Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Evaluating the Performance of Copula-Based Item Response Theory Models for Interpretable Assessment

Eduardo GUZMÁN & Eva MILLÁN*
Departamento de Lenguajes y Ciencias de la Computación,
E.T.S. Ingeniería Informática, Universidad de Málaga, Spain
*emillan@uma.es

Abstract: This paper describes a study evaluating the performance of copula-based Item Response Theory in real-world settings. To achieve this, we used a dataset containing information about 152 students who took a test on first-degree equations. This dataset had previously been employed to assess the performance of a Bayesian Network model in diagnosing 12 concepts related to first-degree equations. Both copula-based Item Response Theory and Bayesian Networks are explainable techniques that can be utilized for educational assessment. In this study, we compare the results of both data-driven methods against the actual state of knowledge of the students, which is a hidden variable, estimated using an expert-driven approach that involved averaging three independent assessments made by experienced primary school teachers. The results show that both methods can be used to obtain reliable estimations of students' knowledge.

Keywords: Student Modeling, Explainable Assessment

1. Introduction

The design and development of intelligent learning environments is founded on the fundamental hypothesis that their behavior should adapt to the learner. The instruction is adapted considering the state of the student (as represented in the student model) and making appropriate decisions regarding the contents (navigation path across the curriculum, contents, forms of presentation) and the assignments proposed to students (tests, exercises, tasks). The state of the student refers to different dimensions related to his/her knowledge about the subject. Frequently, this knowledge model is enriched with other attributes and attitudes, e.g., *conative* (desires, intentions, learning styles, etc.) or *affective* (values, motivation, and emotions) (Nakic, et al., 2015).

Knowledge modeling remains an active research area. A previous study by Pelánek (2017) reported that the Bayesian Knowledge Tracing model and its variations are among the most used approaches and that the second major approach involves logistic models, such as the Item Response Theory models, and the choice between these models remains unresolved. The next-generation assessments must provide fine-grained feedback for students and teachers. Diagnostic classification (or cognitive diagnostic) models have arisen as advanced psychometric models and identify the attributes that define a learner's mastery at the time of assessment (Ma, et al., 2023).

Recently, neural network-based approaches have been increasingly utilized for student modeling, yielding positive results. For example, Chaplot et al. (2018) report that neural networks can learn accurate cognitive models in ill-structured domains with no data and little to no human knowledge engineering. Another study (Swamy et al., 2018) proposes the use of Deep Knowledge Tracing and trains a separate network for each skill. (Jiang et al., 2018) compare the performance of expert feature-engineering vs. deep neural networks and finds that both models reached similar levels of accuracy. However, all these studies acknowledge that traditional expert engineering approaches to student modeling have their own strengths: the resulting models are more explainable and interpretable from a psychological and educational perspective because they can provide meaningful information. The need for explainable AI in educational contexts even greater than in other fields due to issues such as fairness in assessment processes and support for learners' metacognitive and reflective processes (Khosravi et al., 2022).

To address the challenges in this area, several solutions have been proposed. For example, Yeung (Yeung, 2019) proposes Deep-IRT (a synthesis of the classical IRT model and a knowledge tracing model based on a deep neural network architecture) to enhance the explainability of deep-learning-based knowledge tracing. The experiment shows that this combined model demonstrates strong performance and provides a direct psychological interpretation of students and items, by estimating the likelihood of a student answering a question correctly based on their abilities and the difficulty of the item.

The use of eXplainable AI (XAI) in education represents an emerging approach to enhancing trust AI systems as it provides transparent explanations and justifications for decision-making (Khosravi, et al., 2022). To this end, this research focuses on traditional and explainable expert-based methods for assessment.

Classical IRT models can only measure one latent trait and thus are not suitable for hierarchical domains where several concepts or latent traits need to be measured simultaneously or where assessments involve items related simultaneously to more than one latent trait, i.e., there is dependency among items and latent traits. For this reason, our approach is based on the use of *copula* functions. In probability theory, a copula is a multivariate cumulative distribution function that can be used to describe the dependencies between random variables. We will utilize copula-based IRT models (Braeken, 2011), a new generation of IRT-based models that consider the dependence among items. In educational assessment environments, the set of items that exhibit local dependence is known as testlets (Waner and Kiely, 1987). A suitable knowledge structure, considering issues such as dependence and relations among concepts and between concepts and items, strongly influences the model’s fitness and the bias of the estimations (Kadhem & Nikoloulopoulos, 2023b). Traditionally, dependency among items was approached through testlet-based IRT with models such as the Rasch testlet model (Wang & Wilson, 2005). However, in recent years, the greater flexibility, applicability, and modeling power of copulas has led to the increased development and use of copula-based IRT models (e.g., Kadhem & Nikoloulopoulos, 2023a; 2023b) over testlet-based models in the field of IRT. Copulas are prevalent in fields such as statistics, actuarial science, finance, reliability hydrology, etc. This success can be explained by the copula’s ability to summarize the full dependence structure between random variables (Rémillard & Scaillet, 2009). Copulas allow a wide variety of dependencies between items to be modeled, including complex and non-linear dependencies. They can capture dependency structures that are not limited to the grouping of items in testlets and can handle both local and global dependencies (Berghaus & Bücher, 2017).

The primary objective of this study is to evaluate the performance of such models in real settings. For this purpose, we will compare the performance of copula-based IRT models with the performance of a widely used model for explainable assessment: Bayesian networks (BNs). This objective has been specified as two research questions:

RQ_1. Are copula-based IRT models capable to provide accurate estimations of students’ knowledge level from their answers to an exam?

RQ_2. How do copula-based IRT models compare to BN student models (regarding accuracy and development costs)?

Our research builds upon a previous study presenting the development and evaluation of a student model based on BNs for first-degree equations (Millán et al., 2013). The BN student model was constructed by experts and contains 12 concepts at 3 different levels of granularity (knowledge nodes in the BN). The accuracy of the BN assessment method was evaluated with the data obtained in an experiment involving 152 ninth-grade students (14–15 years old) from six different groups in two private schools in Figueira da Foz district (Portugal). To this end, the students took a written exam. Each exam was graded independently by the 3 different teachers who collaborated on the study. Given the high inter-rater agreement, the average of these grades was used as a reliable estimation of the hidden variable (the student’s knowledge level).

In the study presented here, several elements from our previous research were utilized, including the rater average as the gold standard for evaluating the performance of the copula-based IRT model, the nodes and relationships of the BN student model, and the dataset comprising 152 students. Unlike BN models, IRT models do not require conditional probabilities, thereby reducing the complexity and burden associated with providing these estimations.

To construct an IRT-based student model for comparison, we addressed several challenges. Primarily, IRT is traditionally applied to test-based assessments rather than exams. However, previous studies by Gálvez et al. (2009, 2016) and Hernando et al. (2013) have demonstrated the feasibility of using IRT in problem-solving environments.

A second challenge concerns the independence assumption inherent in IRT, which requires that exam questions or items be independent of one another. Specifically, information provided by one item should not be used to solve another. To address this, we employed copula-based IRT models, which can handle dependencies. In our approach, items within an exercise are grouped under a copula, and knowledge estimation is computed accordingly.

The rest of this paper is structured as follows: Section 2 provides a brief introduction to the fundamentals of IRT and copula-based IRT models. Section 3 delineates the methodology used to construct the IRT-based student model, addressing the aforementioned challenges. This section also elucidates the procedure for comparing models. Section 4 describes the experimental procedures, techniques employed, and results obtained. Section 5 provides an interpretation of such results, while section 6 summarizes the key findings, discusses their implications, and proposes future lines of research.

2. Item Response Theory

2.1 Classical IRT

Item Response Theory (Hambleton et al., 1991) is a well-known psychometric framework used for test-based assessments. It provides a sophisticated statistical model to perform reliable assessments through test items. IRT is frequently employed in large-scale educational tests such as the Graduate Record Examination (GRE) and the Scholastic Aptitude Test (SAT). One of the primary strengths of IRT is the invariance of its results, meaning that the knowledge values obtained are independent of the measurement tool used.

In IRT, diagnosis is derived from the evidence provided by students through their performance on a set of items. The theory is based on two principles: (a) a student’s performance on a test can be explained by a single trait, typically the knowledge level, which is measured as an unknown numerical value; and (b) the performance of a student with an estimated trait level on an item can be probabilistically predicted and modeled by a function called the Item Characteristic Curve (ICC). The ICC expresses the probability that a student with a certain trait level (θ) will answer the item correctly. The higher the student's trait level, the greater the probability of a correct response. Each item has its own ICC, and several functions can be used to construct it. One of the most widely used functions is the logistic model, which defines the probability of a correct answer to a test item (i) as $P(X_i = 1 \mid \theta)$, as given by the function in Equation 1.

$$P(X_i = 1 \mid \theta) = c_i + (1 - c_i)\,\frac{1}{1 + e^{(-1.7a_i(\theta - b_i))}} \tag{1}$$

The shape of the Item Characteristic Curve (ICC) is determined by three key parameters:

- Discrimination factor (a_i): This parameter is proportional to the slope of the curve. High values indicate a steep slope, suggesting that students with trait levels above the item difficulty have a high probability of success. This parameter effectively differentiates between students of varying ability levels.
- Difficulty index (b_i): This parameter corresponds to the trait level at which the probability of a correct answer equals that of an incorrect answer. The range of values for this

parameter is consistent with the range of trait levels. It represents the point on the trait scale where the item functions optimally in discriminating between students.

- Guessing factor (*c_i*): This parameter represents the probability that a student with minimal knowledge answers the item correctly through random selection. It accounts for the possibility of correct responses due to chance rather than ability, particularly in multiple-choice items.

Three different models can be derived from the previous formula, in terms of how the three parameters above are used: (a) Three-parameter logistic (3PL) model, that incorporates all three parameters; (b) Two-parameter logistic (2PL) model, that sets the guessing parameter to zero, and (c) One-parameter logistic (1PL) or Rasch model, based on 2PL and setting the discrimination parameter to one. The ICCs The Item Characteristic Curves (ICCs) are inferred through a statistical process called calibration. This data-driven procedure utilizes the performance of students who previously took the test as input. The calibration yields both the ICCs and estimates of students' latent trait values. To estimate a student's trait, the following calculation is performed: the product of the ICCs for items answered correctly is multiplied by the product of the complement of ICCs for items answered incorrectly.

The estimation of ICCs during calibration is commonly accomplished using the Expectation-Maximization (EM) algorithm, which employs maximum likelihood estimation (MLE). The latent ability (*θ*) estimates are obtained by applying the parameters from this phase to the examinees' responses.

Classical IRT relies on two crucial assumptions: (1) Item independence: an item must not provide any hints or assistance to the student in correctly answering another item. Consequently, there should be no relations between items in the same test; and (2) constant student knowledge, i.e. no learning occurs during the test. Given these assumptions, the distribution of a student's knowledge in a particular concept is computed by multiplying and normalizing the characteristic curves of the responses to n items, as indicated in Equation 2.

$$P(\theta|X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i|\theta) \qquad (2)$$

Where for *each i =1, ..., n*, *X_i* is the answer given to item i, *$P(X_i|\theta)$* is the ICC associated with answer *X_i* (in the dichotomous case, there is a single curve related to the correct answer and the opposite associated with the incorrect one), and *$P(\theta|X_1, X_2,\dots, X_n)$* is the knowledge distribution given answers *$X_1, X_2,\dots, X_n$*.

2.2 Copula IRT models

Copula models represent a class of models within IRT designed to address local item dependencies by incorporating copulas to model the association structure between items. These models combine dependent items into a copula function, which is a multivariate cumulative distribution function with uniform *U(0, 1)* margins (Kadhem & Nikoloulopoulos, 2023a). The knowledge level, denoted as *θ*, can be estimated using Equation 3, which replaces the traditional method of computing knowledge level.

$$P(\theta|X_1, X_2, \dots, X_n) = \prod_{i=1}^{m} \varsigma_i \left[P(X_{i_1}|\theta), P(X_{i_2}|\theta), \dots, P\left(X_{i_{|\varsigma_i|}}\middle|\theta\right);\ \theta\right] \qquad (3)$$

In copula models, items exhibiting dependence among them are grouped in a copula (or testlet) described by a copula function ς_i. This function is a multivariate, cumulative, distribution function, that considers the common background for items within the same group by combining their ICCs. Equation 3 demonstrates that student knowledge distribution for a specific concept is now calculated by multiplying the set of all m copula functions of the dependent item groups identified to assess the corresponding concept. These copula functions integrate the set of dependent ICCs into a single function. Thus, each copula function ς_i combines information from the items belonging to its group $X_{i_1}, X_{i_2}, \dots, X_{i_{|\varsigma_i|}}$, where $|\varsigma_i|$ represents the set of items in that copula and $i_1, i_2, \dots, i_{|\varsigma_i|}$ are the indexes of the items within the copula. This approach addresses the problem of local dependency among items, making it particularly suitable for hierarchical domains where an item may assess multiple latent traits. It's important to note that when an item is independent of others, i.e., $|\varsigma_i| = 1$, its copula function is equivalent to its ICC, as shown in Equation 4:

$$\xi_i[P(X_i|\theta)] = P(X_i|\theta) \tag{4}$$

3. The student model

As previously explained, we employed the same student model for first-degree equations in the prior study. This student model comprises knowledge variables, defined at three levels of granularity, and evidential variables, which represent the questions on the written exam. Let us describe these components in more detail.

Knowledge variables represent knowledge about first-degree equations at several levels of granularity. At the elementary level, there are eight binary variables (with values *known/not known*): C_1 (terms and coefficients); C_2 (solution); C_3 (equation); C_4 (members); C_5 (classification); C_6 (add); C_7 (multiply); and C_8 (modeling). To model granularity, some of these concepts are then grouped to form compound concepts: at the first level of granularity, we have concept C_9 (terminology and concepts), which includes concepts C_1 to C_4, and C_{10} (equivalence), which includes concepts C_6 and C_7. At the second level, concept C_{11} (resolution) encompasses concepts C_5 and C_7. Finally, at the higher level of granularity, concept C_{12} represents the global knowledge about first-degree equations (i.e., C_{12} represents θ in the IRT model). The relative importance of each topic in the more general topic was modeled using weights, that were estimated by the participant teachers.

Evidential variables represent student's responses to questions in the written exam. The exam consists of 14 questions, some with multiple sections. Each evidential variable is labelled with the question number and the section number. For example, Q3_1 represents the student response to the 1st section in question 3. Some questions comprise several *true or false* items, with evidence for each item introduced separately. In total, there are 22 responses per student. A human expert evaluated the correctness of these 22 responses. Some items (e.g., true/false questions) were assessed as either correct or incorrect, while others (e.g., exercises) were evaluated with a percentage. In our model, responses with a percentage below 50% were classified as incorrect, and those at or above 50% as correct. Consequently, all evidential variables in our model are binary, with values of correct or incorrect. The model utilizes this evidence to estimate the student's knowledge level at each granularity level.

Relationships among variables are represented by causal links, both for the granularity levels and to model the concepts required for correctly answering each item on the written exam. Figure 1 provides a graphical description of the structural model:

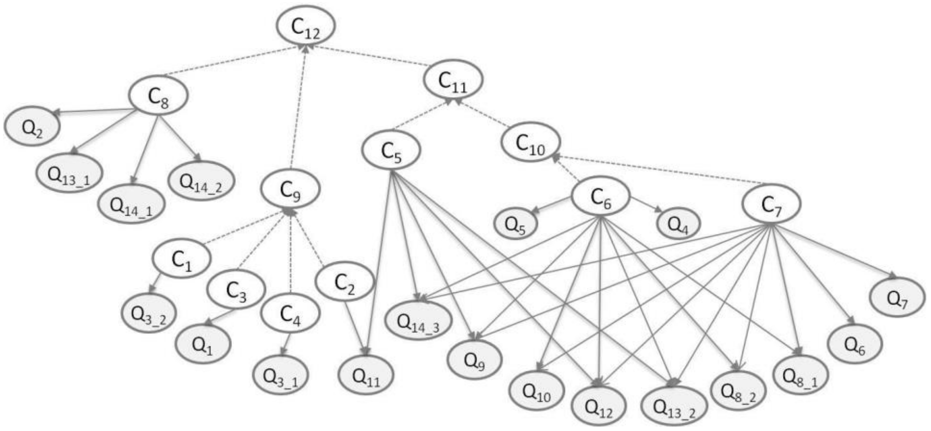


Figure 1. Structure of the student model for first-degree equations

Regarding the parameters, in the previous study (model based on BNs), all relationships were modeled with conditional probabilities, which were computed from parameters specified by experts (please refer to the original source for a detailed description). In this new study (copula-based IRT model), we have modeled the relationships among questions and knowledge nodes using Item Response Theory functions for simple concepts and computing weighted averages for compound concepts, using the same weights as in the BN model. This approach mitigates the need for expert estimations of the conditional probabilities for each question given their parent nodes, which would otherwise require a number of parameters that is exponential in the number of parents of each node.

For the construction of the IRT-based student model, we addressed some challenges. In many cases, exercise sections were related among them, and this fact violated the principle of local independence required by IRT. Local dependence, if not correctly handed, can negatively impact the estimation and reliability of ICCs parameters (Braeken, 2011). To overcome this problem, we used copula-based IRT models to deal with dependency, i.e., the hidden state of knowledge of the student was computed using Equation 5.

$$P(\theta_C | Q_{i_1}, Q_{i_2}, .., Q_{i_C}) = \prod_{i=1}^{m} \varsigma_i \left[P(Q_{i_1}|\theta), P(Q_{i_2}|\theta), \ldots, P\left(Q_{i_{|\varsigma_i|}} \middle| \theta\right);\ \theta \right] \tag{5}$$

where θ_C is the concept to be assessed; $Q_{i_1}, Q_{i_2}, .., Q_{i_C}$ the set of items evaluating that concept; *m* is the number of copulas; and $Q_{i_1}, Q_{i_2}, \ldots, Q_{i_{|\varsigma_i|}}$ the set of items involved in copula *i* whose function is ς_i. In this way, knowledge estimation for concept θ_C is calculated in terms of the copula functions that group the set of items evaluating θ_C. This solves the problem of local dependency among the parts of an exercise, which can now be evaluated properly.

4. Material and Methods

4.1 Experimental Settings and Preprocessing

The experiment utilizes a dataset comprising 152 ninth-grade students, aged 14–15 years, from six different groups across two private schools in the Figueira da Foz district, Portugal. The code for this study is available on GitHub (Guzmán, 2024).

The initial phase of the study involved assessing the dataset's quality, specifically its reliability. To achieve this, a separate dataset was constructed, focusing on students' performance in exercises evaluating each concept. Concepts 1 to 4 were excluded from this analysis due to insufficient evidence, as they were each associated with only one question (see Figure 1), which was inadequate for computing students' knowledge levels.

Following the exclusion of concepts 1-4 and 9, seven distinct datasets were created, each corresponding to an evaluated concept. The reliability of these datasets was assessed using Cronbach's alpha (Cronbach, 1951), a measure of internal consistency for a set of items. Cronbach's alpha ranges from 0 to 1, with values of 0.70 or higher considered acceptable. Table 1 presents the number of exam items associated with each concept and the corresponding Cronbach's alpha values.

Table 1. *Cronbach’s alpha, and the number of items per concept.*

Concepts	12	11	10	9	8	7	6	5
Cronbach’s alpha	0.805	0.718	0.680	0.425	0.720	0.670	0.658	0.723
# items	22	11	10	9	4	8	9	6

Results show that the dataset exhibits an acceptable degree of reliability for concepts 12, 11, 8, and 5. For concepts 10, 7 and 6 the reliability is close to 0.7, while it is low for concept 9 (0.425). In fact, concept 9 is just an aggregation of concepts 1 to 4, which were removed from the analysis due to having only one question associated with them. For this reason, concept 9 was removed from the analysis too.

4.2 Methods and results

As in the original study, the average of the grades assigned by the three different teachers was used as the gold standard for estimating students' knowledge levels. Both the Bayesian Network and copula-based Item Response Theory models were fed with the same evidence: the correctness of students' responses to each of the 22 exam items. Figure 2 presents a box and whiskers plot comparing the BN and copula-based IRT models (referred to simply as IRT in the graphs and tables) to the average of the experts' estimations.

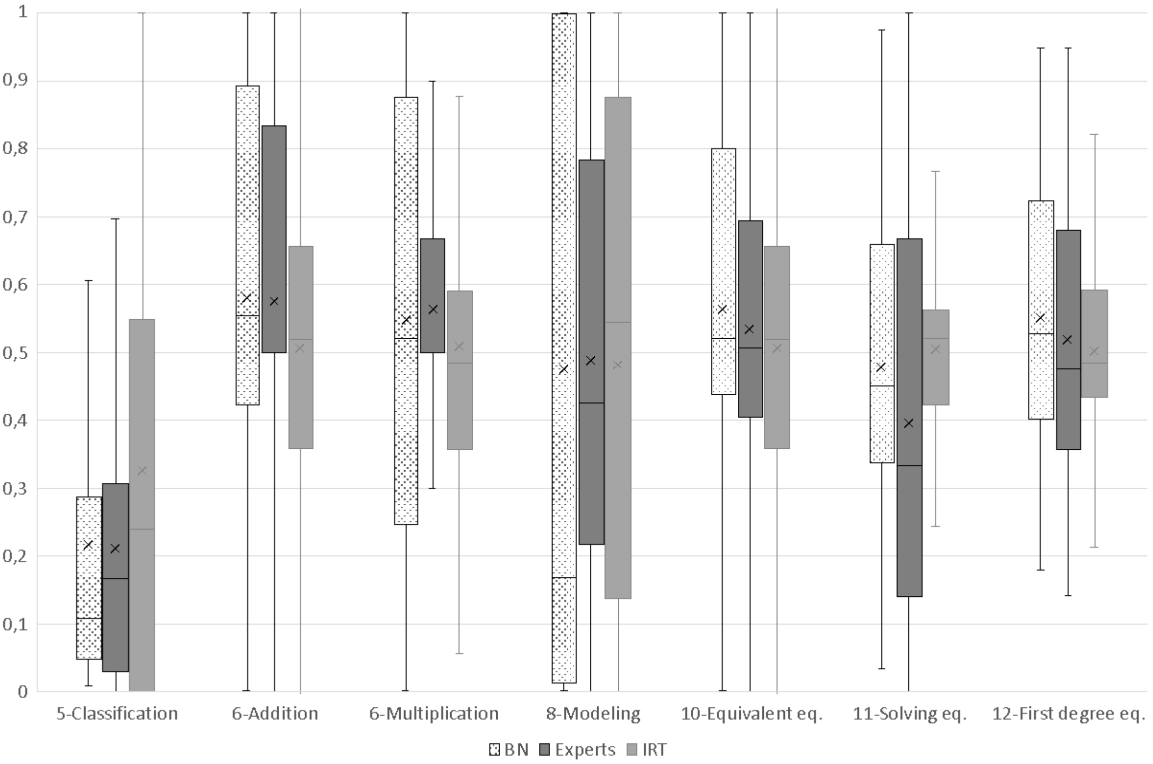


Figure 2. Box and whiskers plot for selected concepts.

Subsequently, the three techniques for constructing student models were compared in terms of concept knowledge levels. To achieve this, two methods were employed: correlation and linear regression analysis. Additionally, scatterplots were used to visually assess the degree of agreement among the methods. Tables 2 and 3 present the results of the correlation and linear regression analyses, respectively, when comparing students' performance as computed by the three different approaches.

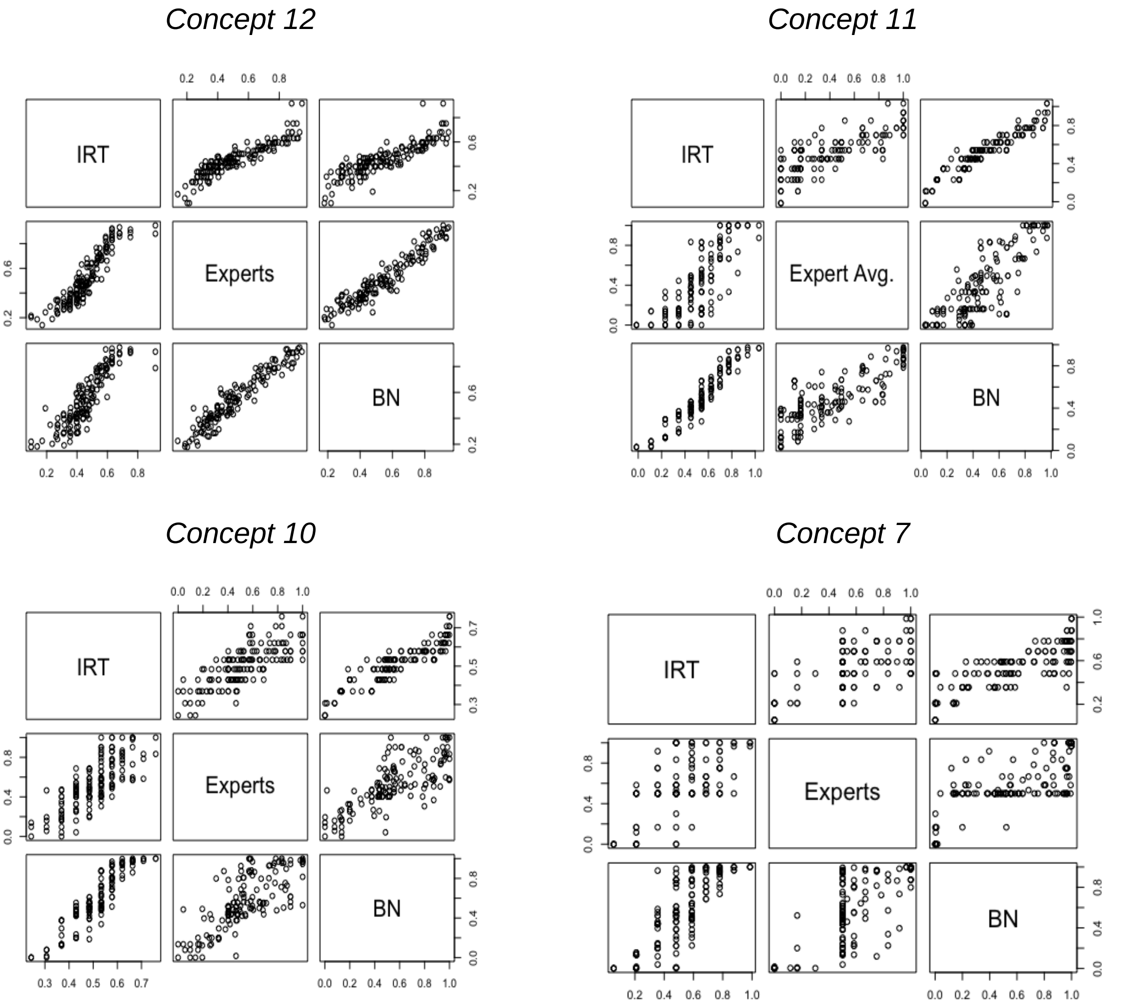
Table 2. *Pearson's correlation coefficient (p < .05).*

Concept	BN-Experts	IRT-Experts	IRT-BN
12	**0.9183**	0.9090	0.9090
11	0.8060	0.7983	**0.9469**
10	0.7560	0.7601	**0.9100**
8	**0.8408**	0.5409	0.4198
7	0.7299	0.6082	**0.8093**
6	**0.7799**	0.6445	0.7565
5	0.5742	0.6464	**0.8075**
Average	0.7722	0.7011	**0.7941**

Table 3. *Regression Analysis: adjusted R-squared and F statistical values (p<.01).*

Concept	BN-experts	IRT-experts	BN-IRT
12	R^2=.91, F=1521	R^2=.82, F=713.7	R^2=.75, F=461.5
11	R^2=.65, F=278.2	R^2=.65, F=263.6	R^2=.90, F=1302
10	R^2=.57, F=200.1	R^2=.56, F=204.8	R^2=.83, F=713.5
8	R^2=.81, F=646.5	R^2=.81, F=646.5	R^2=.89, F=1225
7	R^2=.80, F=618.3	R^2=.37, F=88.05	R^2=.65, F=284.7
6	R^2=.61, F=232.8	R^2=.41, F=106.6	R^2=.57, F=200.7
5	R^2=.33, F=73.78	R^2=.41, F=107.7	R^2=.65, F=281

Figure 3 shows a selection of four scatterplot diagrams corresponding to concepts C_{12}, C_{11}, C_{10}, and C_7. The selection was made to account for different degrees of agreement.



Concept 12

Concept 11

Concept 10

Concept 7

Figure 3. Scatterplot for concepts 12, 11, 10, and 7

Next, we will analyze these results.

5. Discussion

The box and whisker plot in Figure 2 illustrates that the three estimations of student knowledge (BN, copula-based IRT, and expert evaluations) yield similar average values for each concept's knowledge level. However, the ranges of variation differ significantly. For some concepts, the BN model provides the widest range (e.g., concept 8), while for others, the expert average (concept 11) or the IRT model (concept 5) shows the greatest variation. This variability likely stems from the number of samples per class. The differences in averages do not conclusively indicate whether any model consistently over- or underestimates compared to the others.

For the regression analyses, adjusted R^2 indicators were calculated to account for the number of terms in each model. As shown in Table 3 and Figure 2, IRT models perform better in comparison to BN as the number of items and Cronbach's alpha value increase. The highest average correlation is between IRT and BN, while the lowest is between IRT and expert estimations. The same pattern is observed with the average adjusted R^2. However, neither the number of items nor Cronbach's alpha appear to influence the agreement between expert and BN estimations.

These findings suggest that both the quantity and quality of students' knowledge evidence significantly impact copula-based IRT estimations. In optimal scenarios, there is typically a high level of agreement between these estimations and those produced by the BN.

Regarding the relationship with expert estimations, the BN model demonstrated the highest agreement rate. The average correlation for the BN model is 0.77, compared to 0.70 for the IRT model. R^2 indicators corroborate this: the average adjusted R^2 values are 0.67 (expert-BN), 0.58 (expert-IRT), and 0.75 (IRT-BN).

Addressing our research questions, we conclude that the strategy of dividing exercises into distinct knowledge evidence units for student assessment using an IRT-based model is viable. We have addressed the item dependency limitation using copula models. The experimental results suggest that IRT copula-based IRT models can provide accurate, data-driven estimations of students' knowledge levels based on exam performance (RQ1). Both BNs and copula-based IRT models show similar performance in diagnosing students' knowledge. However, the BN approach requires greater engineering effort in structure construction and higher computational costs compared to the copula-based IRT model (RQ2).

6. Conclusions

Bayesian Networks are a systematic method for modeling knowledge with relationships, whereas copula-based IRT models are logistic models. This paper addresses the unresolved issue of selecting the appropriate model (Pelánek, 2017). Both models can accurately estimate the student model for concepts with a reasonable number of associated questions. However, our study shows that the BN model performs better. Both models are interpretable: the BN graph elucidates the interactions among variables and IRT parameters offer insights into item characteristics and respondent abilities, despite lacking an explicit graphical representation.

An experiment involving real students (n=152) indicates that copula-based IRT models can measure performance in complex tasks, such as exercises or problems, by decomposing them into knowledge evidence provider units, treated as items. These models address the issue of dependency among items in testing, which limits the validity of conventional IRT models when an item's answer directly depends on responses to other items. The copula model is used for assessing complex tasks by breaking them into smaller units that provide evidence of students' knowledge. Interdependency among these units necessitates assessment models capable of handling this problem.

Regarding the expert-engineering effort required for student knowledge assessment using copula-based IRT models, it is relatively minimal compared to other techniques such as BNs. BNs typically require significant effort in network structure elicitation and parameter estimation strategy selection. Conversely, IRT is a data-driven model that can automatically compute individual knowledge estimations using students' performance data logs, resulting in substantially lower engineering effort and computational costs. Additionally, IRT-based models have been successfully used in large-scale testing assessments.

Statistical analyses from this study suggest that BN performs slightly better than human expert estimations, particularly when the number of items is small. However, the agreement between BN and IRT results is high and even exceeds the agreement between BN and expert-based estimations. After all, expert-based estimations are merely another method to infer the hidden variable (the true state of the student's knowledge). Further studies with larger sample sizes are needed to confirm these findings. Future work will extend these results to other task types and domains and evaluate them in larger real-world settings.

Acknowledgements

This research has been supported by the Spanish Ministry of Science and Innovation through the research project with reference TED2021-129956B-I00.

References

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. Psychometrika, 76, 57–76.

Berghaus, B., & Bücher, A. (2017). Goodness-of-fit tests for multivariate copula-based time series models. Econometric Theory, 33(2), pp. 292-330.

Chaplot, D., MacLellan, C., Salakhutdinov, R., & Koedinger, K. (2018). Learning Cognitive Models Using Neural Networks. Proceedings of AIED 2018 (pp. 43–56). LNAI 10947.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334.

Gálvez, J., Guzmán, E., Conejo, R., & Millán, E. (2009). Student Knowledge Diagnosis Using Item Response Theory and Constraint-Based Modeling. Proceedings of the 2009 conference on Artificial Intelligence in Education (pp. 291-298). IOS Press.

Gálvez, J., Guzmán, E., Conejo, R., Mitrovic, A., & Mathews, M. (2016). Data calibration for statistical-based assessment in constraint-based tutors. Knowledge-Based Systems, 97, 11-23.

Guzmán, E. (2024). Simplified source code of item calibration and assessment value computation. Available at https://github.com/eduardoguzman/icce2024/

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Press.

Hernando, M., Guzmán, E., & Conejo, R. (2013). Measuring Procedural Knowledge in Problem-solving Environments with Item Response Theory. Proceedings of the 2013 conference on Artificial Intelligence in Education (AIED 2013) (pp. 653-656). Berlin Heidelberg: Springer.

Jiang, Y., Bosch, N., Baker, R., Paquette, L., Ocumpaugh, J., Andres, J., Biswas, G. (2018). Expert Feature-Engineering vs. Deep Neural Networks: Which Is Better for Sensor-Free Affect Detection? Proceedings of AIED 2018 (pp. 198–211). LNAI 10947.

Kadhem, S.H., Nikoloulopoulos, A.K. Factor Tree Copula Models for Item Response Data. Psychometrika 88, 776–802 (2023).

Kadhem, S.H., Nikoloulopoulos, A.K. Bi-factor and Second-Order Copula Models for Item Response Data. Psychometrika 88, 132–157 (2023.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J. & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074.

Ma, W., Wang, C., & Xiao, J. (2023). A Testlet Diagnostic Classification Model with Attribute Hierarchies. Applied Psychological Measurement, 47(3), 183-199.

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

Nakic, J., Granic, A., & Glavinic, V. (2015). Anatomy of student models in adaptive learning systems: a systematic literature review of individual differences from 2001 to 2013. Journal of Educational Computing Research, 51(4), 459-489.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. User Modeling and User-Adapted Interaction, 313-350.

Rémillard, B., & Scaillet, O. (2009). Testing for equality between two copulas, Journal of Multivariate Analysis, 100(3), pp. 377-386.

Swamy, V., Guo, A., Lau, S., Wu, W., Wu, M., Pardos, Z., & Culler, D. (2018). Deep Knowledge Tracing for Free-Form Student Code Progression. Proceedings of AIED 2018 (pp. 348–352). LNAI 10948.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement*, *24*(3), 185-201.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. Applied Psychological Measurement, 29(2), 126-149.

Yeung, C.K. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. Proceedings of the 12th international conference on educational data mining (2019), pp. 683-686