning paths for learners (e- learning system) | [Chen et al., 2005] |
| The solution of the "cold start problem" in e-Learning (Assessment) | Detect the starting difficulty of the content (LMS) | [Abbakumov, 2014] |
| Item difficulty estimation: An auspicious collaboration between data and judgment (Assessment) | Estimate the difficulty level of items, the solution of the "cold start problem" (University) | [Wauters et al., 2012] |

| | | |
|---|--|-------------------------------|
| Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability (Assessment) | Estimate skills and grade from raters' grades on written essays (Not specific) | [Uto, 2019] |
| A model to characterize exercises using probabilistic methods (Assessment) | Estimate exercise parameters such as: grade, time spent and number of attempts using machine learning algorithms (Simulations) | [Jiménez-Macias et. al.,2021] |
| The accurate measurement of students' learning in e-learning environment (Video) | Analyze video sentiments using voices and texts for a course in a MOOC environment (MOOC) | [Yang et al., 2021] |
| Semantic Topic Chains for Modeling Temporality of Themes in Online Student Discussion Forums (Discussion forum) | Analyze messages posted by university students in online discussion forums to capture themes and temporal progression (University) | [Chopra et al., 2023] |
| Improving Automated Scoring of Student Open Responses in Mathematics (Assessment) | Predict grades from student responses to open-ended questions in mathematics using a deep learning model (E-learning platform) | [Baral et al., 2021] |
| Predicting Item Response Theory Parameters Using Question Statements Texts (Assessment) | Predict the difficulty parameter of multiple-choice questions (High School Education) | [Marinho et al., 2023] |
| What You Do Predicts How You Do: Prospectively Modeling Student Quiz Performance Using Activity Features in an Online Learning Environment (Assessment) | Explore the possibility of prospectively predicting students' success in an assessment (E-learning platform) | [Jensen et al., 2021] |

| | | |
|---|---|-------------------------------|
| An evaluation of assessment stability in a massive open online course using item response theory (Assessment) | Evaluate the stability of item difficulty and discrimination parameters in different courses (MOOC) | [Gershon et al., 2023] |
| Recreation of different educational exercise scenarios for exercise modeling (Assessment) | Propose different scenarios of an educational exercise using simulated students by analyzing their behavior (Simulations) | [Jiménez-Macias et al., 2023] |
| Use and Interpretation of Item Response Theory Applied to Machine Learning (Assessment) | Examine the effect of increasing the number of training examples on IRT rates and item difficulty (Not specific) | [Dias et al., 2021] |

Table 8: Summary of accepted studies