Journal of Education and Development; Vol. 6, No. 4; September, 2022
ISSN 2529-7996 E-ISSN 2591-7250
Published by July Press

Diagnostically Cognitive Assessment of Complex Semantic Explanations and Learning: A Bayesian Network Approach

Zhidong Zhang[1]

[1] The College of Education and P-16 Integration, The University of Texas-Rio Grande Valley, Texas, USA

Correspondence: Zhidong Zhang, The College of Education and P-16 Integration, The University of Texas-Rio Grande Valley, Texas, USA. E-mail: zhidong.zhang@utrgv.edu

Received: July 11, 2022 Accepted: July 30, 2022 Online Published: August 6, 2022

doi:10.20849/jed.v6i4.1255 URL: https://doi.org/10.20849/jed.v6i4.1255

Abstract

This study explored a diagnostically cognitive assessment model for the ANOVA score model emphasizing semantic explanations. The study used the mixed methods designs, in which the ANOVA score model was decomposed into measurable components. This consists of the proficiency student model. Such kinds of data were transferred to a quantitative representation via the Bayesian network model of the ANOVA score model and semantic explanation assessment. This diagnostically cognitive assessment consists of 28 variables hierarchically, which are explanatory variables and evidence variables. Nine variables are explanatory variables that are latent. Nineteen variables are evidence variables that collect students' learning information and propagate the information to the explanatory variables. The data are simulated data; the semantic explanations from twelve students were recorded and input into the nineteen evidence variables. Semantic explanations indicate 3 levels: lower level, medium level and high level. The score should be more than 82 points, which indicates a mastery level. The study also suggests that if a student achieves a high score in a module, the student has a better chance of achieving a high score in the overall assessment model.

Keywords: ANOVA score model, semantic explanations, Bayesian network model, diagnostically cognitive assessment, proficiency model, and mixed methods design

1. Introduction

1.1 Introduction to the Focus

The objective of obtaining an informative and practical cognitive assessment of students' developing knowledge and problem-solving proficiency in complex educational domains has been recognized to be an attainable yet still elusive goal (Pellegrino, 2014; Pellegrino, Glaser, Chudowsky, 2001). Research on the cognitive analysis and modeling of semantic knowledge and methods novices and experts use to understand and solve problems in complex domains has advanced to the point where it is possible to specify many of the components of knowledge and problem-solving skills that need to be assessed in such domains.

1.2 Conceptual Frameworks for Assessment

Conceptual frameworks for assessment such as Mislevy's evidence-centered design (ECD) framework provide a systematic conceptual basis for designing assessment to support an inference about the cognitive components of an individual's knowledge and proficiency in solving problems and performing other demanding tasks on the basis of evidence obtained from students' performance of such tasks (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; Mislevy, Almond, & Lukas, 2004; Mislevy, Steinberg, & Almond, 2000; Pretz, et al., 2016).

Within the ECD framework, statistical evidence models have been developed and applied to enable inferences from evidence variables (based on task performance scores) to explanatory assessment variables, which are mapped to components of a cognitive proficiency model. Evidence models may consist of a statistical item response model for measuring proficiency or component skills in a domain (particular for evidence variables based on item-based tasks), or they may involve Bayesian probability networks (Jackman, 2009) for inferring knowledge and skill components (particularly for evidence variables based on scored observations of task performance).

Despite these advances, challenges remain in designing practical cognitive assessments that can provide a valid and useful diagnostic assessment of students' knowledge. Similarly, this includes problem-solving proficiency, and learning in real-life situations of education and practice. First, an assessment should be able to assess students' overall proficiency in complex educational domains as well as severe proficiency in specific areas of knowledge and problem-solving ability. Such assessment may be implemented by means of statistical evidence models that consist of a single or multidimensional item response model, or they may be implemented by means of Bayesian Network (BN) models consisting of Bayesian probability networks that include higher-order explanatory nodes thus enabling abductive inferences from evidence variables to states of these higher-order explanatory nodes (Culbertson, 2016; Koller, & Friedman, 2009; Koski, Nobel, 2009).

2. Cognitive Assessment, Performance Assessment and Diagnostic Assessment

Cognitive assessment should be diagnostic in the sense that, based on a student's performance of representative tasks in a domain (Javidanmehr, & Sarab, 2017). It can provide valid information about a student's mastery of a specific component of declarative and procedural knowledge in the domain, and about the student's ability to apply such knowledge to reason, solve problems, select among alternative methods or strategies, or perform the actions required to successfully complete task in the domain (Ayala, Ayala, & Shavelson, 2000). Such assessments usually have been implemented using Bayesian networks containing nodes that correspond to a particular component of skills or knowledge (Mislevy, 1995) or by item response measurement models applied to scores on items that have been designed to assess particular skills or a subset of skills (Tatsuoka, 1983, 1995; Gierl, Leighton, & Hunka, 2000)

To assess students' knowledge and skill in solving problems in complex domains of expertise, the assessment should be based on the performance of authentic tasks. These should be representative of those which typically occur in situations of expert performance education, and training in the domain of competency or expertise being assessed (Ignizio, 1991). Assessments using task models that are dependent on item-based task formats. These formats are relatively limited in their ability to include authentic and extended task formats are relatively limited in their ability to include authentic and extended tasks such as occur in expert domains of learning and performance (Evans, 2019).

2.1 Cognitive Assessment and Performance Assessment

Existing technologies can facilitate the collection, recording and analysis of extensive performance data in natural situations and domains of performance. Such data are likely to arise in complex knowledge-intensive domains such as a founded in medic medicine, engineering or statistics. However, performance assessments in such situations require evidence rules for coding and scoring components of complex performance, either in real-time or based on records or products of problem-solving(Evans, 2019; Firestone, Mayrowetz, & Fairman (1998).

Methods based on semantic and task analysis can be used to develop cognitive models of semantic and procedural knowledge (respectively) in complex domains. These models can be used to develop techniques for coding analysis and scoring of such records or procedures of performance (Hollnagel, 2003). In principle, both item response and Bayesian evidence models (Zhang, 2022) can be applied to evidence variables based on such coding of a performance. Such models can support inferences about overall proficiency, and diagnostic inferences about specific knowledge, reasoning, and procedures that are involved in and underlie cognitive expertise in a domain (Alexander, 2003)

2.2 Cognitive Assessment and Dynamic Assessment

Cognitive assessment of learning also must be dynamic allowing one to trace a student of progress in developing components of knowledge and problem-solving competence over time. Dynamic assessment can be accomplished by successively updating assessments of a student's knowledge and competency in a domain based on changes in the student's performance over a series of problem-solving tasks which occur over the extended period of learning and development of expertise (Lajoie, Lesgold, 1992).

If item response models are used as evidence models to assess the changes over learning, assessments can be made based on student's performance of the tasks at different times. Learning can be assessed through analysis for changes in these assessment measures over time. If Bayesian networks are used to support inferences about changes in the component of an individual's knowledge and problem-solving skill, the Bayesian probability network can be updated regularly using evidence from each new performance to access changes in estimates in the posterior probabilities that are associated with changes in a student's mastery of components of knowledge and competency (Zhang, 2016; Zhang, 2018; Zhang & Zhang, 2020).

2.3 Cognitive Assessment and Learning Contexts

Finally, it should be possible to obtain diagnostic assessments within natural situations of teaching and learning. Assessment should be appropriate for use in learning contexts such as classrooms or other educational contexts in which knowledge and proficiency are developed. For example, diagnostic assessment could be based on practice tasks that occur within a classroom or technology-enhanced learning environments to chase students' development of knowledge and problem-solving competency, and they could include the use of self-assessment to facilitate students' learning and development of skills in critically evaluating their own problem-solving performance (Bao, Redish, 2022).

3. Study Objectives

In this paper, we report the results of the investigation of a BN approach to diagnostically cognitive assessment that we believe can contribute to meeting there are challenges (Zhang, 2022). We applied Mislevy's evidence-centered assessment design (ECD) framework to develop a diagnostic cognitive assessment model in which Bayesian probability network (BPN models) were developed and used to make inferences about the components of students' knowledge and proficiency in a domain for problem-solving in intermediate statistics (Mislevy, Steinberg, Almond, Breyer, & Johnson 2001; Zhang, 2007; Zieky, 2014).

The problem-solving task used in the assessment involved applying a statistical model to analyze an educational research data set using analysis of variance (ANOVA). Working within this particular domain, our objective was to evaluate the assessment of knowledge and proficiency in solving these problems that were obtained by applying the BN assessment model to scored performance data of the students in our sample. We also investigated how the model could be updated to assess student development of knowledge and competency, and how assessments were affected by specific changes in response patterns (scores on evidence variables) (Mislevy, Steinberg, & Almond, 2000). We were particularly interested in practical assessment models that potentially could be embedded in instructional environments (Frederiksen, & Donin, 1999). Thus, the assessment models we studied were designed to that they could be embedded within a previously developed computer-based coaching and a practice environment that was designed to support students' learning of ANOVA.

A second objective was to use the BN evidence models to obtain an assessment of students' overall proficiency in the domain (Baghaei, 2012). We expected the BN models to be particularly appropriate for diagnostic cognitive assessment in complex performance or domains.

4. The Construction of a BN Assessment Model

4.1 Proficiency Model and Task Model

In the present paper, the proficiency model (cognitive model) (Xue, & von Davier, 2014) in assessment design consists of a hierarchical network of nodes corresponding to components of declarative knowledge (semantic) and procedural knowledge in a particular domain of a statistics (analysis of variance: ANOVA). The task model consists of tasks that require students to respond to questions that require the production of solutions to subtasks involved in solving a statistical data analysis problem using ANOVA and explanations of solutions to subtasks.

4.2 BN Evidence Model

The evidence model consists of a hierarchical Bayesian Network (BN) in which explanatory nodes correspond to components of semantic and procedural knowledge models, and evidence variable nodes correspond to scores on observed responses to questions associated with each subtask (Conrady, & Jouffe, 2022). The approach taken to diagnostic assessment was to update the BN network based on students' responses to obtain estimates of students' progress towards mastery (i.e., their likelihood of mastery) of components of knowledge and skill at different levels in the hierarchy. Overall mastery is reflected in the top-level explanatory mode in the network (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; Zhang, & Zhang, 2020).

5. Research Method

5.1 Establishment of an Assessment Structure and Model

An assessment structure is a framework for representing an arrangement of knowledge components in a hierarchical network. The network structure will define links among potential explanatory variables (constructs), which will be used to evaluate and interpret student mastery of these learning objects (constructs) and changes in student mastery over the course of learning. These knowledge and skill components cannot be observed directly and it may be necessary for decomposed them into more fine-grained components for given assessment purposes. They must also be linked to evidence variables derived from observations of student performance.

Normally, assessment structure can be established through a semantic analysis of the content of verbal problem-solving or tutoring protocols, combined with the cognitive task analysis of the problem-solving. Assessment purposes and the desired "grain" of analysis influence the precision of the analysis carried out to build procedural and semantic models of required problem-solving knowledge. For the ANOVA score model, (a) writing the ANOVA score model component, and (b) explaining what these components refer to semantically, constitute two different aspects of knowledge and skill. Hence, the assessment structure consists of two submodels: (1) writing an ANOVA score model (a procedural model) and (2) explaining the ANOVA score model (a semantic model). Thus, an ANOVA score model can be decomposed into several component procedures which become the basis for assessment model development:

$$Yi(jk) = \mu + \alpha j + \beta k + \gamma jk + ei(jk)$$

$Y_{i(jk)}$ is the score of individual i in group jk. To the right of the equal sign, there are five components: Grand mean μ, the main effect for factor α_j, the main effect for factor β_k, the interaction effect γ_{jk}, and residual score $e_{i(jk)}$.

5.2 Hierarchical ANOVA Score Model Built-in BN

Figure 1 presents a hierarchical frame representing the knowledge required to complete an ANOVA score model (for a two-way classification) and will be referred to as the "ANOVA Score Model (2 way)" Frame. Learner tasks have to be decomposed into fine-grained cognitive components.

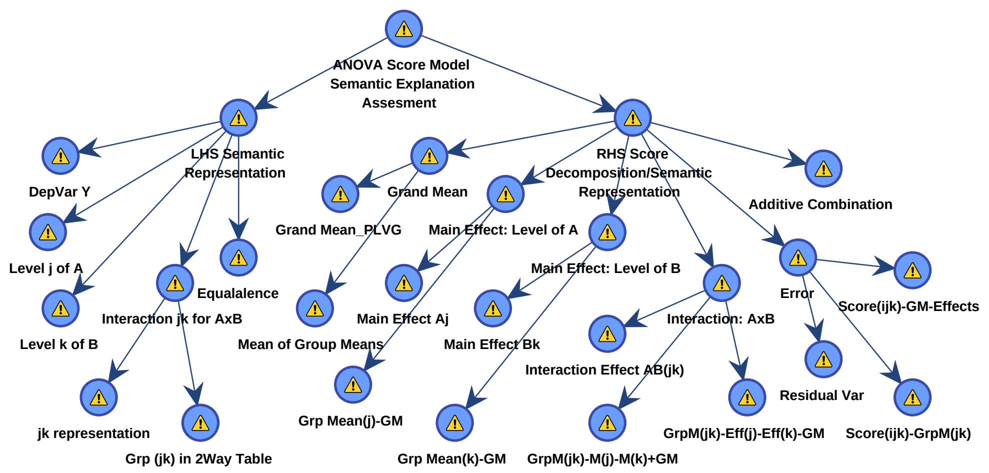

Figure 1. ANOVA score model and semantic explanations with BN representation

ANOVA score model consists of 28 nodes: 9 explanatory variables and 19 evidence variables. The explanatory variable cannot collect any learning evidence/data in this assessment; the evidence variable can be used to collect students' progress evidence. The Full explanations of these variables are referred to in Appendix A.

5.3 Initializing the BN Model and Joint Probability

The author has collected limited evidence of each component in the assessment model. It is normal that any Bayesian network needs to be initialized with data that can be artificial. However, doing so does not bother the subsequent process. Therefore the value of the top component, ANOVA Score Model Semantic Explanation Assessment, is set as .7 as a successful semantic explanation.

A joint probability is the probability of two events occurring simultaneously. If they were events A and B, the probability of the interaction of events A and B may be written p(A ∩ B). For example, if we focus on the joint probability of the interaction of these variables in LHS Semantic Representation; there are eight variables are included. Figure 2 presents the joint probability of these variables. As shown in Figure 2, there are 6 evidence variables, which receive semantic explanation evidence: Dep-Var Y, Level j of A, Level k of B, jk Representation, Grp(jk) in 2Way Table, and Equivalence. After these evidence variables received the data the explanatory variable LHS Semantic Representations has been updated to 98.63%. This is from a part of the network, LHS Semantics Representation. Stated differently, LHS Semantic Representation receives 98.63/100 points if the students correctly respond to the 6 evidence variables above.

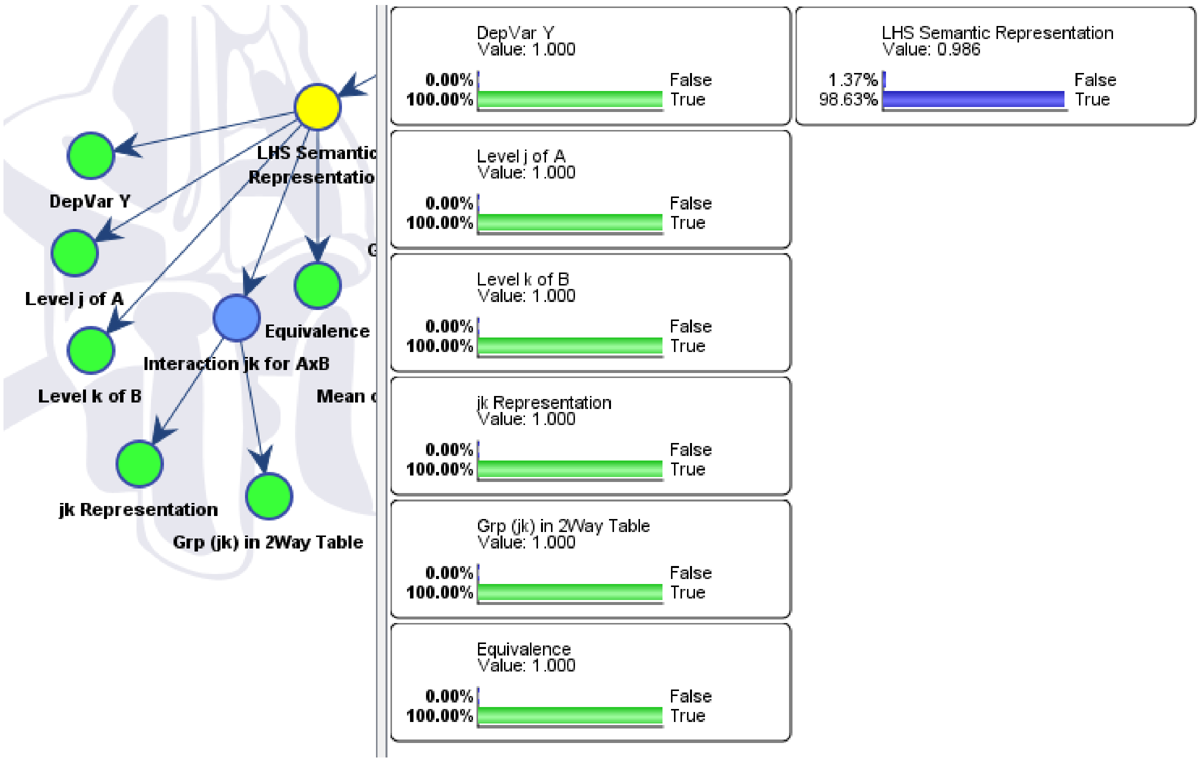

Figure 2. The initialized Bayesian network model for LHS Semantic Representation

5.4 Bayesian Network Theory and Models

Bayesian networks (BNs) are known as belief networks, which are represented in a directed acyclic graph to model an assessment model (Zhang, 2022). Kwan, Chow, Law and Lai (2008) describe the Bayesian network:

> The Bayesian network uses probability theory and graph theory to construct probabilistic inference and reasoning models. It is defined as a directed acyclic graph with nodes and arcs. Nodes represent variables, events or evidence. An arc between two nodes represents a conditional dependency between the nodes. Arcs are unidirectional and feedback loops are not permitted. Because of this feature, it is easy to identify the parent-child relationship or the probability dependency between two nodes (pp. 275-289).

Koller and Friedman (2009) state that, assuming there is a class of variables that can be designated by x_1, x_2, … x_n and C. The structure can be seen in figure 3. In such an example variables x_1, x_2, … x_n are observed. The upper-level variable C means a class. It represents a concept, which is supported by all of these observed/evidence variable x_s. All of these variable x_s are conditionally independent on the variable Class. Thus, the relationship of all of these variables can be described in a model which factorizes as:

$$\mathrm{P}(c, x_1, x_2 \cdots x_n) = P(c)\, \Pi^{n}_{i=1} P(X_i \mid c)$$

This model represents the joint distribution by using a group of variables, which include a prior distribution P(C) and a set of P (Xi)| C). The model indicates that a concept, which is usually an explanatory variable, can be supported by a group observed variable x_s.

5.5 ANOVA Score Model Semantic Explanation Assessment With Subjective Probabilities

The values of variables, regardless of the explanatory variable or the evidence variables, are also set at .7 as successful semantic explanation. The assumption is that there is no evidence to indicate that many students score higher than in the Semantic Explanation of the ANOVA score model, so the value of .7. It is believed that students can master each component represented in both explanatory and evidence variables at above 70% of chance after they practice the ANOVA model semantic explanation with several examples. As shown in Figure 3, all variables also called nodes in BN, are initialized with initial values.

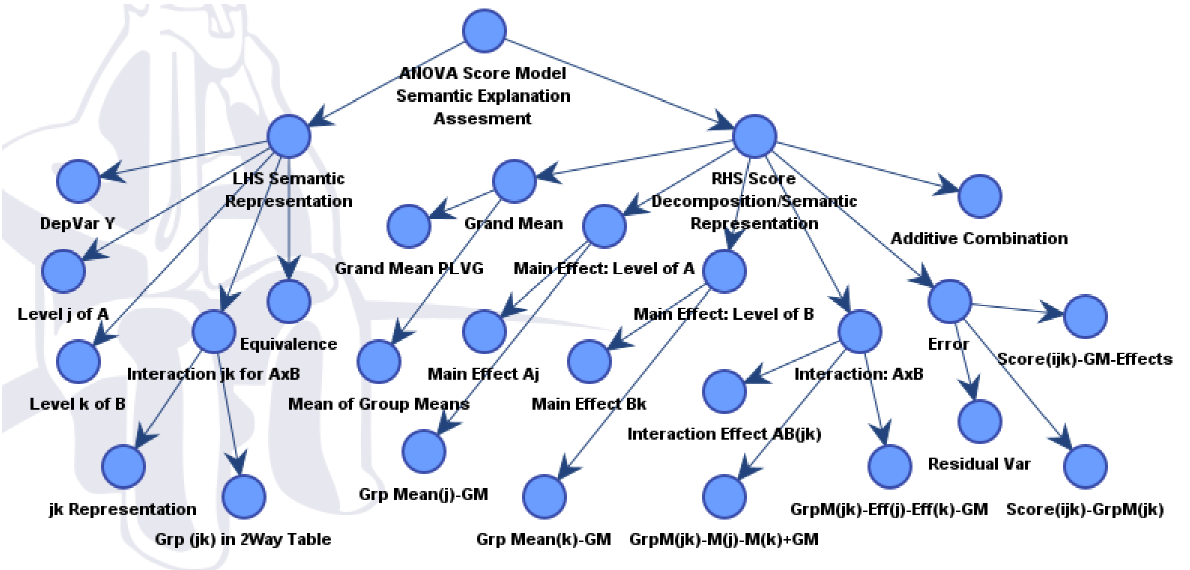

Figure 3. The Bayesian network assessment for the ANOVA Semantic Explanation

As shown in Figure 3, all of these nodes including explanatory variables and evidence variables received the initialized values. As any variable is selected and double-clicked, the joint probability can be read. The RHS Score Decomposition Semantic Representation was selected, and the read value is 53.20% of the successful semantic explanation. We noted that this is the "basic status," which means there is not any data as evidence to be input.

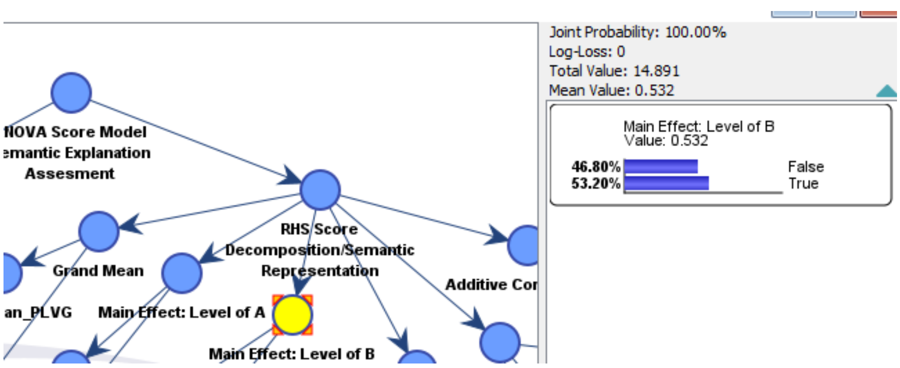

Figure 4. The joint probability of the RHS score decomposition semantic representation

5.6 Inputting Data into the Evidence Variables

The data were simulated from a group of students who studied applied statistics. It was assumed that there were 12 students. We know that there are 19 evidence variables. There are 6 evidence variables in the Left Hand-Side Semantic Representation; there are 13 evidence variables in the Right Hand-Side Score Decomposition Semantic Representation.

Table 1. Students' Semantic Explanations of the ANOVA Score Model

Student	ANOVA Score Model Semantic Explanation Assessment	LHS Semantic Representation	RHS Score Decomposition/ Semantic Representation	Total Evidence Number	The Number of LHS Evidence Variable	The Number of RHS Evidence Variable	Assessment by Range
S1	53.89	39.72	25.50	8	2	6	L
S2	90.78	94.31	97.39	17	5	12	H
S3	92.02	98.11	98.93	18	5	13	H
S4	87.11	89.78	87.11	13	5	8	H
S5	82.07	97.52	56.74	12	5	7	H
S6	82.16	88.65	68.62	11	4	7	H
S7	90.54	98.02	92.64	14	5	9	H
S8	92.21	98.92	98.93	19	6	13	H
S9	92.09	98.91	98.43	18	6	12	H
S10	81.43	95.72	56.49	11	4	7	H
S11	69.44	96.78	3.17	5	5	0	M
S12	48.14	37.58	9.93	5	2	3	L

6. Results and Findings

This study developed the ANOVA Score Model for semantic explanations. The two-way ANOVA model was decomposed into fine-grained terms, which were given semantic explanations. Thus, the ANOVA Score Model is both a learning and assessment platform. The ANOVA Score Model provides students and instructors a scaffolding tool to acquire knowledge, develop problem-solving skills, and receive diagnostically cognitive assessment information. The components of evidence-centered assessment design (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; Mislevy, Almond, & Lukas, 2004) provides assessors with important assembling components for variations of diagnostically cognitive assessment. The cognitive task is the ANOVA Score Model: $Yi(jk) = \mu + \alpha j + \beta k + \gamma jk + ei(jk)$. With the consideration of semantic explanations, the terms of the ANOVA Score Model were decomposed into fine-graded components, which consisted of a proficiency student model (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015), and then was represented in Bayesian Network Model as shown in Figure 3.

6.1 Description of the Bayesian Network Assessment Model for the ANOVA Semantic Explanation

There were two types of variables in the Bayesian network: Explanatory variable and evidence variable. The explanatory variable cannot be directly observed. In other words, there was at least one evidence variable as a child node attached to the explanatory variable in the network. The evidence variable can be directly observed. The data was input into the variable node. As shown in Figure 2, in the Left-hand Side Semantic Representation branch, there are 6 evidence variables with data inputs: Dep Var Y, Level j of A, Level k of B, jk Representation, Grp(jk) in 2Way Table, and Equivalence.

There were 9 explanatory variables and 19 evidence variables in the Bayesian network model. There were four layers of the Bayesian network model. The author took one strand as an example which consisted of a four-level trajectory hierarchically: ANOVA Score Model Semantic Explanation Assessment (1st level), LHS Semantic Representation (2nd level), Interaction k for AxB (3rd level), and jk Representation (4th level). The author only focused on ANOVA Score Model Semantic Explanation Assessment, LHS Semantic Representation and RHS Score Decomposition/Semantic Representation these three explanatory variables, which were latent variables to describe and represent the levels of students' semantic explanations

6.2 Data Input and Students' Semantic Explanation Score

As shown in Table 1, there were 12 students who practiced the ANOVA Score Model for the semantic explanation of all the components of the score model. Their semantic explanations were diverse and indicated that the range of the students received scores on the top explanatory variables was between 48.14 and 92.21. The author wanted to classify the levels of students' semantic explanations into 3 levels: low level, mediate level and high level. The range of students' scores was the difference of 92.21 and 48.14, which is 44.07. The range was divided into three equal parts, 44.07/3=14.69. Thus, the low interval of the student semantic explanation was between 48.14 and 62.83 (48.14+14.69), the medium level was between 62.84 and 77.52, and the high level was between 77.53 and 92.21. The right column of Table one indicated that there are 2 students at the low level of semantic explanations; only one student was at the medium level, and 9 students were at the high level. The model-based assessment was different from the non-model-based assessment. As shown in Table 1, Student 11 and student 12 both scored 5/19, but their score scatters were different (refer to Appendix B). Student 11's 5 scores were clustered on the left—LHS Semantic Representation, which meant this student understands the Left-hand side semantic explanation very well, and almost knew nothing about the RHS Score Decomposition Semantic Representation. This student scored 69.47 at a medium level. Student 12's score was also 5, but it scattered diversely. This student scored 48.14, which was a low level. The fact informs us that clustered score in a model indicates a higher mastery level.

7. Discussion and Conclusion

This study developed the model of a diagnostically cognitive assessment represented by using a Bayesian network model. This assessment model described the rationale and steps to illustrate how initiatively to build an alternative assessment model focusing on both learning processes and outcomes in a complex learning domain. Meantime, the model provided students with diagnostic information about their learning. The students themselves could know what aspects of the content knowledge should be further improved.

7.1 Design of the Diagnostically Cognitive Assessment

The assessment took the evidence-centered assessment design (ECD) as a main framework (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015). Since Schum (1994) published the paper entitled "*The Evidence Foundations of Probabilistic Reasoning*," evidence-centered assessment design received increasing development. Pearl (2009) explored the model and inference; Conrady and Jouffe (2022) mapped reasoning, graphical models and quantitative representation by developing the Bayesian network graph model. All these academic work aids researchers in alternative assessment to develop variations of the diagnostically cognitive assessment models.

In fact, this was a mixed methods design. The qualitative research dimension was about the ANOVA model and term decomposition into fine-grained semantic knowledge, which consisted of a task model. When the tasks were delivered to a student the student proficiency model can be developed. The proficiencies with cognitive tasks were represented in a Bayesian network model. From the research method and design perspective, a Bayesian network model was the best platform to fusion the qualitative phase and quantitative phase into one unit. Through a Bayesian network structure, the qualitative data can be transferred to a quantitative representation.

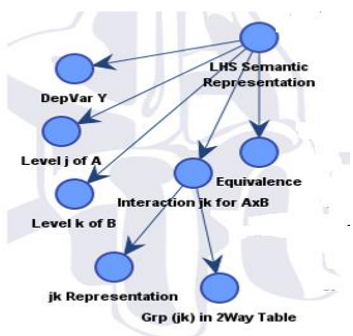

Figure 5. The Left Part of the ANOVA Score Model Semantic Explanation

7.2 Hierarchical Components and Propagation From the Evidence to Explanatory Variables

ANOVA Score Model Semantic Explanation model described a cognitive assessment process. The model was divided into two parts, based on the statistical score model: $Yi(jk) = \mu + \alpha j + \beta k + \gamma jk + ei(jk)$, and then obtained the graphical model as shown in Figure 5. There were two explanatory variables on the left part, which were LHS Semantic Representation and Interaction jk for AXB. These two variables were latent variables, which received students' semantic explanation information. There were six evidence variables in the left part of this model. The students responded to the evidence variable, and the information was propagated to these explanatory variables.

As shown in Figure 6, there are six explanatory variables on the right part of the ANOVA Score Model Semantic Explanation model. These explanatory variables are RHS Score Decomposition Semantic Representation, Grand Mean PLVG, Main Effect: Level of A, Main Effect: Level of B, Interaction: AXB, Error and Additive Combination. The other thirteen variables are evidence variables. When the students responded to the evidence variables, the information was propagated to these explanatory variables.

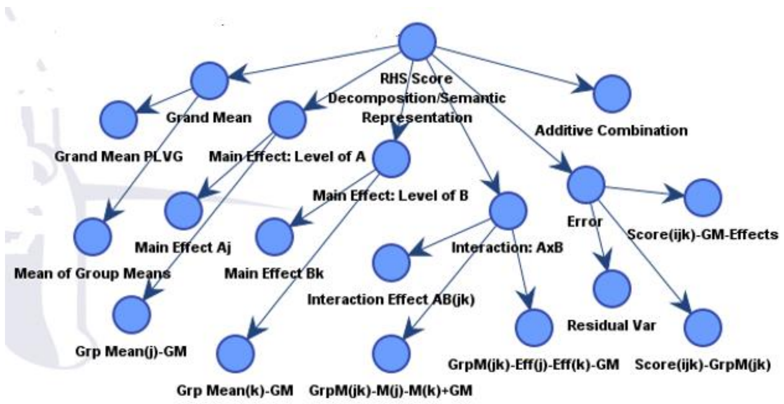

Figure 6. The Right Part of the ANOVA Score Model Semantic Explanation

The cognitive assessment process described students' semantic explanation through the ANOVA Score Model Bayesian network. The author emphasized the top three explanatory variables: ANOVA Score Model Semantic Explanation Assessment, LHS Semantic Representation, and RHS Score Decomposition Semantic Representation.

This assessment model provided moment-by-moment differential information for both students and cognitive feature categories. The cognitive Bayesian network will be more robust in differentiating different student groups and cognitive feature categories with updating evidence by the Bayesian network learning.

There were 3 assessment patterns based on the students' semantic explanations. In the L group, students' semantic explanations were at a lower level. These students were S1 and S2, and their top-level scores were 53.89 and 48.14 separately. There was only one student in the M group and the semantic explanations were at a medium level. The student was S11, and the top-level score was 69.44. There were nine students in the H group and their semantic explanations were at a high level. The range of the top-level scores was from 81.43 to 92.21. Stated differently, a student can do semantic explanations at a high level in this learning model if the student obtained 82 points or above 82 points.

Briefly, The ANOVA Score Model Semantic Explanation was cognitive assessment model because the model embodied a cognitive progress. This model was also a diagnostic assessment model because it reported students' progress and mistakes in the learning process. This model was also a dynamic assessment model because the students and assessors can observe the Semantic Explanation score.

In addition, if a student achieved a high score in a module, the student had a better chance of achieving a high score in the overall assessment model. This was related to mastery learning theory, which can be further discussed.

8. Limitations

This study used data from 12 students. The findings and analyses have limited generalizations. The proficiency student model can be varied based on different expertise. The initialized values of the variables are subjective. The Bayesian network model becomes more robust with the increase of the sample size. Thus, the mastery level of the semantic explanation is relative.

References

Alexander, P. A. (2002). The development of expertise: The journey from acclimation to proficiency. *Educational Research*, *32*(8), 10-14.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessments.* Springer.

Ayala, C. C., Ayala, M. A., & Shavelson, R. (2000, April). *New dimensions for performance assessments.* Paper presented at AERA, New Orleans, LA.

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology*, *10*(1), 233-252.

Bao, L., & Redish, E. F. (n. d.). *Educational assessment and underlying models of cognition*. Retrieved July 2, 2022, from https://www.asc.ohio-state.edu/bao.15/ modelanalysis/ workshop/ Papers/Bao_Redish_3.pdf

Conrady, S., & Jouffe, L. (2022). *Bayesian networks and BayesiaLab: A practical introduction for researchers.* Retrieved from https://library.bayesia.com/articles/#!bayesialab-knowledge-hub/book

Culbertson, M. J. (2016). Bayesian networks in educational assessment: The state of the field. *Applied Psychological Measurement*, *40*(1), 3-21. https://doi.org/10.1177/0146621615590401

Evans, C. M. (2019). Effects of New Hampshire's innovative assessment and accountability system on student achievement outcomes after three years. *Education Policy Analysis Archives*, *27*(10), 1-36.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, *20*(2), 95-113.

Frederiksen, C., & Donin J. (1999). Cognitive assessment in coached learning environments. *Alberta Journal of Educational Research*, *45*(4), 392-408.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2000). An NCME instructional model on exploring the lopgic of Tatsuoka's rule-space model for test development and analysis. *Education Measurement: Issues and Practices*, *19*(3), 34-44.

Hollnagel, E. (2003). Prolegomenon to cognitive task design. In E. Hollnagel (Ed.), *Handbook of cognitive task design* (pp. 3-15). Lawrence Erlbaum. https://doi.org/10.1201/9781410607775

Ignizio, J. P. (1991). *Introduction to expert systems: The development and implementation of rule-based expert system.* McGraw-Hill. Inc.

Jackman, S. (2009). *Bayesian analysis for the social science.* Wiley.

Javidanmehr, Z., & Sarab, M., R., A. (2017). Cognitive diagnostic assessment: Issues and considerations. *International Journal of Language Testing*, *7*(2), 73-98.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: The MIT Press.

Koski, T., & Nobel, J. M. (2009). *Bayesian networks: An introduction*. Wiley.

Kwan, M., Chow, K-P., Law, F., & Lai, P. (2008). Reasoning about evidence using Bayesian network. In D. Ray, & S. Shenoi (Eds.), *Advances in digital forensics IV, IFIP international federation for information processing* (Vol. 285, pp. 275-289). Springer.

Lajoie, S. P., & Lesgold, A. M. (1992). Dynamic assessment of proficiency for solving procedural knowledge tasks. *Educational Psychologist*, *27*(3), 365-384.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Lawrence Erlbaum.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Rep. 632). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2000). *Evidence-centered assessment design (A submission for the NCME award for technical or scientific contribution to the field of educational measurement)*. Unpublished manuscript, Educational Testing Service in Princeton.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Breyer, F. J., & Johnson, L. (2001, March). *Making sense of data from complex assessment* (CSE Tech Rep. 538). Educational Testing Service.

Pearl, R. J. (2009). *Causality: models, reasoning, and inference.* Cambridge University Press.

Pellegrino, J. W. (2014). A learning sciences perspective on the design and use of assessment in education. In K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 233-252). Cambridge University Press.

Pellegrino, J. W., Glaser, R., & Chudowsky, N. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Pretz, D. R., Kean, J., Heinemann, A. W., Kozlowski, A. J., Bode, R. K., & Gebhardt, E. (2016). A multidimensional Rasch analysis of the functional independence measure based on the National Institute on Disability, independent living, and rehabilitation research traumatic brain injury model systems national database. *Journal of Neurotrauma, 33*(14), 1358-1362.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Wiley.

Tatsuoka, K. K. (1893). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Lawrence Erlbaum.

Xue, X., & von Davier, M. (2014, August). *Cognitive diagnosis for NAEP proficiency data.* https://doi.org/10.1002/j.2333-8504.2006.tb02014.x

Zhang, Z. (2007). Cognitive assessment in a computer-based coaching environment in higher education: Diagnostic assessment of development of knowledge and problem-solving skill in statistics. *Dissertation Abstract International: Section A. Humanities and Social Sciences, 68*(10-A), 4271.

Zhang, Z. (2016). Assessment of matrix multiplication learning with a rule based analytical model-A Bayesian network representation. *International Education Studies, 9*(12), 182-193.

Zhang, Z**.** (2018). Designing cognitively diagnostic assessment for algebra content knowledge and thinking skills. *International Education Studies, 11*(2), 106-117.

Zhang, Z. (2022). A mixed methods design for assessing physics learning in the online learning environment. *Journal of Education and Development*, *6*(2), 28-36.

Zhang, Z., & Zhang, Y**.** (2020). Alternatively assessing conceptual learning in an emergency clinical environment: A mixed methods design. *Journal of Education and Development*, *4*(2), 27-37

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*, 79-87.

Appendix A. Full Descriptions of Evidence and Explanatory Variables of the ANOVA Semantic Explanations

Variable Name in the Bayesian Network	Full Variable Name	Definition and Description of the Variable	Type of the Variable
1. ANOVA Score Model Semantic Explanation Assessment	ANOVA Score Model Semantic Explanation Assessment	This is the top node/variable, representing the mastery level of the student learning in Semantic Explanation	Explanatory variable
2. LHS Semantic Representation	Left hand side of the ANOVA score model about the Semantic representation	This is the second level explanatory variable in the ANOVA score model, which represents the left-hand side terms	Explanatory variable
3. RHS Semantic Representation	Right hand side of the ANOVA score model about the Semantic representation	This is the second level explanatory variable in the ANOVA score model, which represents the right-hand side terms	Explanatory variable
4. Dep Var Y	Dependent variable Y	This is the dependent variable	Evidence

		in the ANOVA score model on the left-hand side of the equal sign	variable
5. Level j of A	Levels of group levels of A	j refer to level j of the independent variable A	Evidence variable
6. Level k of B	Level of group levels of B	k refer to level k of the independent variable B	Evidence variable
7. Interaction jk for AxB	Interaction jk for dependent variables A and B	jk are variable A levels across variable B levels	Explanatory variable
8. jk representation	jk presentation/meaning	jk refers to the cross classification cell of the table of subjects.	Evidence variable
9. Grp (jk) in 2Way Table	Grp (jk) in 2Way Table	jk refers to the cross-classification cell of the table of subjects where each officer is classified into a category for each combination of Factor A j and Factor B k.	Evidence variable
10. Equivalence	Equivalence of score and score components	Equivalence means that the expressions (sum of term) reflecting the decomposition of the score on the right side of the equal sign is equivalent to the individual's score on the dependent variable (the left side of the equal sign)	Evidence variable
11. Grand Mean	Grand Mean	Grand mean represent the average level of the dependent variable values	Explanatory variable
12. Grant Mean_PLVG	Grand mean refers to the pooled mean (GMRef)	Grand mean refers to the pooled mean of all scores (pulling over factor A and B) in the population	Evidence variable
13. Mean of Group Means	Mean of Group Means	The Grand Mean is the average of the group means	Evidence variable
14. Main Effect: Level of A	Main Effects: Level of A	A general concept of the main effect A	Explanatory variable
15. Main Effect A_j	Main Effect A_j	α_j refer to the main effect of the independent variable group on the dependent variable, independent of group in the population.	Evidence variable
16. Grp Mean(j)-GM	Main Effect A_j: Grp Mean(j)-GM	$\alpha_j = \mu_j - \mu$. This term means main effect A can be written as a difference between in the population.	Evidence variable
17. Main Effect: Level of B	Main Effects: Level of B	A general concept of the main effect B	Explanatory variable
18. Main Effect B_k	Main Effect B_k	β_k refers to the main effect of the independent variable B	Evidence variable
19. Grp Mean(k)-GM	Main Effects B_k Grp Mean($_k$)-GM	$\beta_k = \mu_k - \mu$. This term means main effect B can be written as a difference between in the population.	Evidence variable
20. Interaction: AxB	Interaction between A and	A general concept of the	Explanatory

	B	interaction between factor A and factor B	variable
21. Interaction Effect AB $(_{jk})$	Interaction Effect $AB(_{jk})$	Y(jk) refers to the interaction effect of the combination of Factor A j and Factor B k on the subjects' score—that is, a value of the dependent variable	Evidence variable
22. GrpM(jk)-M(j)-M(k)+GM	GrpM(jk)-M(j)-M(k)+GM	$Y(_{jk})= \mu_{jk}-\mu_{j-}\mu_{k+}\mu$. The interaction effect $Y(_{jk})$ is the mean of the combination of (factor A) group j with Factor B k minus the pooled (marginal) mean of Factor A j and the pooled (marginal) mean of duration k plus grand mean (population values)	Evidence variable
23. GrpM(jk)-Eff(j)-Eff(k)-GM	GrpM(jk)-Eff(j)-Eff(k)-GM	$Y(_{jk})= \mu_{jk}-\alpha_j-\beta_k+\mu$. The interaction effect may also be written as the mean of the combination of Factor A j with Factor B k (cell mean) minus the main effect of factor A (α_j) minus the grand mean	Evidence variable
24. Error	Error	A general concept	Explanatory variable
25. Error	Residual Var	$e_i(_{jk})= Y_i(_{jk}) -\mu-\alpha_j-\beta_k+\mu.- Y_{jk}$ the error term is a variable that refers to the residual portion (part) of a subject i's score on Y after all of the effects and the grand mean have been subtracted out	Evidence variable
26. Score (ijk)-GrpM(jk)	Score (ijk)-GrpM(jk)	$e_i(_{jk})= Y_i(_{jk}) - \mu_{jk}$ The error term is the difference between a subject's score on Y and the subject's cell mean	Evidence variable
27. Score (ijk)-GM-Effects	Score (ijk)-GM-Effects	$e_i(_{jk})= (Y_i(_{jk}) -\mu)-(\alpha_j+\beta_k+(\alpha_j\beta_k))$. The error score can be interpreted as that portion of a subject's observed score on the dependent variable (expressed as a deviation from the general mean), which is not predictable from the effects of the individual's particular combination of Factor A and Factor B.	Evidence variable
28. Additive Combination	Additive components	$(\mu+\alpha_j+\beta_k+Y_{(jk)} + e_{i\,(jk\,)})$ The score decomposition consists of a sum of five components: main effect of Factor A, main effect of Factor B, interaction of A and B, and error.	Evidence Variable

Appendix B. Twelve Students' Scores on 19 Evidence Variables*

Evidence variable and number	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
1. Dep Var Y		V	V	V	V	V	V	V	V	V	V	V
2. Level j of A	V	v	V	V	V	V	V	V	V	V	V	
3. Level k of B	V	v	V	V	V		V	V	V	V	V	
4. jk Representation		V	V	V	V	V	V	V	V		V	
5. Grp (jk) in 2Way Table		V						V	V			
6. Equivalence			V	V	V	V	V	V	V	V	V	V
7. Grand Mean RLVG		V	V	V	V	V	V	V	V	V		V
8. Mean of Group Means		V	V					V	V			
9. Main Effect Aj	V	V	V	V	V	V	V	V	V	V		
10. Grp Mean(k)-GM	V	V	V				V	V	V			
11. Main Effect Bk	V	V	V	V	V	V	V	V	V	V		V
12. Grp Mean(k)-GM	V	V	V	V	V	V	V	V	V	V		
13. Interaction Effect AB(jk)	V	V	V	V	V	V	V	V	V			
14. GrpM(jk)-M(j)-M(k)+GM		V	V					V	V	V		V
15. GrpM(jk)-Eff(j)-Eff(k)-GM		V	V					V				
16. Residual Var		V	V	V	V	V	V	V	V	V		
17. Score(ijk)-GrpM(jk)		V	V					V	V			
18. Score(ijk)-GM-Effects			V	V	V	V	V	V	V	V		
19. Additive Combination	V	V	V	V			V	V	V			
Total	8	17	18	13	12	11	14	19	18	11	5	5

* V: the student scores the variable successfully.