education
sciences

MDPI

Article

Educational Stakeholders' Independent Evaluation of an Artificial Intelligence-Enabled Adaptive Learning System Using Bayesian Network Predictive Simulations

Meng-Leong HOW * and Wei Loong David HUNG

National Institute of Education, Nanyang Technological University Singapore, Singapore 639798, Singapore; david.hung@nie.edu.sg

* Correspondence: mengleong.how@nie.edu.sg

Received: 6 March 2019; Accepted: 15 May 2019; Published: 20 May 2019

Abstract: Artificial intelligence-enabled adaptive learning systems (AI-ALS) are increasingly being deployed in education to enhance the learning needs of students. However, educational stakeholders are required by policy-makers to conduct an independent evaluation of the AI-ALS using a small sample size in a pilot study, before that AI-ALS can be approved for large-scale deployment. Beyond simply believing in the information provided by the AI-ALS supplier, there arises a need for educational stakeholders to independently understand the motif of the pedagogical characteristics that underlie the AI-ALS. Laudable efforts were made by researchers to engender frameworks for the evaluation of AI-ALS. Nevertheless, those highly technical techniques often require advanced mathematical knowledge or computer programming skills. There remains a dearth in the extant literature for a more intuitive way for educational stakeholders—rather than computer scientists—to carry out the independent evaluation of an AI-ALS to understand how it could provide opportunities to educe the problem-solving abilities of the students so that they can successfully learn the subject matter. This paper proffers an approach for educational stakeholders to employ Bayesian networks to simulate predictive hypothetical scenarios with controllable parameters to better inform them about the suitability of the AI-ALS for the students.

Keywords: evaluation of artificial intelligence educational systems; intelligent adaptive learning; intelligent tutoring systems; Bayesian; nonparametric data

1. Introduction

Adaptive learning systems [1,2] were studied by researchers since the 1950s, drawing on ideas from artificial intelligence (AI) and education. AI is increasingly being deployed in education to enhance teaching practices and address the learning needs of students [3]. In the field of artificial intelligence in education (AIED), significant endeavors were made toward the goal of creating systems that approach the quality of human one-on-one tutoring [4]. Many researchers and developers of software systems that provided interactive learning environments for students published papers to report on the improvements in learning gains, and the efficiencies of learning similar amounts of subject content in reduced amounts of time [5]. Typically, this involves students working individually with a computer to learn new concepts by solving problems that are focused on domain-level knowledge, such as mathematics or science [6]. An AI-enabled adaptive learning system (AI-ALS) might utilize, for example, Bayesian knowledge tracing (BKT) [7], or some other secret proprietary algorithm to make "adjustments in an educational environment in order to accommodate individual differences" [8] to provide a personalized learning experience for each student. A typical sequence which depicts how

Educ. Sci. **2019**, *9*, 110; doi:10.3390/educsci9020110 www.mdpi.com/journal/education

the AI-ALS interacts with the student would be as follows: (1) allow the student to choose a topic or sub-topic to learn; (2) allow the student to view text-based or video-based demonstrations which illustrate the concepts; (3) initiate a short test for each sub-topic within the AI-ALS for the student. If the student could consecutively correctly answer a few questions about the same topic or sub-topic, the AI-ALS would deem that the student passed (which would be indicated as "math_topic_passed" in the dataset) the learning objective for that topic or sub-topic. Otherwise, the student would be remediated by the AI-ALS until that goal is achieved. Finally, upon passing the test for a topic or sub-topic, the AI-ALS would unlock more topics or sub-topics which it deems the student is "ready to learn" (which would be indicated as "math_topic_ready_for_learning" in the dataset).

2. Research Problem

As imperative as it is for educational stakeholders to understand more about the pedagogical characteristics of the AI-ALS, the educational technology vendor would understandably be reticent about divulging the exact pedagogical motif. Even if they are willing to share the information, that pedagogical motif might be dynamically generated by the AI-ALS on the fly; only the developer of the AI-ALS would know the details.

Beyond simply reading and believing in the information provided by the vendors, there arises a need for educators and education policy-makers to independently understand more about the pedagogical characteristics underlying each AI-ALS. Laudable efforts were made by researchers to engender frameworks for the evaluation of adaptive learning systems (ALS) in the first decade of the 21st century [9]; however, those techniques were often highly technical. There remains a dearth in the extant literature for a more intuitive and practical way for educational stakeholders—rather than computer scientists—to carry out the independent evaluation of an AI-ALS to gauge its suitability for students in their own schools.

Potential Issues That Education Researchers Might Encounter

When educational stakeholders such as educational policy-makers or researchers invite a school to participate in a research study which involves an educational technology such as an AI-ALS, the leaders of the school might only allow a small-scale pilot study to be conducted. A possible scenario that researchers might encounter could be that the school might only be willing to provide a class of students as the treatment group. The reason for the school being unable to provide a control group could be that parents might not be willing to let their children participate in a research study group with "placebo" treatment, just to become a so-called baseline group for comparison with the treatment group, with supposedly does not benefit from any intervention. Even if a control group was provided by the school, it might not be easy to do fair comparisons between the treatment and control group, as the teacher of the control group might be very experienced and highly skilled in teaching, or vice versa. Some students from the treatment group or control group might be attending extra paid tuition classes outside of school. In short, there are myriad potential confounding factors that would be difficult to account for, in order to do fair comparisons between the students who utilized the AI-ALS to learn mathematics, and the control group which did not. Against this backdrop, one more problem faced by researchers who consider using null hypothesis significance testing (NHST) frequentist approaches is that it might not yield statistically significant results from a low number of participants, as the data distribution collected from the small number of students is often not parametric [10].

3. Methods

3.1. Rationale for Using the Bayesian Approach

Bayesian statistics [11] became more commonly used in social and behavioral science research [12] in recent years. Bayesian networks (BN) [13–15] are well suited for analyzing data with small sample sizes, because, unlike null hypothesis significance testing frequentist methods, it does not assume or require normal distributions as the underlying parameters of a model [10,16,17]. The Bayesian paradigm offers a more intuitive way to do hypothesis testing by enabling researchers to include prior background knowledge into the analyses. There is no need to perform multiple rounds of null hypothesis testing repeatedly [18–20].

The Bayesian approach was utilized by researchers in education, such as Kaplan [21], Levy [22], Mathys [23], and Muthén and Asparouhov [24]. In the field of educational technology, the Bayesian approach was used by Bekele and McPherson [25], and by Millán, Agosta, and Cruz [26], as it could be used to model and analyze information gain, as espoused in Shannon's information theory [27], which conceptually could be used to depict the notion of information gain (learning) by the students.

3.2. The Bayesian Theorem

The current paper attempts to provide a very brief introduction to the Bayesian theorem and BN. Admittedly, it is well beyond the scope of the current paper to cover the corpus of BN. Interested readers who wish to learn more about research in BN are strongly encouraged to consider perusing the works of Cowell, Dawid, Lauritzen, and Spiegelhalter [28], Jensen [29], and Korb and Nicholson [30]. The mathematical formula (see Equation (1)) on which BN was based was developed and first mentioned in 1763 by the mathematician and theologian, Reverend Thomas Bayes [11].

$$P(H|E) = \frac{P(E|H)\cdot P(H)}{P(E)} \tag{1}$$

In Equation (1), H represents a hypothesis, and E represents a piece of given evidence. $P(H|E)$ is known as conditional probability of the hypothesis H, that is, the likelihood of H occurring given the condition that the evidence E is true. This is also referred to as the posterior probability, that is, the probability of the hypothesis H being true after taking into consideration how the evidence E influences the occurrence of the hypothesis H.

$P(H)$ and $P(E)$ represent the probabilities of observing the likelihood of the hypothesis H occurring, and of the likelihood of the evidence E occurring, respectively, independent of each other. This is referred to as the prior or marginal probability—$P(H)$ and $P(E)$, respectively. $P(E|H)$ represents the conditional probability of the evidence E, that is, the likelihood of E occurring, given the condition that the hypothesis H is true. The quotient $P(E|H)/P(E)$ represents the support which the evidence E provides for the hypothesis H.

3.3. The Research Model

The aim of the research is to simulate how much (or how little) the learning of mathematics can be improved for students who used an AI-ALS. To this end, probabilistic reasoning techniques were used based on BN. Within the implementation of BN, the concept of the Markov blanket [31], in conjunction with response surface methodology (RSM) [32–35], is utilized, as they are suitable for examining the optimization of variables in the relationships between theoretical constructs, even if they are not physically related.

The Bayesian paradigm was selected as it is a sound methodology for modeling students' performance and knowledge, and it was used in research even before the 21st century by the developers of adaptive learning applications such as Collins, Greer, and Huang [36], Conati, Gertner, VanLehn, and Druzdze [37], Jameson [38], and VanLehn, Niu, Siler, and Gertner [39]. As these researchers were also the developers of their respective adaptive learning systems, they most probably had privileged access to knowledge into the inner workings of the algorithms. In contrast, the approach outlined in the current paper enables educational stakeholders to perform independent descriptive analytics, as well as predictive simulations, using the performance data downloaded from the back-end reports of an AI-ALS. This paper presents a detailed BN model of the students' knowledge that can inform educational stakeholders about the specific mathematics topics the students are ready to learn, and the topics they already passed. Teachers may use this vital information depicted by the relationships between the nodes/variables in the BN to provide remediation for the students who are struggling in their studies, and to help students who are attaining average-level scores in the AI-ALS to improve their learning experiences so that they could hopefully be more engaged and achieve high-level scores (higher percentage in the mathematics topics passed).

The model in this study comprises a paper-based pre-test, learning using the AI-ALS, a survey, and finally, a paper-based post-test. The current paper identifies the relationships within the BN as the constructs in the paper-based pre-test, the mediator (which is the AI-ALS), the paper-based post-test, and the noncognitive constructs (e.g., motivation, engagement, interest, self-regulation, etc.) in the survey. When researchers and stakeholders evaluate an AI-ALS, an understanding of these relationships is essential to determine if the interventions would be beneficial to the students. Therefore, the current paper proposes a practical Bayesian approach to demonstrate how educational stakeholders—rather than computer scientists—could analyze data from a small number of students, in order to explore the pedagogical motif of the AI-ALS using the following two segments of analytics, which are presented subsequently in Sections 4 and 5.

> Section 4: *"What has already happened?" Descriptive Analytics*
>
> Purpose: To use descriptive analytics to discover the pedagogical motif in the collected data.
>
> For descriptive analytics, BN modeling (in Sections 4.5–4.7) utilizes the parameter estimation algorithm to automatically detect the data distribution of each column in the dataset. In Section 4.8, further descriptive statistical techniques are employed to understand more about the current baseline conditions of the students including quadrant analysis, curves analysis, and Pearson correlation analysis.
>
> Section 5: *"What If?" Predictive Analytics*
>
> Purpose: To use predictive analytics to perform in silico experiments with fully controllable parameters from the pre-test to the mediating intervention to the post-test to predict future outcomes. Beyond just simply measuring gains by subtracting the students' post-test scores from the pre-test scores, this paper proffers a probabilistic Bayesian approach which could simulate various scenarios to better inform educators and policy-makers about the pedagogical characteristics of the AI-ALS that is being evaluated.
>
> For predictive analytics, counterfactual simulations (in Section 5) will be employed to explore the pedagogical motif of the AI-ALS. In Section 6, the predictive performance of the BN model is evaluated using tools that include the gains curve, the lift curve, and the receiver operating characteristic curve, as well as by statistical bootstrapping of the data inside each column of the dataset (which is also the data distribution in each node of the BN model) 1000 times to generate a larger dataset to measure its precision, reliability, Gini index, lift index, calibration index, the binary log-loss, the correlation coefficient R, the coefficient of determination R^2, root-mean-square error (RSME), and normalized root-mean-square error (NRSME).

4. Descriptive Analytics: "What Has Already Happened?"

In this section, the procedures taken in descriptive analytics to make sense of "what has already happened?" in the collected dataset are presented. To deliberately illustrate the capabilities of BN in handling nonparametric data from a small number of participants, the dataset comprising 16 students (all of whom were about 13–14 years old) who used the AI-ALS was imported into Bayesialab [40]. The purpose is to discover, via machine learning, the informational "pedagogical motif" (coined by the first author) of the learning intervention generated by the AI-ALS. In the context of this study, the notion of "pedagogical motif" is conceptually defined as the pattern, motif, and characteristics with which the AI-ALS pedagogically interacts with the students.

4.1. The Dataset Procured from the Reports Generated by AI-ALS

The supplementary files of the data, "AI-ALS-synthetic-data.csv", as well as the codebook describing the data, "ai-als-data_codebook.txt", can be downloaded from https://figshare.com/articles/datacodebook_zip/7795694.

4.2. Codebook of the Dataset

The dataset was originally procured from the server of the AI-ALS. Each column in the dataset (see Table 1) was utilized as a node in the BN. It was assumed that higher values in the data of both "math_topic_passed" (appended with the letter "P") and "math_topic_ready_for_learning" (appended with the letters "RL") could be considered to be indicators of better performance, and vice versa.

Table 1. Code book of the data downloaded from the artificial intelligence-enabled adaptive learning system (AI-ALS) server, each of which becomes a Bayesian network (BN) node.

Node Name	Description
student_id	Student identifier (ID)
hours	Number of hours spent by student using the AI-ALS
topics_350	Number of topics out of a total of 350 completed by the student in the AI-ALS
topics_percent	Percentage of topics out of a total of 350 completed by the student in the AI-ALS
Arithmetic readiness (AR)	
AR_FMEF_P	AR_Factors_Multiples_Equivalent_Fractions_Passed
AR_FMEF_RL	AR_Factors_Multiples_Equivalent_Fractions_Ready_For_Learning
AR_ASF_P	AR_Addition_Subtraction_with_Fractions_Passed
AR_ASF_RL	AR_Addition_Subtraction_with_Fractions_Ready_for_Learning
AR_MD_P	AR_Multiplication_Division_with_Decimals_Passed
AR_MD_RL	AR_Multiplication_Division_with_Decimals_Ready_for_Learning
AR_MN_P	AR_Mixed_Numbers_Passed
AR_MN_RL	AR_Mixed_Numbers_Ready_for_Learning
AR_RONL_P	AR_Rounding_Number Line_Passed
AR_RONL_RL	AR_Rounding_Number Line_Ready_for_Learning
AR_ASD_P	AR_Addition_Subtraction_with_Decimals_Passed
AR_ASD_RL	AR_Addition_Subtraction_with_Decimals_Ready_for_Learning
AR_MDD_P	AR_Multiplication_Division_with_Decimals_Passed
AR_MDD_RL	AR_Multiplication_Division_with_Decimals_Ready_for_Learning
AR_CBFD_P	AR_Converting_Between_Fractions_Decimals_Passed
AR_CBFD_RL	AR_Converting_Between_Fractions_Decimals_Ready_for_Learning
AR_RUR_P	AR_Ratios_Unit_Rates_Passed
AR_RUR_RL	AR_Ratios_Unit_Rates_Ready_for_Learning
AR_PDF_P	AR_Percents_Decimals_Fractions_Passed
AR_PDF_RL	AR_Percents_Decimals_Fractions_Ready_for_Learning
AR_IPA_P	AR_Intro_Percent_Applications_Passed
AR_IPA_RL	AR_Intro_Percent_Applications_Ready_for_Learning
AR_UM_P	AR_Units_Measurement_Passed
AR_UM_RL	AR_Units_Measurement_Ready_for_Learning
Real numbers (RN)	
RN_PLOT_P	RN_Plotting_Ordering_Passed
RN_PLOT_RL	RN_Plotting_Ordering_Ready_for_Learning
RN_OSN_P	RN_Operations_Signed_Numbers_Passed
RN_OSN_RL	RN_Operations_Signed_Numbers_Ready_for_Learning
RN_EOO_P	RN_Exponents_Order_Operations_Passed
RN_EOO_RL	RN_Exponents_Order_Operations_Ready_for_Learning
RN_EE_P	RN_Evaluation_Expressions_Operations_Passed
RN_EE_RL	RN_Evaluation_Expressions_Ready_for_Learning
RN_VDSRN_P	RN_Venn_Diagrams_Sets_Real_Num_Passed
RN_VDSRN_RL	RN_Venn_Diagrams_Sets_Real_Num_Ready_for_Learning
RN_PROP_O_P	RN_Properties_Operations_Passed
RN_PROP_O_RL	RN_Properties_Operations_Ready_for_Learning
RN_OSLE_P	RN_One_Step_Linear_Equations_Passed
RN_OSLE_RL	RN_One_Step_Linear_Equations_Ready_for_Learning

Table 1. *Cont.*

Node Name	Description
Linear equations (LE)	
LE_MSLE_P	LE_Multi_Step_Linear_Equations_Passed
LE_MSLE_RL	LE_Multi_Step_Linear_Equations_Ready_for_Learning
LE_WEE_P	LE_Writing_Expressions_Equations_Passed
LE_WEE_RL	LE_Writing_Expressions_Equations_Ready_for_Learning
LE_ALE_P	LE_Applications_Linear_Equations_Passed
LE_ALE_RL	LE_Applications_Linear_Equations_Ready_for_Learning
LE_SVDA_P	LE_Solving_Variable_Dimensional_Analysis_Passed
LE_SVDA_RL	LE_Solving_Variable_Dimensional_Analysis_Ready_for_Learning
LE_PROP_P	LE_Proportions_Passed
LE_PROP_RL	LE_Proportions_Ready_for_Learning
LE_MP_P	LE_More_Percents_Passed
LE_MP_RL	LE_More_Percents_Ready_for_Learning
LE_PFL_P	LE_Personal_Financial_Literacy_Passed
LE_PFL_RL	LE_Personal_Financial_Literacy_Ready_for_Learning
Linear inequalities (LI)	
LI_WGI_P	LI_Writing_Graphing_Inequalities_Passed
LI_WGI_RL	LI_Writing_Graphing_Inequalities_Ready_for_Learning
Functions and lines (FL)	
FL_TGL_P	FL_Tables_Graphs_Lines_Passed
FL_TGL_RL	FL_Tables_Graphs_Lines_Ready_for_Learning
FL_IF_P	FL_Introduction_Functions_Passed
FL_IF_RL	FL_Introduction_Functions_Ready_for_Learning
FL_AS_P	FL_Arithmetic_Sequences_Passed
FL_AS_RL	FL_Arithmetic_Sequences_Ready_for_Learning
Exponents and exponential functions (EEF)	
EEF_PPQR_P	EEF_Product_Power_Quotient_Rules_Passed
EEF_PPQR_RL	EEF_Product_Power_Quotient_Rules_Ready_for_Learning
EEF_IR_P	EEF_Intro_Radicals_Passed
EEF_IR_RL	EEF_Intro_Radicals_Ready_for_Learning
Polynomials and factoring (PE)	
PE_PM_P	PE_Polynomial_Multiplication_Passed
PE_PM_RL	PE_Polynomial_Multiplication_Ready_for_Learning
PF_FGCF_P	PE_Factoring_Greatest_Common_Factor_Passed
PF_FGCF_RL	PE_Factoring_Greatest_Common_Factor_Ready_for_Learning
PF_FQT_P	PE_Factoring_Quadratic_Trinomials_Passed
PF_FQT_RL	PE_Factoring_Quadratic_Trinomials_Ready_for_Learning
PF_FSP_P	PE_Factoring_Special_Products_Passed
PF_FSP_RL	PE_Factoring_Special_Products_Ready_for_Learning
Quadratic functions and equations (QFE)	
QFE_SQEF_P	QFE_Solving_Quadratic_Equations_Factoring_Passed
QFE_SQEF_RL	QFE_Solving_Quadratic_Equations_Factoring_Ready_for_Learning
QFE_SRP_P	QFE_Square_Root_Property_Passed
QFE_SRP_RL	QFE_Square_Root_Property_Ready_for_Learning
Pre-test (PRETEST)	Synthetic data for pre-test questions 1–10
Post-test (POSTTEST)	Synthetic data for post-test questions 1–10
Noncognitive (NONCOG)	Synthetic data for noncognitive survey questions 1–10

4.3. *Software Used: Bayesialab*

The software used was Bayesialab version 8.0. The 30-day trial version can be downloaded from http://www.bayesialab.com.

As a strongly recommended pre-requisite activity, before proceeding with the exemplars shown in the rest of this paper, it would be greatly beneficial to the reader to become familiar with Bayesialab by downloading and reading the free-of-charge user guide from http://www.bayesia.com/book/, as it contains the descriptions of the myriad tools and functionalities within the Bayesialab software, which are too lengthy to include in the current paper.

4.4. *Pre-Processing: Checking for Missing Values or Errors in the Data*

Before using Bayesialab to construct the BN, the first step is to check the data for any anomalies or missing values. In the dataset used in this study, there were no anomalies or missing values.

However, should other researchers encounter missing values in their datasets, rather than discarding the row of data with a missing value, the researchers could use Bayesialab to predict and fill in those missing values. Bayesialab would be able to perform this by machine-learning the overall structural characteristics of that entire dataset being studied, before producing the predicted values.

4.5. Overview of the BN Model

Bayesian networks (BN), also referred to as belief networks, causal probabilistic networks, and probabilistic influence diagrams are graphical models which consist of nodes (variables) and arcs or arrows. Each node contains the data distribution of the respective variable. The arcs or arrows between the nodes represent the probabilities of correlations between the variables [41].

Using BN, it becomes possible to use descriptive analytics to analyze the relationships between the nodes (variables) and the manner in which initial probabilities, such as the number of hours spent in the AI-ALS and/or topics passed/ready to learn, and/or noncognitive factors, might influence the probabilities of future outcomes, such as the predicted learning performance of the students in the paper-based post-test.

Further, BN can also be used to perform counterfactual speculations about the initial states of the data distribution in the nodes (variables), given the final outcome. In the context of the current paper, exemplars are presented in the predictive analytics segment (in Section 5) to illustrate how counterfactual simulations can be implemented using BN. For example, if we wish to find out the conditions of the initial states in the nodes (variables) which would lead to high probability of attaining high-level scores in the post-test, or if we wish to find out how to prevent students from attaining low scores or failing in the paper-based post-test, we can simulate these hypothetical scenarios in the BN.

The relationship between each pair of connected nodes (variables) is determined by their respective conditional probability table (CPT), which represents the probabilities of correlations between the data distributions of the parent node and the child node [42]. In the current paper, the values in the NPT are automatically machine-learned by Bayesialab according to the data distribution of each column/variable/node in the dataset. Nevertheless, it is possible but optional for the user to manually enter the probability values into the CPT, if the human user wishes to override the machine-learning software. In Bayesialab, the CPT of any node can be seen by double-clicking on it.

The BN model displayed the distribution of the students' score clusters (see Figure 1) for the mathematics learning in the AI-ALS in terms of the following topics: arithmetic readiness, real numbers, linear equations, linear inequalities, functions and lines, exponents and exponential functions, polynomials and factoring, and quadratic functions and equations. These score clusters were generated via machine learning by the Bayesialab software. By generating this model from the data which contained varying levels of performance of the students (even if it was just 16 participants), we could obtain a "pedagogical motif" of the AI-ALS, which meant that we could then perform simulations in this computational model to study how it could behave under certain conditions. This is elaborated on and presented later in Section 5.

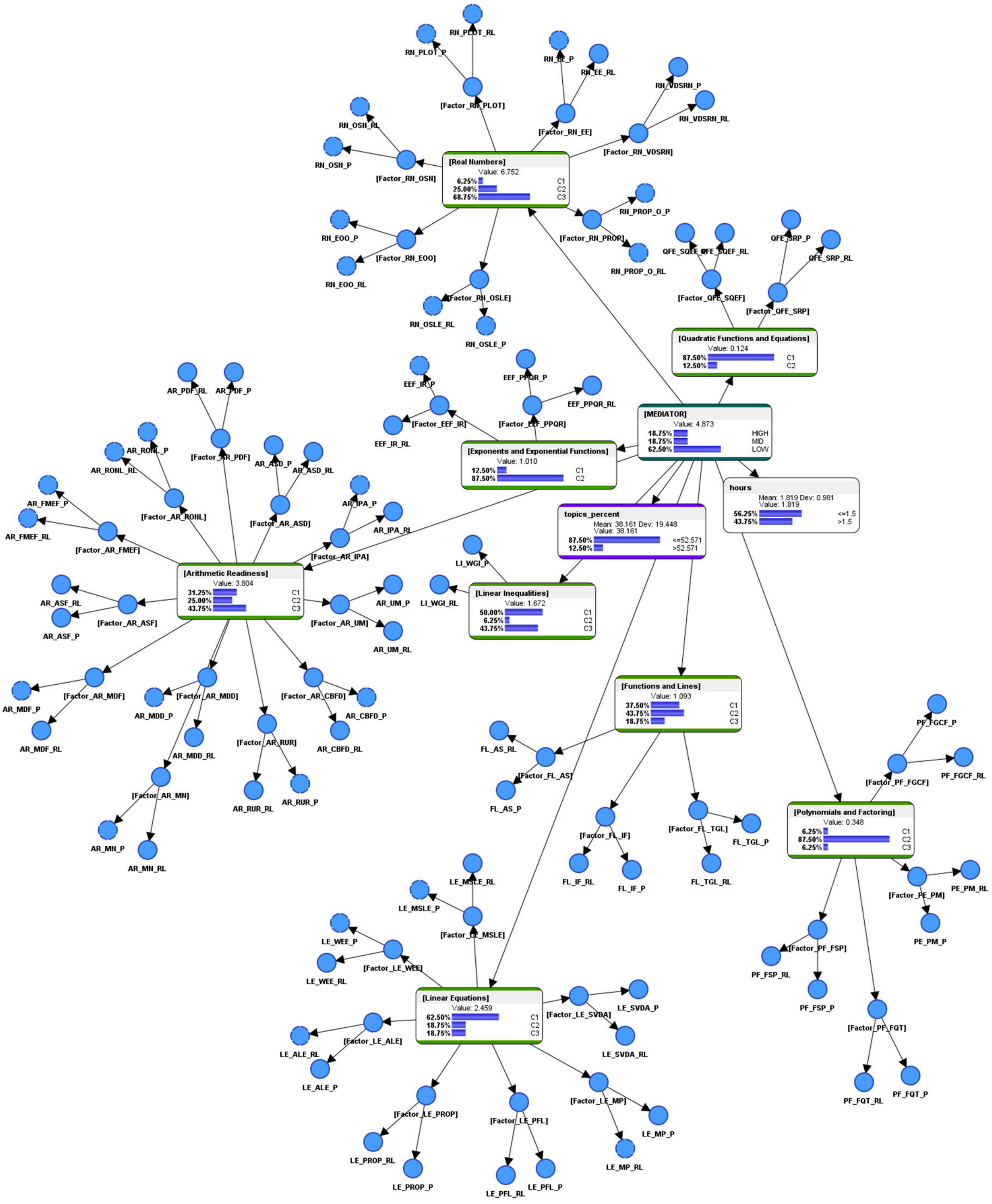


Figure 1. A Bayesian network (BN) representing the "profile" of the AI-ALS was generated via machine learning to understand more about its "pedagogical motif".

4.6. Initial Hypothetical Conjecture of the Researcher

The initial hypothetical conjecture of the researcher was to assume that an AI-ALS might have been designed by its developers to push the higher-performing students a little harder, and conversely, to go easy on the relatively lower-performing students. Therefore, it would not be unreasonable to imagine that a student who performed poorly in the AI-ALS might have experienced having his or her weaknesses being educed by the system. Subsequently, after a personal reflection of those problems via vicarious trial and error (VTE) [43], the student could become cognizant of those weaknesses and could avoid similar predicaments during problem-solving in the paper-based post-test. Conversely, a student who performed well in the AI-ALS might not have experienced having his or her weaknesses educed and, thus, might lack the personal reflections or the VTE to learn from those experiences. Consequently, he or she might perform poorly in the post-test. Regardless of whether a student scored high or low within the AI-ALS, this BN purely profiled its informational pattern (its motif). In other words, it does not affect the calculation of the "gains" attributed to the AI-ALS, as it is not simply a subtraction of the results of the paper-based post-test from the paper-based pre-test.

In this computational model (see Figure 2), the pre-test results (from a paper-based math worksheet) were connected to the profile of the AI-ALS and, subsequently, the profile of the AI-ALS was also connected to the post-test results (from another paper-based math worksheet). This enabled the probabilities of the AI-ALS as a mediator of the students' performance to be calculated, subsequently allowing us the ability to simulate hypothetical scenarios (presented later in Section 5).

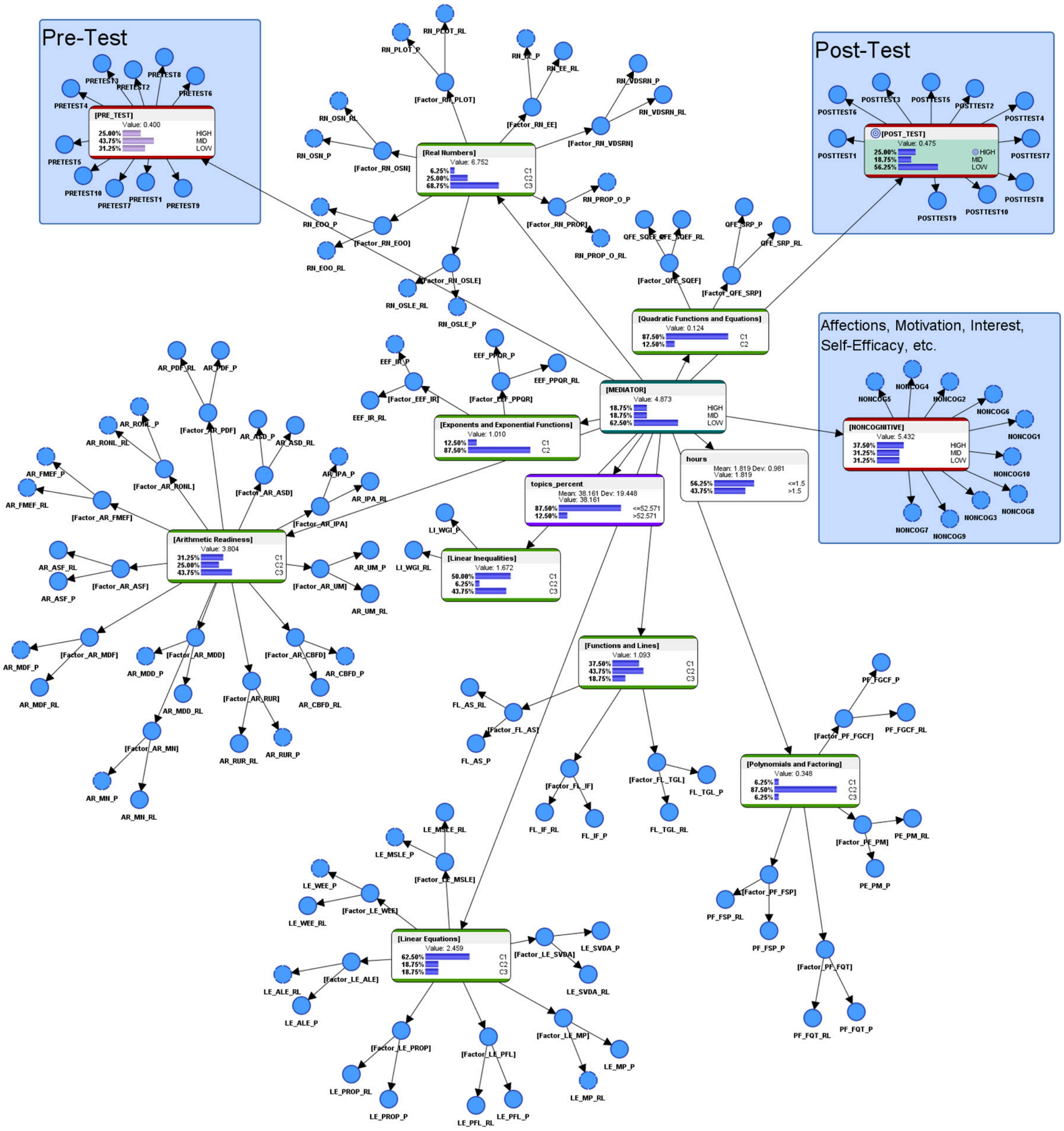


Figure 2. The combined unified impact analysis model, with nodes fully connected, loaded with data from the pre-test, the mediator (the AI-enabled adaptive learning system), the survey (noncognitive factors), and the post-test.

That said, however, it would be contrived to measure "gains" only in terms of cognitive dimensions using the pre-test and post-test, as there might be noncognitive benefits for the students too. Hence, a survey which could be used to understand more about the noncognitive aspects of their learning experiences could also be administered to the students upon completion of their learning process in the AI-ALS. Some of the possible noncognitive instruments that could be utilized by educational stakeholders include those offered by researchers such as Al-Mutawah and Fateel [44], Chamberlin, Moore, and Parks [45], Egalite, Mills, and Greene [46], Lipnevich, MacCann, and Roberts [47], and Mantzicopoulos, Patrick, Strati, and Watson [48].

4.7. Detailed Descriptions of the BN in the Current Paper

Nodes (both the blue round dots and the round-cornered rectangles showing the data distribution histograms) represent variables of interest, for example, the score of a particular mathematics topic (connected to nodes with scores from their corresponding sub-topics), the number of hours spent by a student in the AI-ALS, the percentage of mathematics topics which a student passed in the AI-ALS, or the rating of a particular noncognitive factor (e.g., motivation of a student). Such nodes can correspond to symbolic/categorical variables, numerical variables with discrete values, or discretized continuous variables. Even though BN can handle continuous variables, we exclusively discuss BN with discrete nodes in the current paper, as it is more relevant to helping educational stakeholders categorize students into high-, mid-, and low-achievement groups, allowing teachers to utilize differentiated methods to better address the students' learning needs.

Directed links (the arrows) could represent informational (statistical) or causal dependencies among the variables. The directions are used to define kinship relations, i.e., parent–child relationships. For example, in a Bayesian network with a link from X to Y, X is the parent node of Y, and Y is the child node. In the current paper, it is important to note that the Bayesian network presented is the machine-learned result of probabilistic structural equation modeling (PSEM); it is not a causal model diagram, and, for this reason, the arrows do not represent causation; they merely represent probabilistic structural relationships between the parent node and the child nodes.

4.8. Descriptive Statistical Analysis of the Dataset

In this scenario (see Figure 3), data from pre-test and post-test were connected to the pedagogical motif of the AI-ALS. It could be observed from this analysis that the students initially performed within normal expectations in the pre-test, with 25% scoring at the high level, 43.75% scoring at the mid-level, and 31.25% scoring at the low level. At the mediator stage (which is the AI-ALS), 18.75% scored at the high level, 18.75% scored at the mid-level, and 62.5% scored at the low level. Presumably, the AI-ALS might have exposed many areas of weaknesses of the students, resulting in a high percentage of low scores. In the post-test, a result that was opposite from the researcher's initial hypothetical assumption occurred. Instead of improved scores in the post-test, the students performed poorly; only 25% of the students scored at the high level (unchanged from the pre-test), 18.75% of the students scored at the mid-level (a decrease from 43.75% in the pre-test), and 56.25% of the students scored at the low level (an unfavorable increase from 31.25% in the pre-test).

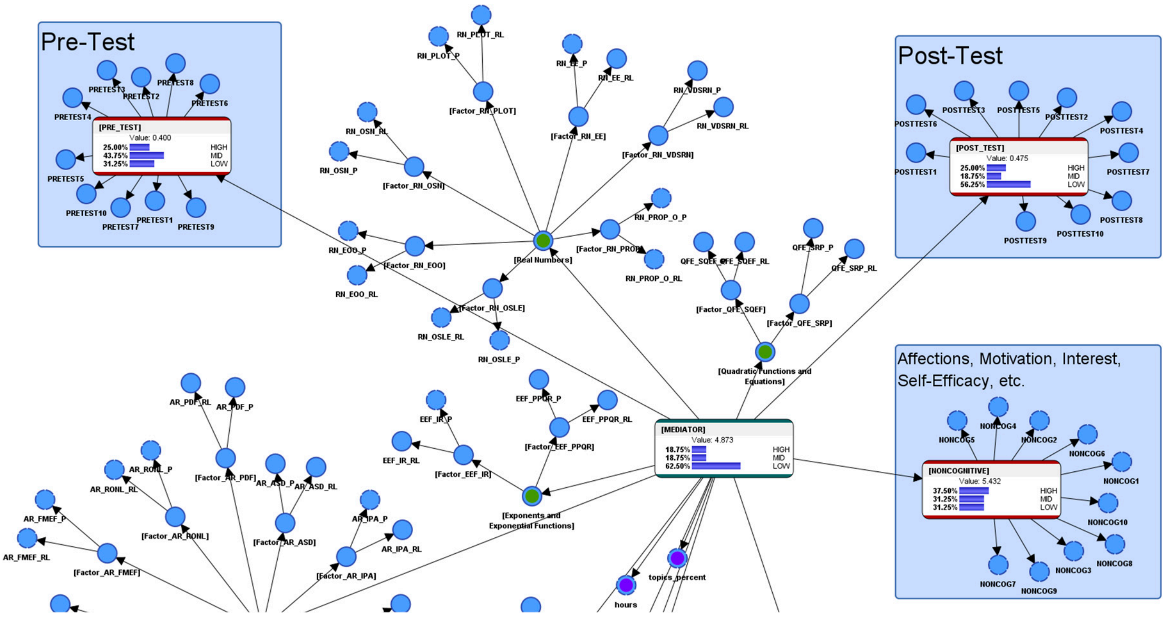


Figure 3. "What already happened?" scenario of students using the AI-ALS.

To investigate the possible relationships between the total effect of each mathematics topic on the target (post-test), a series of techniques utilizing quadrants analysis (in Section 4.8.1), curves analysis (in Section 4.8.2), and Pearson correlations (in Section 4.8.3) were used.

4.8.1. Descriptive Analytics: Quadrant Analysis

Quadrant analysis can be activated in Bayesialab via the following steps: Bayesialab (in validation mode) > Analysis > Report > Target > Total Effects on Target > Quadrants.

The chart of the quadrant analysis generated by Bayesialab (see Figure 4) is divided into four quadrants. The variables' means (of each mathematics topic) are represented along the *x*-axis. The mean of the standardized total effect on the target (the paper-based post-test) is represented along the *y*-axis. As a suggestion, the quadrants could be interpreted as described below.

Top right quadrant (high volume, high impact on target node): This group contains the important variables with greater total effect on the target than the mean value. These mathematics topics are important to the success of the students in the paper-based post-test, and the educational stakeholders should explore ways to help all the students to understand and learn the concepts well in these mathematics topics.

Top left quadrant (low volume, high impact on target node): This group of mathematics topics might already be mastered by the high-performing students, but not yet mastered by the mid- or low-performing students.

Bottom right quadrant (high volume, low impact on target node): This group of mathematics topics might not yet be mastered by many students; thus, educational stakeholders should consider providing remediation of these topics to the students to help bridge the gap that the AI-ALS could not achieve for the students.

Bottom left quadrant (low volume, low impact on target node): This group has a relatively low number of students which mastered these mathematics topics, and these topics also have a relatively lower impact on the target node (the paper-based post-test). The educational stakeholders may consider thinking of ways to improve the students' performances in these topics after their performances in the other topics are already improved.

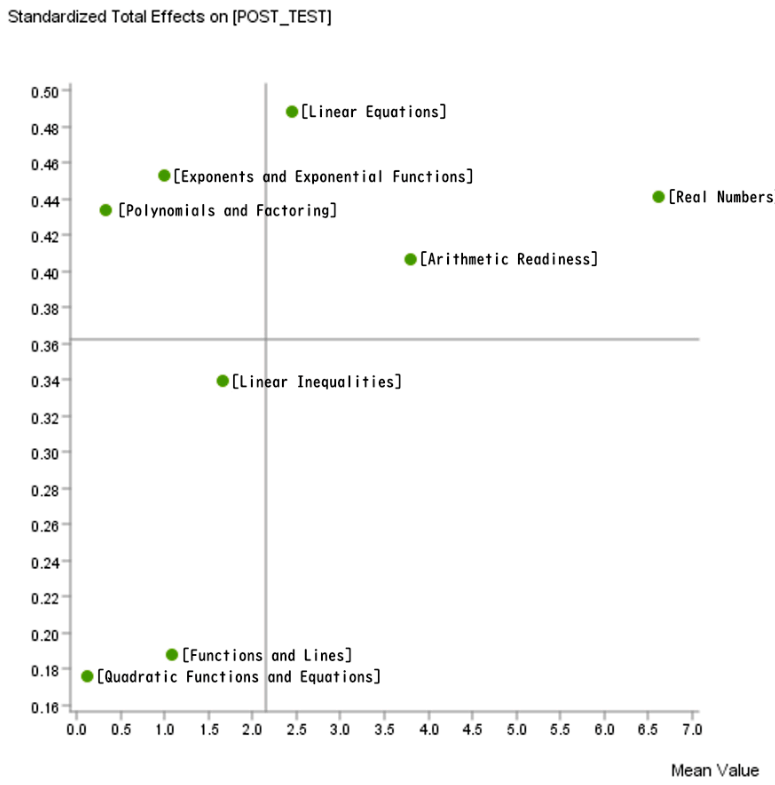


Figure 4. Quadrant analysis of the total effects of the various mathematics topics on the target node.

4.8.2. Descriptive Analytics: Curves Analysis

Another way to visualize the influence of students' mastery of the various mathematics topics on their paper-based post-test is by using this tool in Baysialab via the following steps on the menu bar: *Bayesialab (in validation mode)* > *Analysis* > *Visual* > *Target* > *Target's Posterior* > *Curves* > *Total Effects*.

As observed in Figure 5, the plots of the total effects of the various mathematics topics on the target node (the paper-based post-test) suggest that, with the exception of real numbers, there are positive linear or curvilinear relationships between the students' performances in the mathematics topics in the AI-ALS and the total effect on the paper-based post-test. The results imply that the total effects of mastering the foundational real numbers in the AI-ALS on the paper-based post-test declines as the students spend more effort and time on learning the other mathematics topics. In other words, it could be interpreted that, as the student progresses in the AI-ALS, the marginal utility (to borrow a well-known term from economics) of mastering real numbers decreases, while the marginal utilities of the other mathematics topics continue to increase.

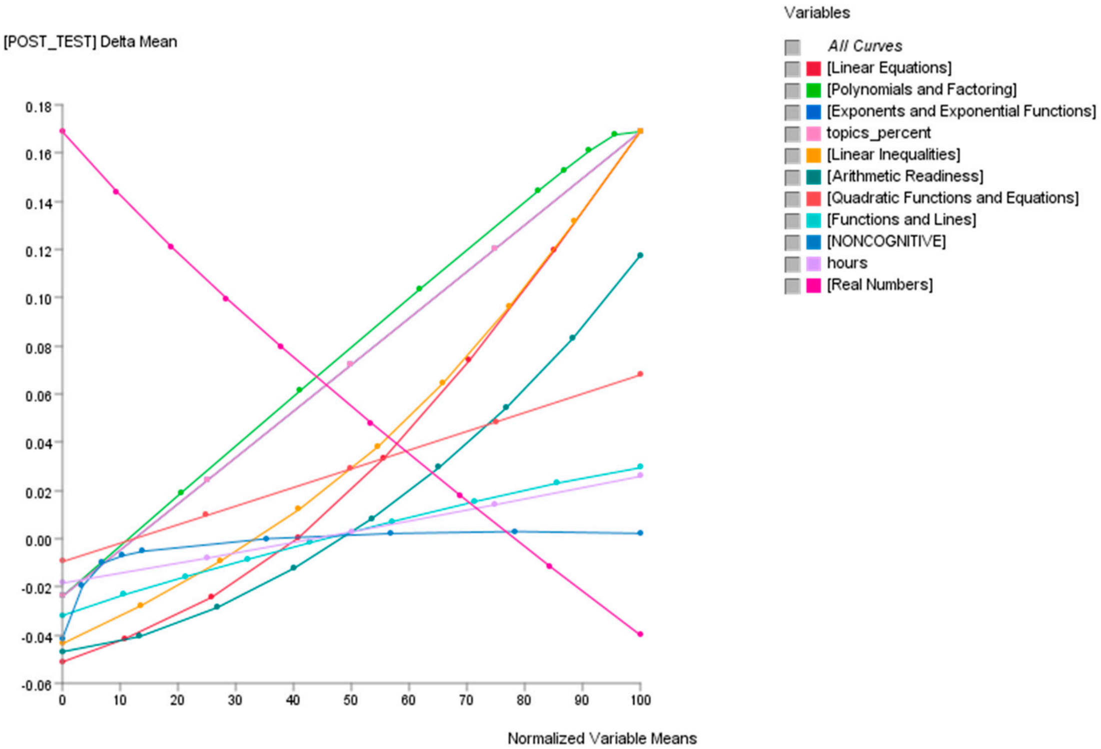


Figure 5. Curves analysis tool, which could be accessed via Bayesialab (in validation mode) > Analysis > Visual > Target > Target's Posterior > Curves > Total Effects.

4.8.3. Descriptive Analytics: Pearson Correlation Analysis

Descriptive analytics can also be performed by using the Pearson correlation analysis tool in Bayesialab. It can be used for corroboration of the relationship analyses between the students' learning performances in the AI-ALS and their corresponding performance in the paper-based post-test. The visualizations of the Pearson correlations can be presented so that it is easier to see the positive correlations highlighted in blue (see Figure 6) and negative correlations highlighted in red (see Figure 7). These two visualizations are based upon the Pearson's correlation analysis of the dataset (a small portion of which is presented in Table 2). The tool can be activated in Bayesialab via the following steps on the menu bar: *Analysis* > *Visual* > *Overall* > *Arc* > *Pearson Correlation* > *R+ (Positive Correlations)*.

One suggestion for interpretation of the negative Pearson correlations could be that the red lines and nodes might represent the regions where the weaknesses of the students were exposed or educed by the AI-ALS. It might not necessarily be an undesirable situation, provided remediation could be provided by the teacher to the students so that the gaps which the AI-ALS could not bridge for the students (for example, if the AI-ALS could not read the students' workings to pinpoint where the mathematical calculation mistakes were for the students) were addressed. The tool can be activated in Bayesialab via the following steps on the menu bar: *Analysis* > *Visual* > *Overall* > *Arc* > *Pearson Correlation* > *R− (Negative Correlations)*.

A portion of the output of the Pearson correlation table generated in Bayesialab (via *Analysis* > *Report* > *Relationships*) is presented in Table 2.

In this section, the descriptive analytics used in Bayesialab were presented. To understand more about how the existing data could be used to forecast the conditions needed to improve the students' performances, we use Section 5 to perform in silico experiments with fully controllable parameters to make sense of what happened.

For now, in the current segment, we temporarily avert focus from the results of the noncognitive survey (affections, motivation, interest, self-efficacy, etc.) as nothing seemed to be out of the ordinary at this point; however, they are also included in Section 5 in the counterfactual "what if?" analyses.

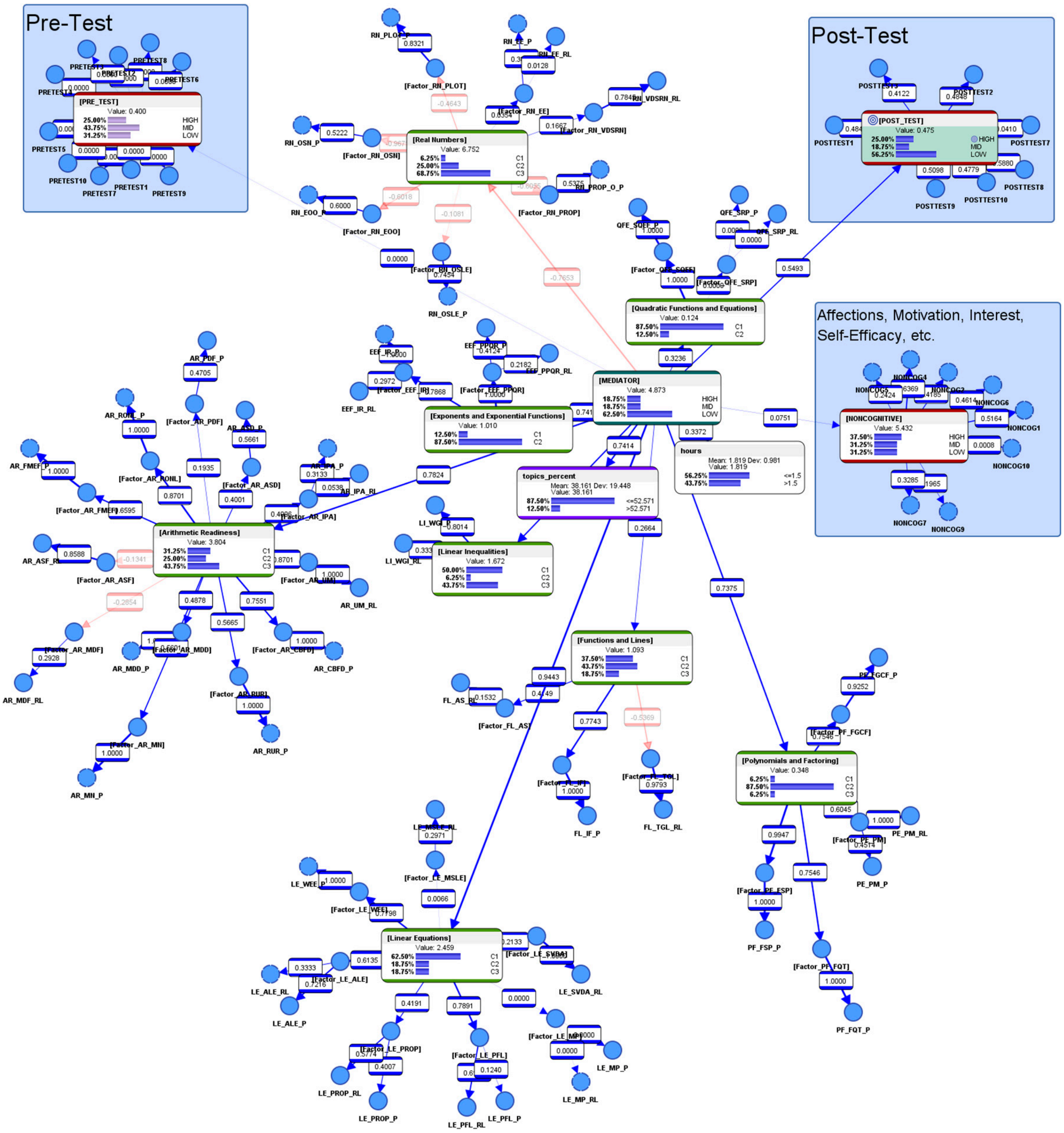


Figure 6. Positive correlations of student learning in AI-ALS and their performance in the post-test.

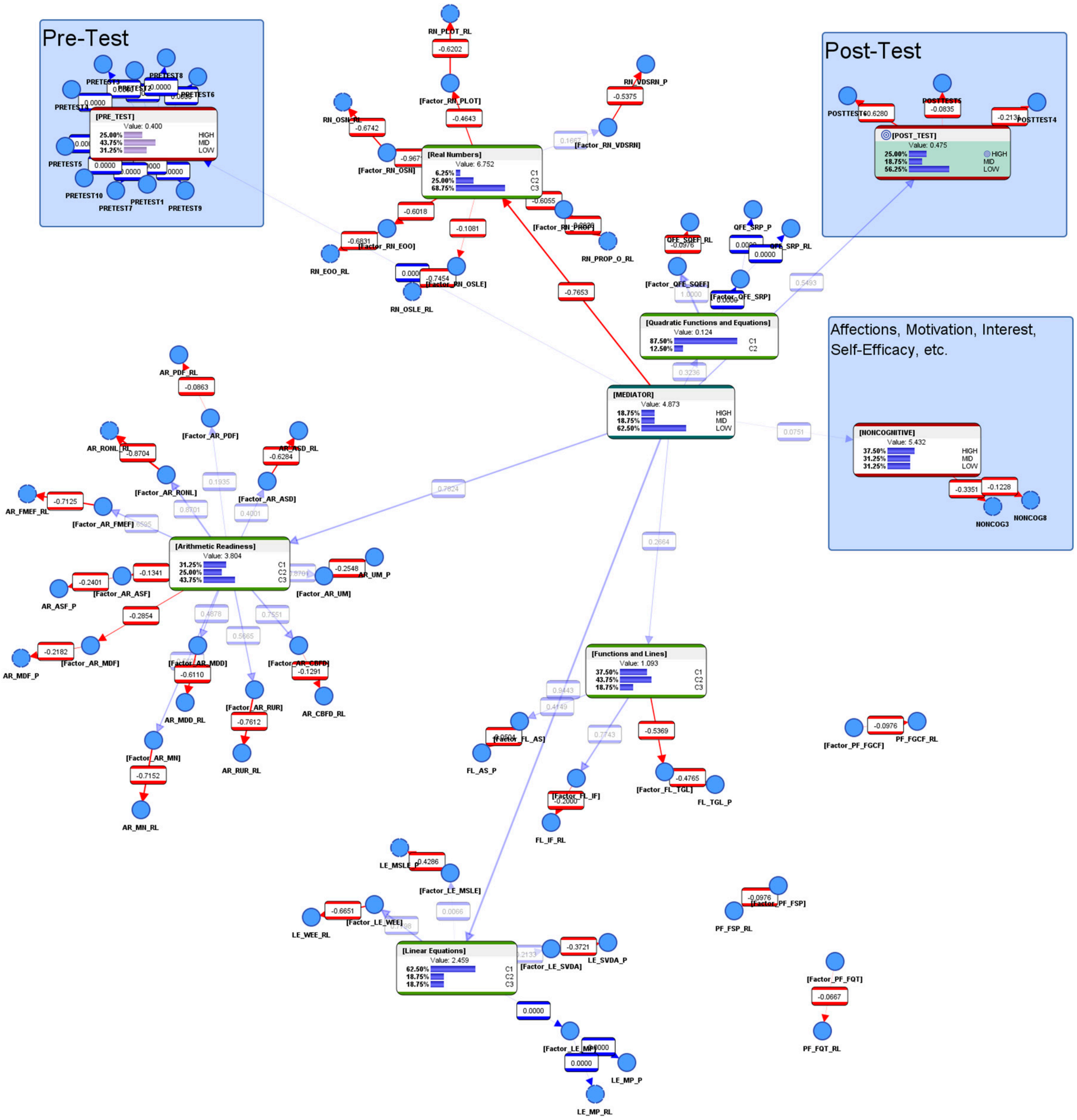


Figure 7. Negative correlations of student learning in AI-ALS and their performance in the post-test.

Table 2. A portion of the Pearson's correlation table generated by Bayesialab.

Parent	Child	Overall Contribution	G_{KL} Test	df	*p*-Value	Pearson's Correlation
[Factor_FL_TGL]	FL_TGL_RL	2.2709%	34.2954	4	0.0001%	0.9793
[MEDIATOR]	[Linear Equations]	1.9526%	29.4878	4	0.0006%	0.9443
[Linear Inequalities]	LI_WGI_RL	1.8679%	28.2090	4	0.0011%	0.3331
[Functions and Lines]	[Factor_FL_TGL]	1.7651%	26.6572	4	0.0023%	−0.5369
[Factor_AR_ASF]	AR_ASF_RL	1.7598%	26.5768	6	0.0174%	0.8588
[Factor_AR_CBFD]	AR_CBFD_P	1.4687%	22.1807	1	0.0002%	1.0000
[Factor_LE_MSLE]	LE_MSLE_RL	1.4521%	21.9301	3	0.0067%	0.2971
[Factor_AR_IPA]	AR_IPA_RL	1.4018%	21.1700	3	0.0097%	0.0538
[Factor_FL_IF]	FL_IF_P	1.4018%	21.1700	1	0.0004%	1.0000
[Factor_AR_MN]	AR_MN_P	1.4018%	21.1700	1	0.0004%	1.0000
[Factor_FL_AS]	FL_AS_P	1.4018%	21.1700	2	0.0025%	−0.0504
[Factor_LE_SVDA]	LE_SVDA_RL	1.4018%	21.1700	1	0.0004%	1.0000
[Factor_FL_TGL]	FL_TGL_P	1.3849%	20.9156	6	0.1900%	−0.4765
[Factor_AR_ASF]	AR_ASF_P	1.3233%	19.9851	8	1.0393%	−0.2401
[Real Numbers]	[Factor_RN_OSN]	1.3160%	19.8748	2	0.0048%	−0.9671
[Factor_AR_UM]	AR_UM_RL	1.3160%	19.8748	1	0.0008%	1.0000
[Factor_AR_RUR]	AR_RUR_P	1.3160%	19.8748	1	0.0008%	1.0000
[Factor_AR_RONL]	AR_RONL_P	1.3160%	19.8748	1	0.0008%	1.0000
[Arithmetic Readiness]	[Factor_AR_UM]	1.3160%	19.8748	2	0.0048%	0.8701
[Arithmetic Readiness]	[Factor_AR_RONL]	1.3160%	19.8748	2	0.0048%	0.8701

5. "What If?" Analytics to Understand More about the Pedagogical Motif of the AI-ALS

This segment describes the simulations performed to explore the following seven hypothetical scenarios, via potential points of leverage in the computational model.

Hypothetical scenario 1. What would happen in the post-test and in the noncognitive factors if all the students scored at the high level in the AI-ALS?

Hypothetical scenario 2. What would happen in the post-test and in the noncognitive factors if all the students scored at the low level in the AI-ALS?

Hypothetical scenario 3. What would happen in the post-test and in the noncognitive factors if all the students scored at the mid-level in the AI-ALS?

Hypothetical scenario 4. What would happen in the post-test and in the noncognitive factors if all the students could spend more hours in the AI-ALS?

Hypothetical scenario 5. What would happen in the post-test and in the noncognitive factors if a higher percentage of topics could be covered in the AI-ALS?

Hypothetical scenario 6. What would happen in the post-test and in the noncognitive factors if a lower percentage of topics could be covered in the AI-ALS?

Hypothetical scenario 7. Finally, the ultimate question, which might be of interest to educational stakeholders: What needs to happen if we would like to have all the students score only at the high level in the post-test (that is, for all of them to become high-performance students)?

5.1. Hypothetical Scenario 1

What would happen in the post-test and in the noncognitive factors if all the students scored at the high level in the AI-ALS?

As shown in Figure 8, the results in the pre-test were fixed at the original levels (by clicking on the node in Bayesialab and selecting "fix probabilities"), since it was the baseline measurement. After applying hard evidence on the mediator node (which is the AI-ALS) by simulating the scenario in which 100% of the students scored at the high level, it could be observed that, in the post-test, the counterfactual results changed to 0% at the high level (compared to the original 25%), 33.3% at the mid-level (compared to the original 18.75%), and 66.7% at the low level (compared to the original 56.25%). It might be suggested that, perhaps, if the questions were too easy, and the students could score high marks easily in the AI-ALS, they might not do so well in the post-test.

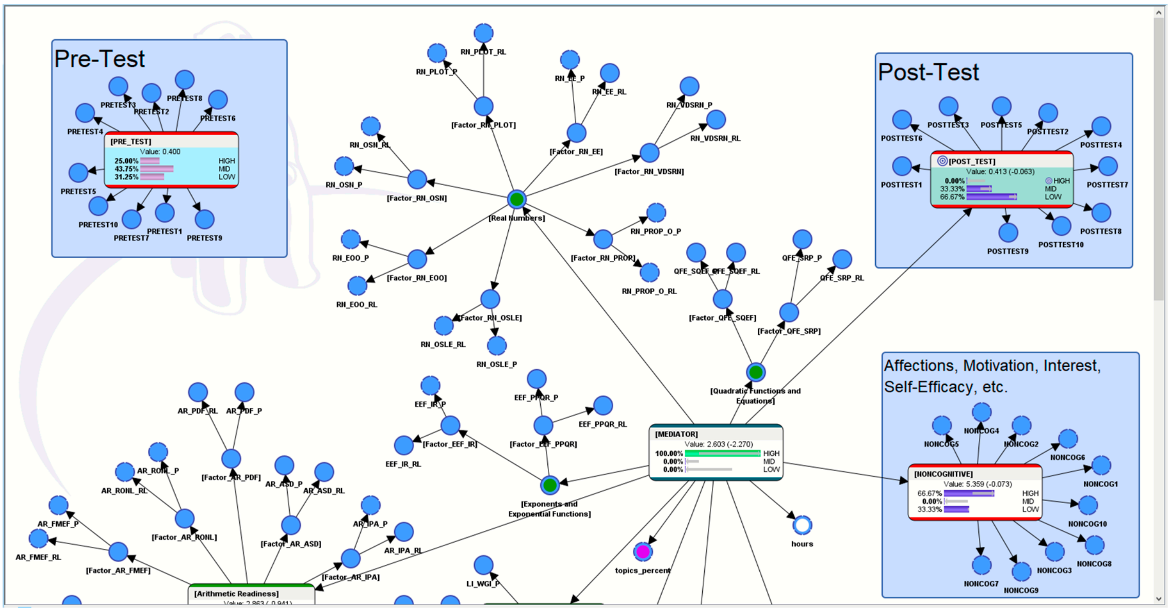


Figure 8. "What if?" simulation of hypothetical scenario 1.

The noncognitive parameters were also changed to 66.67% at the high level (compared to the original 37.50%), 0% at the mid-level (compared to the original 31.25%), and to 33.3% at the low level (compared to the original 31.25%). The disappearance of the mid-level of noncognitive factors might suggest that, perhaps, even though the questions in the AI-ALS were easy, the students were divided into two distinct groups at the high level and low level. This suggested that they were at two extreme ends of the noncognitive parameters, for example, in terms of their interest, attitude toward mathematics, motivation, etc.

Overall, this scenario suggested that making it relatively easy for students to score high marks in the AI-ALS might not lead to a high probability in contributing toward enhancing their performance in the post-test.

5.2. Hypothetical Scenario 2

What would happen in the post-test and in the noncognitive factors if all the students scored at the low level in the AI-ALS?

As shown in Figure 9, the results in the pre-test were fixed at the original levels, since it was the baseline measurement. After applying hard evidence on the mediator node by simulating the scenario in which 100% of the students scored at the low-level, it could be observed that, in the post-test, the counterfactual results changed to 20% at the high level (compared to the original 25%), 10% at the mid-level (compared to the original 18.75%), and 70% at the low level (compared to the original 56.25%). It might be suggested that, perhaps, if the questions were too difficult, and the students could not score high marks easily, they might also not do so well in the post-test.

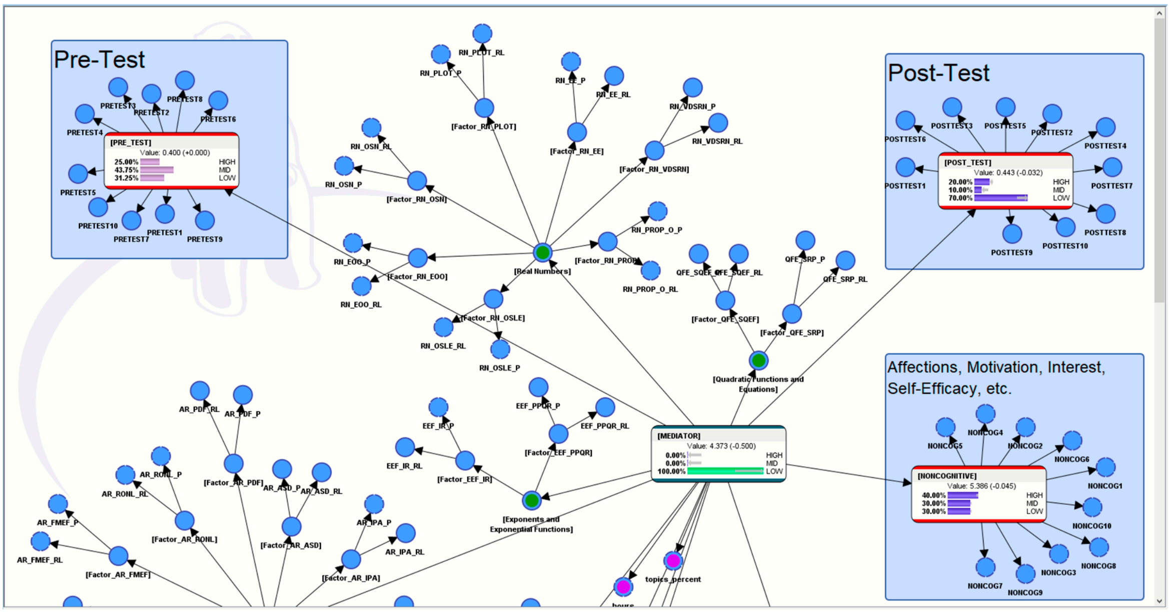


Figure 9. "What if?" simulation of hypothetical scenario 2.

The noncognitive parameters were also changed to 40% at the high level (compared to the original 37.50%), 30% at the mid-level (compared to the original 31.25%), and to 30% at the low level (compared to the original 31.25%). The similarity of the "before" and "after" noncognitive factors might suggest that their noncognitive parameters did not change much, for example, in terms of their interest, attitude toward mathematics, motivation, etc.

Overall, however, this scenario suggested that making it relatively difficult for students to score high marks in the AI-ALS might not be optimal for enhancing their performance in the post-test.

5.3. Hypothetical Scenario 3

What would happen in the post-test and in the noncognitive factors if all the students scored at the mid-level in the AI-ALS?

As shown in Figure 10, the results in the pre-test were fixed at the original levels, since it was the baseline measurement. After applying hard evidence on the mediator node by simulating the scenario in which 100% of the students scored at the mid-level, it could be observed that, in the post-test, the counterfactual results changed to 66.67% at the high level (compared to the original 25%), 33.33% at the mid-level (compared to the original 18.75%), and 0% at the low level (compared to the original 56.25%). It might be suggested that, perhaps, if the questions were not too difficult, and also not too easy, the students might achieve the best performance in the post-test.

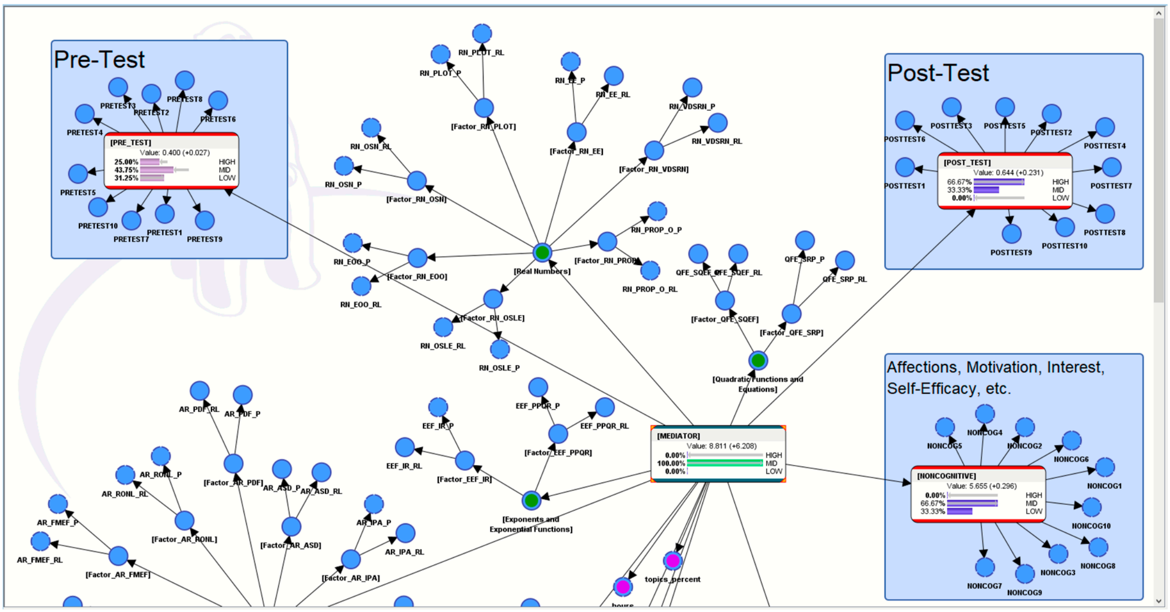


Figure 10. "What if?" simulation of hypothetical scenario 3.

The noncognitive parameters were also changed to 0% at the high level (compared to the original 37.50%), 66.67% at the mid-level (compared to the original 31.25%), and to 33.33% at the low level (compared to the original 31.25%). The disappearance of the high level of noncognitive factors might suggest that, perhaps, the questions in the AI-ALS were not too easy, and also not too difficult; thus, even students at the mid-level and low level of the noncognitive parameters (for example, in terms of their interest, attitude toward mathematics, motivation, self-efficacy, etc.) could still benefit from using this AI-ALS.

Overall, this scenario suggested that making it not too difficult and, also, not too easy for students to score high marks in the AI-ALS might be optimal for enhancing their performance in the post-test.

5.4. Hypothetical Scenario 4

What would happen in the post-test and in the noncognitive factors if all the students could spend more time (or, conversely, less time) in the AI-ALS?

As observed in the original state, the results (see Figure 11) showed that 56.25% of the students spent less than or equal to 1.5 h in the AI-ALS, while 43.75% spent more than 1.5 h in the AI-ALS. The original values in the post-test were 25% at the high level, 18.75% at the mid-level, and 56.25% at the low level. The original values in the noncognitive dimensions (e.g., affections, motivation, interest, self-efficacy, etc.) were 37.5% at the high level, 31.25% at the mid-level, and 31.25% at the low level.

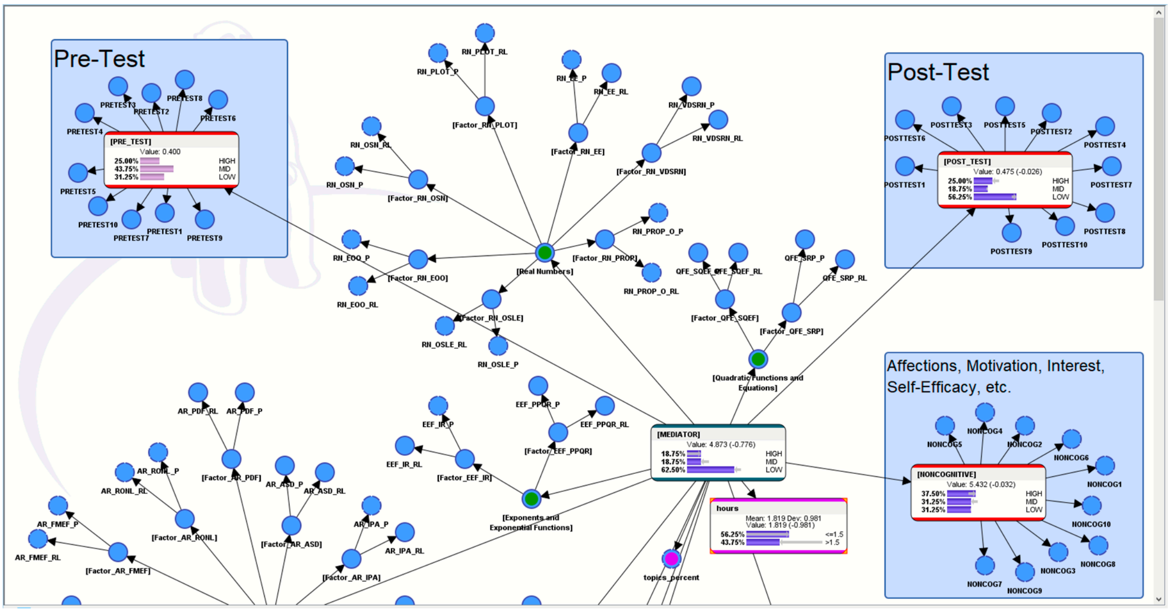


Figure 11. "What if?" simulation of hypothetical scenario 4, part 1.

When hard evidence was applied in this computational model (see Figure 12) to simulate that 100% of the students spent less than or equal to 1.5 h in the AI-ALS, the counterfactual results showed that the values in the post-test became 19.14% at the high level (originally 25%), 19.87% at the mid-level (originally 18.75%), and 60.99% at the low level (originally 56.25%). The values in the noncognitive dimensions (e.g., affections, motivation, interest, self-efficacy, etc.) counterfactually became 43.68% at the high level (originally 37.5%), 24.91% at the mid-level (originally 31.25%), and 31.41% at the low level (originally 31.25%). This finding suggested that, if the students spent less than or equal to 1.5 h in AI-ALS, it might lead to a decrease in performance in the post-test, but only slight changes in the noncognitive parameters.

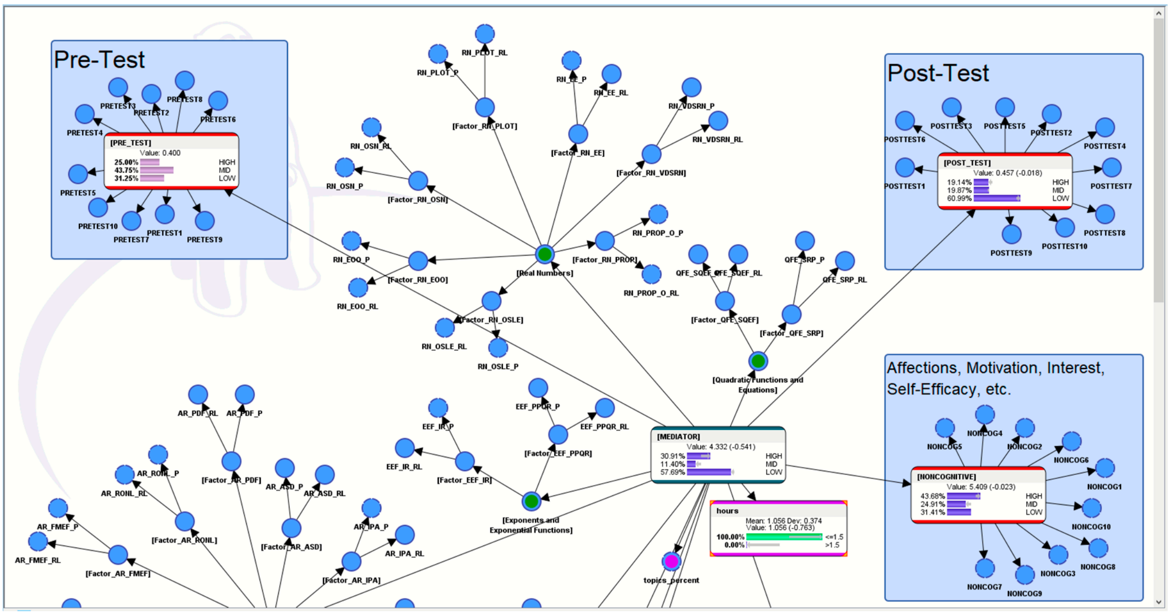


Figure 12. "What if?" simulation of hypothetical scenario 4, part 2.

Conversely, when hard evidence was applied in this computational model (see Figure 13) to simulate that 100% of the students spent more than 1.5 h in the AI-ALS, the counterfactual values in the post-test became 33.42% at the high level (originally 25%), 16.71% at the mid-level (originally 18.75%), and 49.88% at the low level (originally 56.25%). The values in the noncognitive dimensions (e.g., affections, motivation, interest, self-efficacy, etc.) counterfactually became 28.50% at the high level (originally 37.5%), 40.54% at the mid-level (originally 31.25%), and 30.96% at the low level (originally 31.25%). This finding suggested that, if the students spent more than 1.5 h in the AI-ALS, it might lead to a slight increase in performance in the post-test, but only slight changes in the noncognitive parameters. Overall, the outcomes of hypothetical scenario 4 were well within the researcher's expectations.

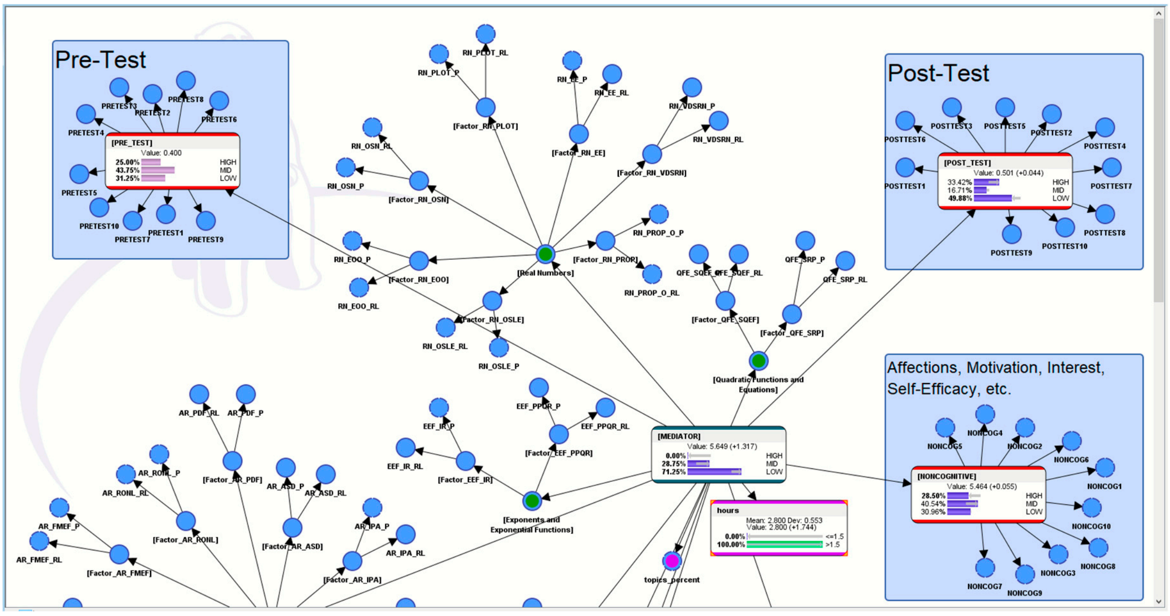


Figure 13. "What if?" simulation of hypothetical scenario 4, part 3.

5.5. Hypothetical Scenario 5

What would happen in the post-test and in the noncognitive factors if a higher (or, conversely, lower) percentage of topics could be covered for the students in the AI-ALS?

As observed in the original state, the results (see Figure 14) showed that less than or equal to 52.57% of the topics were covered for 87.5% of the students in the AI-ALS, while more than 52.57% of the topics were covered for 12.5% of the students in the AI-ALS. The original values in the post-test were 25% at the high level, 18.75% at the mid-level, and 56.25% at the low level. The original values in the noncognitive dimensions (e.g., affections, motivation, interest, self-efficacy, etc.) were 37.5% at the high level, 31.25% at the mid-level, and 31.25% at the low level.

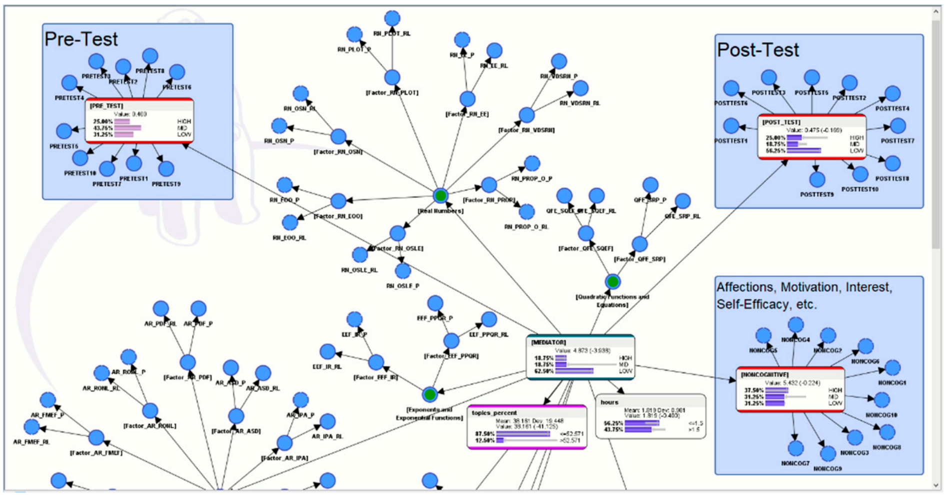

Pre-Test
Post-Test
Affections, Motivation, Interest, Self-Efficacy, etc.
[PRE_TEST]
[POST_TEST]
[MEDIATOR]
[NONCOGNITIVE]
topics_percent
hours
[Real Numbers]
[Quadratic Functions and Equations]
[Exponents and Exponential Functions]
HIGH
MID
LOW

Figure 14. "What if?" simulation of hypothetical scenario 5, part 1.

When hard evidence was applied in this computational model (see Figure 15) to simulate that more than 52.57% of the topics were covered for 100% of the students in the AI-ALS, the counterfactual the values in the post-test became 66.67% at the high level (originally 25%), 33.33% at the mid-level (originally 18.75%), and 0% at the low level (originally 56.25%). The values in the noncognitive dimensions (e.g., affections, motivation, interest, self-efficacy, etc.) counterfactually became 0% at the high level (originally 37.5%), 66.67% at the mid-level (originally 31.25%), and 33.33% at the low level (originally 31.25%). This finding suggested that, if more than 52.57% of the topics could be covered for 100% of the students in the AI-ALS, it might lead to an increase in performance in the post-test, and also could result in an increase in mid-level values in the noncognitive parameters (which we could tentatively interpret as being well-balanced values). Further investigation in a future study might be needed.

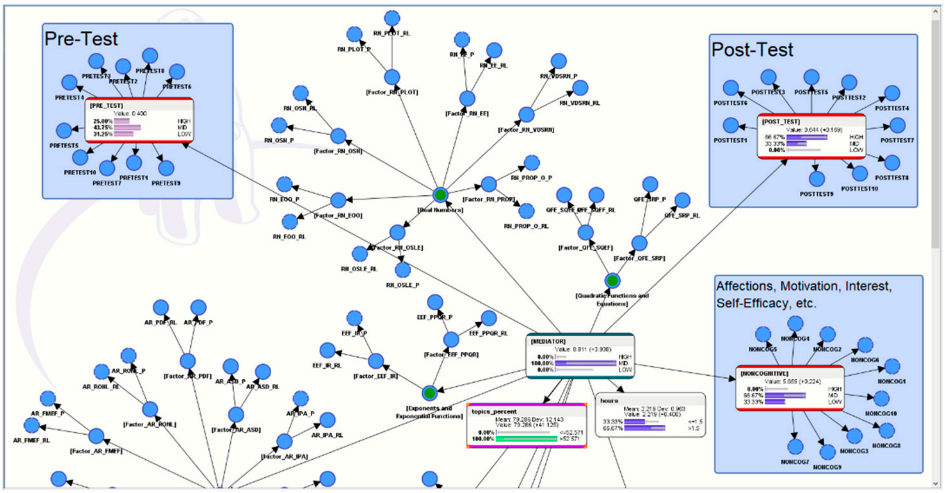
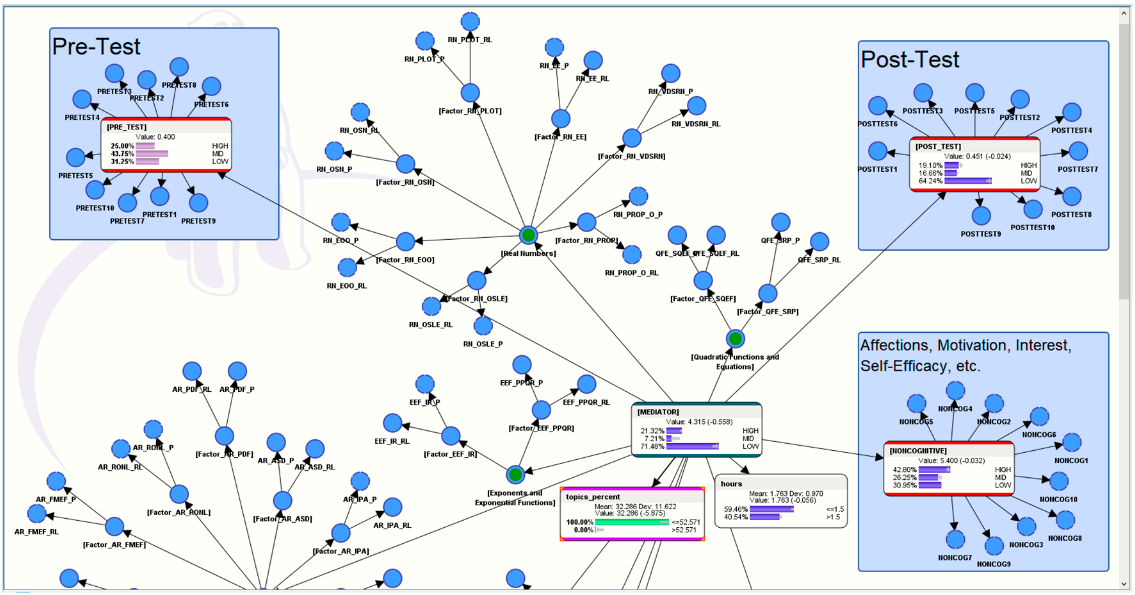

Pre-Test
Post-Test
Affections, Motivation, Interest, Self-Efficacy, etc.
[PRE_TEST]
[POST_TEST]
[MEDIATOR]
[NONCOGNITIVE]
topics_percent
hours
[Real Numbers]
[Quadratic Functions and Equations]
[Exponents and Exponential Functions]
HIGH
MID
LOW

Figure 15. "What if?" simulation of hypothetical scenario 5, part 2.

5.6. Hypothetical Scenario 6

What would happen in the post-test and in the noncognitive factors if a lower percentage of topics could be covered in the AI-ALS?

In other words, would covering a lower percentage of topics within the same original amount of time lead to better performance in the post-test? Let us examine this simulated scenario. When hard evidence was applied in this computational model (see Figure 16) to simulate that less than or equal to 52.57% of the topics were covered for 100% of the students in the AI-ALS, the counterfactual values in the post-test became 19.10% at the high level (originally 25%), 16.66% at the mid-level (originally 18.75%), and 62.24% at the low level (originally 56.25%). The values in the noncognitive dimensions (e.g., affections, motivation, interest, self-efficacy, etc.) counterfactually became 42.8% at the high level (originally 37.5%), 26.25% at the mid-level (originally 31.25%), and 30.95% at the low level (originally 31.25%). This finding suggested that, if less than or equal to 52.57% of the topics could be covered for 100% of the students in the AI-ALS, it might lead to a decrease in performance in the post-test, and also could result in an increase in the mid-level values in the noncognitive parameters (which we could tentatively interpret as being well-balanced values). Further investigation in a future study might be needed.



[PRE_TEST]
Value: 0.400
25.00% HIGH
43.75% MID
31.25% LOW

[POST_TEST]
Value: 0.451 (-0.024)
19.10% HIGH
16.66% MID
64.24% LOW

[MEDIATOR]
Value: 4.315 (-0.558)
21.32% HIGH
7.21% MID
71.48% LOW

topics_percent
Mean: 32.286 Dev: 11.622
Value: 32.286 (-5.875)
100.00% <=52.571
0.00% >52.571

hours
Mean: 1.763 Dev: 0.970
Value: 1.763 (-0.056)
59.46% <=1.5
40.54% >1.5

[NONCOGNITIVE]
Value: 5.400 (-0.032)
42.80% HIGH
26.25% MID
30.95% LOW

Figure 16. "What if?" simulation of hypothetical scenario 6.

5.7. Hypothetical Scenario 7

Finally, the ultimate question: What needs to happen if we would like to have all the students score only at the high level in the post-test (that is, for all of them to become high-performance students)?

Let us examine this simulated scenario. When hard evidence was applied in this computational model (see Figure 17) to simulate that 100% of the students could score at the high level in the post-test, the values in the noncognitive dimensions (e.g., affections, motivation, interest, self-efficacy, etc.) counterfactually became 20.11% at the high level (originally 37.5%), 48.23% at the mid-level (originally 31.25%), and 31.66% at the low level (originally 31.25%).

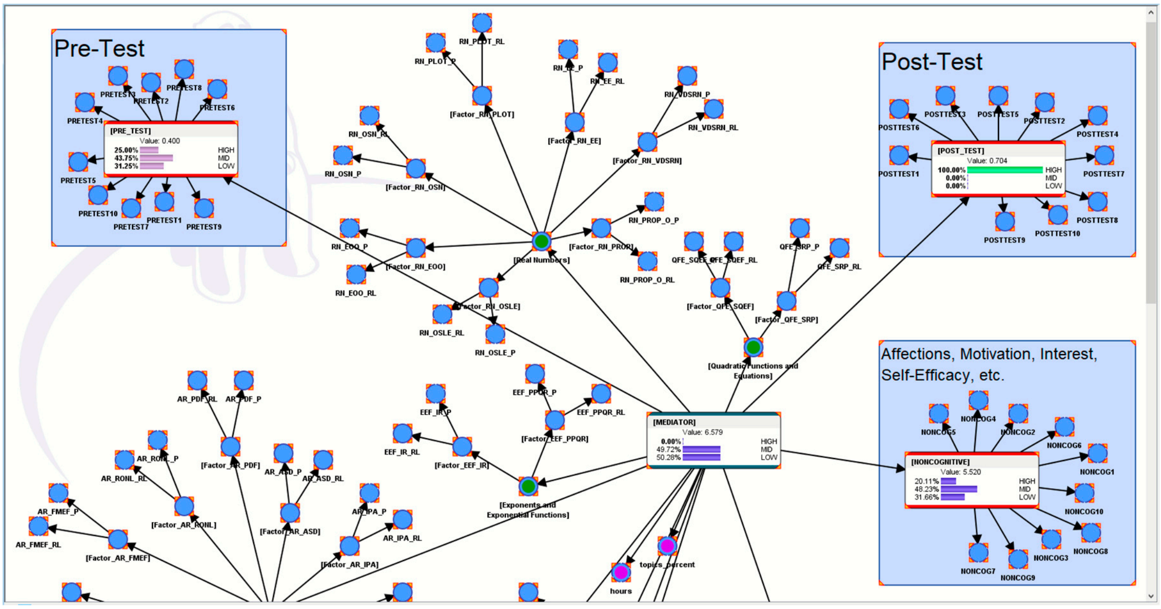


Figure 17. "What if?" simulation of hypothetical scenario 7.

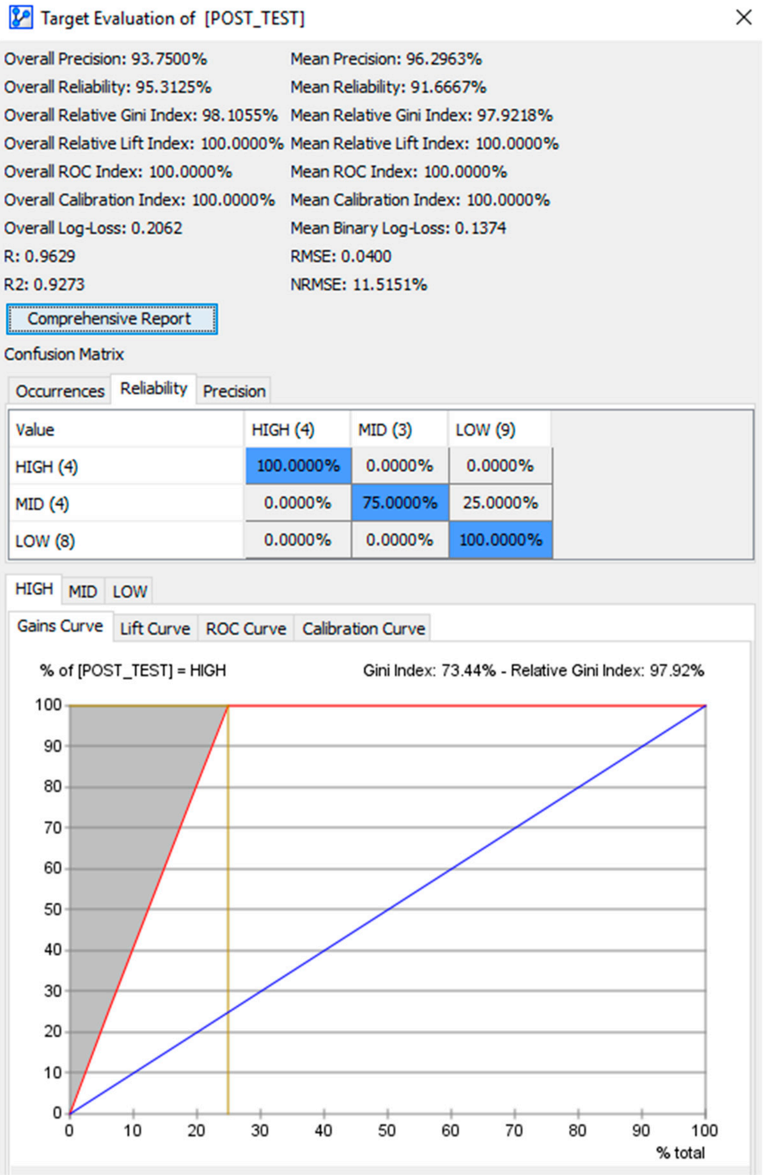
In the Mediator node (which represents the overall effects of the AI-ALS), the simulated counterfactual results suggested that, in order to contribute to optimal performance in the students, it would be ideal if 0% of the students could achieve high-level scores, 49.72% of them could achieve mid-level scores, and 50.28% of them could achieve low-level scores in the AI-ALS. In other words, if the questions in the AI-ALS were easy to solve or at low levels of difficulty, but absolutely not at the easy level of difficulty, that would be ideal for enhancing the performance of the students in the post-test.

Furthermore, a slight increase in the mid-level values in the noncognitive parameters (which we could tentatively interpret as being well-balanced values) might potentially contribute to the students' performance. This hypothetical scenario, however, might need further investigation in a future study.

This section presented the predictive analytics used for simulations of the seven hypothetical scenarios. In the next section, the tools in Bayesialab which could be used for the evaluation of the predictive performance of the BN model in the current paper are presented.

6. Evaluation of the Predictive Performance of the Bayesian Network Model

The predictive performance of a model can be evaluated using measurement tools such as the gains curve (see Figure 18), lift curve (see Figure 19), and the receiver operating characteristic (ROC) curve (see Figure 20), as well as by statistical bootstrapping (see Figure 21).



Target Evaluation of [POST_TEST]

Overall Precision: 93.7500% Mean Precision: 96.2963%
Overall Reliability: 95.3125% Mean Reliability: 91.6667%
Overall Relative Gini Index: 98.1055% Mean Relative Gini Index: 97.9218%
Overall Relative Lift Index: 100.0000% Mean Relative Lift Index: 100.0000%
Overall ROC Index: 100.0000% Mean ROC Index: 100.0000%
Overall Calibration Index: 100.0000% Mean Calibration Index: 100.0000%
Overall Log-Loss: 0.2062 Mean Binary Log-Loss: 0.1374
R: 0.9629 RMSE: 0.0400
R2: 0.9273 NRMSE: 11.5151%

Confusion Matrix

Value	HIGH (4)	MID (3)	LOW (9)
HIGH (4)	100.0000%	0.0000%	0.0000%
MID (4)	0.0000%	75.0000%	25.0000%
LOW (8)	0.0000%	0.0000%	100.0000%

Figure 18. Gains curve.

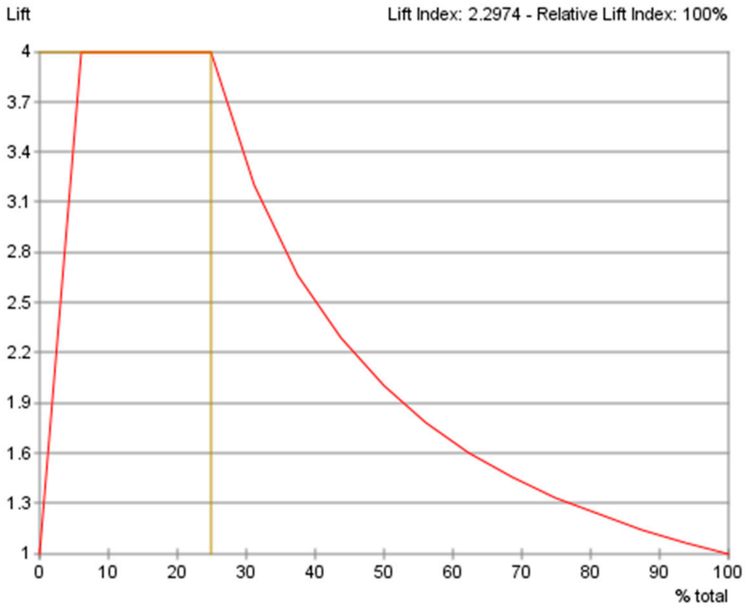


Figure 19. Lift curve.

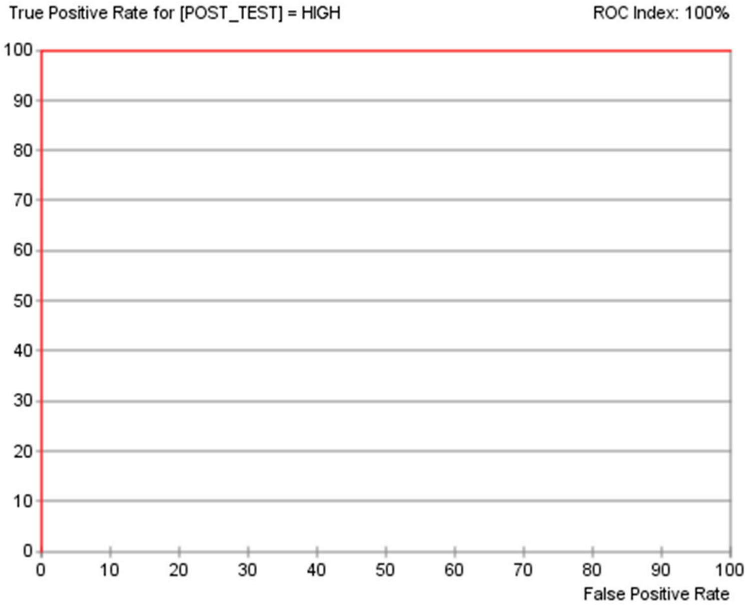
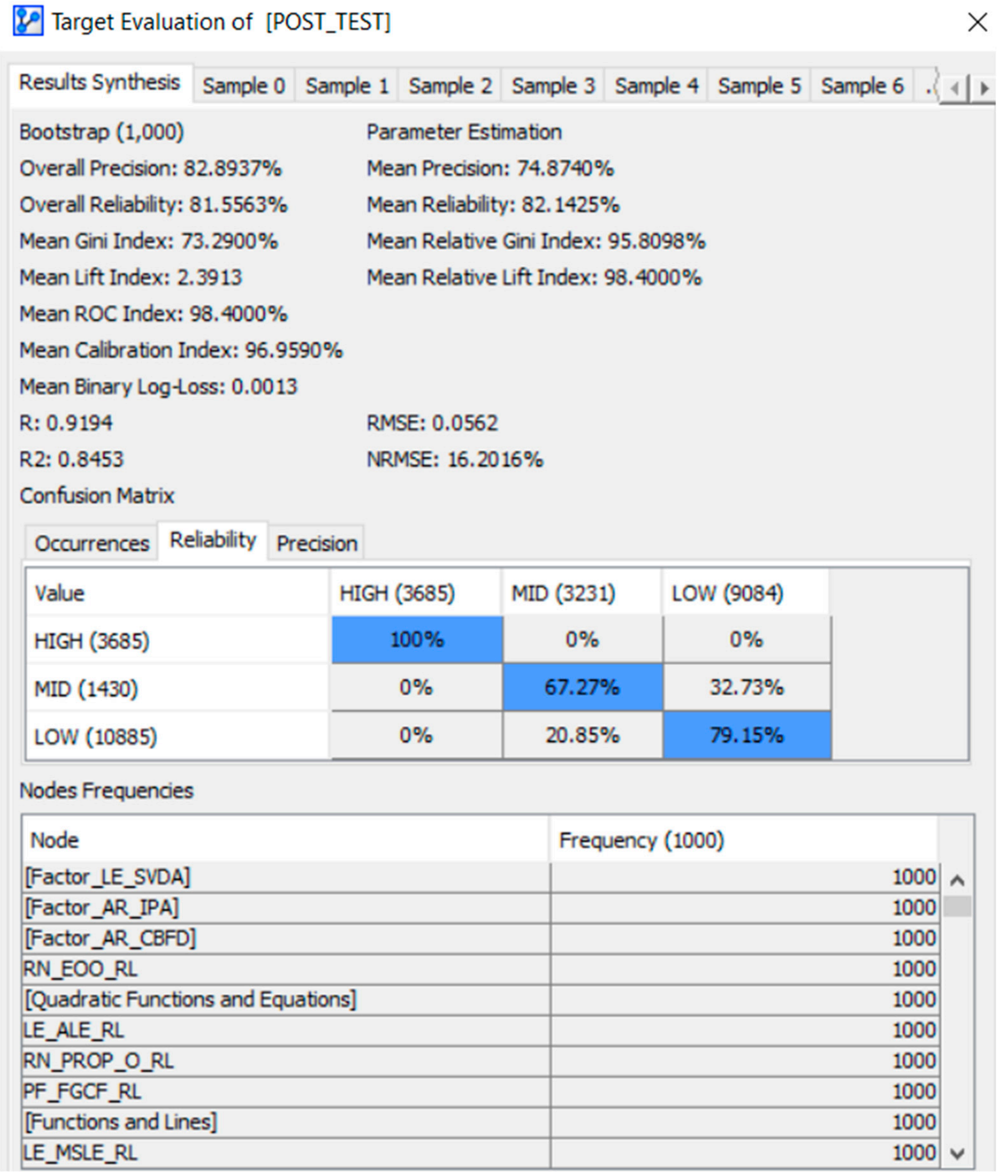


Figure 20. Receiver operating characteristic (ROC) curve of the Bayesian network.



Bootstrap (1,000)
Overall Precision: 82.8937%
Overall Reliability: 81.5563%
Mean Gini Index: 73.2900%
Mean Lift Index: 2.3913
Mean ROC Index: 98.4000%
Mean Calibration Index: 96.9590%
Mean Binary Log-Loss: 0.0013
R: 0.9194
R2: 0.8453
Confusion Matrix

Parameter Estimation
Mean Precision: 74.8740%
Mean Reliability: 82.1425%
Mean Relative Gini Index: 95.8098%
Mean Relative Lift Index: 98.4000%
RMSE: 0.0562
NRMSE: 16.2016%

Value	HIGH (3685)	MID (3231)	LOW (9084)
HIGH (3685)	100%	0%	0%
MID (1430)	0%	67.27%	32.73%
LOW (10885)	0%	20.85%	79.15%

Nodes Frequencies

Node	Frequency (1000)
[Factor_LE_SVDA]	1000
[Factor_AR_IPA]	1000
[Factor_AR_CBFD]	1000
RN_EOO_RL	1000
[Quadratic Functions and Equations]	1000
LE_ALE_RL	1000
RN_PROP_O_RL	1000
PF_FGCF_RL	1000
[Functions and Lines]	1000
LE_MSLE_RL	1000

Figure 21. Evaluation of predictive performance of BN on the target (the scores of the post-test).

6.1. Gains Curve Analysis

As observed in Figure 18, the results were good. The overall precision was 93.7500%; the mean precision was 96.2963%; the overall reliability was 95.3123%; the mean reliability was 91.6667%; the overall relative Gini index was 98.1055%; the mean relative Gini index was 97.9218%; the overall relative lift index was 100%; the mean relative lift index was also 100%; the overall receiver operating characteristic (ROC) index was 100%; the mean ROC index was also 100%. The overall log-loss was 0.2062; the mean binary log-loss was 0.1374, which was quite good as log-loss should be as close to zero as possible, since it measures loss of information in the prediction; the linear correlation coefficient *R* was 0.9629 and could be considered to be very good relationship between the variables (the scores in the AI-ALS and the target scores in the paper-based post-test); the coefficient of determination R^2 was 0.9273 and could be considered to be very good as it measures the strength of the linear association between the scores in the AI-ALS and the target node (the scores of the paper-based post-test); the root-mean-square error (RMSE) was 0.0400 and its low value could be considered to be good since it measures how far the data points were from the regression line; the normalized root-mean-square error (NRMSE) was 11.5151% and could be considered to be acceptably good since it was quite low in value.

In the gains curve (see Figure 18), there were around 25% of participants with the target value of the high level in the paper-based post-test (yellow). The blue diagonal line represents the gain curve of a pure random policy, which is a prediction without using this predictive model. The red lines represent the gains curve using this predictive model. The Gini index of 73.44% and relative Gini index of 97.92% suggest that the gains of using this predictive model vis-à-vis not using it was acceptable.

A confusion matrix is presented in the middle portion of Figure 18. The confusion matrix provides additional information about the computational model's predictive performance. The leftmost column in the matrix contains the predicted values, while the actual values in the data are presented in the top row. Three confusion matrix views are available by clicking on the corresponding tabs. The occurrence matrix indicates the number of cases for each combination of predicted versus actual values. The diagonal shows the number of true positives. The reliability matrix indicates the probability of the reliability of the prediction of a state in each cell. Reliability measures the overall consistency of a prediction. A prediction can be regarded to have a high reliability if the computational model produces similar results under consistent conditions. As observed in Figure 18, the reliability of the model's prediction of the "high" state was 100%, the reliability of the model's prediction of the "mid" state was 75%, and the reliability of predicting the "low" state was 100%. The precision matrix indicates the probability of the precision of the prediction of a state in each cell. Precision measures the overall accuracy which the computational model can perform predictions correctly.

The gains curve, lift curve, and ROC analysis tools can be utilized in Bayesialab via the following steps on the menu bar: *Bayesialab (in validation mode) > Analysis > Network Performance > Target.*

6.2. Lift Curve Analysis

The lift curve (see Figure 19) corresponded to the gains curve (see Figure 18). The value of the best lift was around 25%. The lift index of 2.2974 and relative lift index of 100% suggested that the performance of this predictive model was acceptably good.

6.3. Receiver Operating Characteristic (ROC)

The predictive performance of the Bayesian network model could be evaluated using a receiver operating characteristic curve (ROC) (see Figure 20), which was a plot of the true positive rate (y-axis) against the false positive rate (x-axis). The ROC index indicated that 100% of the cases were predicted correctly with this predictive model.

Together, the gains curve, the lift curve, and the ROC curve indicated that the predictive performance of the Bayesian network model in the current paper was acceptably good.

In addition to the gains curve, lift curve, and ROC, another way to evaluate the predictive model would be to perform bootstrapping (see Figure 21), where the Bayesialab software randomly draws on the data distribution of each node 1000 times to simulate parametric data of 1000 students. This can be done in Bayesialab via the following steps on the menu bar: *Bayesialab (in validation mode) > Tools > Resampling > Target Evaluation > Bootstrap.*

6.4. Statistical Bootstrapping

As observed in the results (see Figure 21) generated by Bayesialab after performing bootstrapping 1000 times on the data distribution of each node in the BN using the parameter estimation algorithm, the overall precision was 82.8937%, the mean precision was 74.8740%, the overall reliability was 81.5563%, the mean reliability was 82.1425%, the mean Gini index was 73.2900%, the mean relative Gini index was 95.8098%, the mean lift index was 2.3913, the mean relative lift index was 98.4000%, the mean ROC index was 98.4000%, the mean calibration index was 96.9590%, the mean binary log-loss was 0.0013, the correlation coefficient R was 0.9194, the coefficient of determination R^2 was 0.8453, the RMSE was 0.0562, and the NRSME was 16.2016%. These results suggested that the predictive performance of the BN model could be considered to be acceptably good.

A confusion matrix (for bootstrapping the data 1000 times in every node) is presented in the middle portion of Figure 21. The confusion matrix provided additional information about the computational model's predictive performance. The leftmost column in the matrix contains the predicted values, while the actual values in the data are presented in the top row. Three confusion matrix views would be available by clicking on the corresponding tabs. The occurrence matrix would indicate the number of cases for each combination of predicted versus actual values. The diagonal shows the number of true positives. The reliability matrix would indicate the probability of the reliability of the prediction of a state in each cell. Reliability measures the overall consistency of a prediction. A prediction could be considered to be highly reliable if the computational model produces similar results under consistent conditions. As observed in Figure 21, the reliability of the model's prediction of the "high" state was 100%, the reliability of the model's prediction of the "mid" state was 67.27%, and the reliability of predicting the "low" state was 79.15%. The precision matrix would indicate the probability of the precision of the prediction of a state in each cell. Precision is the measure of the overall accuracy which the computational model can predict correctly.

6.5. Limitations of the Study

The exploratory nature of predictive analytics in this study using BN analysis renders the simulated counterfactual results suggestive, rather than conclusive. Furthermore, it is only applicable to the BN model which was generated from the current dataset. Therefore, caution must be exercised when interpreting the potential relationships between the variables (nodes) in the BN model.

The current study only utilized 16 students' data. Due to the small sample size of the pilot study, there is limited generalization to other AI-ALS environments. The proposed methodology will be tested in a future knowledge discovery study with a dataset from a much larger number of students. However, in lieu of using parametric statistical methods with a larger sample size, the Bayesian approach delineated in the current paper could still be used as an alternative approach by educational stakeholders in small-scale pilot studies to independently explore the pedagogical motif of any AI-ALS, in order to understand how it could potentially educe the problem-solving abilities of the students.

As in any study which involves simulations, the results are dependent on the dataset which generated the computational model. The Bayesian network model used in the current study was based on the naïve Bayes algorithm, as it is suitable for exploratory studies that do not assume relationships between nodes to be causal in nature. However, educational stakeholders and researchers should be willing to consider alternative models which could better describe the dataset.

In this segment, the tools in Bayesialab which could be used for the evaluation of the predictive performance of the BN, and the limitations of the study were described. In the next section, the discussions and conclusion are presented.

7. Discussion and Conclusions

The current paper demonstrated a Bayesian approach for educational stakeholders to independently explore the underlying pedagogical motif of an AI-ALS, even if the number of participants might be low, and even if there is no control group, because, in the Bayesian implementation of response surface methodology [32–35], individual parameters could be held constant, whilst others were changed to simulate different hypothetical scenarios.

This Bayesian network approach could be used even if the data could only be collected from a small number of participants who used the AI-ALS, because it does not rely on a frequentist paradigm; it is based on the principles of Shannon's information theory [21], which is a suitable framework for analyzing information gain and mutual information between variables in educational settings, and, in the context of the current paper, for the independent exploration of the pedagogical motif of an AI-ALS.

Specific examples with seven hypothetical scenarios were provided to demonstrate how this Bayesian approach could be used to understand more about the underlying pedagogical motif of the AI-ALS. These hypothetical scenarios with fully controllable parameters could be used to better inform educational stakeholders about the AI-ALS's suitability for broader adoption beyond the pilot study.

Beyond the conventional observation of gains in the cognitive pre-test vis-à-vis post-test, this proposed Bayesian approach also generated seven hypothetical scenarios which could inform educational stakeholders about the conditions that might be most suitable for their students, as well as a few hypothetical scenarios which might be of interest for noncognitive researchers to consider in future studies. The implication for education is that the AI-ALS should not be solely relied upon to improve the students' learning of mathematics; rather, the gaps in the learning of mathematical concepts that the AI-ALS could not bridge for the students should and could be addressed by their mathematics teachers. For example, if the student scored low marks in the AI-ALS, but could surprisingly score high marks in the paper-based post-test, it might be due to the opportunities provided by the AI-ALS to the student to experience vicarious trial and error (VTE); hence, active inference [43] in problem-solving of the same mathematics concepts could be successfully accomplished to solve problems in the paper-based post-test. Conversely, if the student could score high-level marks in the AI-ALS for a particular mathematics topic but could not do so for the paper-based post-test, the teacher should interview the student to find out if anyone in the student's family assisted that student to do the problem-solving in the AI-ALS at home, or whether the student was still unable to accomplish active inference from the concepts taught by the AI-ALS to the paper-based post-test.

Moving forward, it would be possible for educational stakeholders and policy-makers to extend the exemplar illustrated in the current paper to perform comparisons of the pedagogical motifs between multiple AI-ALSs provided by multiple vendors at different school sites.

Supplementary Materials: The supplementary files of the data, "AI-ALS-synthetic-data.csv", as well as the codebook describing the data, "ai-als-data_codebook.txt", are available from https://figshare.com/articles/datacodebook_zip/7795694.

Author Contributions: Conceptualization, M.-L.H.; methodology, M.-L.H.; software, M.-L.H.; validation, M.-L.H., W.L.D.H.; formal analysis, M.-L.H.; investigation, M.-L.H., W.L.D.H.; resources, W.L.D.H.; data curation, M.-L.H.; writing—original draft preparation, M.-L.H.; writing—review and editing, M.-L.H., W.L.D.H.; visualization, M.-L.H.; supervision, W.L.D.H.; project administration, W.L.D.H.; funding acquisition, W.L.D.H.

Funding: This research received funding provided by Education Research Funding Programme (ERFP) via the Office of Education Research, National Institute of Education, Nanyang Technological University, Singapore. [grant number: ERFP OOE (R59704120.706022)].

Acknowledgments: The authors sincerely thank the editors, the staff of the journal, and the three anonymous reviewers for helping improve the manuscript. We thank Eva Moo, Lek-Hong Teo, and Shih-Fen Wah for their support, and Ngan-Hoe Lee from the National Institute of Education Singapore for sharing his valuable insights about mathematics education. We are grateful to the Education Technology Division of the Ministry of Education for their guidance about adaptive learning technology, and to Sin-Mei Cheah from Singapore Management University for proofreading the drafts of the manuscript.

Conflicts of Interest: The authors of this manuscript declare no conflicts of interest.

Ethical Considerations for Protection of the Identities of the Students and the AI-ALS Vendor: The identities of the students and the AI-ALS vendor were anonymized. Furthermore, synthetic data were used in the examples in this paper, such that any good (or bad) performances by the students in the AI-ALS, or in the pre- or post-test are not unfairly attributed to the level of quality of the AI-ALS, or to the students.

References

1. Wilson, C.; Scott, B. Adaptive systems in education: A review and conceptual unification. *Int. J. Inf. Learn. Technol.* **2017**, *34*, 2–19. [CrossRef]
2. Nkambou, R.; Mizoguchi, R.; Bourdeau, J. Introduction: What Are Intelligent Tutoring Systems, and Why This Book? In *Advances in Intelligent Tutoring Systems*; Nkambou, R., Mizoguchi, R., Bourdeau, J., Eds.; Studies in Computational Intelligence; Springer: Berlin/Heidelberg, Germany, 2010; Volume 308, ISBN 978-3-642-14362-5.

3. Garrido, A. AI and Mathematical Education. *Educ. Sci.* **2012**, *2*, 22–32. [CrossRef]
4. VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
5. Cen, H.; Koedinger, K.R.; Junker, B. Is Over Practice Necessary?—Improving learning efficiency with the cognitive tutor through Educational Data Mining. *Front. Artif. Intell. Appl.* **2007**, *158*, 511.
6. VanLehn, K. The behavior of tutoring systems. *Int. J. Artif. Intell. Educ.* **2006**, *16*, 227–265.
7. Hawkins, W.J.; Heffernan, N.T.; Baker, R.S.J.D. Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities. In *Intelligent Tutoring Systems*; Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K., Eds.; Springer International Publishing: Cham, Switzerland, 2014; Volume 8474, pp. 150–155. ISBN 978-3-319-07220-3.
8. Magoulas, G.D.; Papanikolaou, Y.; Grigoriadou, M. Adaptive web-based learning: Accommodating individual differences through system's adaptation. *Br. J. Educ. Technol.* **2003**, *34*, 511–527. [CrossRef]
9. Brusilovsky, P.; Karagiannidis, C.; Sampson, D. Layered evaluation of adaptive learning systems. *Int. J. Contin. Eng. Educ. Lifelong Learn.* **2004**, *14*, 402. [CrossRef]
10. Hox, J.; van de Schoot, R.; Matthijsse, S. How few countries will do? Comparative survey analysis from a Bayesian perspective. *Surv. Res. Methods* **2012**, *6*, 87–93.
11. Bayes, T. A Letter from the Late Reverend Mr. Thomas Bayes, F.R.S. to John Canton, M.A. and F. R. S. *R. Soc. Philos. Trans.* **1763**, *53*, 269–271.
12. van de Schoot, R.; Kaplan, D.; Denissen, J.; Asendorpf, J.B.; Neyer, F.J.; van Aken, M.A.G. A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Dev.* **2014**, *85*, 842–860. [CrossRef] [PubMed]
13. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2010; ISBN 978-0-521-89560-6.
14. Pearl, J. Causes of Effects and Effects of Causes. *Sociol. Methods Res.* **2015**, *44*, 149–164. [CrossRef]
15. Pearl, J. Fusion, propagation, and structuring in belief networks. *Artif. Intell.* **1986**, *29*, 241–288. [CrossRef]
16. Lee, S.-Y.; Song, X.-Y. Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivar. Behav. Res.* **2004**, *39*, 653–686. [CrossRef]
17. Button, K.S.; Ioannidis, J.P.; Mokrysz, C.; Nosek, B.A.; Flint, J.; Robinson, E.S.; Munafao, M.R. Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **2013**, *14*, 365–376. [CrossRef]
18. Kaplan, D.; Depaoli, S. Bayesian structural equation modeling. In *Handbook of Structural Equation Modeling*; Hoyle, R., Ed.; Guilford Press: New York, NY, USA, 2012; pp. 650–673.
19. Walker, L.J.; Gustafson, P.; Frimer, J.A. The application of Bayesian analysis to issues in developmental research. *Int. J. Behav. Dev.* **2007**, *31*, 366–373. [CrossRef]
20. Zhang, Z.; Hamagami, F.; Wang, L.; Grimm, K.J.; Nesselroade, J.R. Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* **2007**, *31*, 374–383. [CrossRef]
21. Kaplan, D. Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-Scale Assess. Educ.* **2016**, *4*, 7. [CrossRef]
22. Levy, R. Advances in Bayesian Modeling in Educational Research. *Educ. Psychol.* **2016**, *51*, 368–380. [CrossRef]
23. Mathys, C. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **2011**, *5*, 39. [CrossRef]
24. Muthén, B.; Asparouhov, T. Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychol. Methods* **2012**, *17*, 313–335. [CrossRef] [PubMed]
25. Bekele, R.; McPherson, M. A Bayesian performance prediction model for mathematics education: A prototypical approach for effective group composition. *Br. J. Educ. Technol.* **2011**, *42*, 395–416. [CrossRef]
26. Millán, E.; Agosta, J.M.; Cruz, J.L. Pérez de la Bayesian student modeling and the problem of parameter specification. *Br. J. Educ. Technol.* **2002**, *32*, 171–181.
27. Shannon, C.E. The lattice theory of information. *IRE Prof. Group Inf. Theory* **1953**, *1*, 105–107. [CrossRef]
28. Cowell, R.G.; Dawid, A.P.; Lauritzen, S.L.; Spieglehalter, D.J. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*; Springer: New York, NY, USA, 1999; ISBN 978-0-387-98767-5.
29. Jensen, F.V. *An Introduction to Bayesian Networks*; Springer: New York, NY, USA, 1999; ISBN 0-387-91502-8.

30. Korb, K.B.; Nicholson, A.E. *Bayesian Artificial Intelligence*; Chapman & Hall/CRC: London, UK, 2010; ISBN 978-1-4398-1591-5.
31. Tsamardinos, I.; Aliferis, C.F.; Statnikov, A. Time and sample efficient discovery of Markov blankets and direct causal relations. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'03, Washington, DC, USA, 24–27 August 2003; p. 673.
32. Guoyi, C.; Hu, S.; Yang, Y.; Chen, T. Response surface methodology with prediction uncertainty: A multi-objective optimisation approach. *Chem. Eng. Res. Des.* **2012**, *90*, 1235–1244.
33. Fox, R.J.; Elgart, D.; Christopher Davis, S. Bayesian credible intervals for response surface optima. *J. Stat. Plan. Inference* **2009**, *139*, 2498–2501. [CrossRef]
34. Miró-Quesada, G.; Del Castillo, E.; Peterson, J.J. A Bayesian approach for multiple response surface optimization in the presence of noise variables. *J. Appl. Stat.* **2004**, *31*, 251–270. [CrossRef]
35. Myers, R.H.; Montgomery, D.C.; Anderson-Cook, C.M. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed.; Wiley and Sons, Inc.: New Jersey, NJ, USA, 2009; ISBN 978-0-470-17446-3.
36. Collins, J.A.; Greer, J.E.; Huang, S.H. *Adaptive Assessment Using Granularity Hierarchies and Bayesian Nets*; Springer: Berlin/Heidelberg, Germany, 1996; Volume 1086, pp. 569–577.
37. Conati, C.; Gertner, A.; VanLehn, K.; Druzdzel, M. On-line student modelling for coached problem solving using Bayesian networks. In Proceedings of the UM'97, Sardinia, Italy, 2–5 June 1997; Springer: Berlin/Heidelberg, Germany, 1997; pp. 231–242.
38. Jameson, A. Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling User-Adapt. Interact.* **1996**, *5*, 193–251. [CrossRef]
39. VanLehn, K.; Niu, Z.; Siler, S.; Gertner, A.S. *Student Modeling from Conventional Test Data: A Bayesian Approach without Priors*; Springer: Berlin/Heidelberg, Germany, 1998; Volume 1452, pp. 434–443.
40. Conrady, S.; Jouffe, L. *Bayesian Networks & BayesiaLab: A Practical Introduction for Researchers*; Bayesia: Franklin, TN, USA, 2015; ISBN 0-9965333-0-3.
41. Lauritzen, S.L.; Spiegelhalter, D.J. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc.* **1988**, *50*, 157–224. [CrossRef]
42. Kschischang, F.; Frey, B.; Loeliger, H. Factor graphs and the sum product algorithm. *IEEE Trans. Inf. Theory* **2001**, *47*, 498–519. [CrossRef]
43. Pezzulo, G.; Cartoni, E.; Rigoli, F.; Pio-Lopez, L.; Friston, K. Active inference, epistemic value, and vicarious trial and error. *Learn. Mem.* **2016**, *23*, 322–338. [CrossRef] [PubMed]
44. Al-Mutawah, M.A.; Fateel, M.J. Students' Achievement in Math and Science: How Grit and Attitudes Influence? *Int. Educ. Stud.* **2018**, *11*, 97. [CrossRef]
45. Chamberlin, S.A.; Moore, A.D.; Parks, K. Using confirmatory factor analysis to validate the Chamberlin affective instrument for mathematical problem solving with academically advanced students. *Br. J. Educ. Psychol.* **2017**, *87*, 422–437. [CrossRef] [PubMed]
46. Egalite, A.J.; Mills, J.N.; Greene, J.P. The softer side of learning: Measuring students' non-cognitive skills. *Improv. Sch.* **2016**, *19*, 27–40. [CrossRef]
47. Lipnevich, A.A.; MacCann, C.; Roberts, R.D. Assessing Non-Cognitive Constructs in Education: A Review of Traditional and Innovative Approaches. In *The Oxford Handbook of Child Psychological Assessment*; Oxford University Press: New York, NY, USA, 2013.
48. Mantzicopoulos, P.; Patrick, H.; Strati, A.; Watson, J.S. Predicting Kindergarteners' Achievement and Motivation from Observational Measures of Teaching Effectiveness. *J. Exp. Educ.* **2018**, *86*, 214–232. [CrossRef]