Likelihood Ratio Approach for Identifying Hierarchical Attribute Structure

Research Article

Lokman AKBAY

1 Mehmet Akif Ersoy University, Faculty of Education, Department of Educational Measurement and Evaluation, Burdur, Turkey, ORCID: 0000-0003-4026-5241

To cite this article: Akbay, L. (2019). Likelihood Ratio Approach for Identifying Hierarchical Attribute Structure, International Online Journal of Educational Sciences, 11 (4), 181-195

ARTICLE INFO

ABSTRACT

When attributes are hierarchically structured, modifying the Q-matrix or prior distribution in the estimation process yields more accurate and precise item and person parameter estimates. Modification of the prior distribution and the Q-matrix depend on the assumed attribute hierarchy, as such, identifying the correct hierarchical structure among a set of measured is of the essence. To address the subjectivity in the conventional methods for attribute structure identification (i.e., expert opinions via content analysis and verbal data analyses such as interviews and think-aloud protocols); this study proposes a likelihood-ratio-test based exhaustive empirical search method for identifying hierarchical structures. It further suggests employment of likelihood-ratio-test based model selection approach for choosing the most accurate hierarchical structure among proposed candidates. Results of this study show that the likelihood-ratio-test based exhaustive search produces a reachability matrix that specifies all the true prerequisite relationships among the attributes. Thus, the method is promising and may be used for exploratory purposes for identification of hierarchical attribute structure.

© 2019 IOJES. All rights reserved

Keywords:
CDM, attribute hierarchy, likelihood ratio, hierarchy identification

Introduction

In many educational and psychological tests, examinees are required to use their knowledge, skills, strategies, and cognitive competencies for successful completion of the assessment tasks. Such categorical latent variables representing the knowledge states of examinees are referred to as attributes. These attributes may have a hierarchical structure (Templin & Bradshaw, 2014). Cognitive and educational research suggest that building conceptual understanding requires incorporating novel knowledge to preliminary or more basic knowledge (Linn, Eylon, & Davis, 2004; Smith, Wiser, Anderson, & Krajcik, 2006; Vosniadou & Brewer, 1992). Therefore, acquisition of domain related attributes may proceed sequentially. Hence, disciplinary and
interdisciplinary ideas need to form a meaningful structure in the curriculum that allows teaching steps build upon one another (Schmidt, Wang, & McKnight, 2005).

Cognitive model of task performance based assessments are, in general, referred to as cognitively diagnostic assessment (CDA) (de la Torre & Minchen, 2014). Such assessments aim to identify the attribute mastery profiles of examinees. CDAs serve as formative assessment so that the feedback obtained from the assessment results might be used to modify teaching and learning activities (DiBello & Stout, 2007). The role of cognitive theory must be well articulated in test design; only then, a CDA will be useful. Statistical models extracting diagnostic information from CDA have also been developed. These statistical models are referred to as cognitive diagnosis models (CDMs) or diagnostic classification models (DCMs) (de la Torre & Minchen, 2014). To derive a CDM, (1) attribute interaction in response construction and (2) the attributes required for each item need to be known (Chiu, Douglas, & Li, 2009). To define the relationships between items and attributes, a $J \times K$ item-by-attribute specifications matrix is used. This matrix is referred to as Q-matrix (Tatsuoka, 1983). The Q-matrix is usually a binary matrix of $J$ rows and $K$ columns where $j = 1, ..., J$ indicates the items and $k = 1, ..., K$ represents attributes measured by the test. In a Q-matrix, an element $q_{jk}$ is coded 1 when item $j$ requires attribute $k$; otherwise, it is coded 0. When attributes follow a hierarchical structure, the Q-matrix and the prior distribution employed in CDM estimation can be modified to obtain more accurate and precise item and person parameters (Akbay & de la Torre, 2015).

Modifications in the prior distribution and in the Q-matrix are applied based on the assumed hierarchical structure. Therefore, determining the correct hierarchical structure is of the essence. Because specifying an incorrect sequential relationship between any two attributes may substantially degrade estimation accuracy. As such, the importance of identifying the true hierarchical structure cannot be overemphasized. Current practices for hierarchical structure detection include expert opinions via content analysis and verbal data analyses such as interviews and think-aloud protocols (Akbay, Terzi, Kaplan, & Karaaslan, 2017; Cui & Leighton, 2009; Gierl, Wang, & Zhou, 2008). These procedures may result in disagreements over the prerequisite relationships, which may consequently yield more than one hierarchical structure. Furthermore, emerging hierarchical structures from verbal analysis and expert opinion may not be the same (Gierl et al., 2008). In the literature, there is no model based statistical tests addressing the subjectivity in the conventional methods for detecting hierarchical attribute structure. To address this subjectivity, this study proposes a model-fit based empirical exhaustive search method that can be used for hierarchical relationship detection among a set of attributes. It should be noted here that the proposed method is intended to complement rather than replace the current procedures.

The rest of this manuscript is organized as follows: Background information for the two specific models employed in this research is provided. Then, the empirical search algorithm is presented and simulation studies conducted to test the usefulness of the algorithm. Presentation of simulation results is followed by a numerical example. There will be some concluding remarks in the last section.

**Background: The DINA and DINO Models**

The deterministic input, noisy `and'' gate (DINA; de la Torre, 2009b; Junker and Sijsma, 2001) model has two item parameters (i.e., guessing and slip). This property makes the DINA model one of the most parsimonious and interpretable CDMs (de la Torre, 2009b). The DINA is known to be a conjunctive model (de la Torre & Douglas, 2004) as it assumes that missing one of the required attributes result in the baseline probability that is equal to the probability of answering an item when none of the required attributes is mastered (de la Torre, 2009b; Rupp & Templin, 2008). For a given examinee latent group, $\alpha_x$, and the $j^{th}$ q-vector; an ideal response ($\eta_{ij} = 1$ or 0) for the latent group is produced by the conjunctive condensation function (Maris, 1995, 1999),
\[ \eta_{ij} = \prod_{k=1}^{K} a^{q_{ik}}. \]

Hence, examinees are divided into two distinct groups by the DINA model. The first group is referred to as mastery group, which involves examinees who mastered all required attributes for the item. The second group, which is called non-mastery group, consists of examinees lacking at least one of the required attributes.

Possibility of slipping on an item for examinees in mastery group and guessing on the item for examinees in non-mastery group are allowed by the probabilistic component of the model’s item response function (IRF). The slipping and guessing probabilities on item \( j \) are denoted as \( s_j = P(X_{ij} = 0|\eta_{ij} = 1) \) and \( g_j = P(X_{ij} = 1|\eta_{ij} = 0) \), respectively, where \( X_{ij} \) is the observed response of examinee \( i \) to item \( j \). Given \( s_j \) and \( g_j \), the IRF of the DINA model is written as

\[
P(X_j = 1|\alpha_i) = P(X_j = 1|\eta_{ji}) = g_j^{1-\eta_{ji}} (1 - s_j^{-\eta_{ji}}),
\]

where \( \alpha_i \) is attribute pattern \( i \) among \( 2^K \) possible attributes patterns; \( \eta_{ji} \) is the expected response of an examinee to item \( j \) who possesses attribute pattern \( i \); and \( g_j \) and \( s_j \) are guessing and slip parameters, respectively (de la Torre, 2009a).

The deterministic input, noisy `or’ gate (DINO; Templin and Henson, 2006) model is the disjunctive counterpart of the DINA model. This model assumes that mastering at least one of the required attributes and mastering all of the required attributes results in the same success probability for correctly answering an item (Rupp & Templin, 2008). Due to the disjunctive nature of the model, ideal response of an examinee (i.e., \( \omega_{ji} = 1 \) or 0) in latent group \( \alpha_i \) is given by the function

\[
\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - a_{ik})^{q_{ik}}.
\]

Although the DINO also partition examinees into mastery and non-mastery group, the non-mastery group consisted of examinees missing all the required attributes while the rest of the examinees are classified into the mastery group.

The model parameters are defined as \( s_j = P(X_{ij} = 0|\omega_{ij} = 1) \) and \( g_j = P(X_{ij} = 1|\omega_{ij} = 0) \). Therefore, the success probabilities of the mastery and non-mastery groups on item \( j \) become \( 1-s_j \) and \( g_j \), respectively. The IRF of the DINO model can be written as

\[
P(X_j = 1|\alpha_i) = P(X_j = 1|\omega_{ji}) = g_j^{1-\omega_{ji}} (1 - s_j^{-\omega_{ji}}),
\]

where \( \omega_{ji} \) is the expected response of an examinee with attribute pattern \( i \) to item \( j \); and \( g_j \) and \( s_j \) are guessing and slip parameters for item \( j \), respectively (Templin & Rupp, 2006).

### An Empirical Exhaustive Search for Identifying Hierarchical Attribute Structure

In cases where attribute \( k \) is prerequisite to attribute \( k' \), some of the \( 2^K \) attribute patterns become impermissible. For example, imagine three attributes where \( A1 \) is prerequisite for \( A2 \) while \( A3 \) is independent from \( A1 \) and \( A2 \). Then, theoretically, some of the latent classes (i.e., \( 010 \) and \( 011 \)) do not exist. In this case, one can employ a structured prior distribution in the model estimation that leaves out the impermissible latent classes. Alternatively, an unstructured prior distribution may also be employed in the estimation that allows all \( 2^K \) latent classes in the estimation. Notice that in the first approach, some of the latent class probability parameters are fixed to zero such that there are less parameters to be estimated. Thus, by treating the latter as the null model and the former as the alternative model, a likelihood ratio test (LRT) may be conducted with an expectation of retaining the null hypothesis when attribute \( k \) is a prerequisite attribute for attribute \( k' \). Moreover, Akaike information criterion (AIC) and Bayesian information criterion (BIC) may also be used for model comparison purposes.
Rationale and Search Algorithm

When attribute \( k \) is prerequisite to attribute \( k' \), \( 3(2^{k-2}) \) latent classes are permissible while \( 2^{k-2} \) are impermissible. For example, when six attributes are measured and one attribute (A1) is prerequisite for another attribute (A2); only the subset of attribute patterns conforming to this hierarchical relationship (i.e., 00***, 10***, and 11*** ) becomes permissible. The attribute patterns against this prerequisite relationship (i.e., 01*** ) will not be allowed. Here * stands for either 0 or 1 producing 16 different classes.

When items in a test are sufficiently discriminating, a model estimating only the permissible latent class probabilities is expected to yield a model-fit statistic that is not statistically different from the model estimating all latent class probabilities. Thus, we obtain a null model using a structured prior distribution, whereas the alternative model is obtained by using an unstructured prior distribution. For instance, DINA or DINO model estimates \( 2J+2k-1 \) parameters in the alternative model; whereas either one estimates \( 2J+L-1 \) parameters where \( L \) stands for the number of permissible latent classes.

Because the null model is nested within the alternative model, a likelihood ratio test can be conducted with an expectation of retaining the null hypothesis (i.e., the null model fits the data equally well) when attribute \( k' \) in fact requires attribute \( k \). Therefore, an empirical exhaustive search based on the LRT can be implemented to detect hierarchical relationships among the measured attributes. To accomplish that, \( \chi^2_P = K(K - 1) \) reduced models each of which specifying a distinct hierarchical relationship between attributes \( k \) and \( k' \) need to be set. In other words, all possible pairwise prerequisite relationships between measured attributes must be specified to form reduced models. Then, LRT is conducted between each of the reduced models and the full model. The hierarchical structure identification procedure may be completed in the following steps;

- Step 1: Estimate the model parameters for the alternative model, which incorporates an unstructured prior distribution.
- Step 2: Estimate the model parameters for the null models that incorporate structured prior distributions to fix the impermissible latent class parameters to zero when \( k \) is prerequisite to \( k' \).
- Step 3: Repeat Step 2 for all possible \( \chi^2_P \) pairwise relationships.
- Step 4: Compare the fit of the alternative model against the fit of each of \( \chi^2_P \) null models.
- Step 5: Report the model comparison results in binary outcomes where 0 indicates rejected null hypotheses and 1 stand for retained null hypotheses, respectively.
- Step 6: Fill in the off-diagonals of a \( KxK \) identity matrix with these binary outcomes. Then, this matrix becomes a reachability matrix (R-matrix: Leighton, Gierl, & Hunka, 2004) representing all direct and indirect prerequisite relationships among the attributes.

| Attributes | Hypotheses | 000 | 100 | 010 | 001 | 110 | 101 | 011 | 111 | -2LL | Deviance | p-val. | Rej. |
|------------|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------|--------|-------|-----|
| A1→A2     |            | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | 28699.59 | 0.36   | 0.547 | ✓   |
| A1→A3     | ✓          | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | 28699.23 | 0.01   | 0.966 | ✓   |
| A2→A1     | ✓          | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | 29306.97 | 607.74 | 0.000 | ✓   |
| A2→A3     | ✓          | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | 28700.21 | 0.98   | 0.322 | ✓   |
| A3→A1     | ✓          | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | 29985.20 | 1285.96 | 0.000 | ✓   |
| A3→A2     | ✓          | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | 29493.34 | 794.11 | 0.000 | ✓   |
| Full model | ✓          | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | ✓   | 28699.23 |        |       |     |

*Note.* A1→A2 = A1 is prerequisite to A2; p-val. = p-value obtained from the chi-square test with two degrees of freedom; and Rej. = rejection decision.
Implementation of the Exhaustive Search Algorithm

For implementation purposes of the algorithm, consider a linear hierarchical structure among three attributes (i.e., A1→A2, A1→A3, and A2→A3). First attribute (i.e., A1) is prerequisite for the second and the third attributes, and the second attribute is also prerequisite for the third attribute. The Q-matrix used for this demonstration consisted of 20 items measuring single attributes and combination of the attributes. The DINA model parameters (i.e., guessing and slip parameters) were drawn from a uniform distribution U(0.05, 0.30). Response data were generated for 1000 examinees. Then, 3(23−2) = 6 reduced models specified and estimated. Each of these reduced models specifies a prerequisite relationship among a pair of attributes (e.g., A1→A2). The Full model allowing all latent classes is also estimated.

Table 2. Incorporation of the hypothesis testing results into R-matrix

| Identity matrix | A1 | A2 | A3 | R-matrix | A1 | A2 | A3 |
|-----------------|----|----|----|----------|----|----|----|
| A1              | 1  | 0  | 0  | A1       | 1  | 1  | 1  |
| A2              | 0  | 1  | 0  | A2       | 0  | 1  | 1  |
| A3              | 0  | 0  | 1  | A3       | 0  | 0  | 1  |

The model comparison results are given in Table 1. The results indicated that the reduced models conforming the relationships A1→A2, A1→A3, and A2→A3 fitted to the data as good as the full model such that specified null these hypotheses for model comparison were retained. The hypotheses for the remaining model comparisons were rejected. Then, off-diagonals of a K×K identity matrix, which is created in the same order with the hypotheses, were filled with the binary model selection outcomes. The identity matrix then becomes an R-matrix defining hierarchical attribute structure. This procedure is shown in Table 2. The given R-matrix in the table indicates that A1 is prerequisite for both A2 and A3; and A2 is prerequisite for A3. These prerequisite relationships all together define a linear structure among these attributes.

Hierarchical Structure Selection

Domain experts can identify the attributes and hierarchical structure among them. To do so, experts either rely on the literature and existing theories about cognitive process of human performance (Embretson, 1998; Leighton et al., 2004); or they analyze the examinee response data that are directly collected via interview and think-aloud procedures (Chi, 1997; Leighton et al., 2004). Both approaches may be used iteratively where the former approach is used to identify the attributes and the latter is used to validate them (Akbay, Terzi, Kaplan, & Karaaslan, 2017; Tjo & de la Torre, 2014). However, employment of both approaches together may be time consuming and costly.

Two approaches given above may result in different hierarchical structures among measured attributes (see Gierl et al., 2008). Moreover, in either approach, experts may not have a consensus such that more than one hierarchical structure may be proposed. In such cases the optimum structure must be selected to provide the most accurate information regarding the examinees’ attribute-mastery level. The viability of likelihood-ratio approach for hierarchical structure selection, when multiple structures are proposed, is also examined in this study. In this manuscript, the null and alternative models defined based on the hierarchical structures S\(^0\) and S\(^A\), respectively. S\(^0\) subsumes S\(^A\) such that all direct and indirect prerequisite relationships specified in S\(^A\) are also specified in S\(^0\). This relationship further implies that all permissible latent classes defined by S\(^0\) are also in the set of permissible latent classes defined by S\(^A\) (i.e., L\(^0\) ⊂ L\(^A\)). Then, a model allowing L\(^0\) in the estimation becomes the null model and the model allowing L\(^A\) may be regarded as the alternative model to conduct LRT.
Simulation Studies

Design

The viability of the exhaustive search for attribute structure identification and LRT based hierarchical structure selection were assessed with two simulation studies. In the first simulation study, the three general attribute hierarchy types, namely linear, convergent, and divergent hierarchies, consisting six attributes were considered. These structures have been defined by Leighton et al. (2004) and they are displayed in Figure 1. Permissible latent classes allowed by each of these hierarchies are given in the appendix. An unstructured attributes condition was also considered, in which all possible latent classes were allowed. The DINA and DINO models were the two CDMs employed in the simulations. Two different sample sizes (i.e., \(N = 500\) and \(N = 1000\)), two levels of item quality (i.e., higher and lower), four significance levels (i.e., \(\alpha\)-levels), and three model selection criteria (i.e., LRT, AIC and BIC) were considered.

![Figure 1](image1.png)

Figure 1. The linear, convergent, and divergent hierarchical structures defined in Leighton et al., 2004.

In the second simulation study, three hypothetical hierarchical structures (i.e., \(S^1\), \(S^2\), and \(S^3\)) were employed. These structures are presented in Figure 2 and their corresponding permissible latent classes (i.e., \(L^1\), \(L^2\), and \(L^3\)) are given in Appendix. As can be seen from the appendix, \(L^1\) is a subset of \(L^2\), and \(L^2\) is a subset of \(L^3\). This simulation study aimed to assess the viability of the LRT approach for hierarchical structure selection under various conditions. Factors considered in the simulation include the item quality, generating CDM, sample size, and model selection criterion. The Q-matrix that was used in both simulation studies is given in Table 3.

![Figure 2](image2.png)

Figure 2. Three hypothetical attribute structures for six attributes
Table 3. Q-matrix used in the simulations

| Item | A1 | A2 | A3 | A4 | A5 | A6 | Item | A1 | A2 | A3 | A4 | A5 | A6 |
|------|----|----|----|----|----|----|------|----|----|----|----|----|----|
| 1    | 1  | 0  | 0  | 0  | 0  | 0  | 11   | 0  | 0  | 0  | 0  | 1  | 1  |
| 2    | 0  | 1  | 0  | 0  | 0  | 0  | 12   | 1  | 0  | 0  | 0  | 0  | 1  |
| 3    | 0  | 0  | 1  | 0  | 0  | 0  | 13   | 1  | 1  | 1  | 0  | 0  | 0  |
| 4    | 0  | 0  | 0  | 1  | 0  | 0  | 14   | 0  | 1  | 1  | 1  | 0  | 0  |
| 5    | 0  | 0  | 0  | 0  | 1  | 0  | 15   | 0  | 0  | 1  | 1  | 1  | 0  |
| 6    | 0  | 0  | 0  | 0  | 0  | 1  | 16   | 0  | 0  | 0  | 1  | 1  | 1  |
| 7    | 1  | 1  | 0  | 0  | 0  | 0  | 17   | 1  | 0  | 0  | 0  | 1  | 1  |
| 8    | 0  | 1  | 1  | 0  | 0  | 0  | 18   | 1  | 1  | 0  | 0  | 0  | 1  |
| 9    | 0  | 0  | 1  | 1  | 0  | 0  | 19   | 1  | 0  | 0  | 0  | 0  | 0  |
| 10   | 0  | 0  | 0  | 1  | 1  | 0  | 20   | 0  | 0  | 0  | 0  | 0  | 1  |

Note. A1 through A6 = Measured attributes by the test.

Data Generation and Model Estimation

The two-levels of item quality were combined with the other factors considered in the study. For the higher item quality conditions (HQ), the lowest and highest success probabilities (i.e., \(P(0)\) and \(P(1)\)) were drawn from \(U(0.05, 0.20)\) and \(U(0.80, 0.95)\). These two probability parameters are the guessing and 1-slip parameters, respectively. For the lower item quality (LQ) conditions, the lowest and highest success probabilities were drawn from \(U(0.15, 0.30)\) and \(U(0.70, 0.85)\), respectively. The test length and number of measured attributes were fixed to 20-items and six-attributes.

Examinees' attribute profiles followed a uniform distribution of permissible latent classes. The attributes were generated based on the permissible latent classes defined by the linear, convergent, divergent, and unstructured attribute conditions in the first simulation study. In the second simulation, attribute profiles generated based on the permissible latent classes defined by three hypothetical hierarchies given in Figure 2. In both studies, 100 data sets were generated and analyzed for each condition. Item parameters estimated using marginal maximum likelihood (MMML) estimator via expectation-maximization (EM) algorithm. Attribute estimation was carried out based on expected a posteriori (EAP) estimator. Data generation and model estimation were performed through the Oxmetrics programming language (Doornik, 2011). All factors considered in these two simulation studies are presented in Table 4.

Table 4. Simulation factors

| CDM   | Sample size | True structure | Simulation I | Selection criterion | LRT α-level |
|-------|-------------|----------------|--------------|---------------------|-------------|
| DINA  | 500         | Linear         | Higher quality | LRT                | \(\alpha = .01\) |
| DINO  | 1000        | Convergent     | Lower quality  | AIC                 | \(\alpha = .05\) |
|       |             | Divergent      |              | BIC                 | \(\alpha = .10\) |
|       |             | Unstructured   |              |                     | \(\alpha = .20\) |

| CDM   | Sample size | True str. | Simulation I | Selection criterion | Candidate structure |
|-------|-------------|-----------|--------------|---------------------|---------------------|
| DINA  | 500         | \(S^2\)   | Higher quality | LRT                | \(S^1\)               |
| DINO  | 1000        |           | Lower quality  | AIC                 | \(S^2\)               |
|       |             |           |              | BIC                 | \(S^3\)               |

Note. DINA = deterministic input, noisy "and" gate model; DINO = deterministic input, noisy "or" gate model; LRT = likelihood ratio test; AIC = Akaike information criterion; BIC = Bayesian information criterion.
Results and Discussion

Results of Simulation Study I

To assess the viability of the search algorithm, the empirical Type-I and Type-II error rates (i.e., false positive and false negative) were computed. The complement of Type-I error rate (i.e., true-negative) can be referred to as sensitivity that indicates the proportion of retained null hypothesis when it is true. Likewise, the complement of Type-II error rate (i.e., true-positive) may be referred to as specificity that reports the proportion of rejected null hypothesis when it is wrong. Sensitivity and specificity levels of the search algorithm are presented in Table 5. Given proportions of sensitivity and specificity in the table were obtained by averaging across the true and false hypothesized pairwise relationships in each replication. For example, in linear structure, half of the hypotheses we test are true and the other half of them are wrong. Thus, sensitivity and specificity rates are the average of 15 hypothesis test results across 100 replications.

| CDM | N   | IQ | Sig. level | Linear | Convergent | Divergent | Unstructured |
|-----|-----|----|------------|--------|------------|-----------|-------------|
|     |     |    |            | SEN.   | SPE.       | SEN.      | SPE.        |            |
| DINA| 500 | HQ | α = .001   | 1.000  | .993       | .714      | .961        | 1.000       |
|     |     |    | α = .005   | .999   | .998       | .714      | .979        | 1.000       |
|     |     |    | α = .010   | .999   | .999       | .714      | .982        | 1.000       |
|     |     |    | α = .020   | .998   | .999       | .714      | .988        | 1.000       |
| LQ  | α = .001| 1.000  | .867     | .715      | .906      | 1.000     | .939        | .967        |
|     | α = .005| 1.000  | .903     | .714      | .923      | 1.000     | .962        | .986        |
|     | α = .010| 1.000  | .918     | .714      | .926      | 1.000     | .970        | .993        |
|     | α = .020| 1.000  | .931     | .712      | .934      | .998      | .982        | .995        |
|     | 1000 | HQ | α = .001  | 1.000   | 1.000      | .714      | .997        | 1.000       |
|     | α = .005| 1.000  | 1.000     | .714     | .999      | 1.000     | 1.000       | NA 1.000    |
|     | α = .010| 1.000  | 1.000     | .714     | .999      | 1.000     | 1.000       | NA 1.000    |
|     | α = .020| 1.000  | 1.000     | .714     | .999      | 1.000     | 1.000       | NA 1.000    |
| LQ  | α = .001| 1.000  | .941     | .714      | .940      | 1.000     | .992        | .999        |
|     | α = .005| 1.000  | .959     | .714      | .942      | 1.000     | .995        | 1.000       |
|     | α = .010| 1.000  | .965     | .714      | .949      | 1.000     | .998        | 1.000       |
|     | α = .020| 1.000  | .974     | .714      | .954      | 1.000     | .999        | 1.000       |
| DINO| 500 | HQ | α = .001  | 1.000   | .987      | .918      | .971        | 1.000       |
|     | α = .005| 1.000  | .991     | .874      | .982      | 1.000     | .979        | NA 1.000    |
|     | α = .010| 1.000  | .995     | .842      | .987      | 1.000     | .983        | 1.000       |
|     | α = .020| 1.000  | .997     | .811      | .994      | .999      | .988        | NA 1.000    |
| LQ  | α = .001| 1.000  | .999     | .874      | .999      | .789      | 1.000       | .966        |
|     | α = .005| 1.000  | .906     | .993      | .853      | .999      | .796        | .984        |
|     | α = .010| 1.000  | .921     | .990      | .880      | .998      | .838        | NA 0.992    |
|     | α = .020| 1.000  | .932     | .984      | .907      | .996      | .880        | NA 0.995    |
|     | 1000 | HQ | α = .001  | 1.000   | 1.000     | .763      | .996        | 1.000       |
|     | α = .005| 1.000  | 1.000     | .743     | .999      | 1.000     | .996        | NA 1.000    |
|     | α = .010| 1.000  | 1.000     | .730     | .999      | 1.000     | .997        | 1.000       |
|     | α = .020| 1.000  | 1.000     | .724     | .999      | 1.000     | .999        | NA 1.000    |
| LQ  | α = .001| 1.000  | .939     | .996      | .927      | 1.000     | .915        | NA 1.000    |
|     | α = .005| 1.000  | .956     | .986      | .939      | 1.000     | .940        | 1.000       |
|     | α = .010| 1.000  | .963     | .973      | .947      | 1.000     | .954        | 1.000       |
|     | α = .020| 1.000  | .971     | .947      | .960      | 1.000     | .964        | 1.000       |

Note. N = sample size; IQ = item quality; Sig. level = significance level; SEN. = rate of retaining true null hypothesis; SPE. = rate of rejecting false null hypothesis; HQ = higher quality; LQ = lower quality; and NA = not applicable.
Observed sensitivity rates are, in general, close to 1 for linear and divergent hierarchies. The minimum sensitivity rates (i.e., .997 and .996 for the DINA and DINO) were observed under the significance level of .20, the sample size of 500 and lower item quality. Observed sensitivity rates were even smaller in conditions where larger sample and higher item quality employed. In contrary to the results for the linear and divergent cases, reduced sensitivity rates were obtained for the convergent hierarchical structure case. Observed sensitivity in convergent hierarchy case was .714 for the DINA model conditions; whereas sensitivity rates varied from .724 to .999 in DINO model conditions. These low sensitivity rates in the convergent cases may be due to the fact that A5 can be mastered after mastering either A3 or A4. Notice that Type-I error results in additional latent classes to be allowed in the permissible latent class set. In contrast, Type-II error yields in discarding some of the latent classes that conform to the true hierarchical structure. Therefore, adverse impact of conducting Type-II error on model estimation may be much stronger in comparison to the negative impact of conducting Type-I error.

### Table 6. Hypothesis testing results: AIC and BIC

| N   | Hierarchy | IQ | DINA |   | DINO |   | AIC |   | BIC |   |
|-----|-----------|----|------|---|------|---|-----|---|-----|---|
| 500 | Linear    | HQ | SEN. | .999 | SPE. | .992 | SEN. | .907 | SPE. | .987 |
|     |           | LQ |      | .999 |      | .867 |      | .907 |      | .872 |
|     | Convergent| HQ |      | .714 |      | .961 |      | .913 |      | .976 |
|     |           | LQ |      | .715 |      | .906 |      | .997 |      | .791 |
|     | Divergent | HQ |      | 1.000 |      | 1.000 |      | .981 |      |     |
|     |           | LQ |      | 1.000 |      | .938 |      | 1.000 |      | .714 |
|     | Unstructured| HQ |    | NA   |      | 1.000 |    | NA   |      | 1.000 |
|     |           | LQ |    | NA   |      | .966 |    | NA   |      | .966 |
| 1000| Linear    | H  |      | 1.000 |      | 1.000 |      | 1.000 |      |     |
|     |           | LQ |      | 1.000 |      | 1.060 |      | 1.000 |      | .939 |
|     | Convergent| HQ |      | .714 |      | .996 |      | .763 |      | .996 |
|     |           | LQ |      | .714 |      | .940 |      | .996 |      | .928 |
|     | Divergent | HQ |      | 1.000 |      | 1.000 |      | .995 |      |     |
|     |           | LQ |      | 1.000 |      | .992 |      | 1.000 |      | .915 |
|     | Unstructured| HQ |  | NA   |      | 1.000 |   | NA   |      | 1.000 |
|     |           | LQ |  | NA   |      | .999 |  | NA   |      | 1.000 |

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; N = sample size; IQ = item quality; SEN. = rate of retaining true null hypothesis; SPE. = rate of rejecting false null hypothesis; HQ = high quality; LQ = low quality.

Under the larger sample size conditions, the smallest specificity rates for the DINA and DINO models were .940 and .915, respectively. The corresponding rates for the small sample conditions were .867 and .696, respectively. The specificity rates increased significantly and approached to one under the higher quality item conditions. For the larger sample size cases, the reported specificities for all conditions were about and over .940 and .995 for the lower and higher item quality conditions, respectively. Corresponding rates for the small sample cases were reported as .870 and .960, respectively, when the estimating model was the DINA model. When it was the DINO model, these observed specificity rates were comparable to the DINA results under the linear and convergent hierarchies, but the rates were lower in the divergent hierarchy. In terms of the impact of significance level, empirical Type-I error rates were much smaller than the nominal alpha levels for the linear and divergent hierarchical structure cases. However, it was not the case for the convergent hierarchical structure cases.

Table 6 presents the sensitivity and specificity rates obtained with the AIC and BIC model selection criteria. Comparison of the results of the LRT, AIC, and BIC showed that the results of the AIC criterion were
almost identical to the results of the LRT under significance level of .01. Moreover, when the BIC and LRT results were compared, the BIC significantly increased the specificity rates with a smallest specificity rate of .955 across all conditions. However, in return, sensitivity rates slightly decreased (down to .95). When the BIC was employed as the model selection criterion, both sensitivity and specificity rates were above .95 for all conditions of linear, divergent, and unstructured hierarchies. Therefore, BIC may be slightly more informative criterion for detecting direct prerequisite relationships among the attributes.

### Results of Simulation Study II

Simulation results for structure selection are presented in Table 7. The results were presented in terms of null hypothesis rejection rates across 100 replications. Each of the null hypothesis specifies that the more parsimonious model (model considering a more strict hierarchical structure) fits the data as good as the more general model. True hierarchical structures (i.e., generating hierarchical structures) are given in the columns of the table. The candidate hierarchical structures that are used to restrict the models to follow the hierarchies are given in the rows of the table. For example when generating hierarchical structure among the attributes was $S^1$, fit of the model structured by $S^2$ was compared with the model-fits that are obtained by fitting the models structured by $S^2$ and $S^3$.

#### Table 7. Structure selection results for the DINA and DINO models

| N     | SM   | HS  | Higher Quality Items | Lower Quality Items |
|-------|------|-----|----------------------|---------------------|
|       |      |     | DINA $S^1$ $S^2$ $S^3$ | DINA $S^1$ $S^2$ $S^3$ | DINO $S^1$ $S^2$ $S^3$ | DINO $S^1$ $S^2$ $S^3$ |
| 500   | LRT  | $S^1$ | --- 1.00 1.00 | --- 1.00  .96 | --- 1.00 1.00 | --- .95  .48 |
|       |      | $S^2$ | --- .00 --- | .02 --- 1.00 | --- .00 --- | .03 --- .95 |
|       |      | $S^3$ | --- .00 .01 --- | .00 .00 --- | --- .00 .00 --- | .02 .02 --- |
|       | AIC  | $S^1$ | --- 1.00 1.00 | --- 1.00 1.00 | --- 1.00 1.00 | --- .97 .99 |
|       |      | $S^2$ | --- .01 --- | .06 --- 1.00 | --- .01 --- | .06 --- .97 |
|       |      | $S^3$ | --- .00 .01 --- | .00 .02 --- | --- .00 .03 --- | .02 .03 --- |
|       | BIC  | $S^1$ | --- 1.00 1.00 | --- .96 .86 | --- 1.00 1.00 | --- .58 .23 |
|       |      | $S^2$ | --- .00 --- | .00 --- .93 | --- .00 --- | .00 --- .51 |
|       |      | $S^3$ | --- .00 .00 --- | .00 .00 --- | --- .00 .00 --- | .00 .00 --- |
| 1000  | LRT  | $S^1$ | --- 1.00 1.00 | --- 1.00  .96 | --- 1.00 1.00 | --- 1.00 1.00 |
|       |      | $S^2$ | --- .00 --- | .02 --- 1.00 | --- .00 --- | .00 --- 1.00 |
|       |      | $S^3$ | --- .00 .00 --- | .02 .02 --- | --- .00 .00 --- | .00 .00 --- |
|       | AIC  | $S^1$ | --- 1.00 1.00 | --- 1.00 1.00 | --- 1.00 1.00 | --- 1.00 1.00 |
|       |      | $S^2$ | --- .00 --- | .04 --- 1.00 | --- .00 --- | .03 --- 1.00 |
|       |      | $S^3$ | --- .00 .00 --- | .01 .03 --- | --- .00 .01 --- | .00 .03 --- |
|       | BIC  | $S^1$ | --- 1.00 1.00 | --- 1.00 1.00 | --- 1.00 1.00 | --- .92 .74 |
|       |      | $S^2$ | --- .00 --- | .00 --- 1.00 | --- .00 --- | .00 --- .87 |
|       |      | $S^3$ | --- .00 .00 --- | .00 .00 --- | --- .00 .00 --- | .00 .00 --- |

Note. N = sample size; SM = selection method; HS = hierarchical structure; LRT = likelihood ratio test; AIC = Akaike information criterion; and BIC = Bayesian information criterion.

Under the sample size of 500 and higher quality items, rejection rates of $S^1$ were .00 and .00 in favor of $S^2$ and $S^3$, respectively. In other words, all the null hypotheses were retained such that structured model based on $S^1$ was fitted to the data as good as the structured models consistent with hierarchies $S^2$ and $S^3$. Similarly, when generating hierarchy was $S^2$, fit of the model structured by $S^2$ was compared against the models structured by $S^3$ and $S^3$. Fit of the model based on $S^1$ was rejected 100% of the time in favor of the model consistent with $S^2$. Furthermore, the model conforming $S^2$ was rejected 0% of the time in favor of the model set by $S^3$. Thus, the true model was selected 100% of the time.
Table 7 shows that regardless of the sample size and item quality levels, all three selection-criteria (i.e., LRT, AIC, and BIC) almost always selected the true hierarchy when the various structured versions of the DINA model were fitted. Under the DINO model conditions, all three model-selection methods detected the generating structure at and above 95% of the time under higher sample size and higher item quality conditions. However, performance of the model selection criteria, especially the performance of the BIC, significantly decreased with decrease in the item quality. Results indicated further reduction in detection success of the true hierarchical structure under the smaller sample sizes.

Results given in tables 7 suggest that true hierarchical structure can be detected by all three model-selection methods, especially when items are in higher quality and the sample size is larger than 500 examinees. Results showed that when the generating model was the DINA model, performance of model-selection methods in detection of the true hierarchical structure was quite high even when the item quality was lower. Although the observed results show that hierarchical structure selection for the DINO model cases were not be as accurate, the observed differences under the DINA and DINO models could be due to the Q-matrix used in this study.

**Real Data Analysis**

A numerical example for hierarchy detection and selection was also conducted. The dataset consists of 2922 examinees’ binary responses to the 28 items. These are the examinee responses to the grammar section of the Examination for the Certificate of Proficiency in English (ECPE). The test was developed and administered by the University of Michigan English Language Institute in 2003. The response data and the Q-matrix are available in and obtained from the ‘CDM’ package (Robitzsch, Kiefer, George, & Uenlue, 2014) in R software environment for statistical computing. This test with 28 items measures three attributes that are referred to as A1: lexical rules, A2: cohesive rules, and A3: morphosyntactic rules.

**Table 8. Attribute hierarchy search on ECPE**

| Null hypothesis | A1 | A2 | A3 |
|-----------------|----|----|----|
| Deviance        |     |     |    |
| p-values        |     |     |    |
| AIC             | 85705.47 | 85705.76 | 85705.99 |
| BIC             | 85753.31 | 85753.60 | 85753.62 |

Note. A1 = lexical rules; A2 = cohesive rules; A3 = morphosyntactic rules; AIC = Akaike information criterion; and BIC = Bayesian information criterion.

The exhaustive search algorithm was employed to detect hierarchical relationships among these three attributes. Model selection results obtained through the LRT, AIC, and BIC were summarized in Table 8. The LRT and AIC based model selection results indicated that null models assuming A1 was prerequisite to A2, and A1 was prerequisite to A3 fitted to the data as well as the full model (i.e., unstructured model that allows all latent classes in the estimation). Furthermore, on top of the two null models retained by the LRT and AIC, BIC has retained another null model, which has assumed that A2 was a prerequisite attribute for A3.

Therefore, when the model selection criterion was either LRT or AIC, exhaustive search algorithm resulted in a divergent hierarchical structure such that A1 was prerequisite for both A2 and A3. However, BIC resulted in a linear hierarchical structure A1 is prerequisite attribute for A2, which, in turn, is a prerequisite attribute for A3. In linear and divergent attribute structures, rather than all eight (i.e., $2^3$) latent classes, four and five latent classes are permissible, respectively. The permissible latent classes in the linear hierarchical structure are: 000, 100, 110, and 111. The latent class 101 is also permissible in the divergent hierarchical structure. Then, comparing the likelihoods of these two computing hierarchically structured models may help
us to choose one or the other. When the fit of the models structured by these two attribute hierarchies were compared, the linear hierarchy was rejected in favor of the divergent hierarchy ($p$-value = .000 and degrees of freedom = 1).

**Conclusion**

In conditions where measured attributes hold a hierarchical structure, cognitive diagnosis model estimation may be improved by incorporating the hierarchical attribute structure into the estimation process. To benefit from this incorporation, attribute hierarchy must be correct. Otherwise, considering incorrect assumptions on the hierarchical relationships in the model estimation process may degrade accuracy of obtained information. This study proposes an empirical exhaustive search algorithm to detect hierarchical relationships among the attributes. The viability of the algorithm is also investigated in this study under various conditions. For each of all possible pairwise prerequisite relationships, the search algorithm tests the fits of the hierarchically structured CDMs against the unrestricted CDM. Hierarchically structured CDMs do not allow impermissible latent classes in the estimation such that these CDMs are nested within the unstructured one.

Results of the study indicated that the likelihood ratio test based exhaustive search yields an R-matrix that specifies all the prerequisite relationships among the attributes forming a linear or divergent hierarchy. It only fails to recover a convergent hierarchical structure where two or more attributes are prerequisite for an attribute. In such cases, mastering one of the several prerequisite will allow examinee to master the more complex attribute. In other words, when there is more than one way of being able to master the more complex attribute, then the search algorithm may have hard time detecting these prerequisite relationships. In cases where attributes have a convergent structure, the exhaustive search yields a more liberal structure, in which some truly impermissible latent classes are added on top of true permissible latent class set. Even in such cases, the exhaustive search eliminates many of the non-existing latent classes. It should be noted here that, even if the method is promising, it should be used to complement method rather than to replace the conventional subjective procedures. Because the exhaustive search is computationally intensive it may be possible to develop more efficient algorithms. One way of accomplishing this requires fixing the prerequisite relationship, when found, for the rest of the search. Then, the number of remaining possible pairwise prerequisite relationships to be checked will be reduced.

The secondary purpose of this study was to assess the viability of model-fit based hierarchical structure selection. The study results showed that, through the likelihood model selection criteria, a model structured by the generating attribute hierarchy could be selected accurately when the several candidates are present. Under high item quality conditions, correct hierarchy selection rates of LRT, AIC, and BIC may be % or even higher. In practice, when candidate hierarchies are nested, LRT may be used for model selection purposes. However, when the candidates are not nested, practitioner can always consider the AIC and BIC results to for the final decision.

Although some factors, which may have impact on the performance of the search algorithm, are considered in the current study, more factors such as test lengths and number of measured attributes may also be considered. Thus, considering only a fixed test length, fixed number of measured attributes, and one single Q-matrix is among the limitations of the study. Furthermore, investigation of impact of misspecifications in Q-matrix on the performance of search algorithm and hierarchy selection may also be instructive. Therefore, investigation of the impact of misspecifications in Q-matrices may be the next step.
REFERENCES

Akbay, L., & de la Torre, J. (2015, April). Effect of Q-matrix design under hierarchical attribute structures. Poster presented at the meeting of National Council on Measurement in Education, Chicago, IL - USA.

Akbay, L., Terzi, R., Kaplan, M., & Karaaslan, K. G. (2017). Expert-based attribute identification and validation on fraction subtraction: A cognitively diagnostic assessment application. Journal on Mathematics Education, 8(1), 103-120.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. The journal of the learning sciences, 6, 271-315. https://doi.org/10.1007/s11336-009-9125-0

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. Journal of Educational Measurement, 46, 429-449.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. Applied Psychological Measurement, 33, 163-183. https://doi.org/10.1177/0146621608320523

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. Journal of Educational and Behavioral Statistics, 34, 115-130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. Psychometrika, 76, 179-199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. Psychometrika, 69, 333-353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Minchen, N. (2014). Cognitive diagnostic assessments and the cognitive diagnosis model framework. Psicologia Educativa, 20, 89-97. https://doi.org/10.1016/j.pse.2014.11.001

Doornik, J. A. (2011). Object-oriented matrix programming using Ox (Version 6.20) [Computer software]. London: Timberlake Consultants Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods, 3, 380-396. http://dx.doi.org/10.1037/1082-989X.3.3.380

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. Journal of Technology, Learning, and Assessment, 6(6), 1-50.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. Applied Psychological Measurement, 25, 258-272. https://doi.org/10.1177/01466210122032064

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. Journal of Educational Measurement, 41, 205-237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Linn, M. C., Eylon, B. S., & Davis, E. A. (2004). The knowledge integration perspective on learning. In: M.C. Linn, E.A. Davis, & P. Bell (Eds.), Internet environments for science education (pp. 29-46). Mahvah, NJ: Lawrence Erlbaum Associates.

Maris, E. (1995). Psychometric latent response models. Psychometrika, 60, 523-547. https://doi.org/10.1007/BF02294327

Maris, E. (1999). Estimating multiple classification latent class models. Psychometrika, 64, 187-212. https://doi.org/10.1007/BF02294535

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2019). CDM: Cognitive diagnosis modeling [Computer software]. Retrieved from https://CRAN.R-project.org/package=CDM

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. Measurement, 6, 219-262. https://doi.org/10.1080/15366360802490866

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of US
mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, 37, 525-559.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective*, 4, 1-98.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79, 317-339. https://doi.org/10.1007/s11336-013-9362-0

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, 11(3), 287-305. http://dx.doi.org/10.1037/1082-989X.11.3.287

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, 26, 237-255.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive psychology*, 24, 535-585.
Appendix. Permissible latent classes defined by the hierarchical structures

| Latent class | Attributes | Hierarchical Structures | Permissible Classes |
|--------------|------------|------------------------|---------------------|
| a1           | 0          | 0                      | ✔️                  |
| a2           | 0          | 0                      | ✔️                  |
| a3           | 0          | 0                      | ✔️                  |
| a4           | 0          | 0                      | ✔️                  |
| a5           | 0          | 0                      | ✔️                  |
| a6           | 0          | 0                      | ✔️                  |
| a7           | 0          | 0                      | ✔️                  |
| a8           | 0          | 0                      | ✔️                  |
| a9           | 0          | 0                      | ✔️                  |
| a10          | 0          | 0                      | ✔️                  |
| a11          | 0          | 0                      | ✔️                  |
| a12          | 0          | 0                      | ✔️                  |
| a13          | 0          | 0                      | ✔️                  |
| a14          | 0          | 0                      | ✔️                  |
| a15          | 0          | 0                      | ✔️                  |
| a16          | 0          | 0                      | ✔️                  |
| a17          | 0          | 0                      | ✔️                  |
| a18          | 0          | 0                      | ✔️                  |
| a19          | 0          | 0                      | ✔️                  |
| a20          | 0          | 0                      | ✔️                  |
| a21          | 0          | 0                      | ✔️                  |
| a22          | 0          | 0                      | ✔️                  |
| a23          | 0          | 0                      | ✔️                  |
| a24          | 0          | 0                      | ✔️                  |
| a25          | 0          | 0                      | ✔️                  |
| a26          | 0          | 0                      | ✔️                  |
| a27          | 0          | 0                      | ✔️                  |
| a28          | 0          | 0                      | ✔️                  |
| a29          | 0          | 0                      | ✔️                  |
| a30          | 0          | 0                      | ✔️                  |
| a31          | 0          | 0                      | ✔️                  |
| a32          | 0          | 0                      | ✔️                  |

© 2019 International Online Journal of Educational Sciences (IOJES)