On Validating Cognitive Diagnosis Models for the Arithmetic Skills of Elementary School Students

Hyejung Koh[1], Wonjin Jang[2], Yongseok Yoo[3]*
Uiduk University, Gyeongju 38004, Republic of Korea[1]
Incheon National University, Incheon, 22012, Republic of Korea[2, 3]

***Abstract*—Cognitive diagnosis models (CDMs) have been shown to provide detailed evaluations of students' achievement in terms of proficiency of individual cognitive attributes. Attribute hierarchy model (AHM), a variant of CDM, takes the hierarchical structure of those cognitive attributes to provide more accurate and interpretable measurements of learning achievement. However, advantages of the richer model come at the expense of increased difficulty in designing the hierarchy of the cognitive attributes and developing corresponding test sets. In this study, we propose quantitative tools for validating the hierarchical structures of cognitive attributes. First, a method to quantitatively compare alternative cognitive hierarchies is established by computing the inconsistency between a given cognitive hierarchy and students' responses. Then, this method is generalized to validate a cognitive hierarchy without real responses. Numerical simulations were performed starting from an AHM designed by experts and responses of elementary school students. Results show that the expert-designed cognitive attribute explains the students' responses better than most of alternative hierarchies do, but not all; a superior cognitive hierarchy is identified. This discrepancy is discussed in terms of internalization of cognitive attributes.**

Keywords*—Cognitive diagnosis model; attribute hierarchy model; cognitive hierarchy; model validation

I. Introduction

A. Cognitive Diagnostic Models

Learning analytics has attracted much attention recently, as more data become available for educators and learners [1]. Vast amounts of data are generated in the field of education, due to the widespread adoption of online education systems, such as massive open online courses [2]. By utilizing more data, a more accurate and detailed assessment of learning achievement is possible [3], resulting in enhanced learning experience [4]. For instance, teachers could offer individualized learning strategies tailored to needs of the target students, such as university students [5], under-represented students [6] or foreign language learners [7].

Cognitive diagnosis models (CDMs) have been actively studied as a useful tool for assessing students' knowledge states in terms of multiple cognitive attributes [8]. CDMs incorporate multiple cognitive attributes that are required to understand a concept or to perform a task. Items that require different combinations of cognitive attributes are developed, and the degree of proficiency for each cognitive attribute is estimated from the students' responses to these items. The quantitative assessment of a student's proficiency of individual cognitive attributes allows a more detailed evaluation of a student's achievement, compared to a total score-based learning diagnosis.

CDMs are largely divided into two groups: compensatory models and non-compensatory models [9]. Compensatory models assume that one cognitive attribute could compensate for another. Thus, even if a particular cognitive attribute is not mastered by the examinee, they may solve an item by using other cognitive attributes. For instance, reading comprehension requires numerous cognitive attributes, such as grammar and vocabulary. Even if there are a few words that a reader is not familiar within a given text, the reader could postulate the meanings of the words from the grammatical structure and solve related items correctly. In such cases, grammar plays a compensatory role for vocabulary.

In contrast, non-compensatory models assume that the lack of a cognitive attribute is not compensated for by other cognitive attributes [10]. Therefore, if one fails to master any of the cognitive attributes required, an item cannot be solved. For example, according to the non-compensatory model, an item requiring the concepts of logarithmic function and algebraic function could only be correctly solved by those who have mastered both concepts. If any one of the two concepts is lacking, they will not be able to correctly solve the item.

B. Related Work

Among early non-compensatory models, the rule-space model (RSM) was explored and established by Tatsuoka [11]. The relationship between an item and the cognitive attributes required to solve the item is represented by a matrix, called the Q-matrix. Each row of the Q-matrix corresponds to an item, while each column corresponds to a cognitive attribute; Q_{ij} is one if the j^{th} cognitive attribute is required to solve the i^{th} item. Otherwise, Q_{ij} is zero. The RSM was developed to estimate students' knowledge states from their item responses and a corresponding Q-matrix.

In this study, we adopt a variant of the RSM, called the attribute hierarchy model (AHM) [12] to quantify elementary school students' learning achievements in arithmetic operations. Similarly to CDMs, the AHM aims to represent the relationship between items and cognitive attributes. Furthermore, the AHM includes hierarchical relationships between the cognitive attributes in the model. This hierarchical structure of cognitive attributes is represented by a graph, in which each node corresponds to a cognitive

*Corresponding Author.

attribute, and an edge between two nodes implies that one cognitive attribute is a prerequisite for the other cognitive attribute. Therefore, the AHM is suitable for evaluating the achievement of mathematics subjects, where the hierarchy of cognitive attributes is important [13].

A precise relationship between cognitive attributes and items is crucial for the accurate measurement of learning achievement. However, in practice, developing a Q-matrix and corresponding items, requires months of intensive collaboration between modeling experts and teachers in the field.

To assist the development of the Q-matrix, several quantitative tools have been proposed. First, de la Torre [14] proposed to validate a given Q-matrix by the degree of agreement with the response data using the EM algorithm [15]. Such quantification provides an objective measure to compare multiple Q-matrices. However, it remains elusive which Q-matrices should be compared. To address this gap, DeCarlo proposed a Bayesian framework to validate individual elements of a given Q-matrix [16]. However, the flexibility of the Bayesian model comes with an increased complexity of the validation procedure because the reliability of each element of the Q-matrix must be provided in advance. Last, Chiu proposed a simpler nonparametric method to identify misspecified entries of a Q-matrix [17].

However, those validation methods are applicable only to RSMs, not to AHMs. The hierarchical structure of the cognitive attributes is not considered for the validation of a given Q-matrix. Therefore, adopting AHMs in practice requires extention of the validation framework and more careful consideration of the structure of the cognitive attributes of interest.

C. Contributions of the Study

We propose quantitative tools for validating the hierarchical structures of cognitive attributes, to aid the development of AHMs. Whereas the current validation methods quantify the validity of each element of the Q-matrix, we explore alternative hierarchical structure of the cognitive attributes. Thus, the search space of out method is a graph rather than a matrix.

Our approach would provide a more natural way to validate AHMs in conjunction with the experts' domain knowledge. For instance, the current methods provide a refined Q-matrix with some elements flipped from the initial Q-matix. However, such a refined Q-matrix may contradict to the hierarchical structure of cognitive attributes developed by the experts. Instead, our method starts from the experts' knowledge in terms of the hierarchical structure and validate individual associations of cognitive attributes.

The rest of this paper is organized as follows: the Methods section describes the AHM model, as well as the data collection and evaluation of alternative hierarchies of cognitive attributes. In the Results section, we show numerical simulations; finally, in the Conclusions and Discussions section, we discuss the results and their implications, with concluding remarks.

II. Methods

A. Modeling and Data Collection

An AHM model for arithmetic operations with natural numbers was designed as follows. First, seven cognitive attributes were chosen based on the elementary school mathematics curriculum – addition (A1), subtraction (A2), multiplication (A3), division (A4), carry (A5), borrow (A6), and '0' in multiplication (A7). Thereafter, the hierarchy of the seven attributes (H_0, Fig. 1A) was designed by experts. In brief, the root node (A1) represents addition, which is the prerequisite for all other cognitive attributes. The descendant nodes of the root node are A2, A3, and A5, which correspond to subtraction, multiplication, and carry, respectively. To understand a leaf node (A4, A6, or A7, corresponding respectively to division, borrow, and "0" in multiplication), one should master all the preceding cognitive attributes up to the root node. For example, A4 (division) requires A1 (addition), A2 (subtraction), and A3 (multiplication).

Based on the expert-designed hierarchical structure (H_0, Fig. 1A), alternative hierarchical structures ($H_1, H_2, \cdots, H_7$) were created by removing an edge from H_0 (Fig. 1B). For example, H_1 was generated by removing the edge between A1 and A2 (dashed line in Fig. 1B) from H_0. Similarly, the removed edges in H_k for $k = 2, 3, \cdots 7$ are indicated by (H_k).

Next, thirty items involving the seven cognitive attributes were developed. Our participant sample comprised 977 fourth graders who participated in the test; their responses to individual items were coded as either correct (1) or incorrect (0).

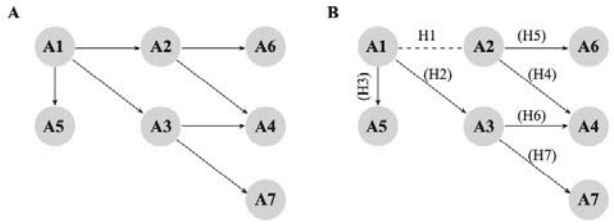

Fig. 1. (A) The Hierarchical Structure (H_0) of the Seven Cognitive Attributes (A1, A2, ⋯, A7) Designed by the Experts. (B) Alternative Hierarchical Structures ($H_1, H_2, \cdots, H_7$) were Created by Removing One Edge from H_0. For Example, the Graph in Panel B shows H_1, Generated by Removing the Edge between A1 and A2 (Dashed Line) from H_0. Similarly, other Hierarchical Structures (in Parentheses) are Generated by Removing the Corresponding Edges.

B. Validating the Hierarchies of Cognitive Attributes

The alternative hierarchies were validated by quantifying the degree to which a given hierarchy is in agreement with students' actual responses. The block diagram in Fig. 2 summarizes the quantification steps, each of which is explained as follows.

First, a different attribute hierarchy (H_k) implies a different structure of students' knowledge states. For the seven cognitive attributes, $2^7 = 128$ combinations of the seven attributes are enumerated. Among all the potential combinations, those that are in conflict with H_k are eliminated (Leighton et al., 2004) and the remaining attribute combinations comprise the set of valid knowledge states S_k.

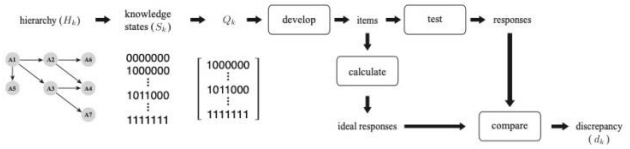

Fig. 2. A Schematic Diagram of the Quantitative Validation of a Hierarchical Structure. For any given Hierarchical Structure H_k, Corresponding Knowledge States S_k (Combinations of Attributes Consistent with H_k) and a Q-matrix Q_k are Generated. Items Corresponding to the Rows of Q_k are Developed and Students' Responses (r) to these Items are Collected. Ideal Responses (r_{ideal}) are Theoretically Calculated from the same Items. Comparing r and r_{ideal} Produces a Discrepancy Value d_k, which Quantifies the Validity of H_k; a Lower d_k Implies Higher Validness.

Next, the Q-matrix Q_k is defined by taking all the valid states in S_k—except the all-zero state (0000000)—as its rows.

Three statistical characteristics of each Q_k-matrix are then measured. First, the sum of each column of Q_k (column sum) indicates the number of items that require the corresponding cognitive attribute. Second, the sum of each row of Q_k (row sum) indicates the number of attributes included in the corresponding item. Third, the sparsity of Q_k is defined by the number of ones divided by the number of elements of Q_k, which corresponds to the frequency of attributes examined in the test set.

Next, we develop an item involving corresponding attributes for each row of Q_k, and students' item responses (r) to the items, which are collected through tests. The set of all the item responses is called R.

The ideal responses (r_{ideal}) to the items are theoretically generated. Here, *ideal* means that the response is solely based on the mastery of the cognitive attribute, excluding guessing or mistakes. For each knowledge state $s \in S_k$, an ideal response r_{ideal} is calculated by Equation 1.

$$r_{ideal} = \sim((\sim s){Q_k}^T), \quad (1)$$

where $\sim$ and T mean the logical NOT operator and the matrix transpose, respectively. The set of ideal responses for all the states in S_k is called I_k.

Last, the discrepancy between the ideal responses (I_k) and the actual responses (R) is calculated as follows. A lower discrepancy value indicates that H_k provides a better account for actual students' responses. First, the discrepancy for each response $r \in R$ is defined by the Hamming distance (the number of mismatches) between r and the closest r_{ideal} in I_k, presented as $h(r, I_k)$. The average discrepancy of responses is normalized by the number of items and defined as the discrepancy for H_k (Equation 2).

$$d(H_k) = \frac{\sum_{r \in R} h(r, I_k)}{N_R N_I}, \quad (2)$$

where N_R and N_I are the numbers of responses and items, respectively.

C. Generating Virtual Responses from an Existing Dataset

Ideally, a different test set would be developed and responses for each hierarchical structure H_k would be collected. However, this would be costly and require too much time. Developing a test set for H_k requires determining the combinations of cognitive attributes based on H_k and developing corresponding items; each of these steps requires repeated feedback from experts. Recruiting test takers for multiple test sets is also costly. Repeating multiple test sets involving the same set of cognitive attributes for a fixed target group would be impractical. Students might learn to recognize patterns during the sequence of similar tests or become bored by the similarity of multiple tests. In either case, collecting unbiased responses for each H_k is a challenging task. Furthermore, if each test is performed on a different group of students, it is unclear whether any difference in item responses is due to the different hierarchical structure or the heterogeneity of the student groups.

To overcome this limitation, we propose to generate virtual responses to each H_k by employing random sampling using a common dataset (Fig. 3). Specifically, for each item, corresponding attributes (each row of Q_k) are used as a query to retrieve responses to items that require attributes similar to the query from the existing database. To simplify, attributes that have the smallest Hamming distance to the query attribute are chosen as candidates first. Thereafter, if there are multiple candidates, one is randomly selected with equal probability. Otherwise, (if there is only one item with the smallest Hamming distance), the candidate response is the virtual response. Repeating the above procedure for all the items (all the rows of Q_k) comprises an iteration of simulation. Subsequently, the average discrepancy is measured for 1,500 iterations for each H_k. Additionally, the statistical significance of the discrepancy is measured by repeating the above calculations, starting from 10 different random seeds.

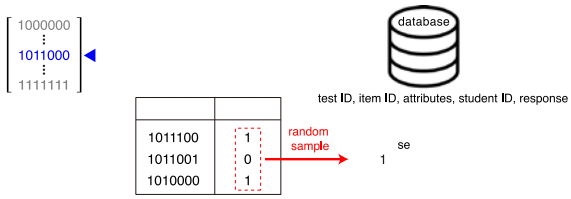

Fig. 3. Virtual Item Responses are Generated from an Existing Database. Each Row of Q_k is used as a Query to Retrieve Candidate Responses to Items with Similar Attributes. among these Candidate Responses, One Response is Randomly Sampled and used as a Virtual Response.

III. Results

A. Statistical Characteristics of the Q-matrix

Even with the same set of cognitive attributes, different hierarchical structures necessitate items with different combinations of cognitive attributes; these result in different numbers of rows in the corresponding Q-matrices. Removing an edge from H_0 increases the number of rows of Q_k (27 ~ 33) for $k = 1, 2, \cdots, 7$ (Fig. 4). This is because more items are needed to determine participants' mastery of independent attributes than what are needed for related attributes. More specifically, Q-matrices corresponding to H_5 and H_7 had the largest number of rows (33). Compared with H_0, H_5 and H_7 have independent nodes A6 and A7, respectively; these nodes had many preceding attributes in H_0.

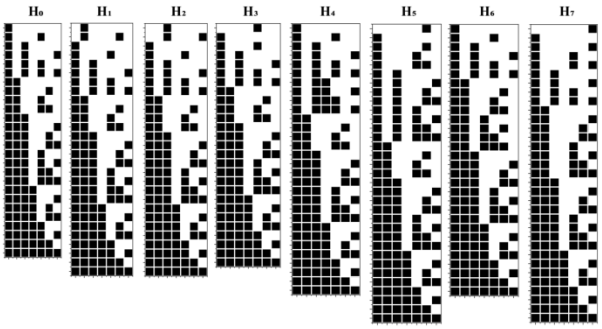

Fig. 4. The Q-matrices (Q_k) for different Hierarchical Structures (H_k). The Rows and Columns of each Q_k are Associated with Items and Attributes, Respectively. Each Row of Q_k Represents Cognitive Attributes Required for each Item in Black.

Even though the number of rows of Q_k differ, the frequencies of ones in Q_k are similar. The average numbers of items per attribute did not differ significantly ($p = 0.970$, one-way ANOVA). Similarly, the average numbers of attributes per item did not differ significantly ($p = 0.998$, one-way ANOVA). Finally, the average sparsity of Q_k was $0.571\ (\pm 0.012)$. Thus, on average, hierarchies (H_k) are homogeneous in terms of the number of items per cognitive attribute and the number of cognitive attributes per item.

B. Comparison of the Discrepancy for Each Hierarchy

The hierarchy designed by the experts (H_0) had a lower discrepancy than all the alternative hierarchies, except for H_2 (Fig. 5); this indicates that H_0 explains students' responses better than most alternative hierarchies, except for H_2. The increase in the discrepancy between the alternative hierarchies implies that the removed edge is important for explaining the actual students' responses.

In contrast, $d(H_2)$ was lower than $d(H_0)$ (Fig. 6). The *t*-test for 10 simulations with different random seed numbers confirmed that there was a significant difference between $d(H_0)$ and $d(H_2)$ ($p = 4.18 \times 10^{-18}$). This result implies that H_2 explains the students' real item responses better than H_0. There is thus a hierarchy of cognitive attributes that better describes responses than the hierarchy designed by the experts.

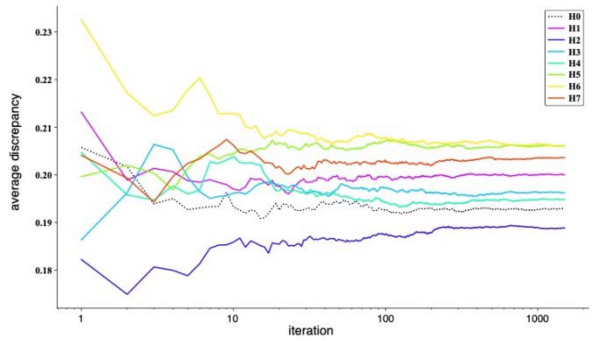

Fig. 5. Average Discrepancy (d_k) of each Hierarchy (H_k) as a Function of Iteration. During Early (< 10) Iterations, d_k Fluctuates Considerably; However, it Stabilizes after about 500 Iterations. After 1,500 Iterations, the Value of d_k was Measured for each H_k. The Hierarchy H_k, which was Designed by the Experts, showed Lower d_k (Black, Dotted) than Most of the other H_k, Except for H_2 (Purple).

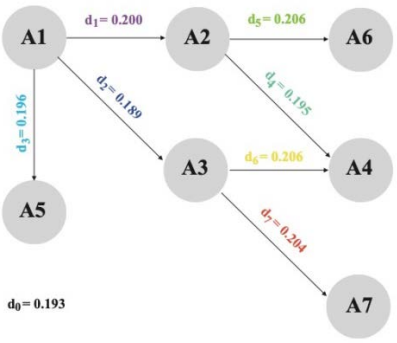

Fig. 6. The Discrepancy (d_k) of each Hierarchy (H_k) is shown on the Removed Edge (for $k > 0$). The Hierarchy without the Edge between A1 and A3 (H_2) showed the Lowest d_k, even Lower than that of the Expert-Designed Hierarchy (d_0).

IV. Conclusion and Discussion

In this study, we propose a method to quantitatively validate the hierarchy of cognitive attributes of a CDM and corresponding Q-matrices. The hierarchy designed by experts (H_0) was compared with alternative hierarchies ($H_1, H_2, \cdots, H_7$) with an edge removed. The discrepancy for each hierarchy was defined by the distance between the real and ideal responses (r and r_{ideal}), the inverse of which is interpreted as a quantitative indicator of how well a hierarchy and the corresponding Q-matrix describe the students' item responses.

Virtual responses were generated from an existing database, rather than directly collecting responses for each hierarchy and its corresponding Q-matrix. After generating a Q-matrix that corresponds to each hierarchy, we selected the items with the closest Hamming distance ($d_{Hamming}$) to each row of Q by comparison with the existing datasets, and one of the responses to these items was randomly selected as a virtual response.

The hierarchy of cognitive attributes designed by the experts (H_0) had generally lower discrepancy than alternative hierarchies; however, one hierarchy (H_2) had a lower discrepancy than H_0. The difference between H_0 and H_2 was the edge between addition (A1) and multiplication (A3), which is present in H_0, but absent in H_2. This implies that the link between addition and multiplication might be weaker than was expected by the experts.

Our interpretation of this gap is that multiplication, once acquired as a separate skill, may not require the concept of addition. The concept of multiplication can be divided into three categories: repetitive addition, multiples, and product set [18]. The first concept of multiplication—repetitive addition—is utilized to teach multiplication to first-time learners. It is therefore reasonable to assume that understanding or performing multiplication also relies on the knowledge of addition, in agreement with the hierarchy designed by the experts (H_0). However, the other aspects of multiplication may play more important roles for students who mastered the concept of multiplication. In other words, after acquiring the knowledge of multiplication, they may perform multiplication as an independent skill, rather than repeating

addition multiple times. Therefore, we posit that the relationship between addition and multiplication may be important for learning, but that it is less so for practicing the actual skill of multiplication.

We propose a scalable validation tool for comparing alternative hierarchies, which could encourage more teachers to utilize CDMs for learning achievement analyses. Selecting relevant cognitive attributes and designing the hierarchy among them requires experts' knowledge and experience, which could hinder wider uses of CDMs. When a user wants to validate a chosen hierarchy or explore alternative hierarchical structures, item responses are sampled from an existing database, without having to develop new items and collect responses for each candidate. The proposed quantitative validation of alternative hierarchies could be used as objective indicators of the validity of established hierarchies.

Our future work will generalize the proposed framework to more complex cases. In this study, we considered only seven attributes with rather simple associations. In general, presence of loop a graph hinders theoretical analysis as well as numerical calculations based on the graph. The hierarchical structure in this study has only one loop. It is of great interest to explore a more complex attributes structure with multiple loops.

ACKNOWLEDGMENT

This study was supported by Incheon National University Research Grant in 2019.

REFERENCES

[1] Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), 2020.

[2] N. B. Shah, J. Bradley, S. Balakrishnan, A. Parekh, K. Ramchandran, and M. J. Wainwright, "Some scaling laws for MOOC assessments," KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014), 2014.

[3] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, and S. N. Liao, "Predicting academic performance: a systematic literature review," Proceedings of the 23rd annual ACM conference on innovation and technology in computer science education, 2018.

[4] K. Mangaroska and M. Giannakos, "Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning," IEEE Transactions on Learning Technologies, 12(4), 516-534, 2018.

[5] H. Aldowah, H. Al-Samarraie, H., and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis." Telematics and Informatics, 37, 13-49, 2019.

[6] A. Cano and J. D. Leonard, "Interpretable Multiview Early Warning System Adapted to Underrepresented Student Populations," IEEE Transactions on Learning Technologies, vol. 12, no. 2, pp. 198-211, 2019.

[7] Bravo-Agapito, J., Bonilla, C. F., & Seoane, I, "Data mining in foreign language learning," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), e1287, 2020.

[8] M. von Davier and Y. S. Lee, Handbook of diagnostic classification models, Springer International Publishing, 2019.

[9] A. A. Rupp and J. L. Templin, "Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art," Measurement, 6(4), 219-262, 2008.

[10] M. Birenbaum, A. E. Kelly, and K. K. Tatsuoka. "Diagnosing knowledge states in algebra using the rule-space model," Journal for Research in Mathematics Education, 24(5), 442-459, 1993.

[11] K. K. Tatsuoka, "Rule space: An approach for dealing with misconceptions based on item response theory," Journal of educational measurement, 345-354, 1983.

[12] J. P. Leighton, M. J. Gierl, and S. M. Hunka, "The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach," Journal of educational measurement, 41(3), 205-237, 2004.

[13] M. J. Gierl, C. Alves, and R. T. Majeau, "Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment," International Journal of Testing, 10(4), 318-341, 2010.

[14] J. de la Torre, "An empirically based method of Q-matrix validation for the DINA model: Development and applications," Journal of Educational Measurement, 45, 343-362, 2008.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Series B, 39(1), 1–38, 1977.

[16] L. T. DeCarlo, "Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model," Applied Psychological Measurement, 36, 447-468, 2012.

[17] C. Y. Chiu, "Statistical refinement of the Q-matrix in cognitive diagnosis," Applied Psychological Measurement, 37, 598-618, 2013.

[18] H. K. Kang, An Alternative Program for the Teaching of Multiplication Concept Based on Times Idea. Journal of Korea Society of Educational Studies in Mathematics, 11(1), 17-37, 2009.