The Impact of Q-Matrix Designs on Diagnostic Classification Accuracy in the Presence of Attribute Hierarchies

Educational and Psychological

Measurement

2017, Vol. 77(2) 220–240

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416645636

journals.sagepub.com/home/epm

Ren Liu¹, Anne Corinne Huggins-Manley¹,
and Laine Bradshaw²

Abstract

There is an increasing demand for assessments that can provide more fine-grained information about examinees. In response to the demand, diagnostic measurement provides students with feedback on their strengths and weaknesses on specific skills by classifying them into mastery or nonmastery attribute categories. These attributes often form a hierarchical structure because student learning and development is a sequential process where many skills build on others. However, it remains to be seen if we can use information from the attribute structure and work that into the design of the diagnostic tests. The purpose of this study is to introduce three approaches of Q-matrix design and investigate their impact on classification results under different attribute structures. Results indicate that the adjacent approach provides higher accuracy in a shorter test length when compared with other Q-matrix design approaches. This study provides researchers and practitioners guidance on how to design the Q-matrix in diagnostic tests, which are in high demand from educators.

Keywords

diagnostic measurement, test design, Q-matrix, hierarchical diagnostic classification model, classification accuracy, attribute structure

¹University of Florida, Gainesville, FL, USA

²University of Georgia, Athens, GA, USA

Corresponding Author:

Ren Liu, Department of Research and Evaluation Methodology, University of Florida, 170 HUB, Gainesville, FL 32611, USA

Email: liurenking@ufl.edu

Introduction

Personalized learning is named as a top priority by the U.S. Department of Education (2014). The U.S. Department of Education emphasizes the need for tailored instructional improvements informed by K-12 assessments that provide personalized feedback on the strengths and weaknesses of students on specific learning objectives, or attributes. Tests that are diagnostic in nature can fill this need by giving teachers and education stakeholders more fine-grained information about students than can be obtained from standardized tests. These types of tests can be developed under a diagnostic classification model (DCM) framework, which entails designing a test for the purpose of classifying examinees into categorical attribute-level variables.

One of the most important aspects of designing a diagnostic test is determining which items measure which attributes. This process can be called Q-matrix design, because a Q-matrix in the DCM framework is a matrix with rows as items, columns as attributes, and elements as binary indicators of which items measure which attributes (Tatsuoka, 1983). Q-matrix designs can be influenced by (a) the number of attributes and items and (b) the number of nonzero entries in the Q-matrix (Madison & Bradshaw, 2015). Three studies that have addressed the effects of Q-matrix design on classification results of diagnostic tests are Chiu, Douglas, and Li (2009), DeCarlo (2011), and Madison and Bradshaw (2015). The first article showed that each attribute needs to be measured by at least one item that does not measure any other attributes in order to obtain acceptable classification accuracy in both the *deterministic-input, noisy-and-gate* (Haertel, 1989) model and the *deterministic-input, noisy-or-gate* (Templin & Henson, 2006) model. The second article investigated the deterministic-input, noisy-and-gate model and suggested that if an attribute is always measured through interaction terms and never measured in isolation, the classification obtained only reflects the prior probabilities. The last article examined the effects of Q-matrix design in the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) and concluded that attributes measured in isolation can help increase classification accuracy when holding constant the number of times an attribute is measured on a test. To sum, there is a consensus that measuring attributes in isolation produces more accurate attribute classifications under conditions in which attributes are not hierarchically structured. It is not known whether the results of these studies generalize to attributes with hierarchical structures.

The attributes measured within one diagnostic test may be dependent on each other in specific ways and can be expected to have some form of a hierarchical relationship. The hierarchy indicates that there are some higher-order attributes that can only be mastered if lower-order attributes are first mastered (Templin & Bradshaw, 2014). Although attribute hierarchies are anticipated in many areas of learning theory (e.g., Dahlgren et al., 2006; Jimoyiannis & Komis, 2001; Simon & Tzur, 2004), approaches for incorporating those hierarchies in the DCM are only recently being studied, and much is unknown. Specific to this study, the effects of Q-matrix design choices on classification results under different attribute structures are unknown.

The purpose of this study is to introduce three approaches of Q-matrix design and examine their effects on diagnostic classification accuracy when attribute hierarchy is present. The primary research inquiry is to discover which approaches can be used to load items on attributes and whether they provide a satisfactory classification result. Consideration is given to the effects that other factors may have on the results of this inquiry, including the type of attribute hierarchy, the number of items, the number of times each attribute is measured, and item quality. This study intends to inform researchers and practitioners of Q-matrix design strategies at two different stages. If one hypothesizes that attributes are hierarchically structured at the beginning of test development, the results from our study can be used to design the Q-matrix. After the initial data are collected, one needs to validate whether the hypothesized hierarchy exists, and the Q-matrix design approaches we introduce are helpful for informing practice once a hierarchy is detected in the data.

The remainder of this article is organized as follows. First, we introduce two mainstays of the Q-matrix design: the attribute structure and the item–attribute alignment (i.e., the Q-matrix). Then, three approaches of Q-matrix design are presented, followed by the DCM we used to examine the three approaches. A simulation study is then conducted to examine the classification accuracy and reliability associated with different Q-matrix designs, where different hierarchical structures interact with different approaches and test lengths. Discussions and practical guidelines are provided in the end.

Background

In diagnostic measurement, two factors that influence the Q-matrix design are the specification of attribute structures and item–attribute alignment. The specification of attribute structure refers to the process of specifying the number of attributes and the relationships among attributes. The specification of item–attribute alignments refers to correctly identifying which items are intended to measure which attributes. In this article, we assume that both the attribute structures and the Q-matrix are correctly specified so that we can isolate the effects of the design of the Q-matrix on classification. The remainder of this section summarizes literature on attribute and Q-matrix specification, as it relates to the Q-matrix designs.

Attribute Structures

When we look into students' learning trajectories, it is not unusual to observe that learning is a sequential process, with successive stages of learning that build on each other. One can view the skills that students learn as hierarchical because the mastering of some skills is a prerequisite for the mastering of other skills. The specified attribute hierarchies in diagnostic tests are formalizations of these attribute dependencies. Leighton, Gierl, and Hunka (2007) proposed four types of attribute hierarchies, *linear*, *divergent*, *convergent*, and *unstructured*, as illustrated in Figure 1. In a

Figure 1. Four hierarchical structures using five attributes.

linear hierarchy (L), all five attributes are sequentially ordered in one single chain. Therefore, examinees who have mastered Attribute 5 are expected to have mastered all the preceding attributes (i.e., Attributes 1-4). More generally, mastery of a higher-order attribute assumes mastery of all lower attributes. In a *divergent hierarchy* (D), multiple branches diverge from a common parent attribute. Thus, examinees who have mastered Attribute 5 are expected to have mastered all the preceding attributes on that specific branch (i.e., Attributes 1 and 2). In a *convergent hierarchy* (C), multiple parent attributes converge to a common attribute. Thus, examinees who have mastered Attribute 5 are expected to have mastered one or more of the preceding attributes (i.e., Attributes 1, 2, and 3; Attributes 1, 2, and 4; or Attributes 1, 2, 3, and 4). In an *unstructured hierarchy* (U), one attribute precedes multiple distinct attributes. Thus, examinees who have mastered Attribute 5 are expected to have mastered the parent attribute (i.e., Attribute 1). For educational assessment data originating from a hierarchical learning structure, traditional nonhierarchical DCMs may overfit the data by allowing examinees to master an attribute without mastering the preceding attributes. For example, it would be problematic to assume that examinees master Attribute 5 without mastering Attribute 1 when any of the attribute hierarchies in Figure 1 are present. Previous study has examined the effects of multiple facets of attribute structures on classification accuracy and found that higher-level attributes have higher classification accuracy because there is more information about those attributes from the hierarchical structure (Liu & Huggins-Manley, 2016).

Q-Matrix Specification

The relationship between each item and single or multiple latent attributes, as specified in the Q-matrix, is the other critical component of a Q-matrix design. Suppose a test intends to measure A attributes (i.e., $a = 1, 2, \dots, A$) with I items (i.e., $i = 1, 2, \dots, I$), the item–attribute associations are indicated in a binary $I \times A$ matrix, $\mathbf{Q} = \{q_{ia}\}$. The q_{ia} entries indicate whether or not the i th item loads on the a th attribute. The

1. Example of Independent Approach						2. Example of Adjacent Approach						3. Example of Reachable Approach					
	a1	a2	a3	a4	a5		a1	a2	a3	a4	a5		a1	a2	a3	a4	a5
Item1	1	0	0	0	0	Item1	1	1	0	0	0	Item1	1	0	0	0	0
Item2	0	1	0	0	0	Item2	0	1	1	0	0	Item2	1	1	0	0	0
Item3	0	0	1	0	0	Item3	0	0	1	1	0	Item3	1	1	1	0	0
Item4	0	0	0	1	0	Item4	0	0	0	1	1	Item4	1	1	1	1	0
Item5	0	0	0	0	1	Item5	0	0	0	0	1	Item5	1	1	1	1	1
Item6	1	0	0	0	0	Item6	1	1	0	0	0	Item6	0	0	0	0	1
Item7	0	1	0	0	0	Item7	0	1	1	0	0	Item7	0	0	0	1	1
Item8	0	0	1	0	0	Item8	0	0	1	1	0	Item8	0	0	1	1	1
Item9	0	0	0	1	0	Item9	0	0	0	1	1	Item9	0	1	1	1	1
Item10	0	0	0	0	1	Item10	0	0	0	0	1	Item10	1	1	1	1	1
...						...						...					

Figure 2. Example of three Q-matrix design approaches under a linear hierarchy.

construction of the Q-matrix is usually handled by domain experts and is subjective in nature, and validity concerns have been addressed by many researchers (e.g., de la Torre, 2008). In fact, most previous research on Q-matrices examined the impact of misspecification on various outcomes (e.g., de la Torre, 2008; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp & Templin, 2008). The specification of the Q-matrix usually reflects a particular representation of cognitive processes and learning theories and therefore encompasses issues of content validity in DCMs (Boorsboom & Mellenberg, 2007).

Theoretical Framework

Three Approaches for Designing Q-Matrices

Researchers have introduced frameworks for conceptualizing the relationships among attributes (Leighton et al., 2004; Liu & Huggins-Manley, 2016), but none has imposed these relationships a priori in the design of the Q-matrix. In this section, we introduce three approaches to designing items to fill the Q-matrix in a way that mirrors their interrelationships: the *independent approach*, the *adjacent approach*, and the *reachable approach*. Figure 2 shows an example of how to use the three approaches in designing a Q-matrix. All the designs presented in this article are seeking to be balanced. A balanced Q-matrix design measures each attribute with the same number and types (i.e., single-attribute items or multiple-attribute items) of items. For the independent approach, each item measures only one attribute. This is an extreme form of isolating attributes, a design we use to serve as a contrast to the other two Q-matrix designs that we introduce, both of which contain items that measure more than one attribute.

To design Q-matrices under the *adjacent* and *reachable approaches*, we first considered the systematic ways in which hierarchical attributes may be interrelated. In any hierarchical structure, some attributes are directly connected, whereas others are indirectly connected through a chain. For example, in Figure 1 in the linear hierarchy,

α_1 and α_2 are directly connected, whereas α_1 and α_3 are indirectly connected. Tatsuoaka (1983, 2009) introduced two frameworks for representing these relationships: the adjacency and the reachability matrices. These matrices are specified with the full list of measured attributes as both rows and columns. The elements in the adjacency matrix (*A*-matrix) represent the direct relationships among attributes, where an entry of 1 as opposed to 0 indicates that the attribute listed in the row is a direct prerequisite of the attribute listed in the column. The elements in the reachability matrix (*R*-matrix) represent both direct and indirect relationships among attributes, so an entry of 1 as opposed to 0 indicates that the attribute listed in the row is a prerequisite for the attribute listed in the column, even if the relationship is indirect rather than direct.

To impose these relationship rules onto the design of a Q-matrix, we introduce the adjacent approach and the reachable approach, as seen in the middle and right tables in Figure 2. Under the adjacent approach, each item is loaded on at most two adjacent (i.e., directly connected) attributes, where one attribute is a prerequisite for the other. While one could possibly allow items to load onto more than two attributes that are directly connected (e.g., to mimic the unstructured hierarchy shown in Figure 1), we did not allow this in our approach so that it could be fully distinguished from the reachable approach. In the reachable approach, each item can measure all attributes that are both directly and indirectly connected with that attribute. The reachable approach extends the adjacent approach by allowing the largest number of attributes one item can measure to be equal to the number of attributes on the longest branch. In this study, we strictly restricted the two approaches to their extreme forms to show their differences (i.e., loading at most two adjacent attributes for the adjacent approach and loading all reachable attributes for the reachable approach). In practice, the adjacent approach is implemented when each item measures at most two attributes, while having some items measure one attribute in isolation.

One salient aspect differentiating the three approaches is that the number of times each attribute can be measured varies as a function of the approach used. For example, in Figure 2 suppose we have five attributes that are linearly structured, where α_1 is at the lowest level and α_5 is at the highest level. Using the independent approach, each attribute can be measured two times within 10 items if a balanced Q-matrix design is used. For example, Items 1 and 6 measure α_1 , and Items 2 and 7 measure α_2 . Using the adjacent approach, four out of the five attributes can be measured four times within 10 items because each item measures at most two attributes. For example, Items 1, 2, 6, and 7 measure α_2 , and Items 2, 3, 7, and 8 measure α_3 . Using the reachable approach, each attribute can be measured around six times within 10 items because each item can measure a maximum of five attributes. For example, Items 1, 2, 3, 4, 5, and 10 measure α_1 , and Items 2, 3, 4, 5, 9, and 10 measure α_2 . It can be seen that within a fixed test length, the reachable approach allows for more times of measurement of the attributes than does the adjacent approach, which in turn allows for more times of measurement of the attributes than does the independent approach.

This study was in part motivated by the belief that loading items on adjacent attributes yields higher classification accuracy than loading items on two randomly chosen

attributes because the hierarchical structure of the attributes helps eliminate some of the mastery status that is unlikely to happen (i.e., mastery of an attribute without concomitant mastery of its prerequisite attribute). For example, suppose we want to load two items on three attributes: α_1 , α_2 , and α_3 . One item is loaded on α_1 and α_2 , and the other item is loaded on α_2 and α_3 . A score of 2 indicates correct answers on both items, a score of 1 indicates one correct answer on one of the two items, and a score of 0 indicates incorrect answers on both items. If attributes are randomly combined, which means α_1 , α_2 , and α_3 are not assumed to be hierarchically related, two possible events can result in a score of 1: The examinee has mastered α_1 and α_2 or has mastered α_2 and α_3 . Four possible events can result in a score of 0: The examinee has mastered only α_1 , has mastered only α_2 , has mastered only α_3 , or has not mastered any of them. However, if the attributes we load on the two items are adjacent, meaning α_1 is a prerequisite for α_2 and α_2 is a prerequisite for α_3 , half of the events can no longer be expected to happen, and we can work this expectation into the design and modeling. An examinee can get a score of 1 only if both α_1 and α_2 are mastered. A score of 0 can mean that an examinee has mastered either α_1 only or none of the attributes.

The Hierarchical Diagnostic Classification Model

To compare the classification results using the three item-loading approaches under the four structures, we use the hierarchical diagnostic classification model (HDCM; Templin & Bradshaw, 2014). HDCM is a general DCM that flexibly models attribute effects and interactions on individual items. It is nested within the full LCDM, where redundant parameters are set to zero (Templin & Bradshaw, 2014). It assumes that the nonhierarchical DCMs overfit the data by allowing for the separation of latent classes that would be combined under a hierarchical structure.

Suppose an item measures α_1 and α_2 , where α_1 is a prerequisite for α_2 . The item response function under the LCDM for an examinee e on item i is

$$P(y_{ei} = 1 | \alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(2)}\alpha_{e2} + \lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2})}{1 + \exp(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(2)}\alpha_{e2} + \lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2})}, \quad (1)$$

and the item response function under the HDCM model is

$$P(y_{ei} = 1 | \alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,2,(2(1))}\alpha_{e1}\alpha_{e2})}{1 + \exp(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,2,(2(1))}\alpha_{e1}\alpha_{e2})}. \quad (2)$$

In both equations, α_c denotes the attributes in profile c ($\alpha_c = [\alpha_1, \alpha_2, \dots, \alpha_A]$) with an entry of 1 for every attribute that examinees in profile c have mastered and an entry of 0 for every attribute that examinees have not mastered. $\lambda_{i,0}$ is an intercept parameter representing the logit of a correct response where all entries in the examinee's α_c equal 0. $\lambda_{i,1,(1)}$ is the main effect associated with α_1 , $\lambda_{i,1,(2)}$ is the main effect associated with α_2 . $\lambda_{i,2,(1,2)}$ and $\lambda_{i,2,(2(1))}$ are the two-way interaction effects associated with α_1 and α_2 in the LCDM and HDCM, respectively. The primary difference between

Equations 1 and 2 is that the HDCM does not have main effects for nested attributes. For example, in this item, the HDCM does not include a main effect for α_2 because that effect would refer to an increase in the log-odds of a correct response for a class of examinees—examinees who have mastered α_2 but not α_1 —that does not exist in the attribute hierarchy being modeled by the HDCM. In the adjacent approach, we limit the maximum number of attributes one item can measure to two. Therefore, only the intercept, the main effect for the parent attribute, and the interaction effect are estimated. We also want to point out that for each additional attribute added to the Q-matrix, the number of model parameters increases exponentially, and it may not be feasible to recover the parameters based on the observed data, in which case one may find that the models are unidentifiable (Xu & Zhang, 2016) under the reachable approach.

Before closing this section, we want to draw readers' attention to the two-step process of detecting attribute hierarchy, as designing the Q-matrix and fitting the HDCM belong to each of the two-steps, respectively. First, if one hypothesizes that an attribute hierarchy exists, one could use the results from this study to inform Q-matrix design and item development. Then, after data are collected, one should fit both the LCDM and the HDCM to compare the model fit (Hu et al., 2016; Sinharay & Almond, 2007; Templin & Bradshaw, 2014). If the more parsimonious HDCM has the better fit, the a priori hypothesized hierarchy can be confirmed. If the HDCM does not fit well to the data, one may consider revising the attribute structure. If the assumed hierarchy does not exist, one should proceed using the results obtained from the LCDM and/or consider redesigning the Q-matrix following the suggestions of Madison and Bradshaw (2015). The Q-matrix design approaches we introduce are helpful for informing practice once a hierarchy is detected in the data.

Method

To explore the effects of Q-matrix design on classification using the three item loading approaches, we introduced 18 Q-matrix designs in a simulation study. For each of these Q-matrix designs, items were loaded on attributes in isolation (the *independent approach*), in combination with adjacent attributes (the *adjacent approach*), or with reachable attributes (the *reachable approach*). The number of attributes was fixed at five, which helped form a hierarchy and remain realistic in test design. Table 1 provides a summary of each simulation condition. Each condition was labeled with the letter of the hierarchical structure (i.e., *L*, *D*, *C*, or *U*), the letter of the loading approach (i.e., *I*, *A*, or *R*), and the test length (e.g., 30). Across the three approaches, we developed two sets of conditions.

In the first set of conditions, the test length was fixed at 30. This test length allowed each of the five attributes to be measured *at least* six times. According to Madison and Bradshaw (2015), each attribute should be measured at least five times to achieve acceptable classification accuracy. Under this set of conditions, we investigated the classification results of three item-loading approaches under four hierarchies while holding the number of test items constant. For example, L-I30 represents the design

Table 1. Q-Matrix Conditions in the Simulation Design.

#	CN	Description	MI
1	L-I30	Load each item on one attribute. To measure each item six times, the minimum number of items needed under each of the four structures is 30 using the independent approach.	30
2	D-I30		
3	C-I30		
4	U-I30		
5	L-A30	Load two attributes that are directly connected in the attribute structure where one attribute is a prerequisite for the other one. Those attributes that have not been measured a certain number of times can be measured with another attribute or in isolation. The minimum number of items needed under the linear hierarchy is 15 using the adjacent approach.	15
6	L-A15		
7	L-R30	Each item takes all attributes that are both directly and indirectly connected with one attribute. We start to load items on attributes from α_1 to α_5 , and from α_5 to α_1 . The minimum number of items needed under the linear hierarchy is 10 using the reachable approach.	10
8	L-R10		
9	D-A30	Load two attributes that are directly connected in the attribute structure where one attribute is a prerequisite for the other one. Those attributes that have not been measured a certain number of times can be measured with another attribute or in isolation. The minimum number of items needed under the divergent hierarchy is 22 using the adjacent approach.	22
10	D-A22		
11	D-R30	Each item takes all attributes that are both directly and indirectly connected with one attribute. We start to load items on attributes from α_1 to α_5 and from α_5 to α_1 . For those attributes that have not been measured six times, they are measured with reachable attributes or in isolation. The minimum number of items needed under the divergent hierarchy is 16 using the reachable approach.	16
12	D-R16		
13	C-A30	Load two attributes that are directly connected in the attribute structure where one attribute is a prerequisite for the other one. Those attributes that have not been measured a certain number of times can be measured with another attribute or in isolation. The minimum number of items needed under the convergent hierarchy is 18 using the adjacent approach.	18
14	C-A18		
15	C-R30	Each item takes all attributes that are both directly and indirectly connected with one attribute. For those attributes that have not been measured six times, they are measured with reachable attributes or in isolation. The minimum number of items needed under the convergent hierarchy is 13 using the reachable approach.	13
16	C-R13		
17	U-A30	Load two attributes that are directly connected in the attribute structure where one attribute is a prerequisite for the other one. The minimum number of items needed under the unstructured hierarchy is 24 using the reachable approach.	24
18	U-A24		

Note. CN = condition name; MI= minimum number of items.

under a linear structure where the independent approach is used and the number of items is fixed at 30, and D-A30 represents the design under a divergent structure where the adjacent approach is used and the number of items is fixed at 30.

In the second set of conditions, we fixed the number of times each attribute is measured to six. Controlling the number of times that each attribute was measured resulted in a minimum number of items required under each condition. For example, the minimum number of items required under the linear hierarchy was 15 when using the adjacent approach and 10 when using the reachable approach. In this set of conditions, the independent approach was not considered because the number of items required for measuring each attribute six times was equal to 30, the same as in the first set of conditions. There is a special case in condition U-A30 where α_1 was measured more than six times because of the nature of the unstructured hierarchy. The conditions with a smaller number of items were included in this study as an extreme situation for which to compare others. We aim to (a) keep the number of times each attribute is measured constant to make a fair comparison and (b) also look at the impact of the test length on different designs and hierarchical structures.

In total, we specified 18 Q-matrix designs using the three approaches. Designs 1 to 4 (L-I30, D-I30, C-I30, and U-I30) explored the effects of the independent approach when the attribute hierarchy was linear, divergent, convergent, and unstructured, respectively. Designs 5 to 16 were constructed to examine the adjacent approach and reachable approach under linear, divergent, and convergent structures. Designs 17 and 18 were constructed to examine the adjacent approach under the unstructured hierarchy. Two designs were sufficient for this because the loading items on adjacent or reachable attributes were identical under this hierarchy. In sum, under each of the linear, divergent, and convergent hierarchies, five conditions were examined, and under the unstructured hierarchy, three conditions were examined. Due to space limitations, the specific Q-matrices we used under each condition are not presented in this article, but they are available on request from the first author.

In addition to manipulating the Q-matrix designs, item quality was manipulated at two levels through the magnitude of attribute effects. An item pool of low quality was simulated by generating the probability of the correct response interval for examinees who have mastered none of the required attributes on an item $P(1|\alpha_c = 0)$ from $U(0.25, 0.40)$ and the probability of the correct response interval for examinees who have mastered all the required attributes on an item $P(1|\alpha_c = 1)$ from $U(0.60, 0.75)$, respectively. An item pool of high quality was simulated by generating $P(1|\alpha_c = 0)$ and $P(1|\alpha_c = 1)$ from $U(0.10, 0.25)$ and $U(0.75, 0.90)$, respectively, as in de la Torre (2009). These item parameter differences define the lower-quality items as those that have higher chances of slipping and guessing. The tetrachoric correlations among attributes were fixed at .70, as in Cui, Gierl, and Chang (2012) and Bradshaw and Templin (2014). We did not use a lower correlation in this study because (a) .70 falls in a reasonable range as observed in educational contexts (Sinharay, Puhane, & Haberman, 2011), and thus the results obtained are expected to be informative for practitioners; (b) if we assume that attributes are hierarchically related, the tetrachoric correlation among them should not

be low by definition; and (c) a low tetrachoric correlation (e.g., .30) among attributes may result in slower convergence and/or nonconvergence. We generated data and estimated the HDCM in R 3.2 (R Core Team, 2015). HDCM was the true generating model, which included the main effect and higher-order interaction effect parameters for all conditions. Sample size was fixed at $N = 2,000$ to avoid confounding the results on account of the limited number of examinees. Attribute profiles were assigned using maximum a posteriori estimates. Each condition was replicated 1,000 times to provide stable results.

Three indices of classification accuracy and reliability were used to evaluate the simulation results. The accuracy of classification refers to the degree of agreement between the examinees' estimated mastery of attributes and their true profile. It was evaluated by the attribute-wise classification accuracy (ACA) and the profile classification accuracy (PCA), defined as

$$ACA = \sum_{e=1}^N \sum_{a=1}^A \frac{E[\hat{a}_{ea} = a_{ea}]}{NA} \quad (3)$$

and

$$PCA = \sum_{e=1}^N \frac{E[\hat{a}_e = a_e]}{N}, \quad (4)$$

where $\hat{a}_{ea}$ is the estimated mastery status for examinee e on the a th attribute, and $\hat{a}_e$ is the estimated mastery pattern for examinee e . The reliability of classification refers to the agreement of examinee classifications in two administrations (Templin & Bradshaw, 2013). The computation of the reliability index uses the tetrachoric correlation coefficient, which mirrors reliability as it is understood in item response theory.

Results

We begin by reporting the results in tables and figures and then describe our findings under each attribute structure. Tables 2 and 3 present the ACA for the low- and high-quality item banks, respectively. Figure 3 presents the PCA for the low- and high-quality item banks. Figure 4 presents the reliability of classifications for the low- and high-quality item banks. Overall, item quality is an important factor in classification results as the classification accuracy and reliability in the high-quality bank were consistently higher than in the low-quality bank regardless of the item-loading approach, the attribute structure, or the number of items.

Linear Hierarchy

When the number of items was fixed at 30, the adjacent approach (L-A30) resulted in higher mean ACA and higher PCA than the other two approaches. Specifically,

Table 2. Accuracy of Classifications for Low-Quality Item Bank.

Attribute structure	Condition	α_1	α_2	α_3	α_4	α_5	Mean
Linear	L-I30	0.959	0.972	0.961	0.948	0.898	0.948
	L-A30	0.950	0.906	0.954	0.977	0.980	0.953
	L-R30	0.855	0.951	0.900	0.903	0.927	0.907
	L-A15	0.860	0.805	0.829	0.881	0.925	0.860
	L-R10	0.678	0.761	0.739	0.759	0.845	0.756
Divergent	D-I30	0.958	0.945	0.859	0.899	0.931	0.918
	D-A30	0.895	0.887	0.917	0.911	0.936	0.909
	D-R30	0.502	0.742	0.962	0.869	0.869	0.789
	D-A22	0.852	0.820	0.817	0.876	0.850	0.843
	D-R16	0.492	0.762	0.799	0.802	0.736	0.718
Convergent	C-I30	0.966	0.974	0.927	0.937	0.864	0.934
	C-A30	0.958	0.957	0.957	0.965	0.900	0.947
	C-R30	0.415	0.838	0.938	0.951	0.982	0.825
	C-A18	0.883	0.848	0.874	0.851	0.834	0.858
	C-R13	0.422	0.778	0.757	0.842	0.881	0.736
Unstructured	U-I30	0.951	0.916	0.854	0.880	0.780	0.876
	U-A30	0.853	0.892	0.893	0.897	0.868	0.881
	U-A24	0.703	0.819	0.832	0.858	0.842	0.811

Note. Bold values are smaller than 0.8.

Table 3. Accuracy of Classifications for High-Quality Item Bank.

Attribute structure	Condition	α_1	α_2	α_3	α_4	α_5	Mean
Linear	L-I30	0.996	1	1	0.999	0.989	0.997
	L-A30	0.988	0.998	1	1	0.999	0.997
	L-R30	0.989	0.995	0.997	0.998	1	0.996
	L-A15	0.917	0.926	0.988	0.998	0.991	0.964
	L-R10	0.921	0.914	0.926	0.914	0.998	0.935
Divergent	D-I30	0.994	0.999	0.997	0.992	0.999	0.996
	D-A30	0.964	0.985	0.999	1	0.998	0.989
	D-R30	0.787	0.857	1	0.998	0.998	0.928
	D-A22	0.918	0.951	0.987	0.993	0.996	0.969
	D-R16	0.714	0.849	0.993	0.974	0.983	0.903
Convergent	C-I30	0.996	1	0.999	0.999	0.984	0.996
	C-A30	0.999	0.999	1	1	0.977	0.995
	C-R30	0.797	0.901	0.999	0.999	1	0.939
	C-A18	0.986	0.986	0.996	0.997	0.938	0.981
	C-R13	0.699	0.852	0.961	0.992	0.997	0.900
Unstructured	U-I30	0.989	0.998	0.997	0.996	0.928	0.982
	U-A30	0.982	0.999	1	0.998	0.996	0.995
	U-A24	0.901	0.995	0.989	0.995	0.996	0.975

Note. Bold values are smaller than 0.8.

Figure 3. Accuracy of profile classifications in low- and high-quality item banks.

Figure 4. Reliability of attribute classifications for low- and high-quality item banks.

when the quality of items was low, the PCA under the independent, adjacent, and reachable approaches was .74, .79, and .72, respectively. When the quality of items was high, the difference in PCA diminished, with PCA values of .93, .96, and .95, respectively. Results show that the reachable approach was most strongly affected by the quality of the item bank, as it had the lowest PCA among the loading approaches when item quality was low yet the highest when item quality was high.

When the number of measurement times for each attribute was fixed at six, the independent approach produced a higher PCA of .93, whereas the adjacent and reachable approaches produced PCA values of .89 and .86, respectively, when the high-quality item bank was used. It was expected that the adjacent and reachable approaches would not produce similar PCA values to the independent approach as a matter of design; the adjacent approach used only 15 items, and the reachable approach used only 10 items, whereas the independent approach used 30 items. Similarly, in the low-quality item bank, the independent approach had higher ACA values on each attribute than the adjacent and reachable approaches. However, it stands out that the adjacent approach produced a PCA of .77 when using 15 items, which was higher than the independent approach PCA of .74 when 30 items were used.

Divergent Hierarchy

When the number of items was held constant at 30 in the low-quality item bank, the adjacent approach (D-A30) resulted in a PCA of .75, which was slightly higher than the PCA using the independent approach (.73; D-I30) and the reachable approach (.71; D-R30). In the high-quality item bank, the adjacent approach and the independent approach produced similar PCA values of .945 and .946, respectively.

Comparatively, the reachable approach is less satisfactory, with PCA values of .71 and .89 in low- and high-quality item banks, respectively.

When each attribute was measured six times, the adjacent approach resulted in a PCA of .90 and ACA values $> .91$ for each attribute. Although the classification results were lower as compared with the results from the independent approach, it is worth noting that eight fewer items were used in the adjacent approach. We also found that when the reachable approach was used, the ACA values of the first two attributes were significantly lower than those for the other attributes in both low-quality and high-quality item banks. We further investigated the Q-matrices that we used under D-R30 and D-R16 and found that the first two attributes were never measured in isolation.

Convergent Hierarchy

Under the convergent structure, the adjacent approach (C-A30; .81) also had higher classification accuracy than the independent approach (C-I30; .72) and the reachable approach (C-R30; .70) in both item banks. The adjacent approach was the only one that achieved PCA greater than .80 in the low-quality item bank and the only approach that achieved PCA greater than .96 in the high-quality item bank.

When each attribute was measured six times, we found that that C-A18 had higher PCA (.77) and similar ACA in the high-quality item bank as compared with the analogous condition that had 12 more items (i.e., C-I30). This indicates that even with fewer items, the adjacent approach works well when the attribute structure is convergent. As in the divergent structure, when the reachable approach was used, the first attribute had very low ACA because it was more often measured with three other attributes (i.e., one item measures four attributes) and was never measured in isolation in the Q-matrices we used in C-R30 and C-R13.

Unstructured Hierarchy

When the attribute hierarchy was specified as unstructured, the adjacent approach resulted in higher PCA and mean ACA across attributes regardless of item quality. The PCA was .99 and .75 for U-A30 in the high- and low-quality item banks, respectively. We observed that in the adjacent approach (U-A30 and U-A24), the first attribute had the lowest ACA among all attributes. For example, in U-A30, the ACA of the first attribute was .98, whereas the others had PCA values that were above .99. In U-A24, the ACA of the first attribute was .90, whereas the others had ACA values above .99. When we look into the Q-matrices under the unstructured hierarchy, only four sets of loading combinations can be specified: α_1 with α_2 , α_1 with α_3 , α_1 with α_4 , and α_1 with α_5 ; therefore, α_1 was never measured in isolation. This finding aligns with previous research that isolating attributes increases classification accuracy. We can view the unstructured hierarchy as a special case of the divergent hierarchy as

multiple branches diverge from a common parent attribute in both hierarchies. In both hierarchies, higher-order attributes have higher ACA.

Reliability of Classifications

The reliability across different Q-matrix designs echoes the results of PCA and ACA. We observed that the quality of an item bank is an important factor in determining the reliability of classifications. Specifically, the reliability across all conditions in a high-quality item bank was mostly above .95, whereas the numbers dropped to between .70 and .90 when low-quality item banks were used. We also observed that decreasing the number of items results in decreased reliability, while holding other factors constant. For example, L-A15 has lower reliability than L-A30. Overall, when the number of items is fixed at 30 in both item banks, the adjacent approach has the highest classification reliability and the reachable approach has the lowest across all four structures.

Discussion

Given that classifying examinees is the main goal for end users of diagnostic tests, how a practitioner goes about loading items on attributes is one of the most important questions during the design phase of a test. It has direct implications for classification accuracy and reliability (Chiu et al., 2009; DeCarlo, 2011; Madison & Bradshaw, 2015). However, little is known about how best to load items onto attributes when the attributes form a hierarchical structure.

This study presented three item-loading approaches for Q-matrix designs and then compared their classification results under four hierarchical attribute structures. In this section, we first discuss the implications of the classification results when different Q-matrix design approaches are applied to hierarchical attribute structures. Then we present a flowchart that guides practitioners step-by-step in choosing a design approach. We conclude by discussing other factors that are involved in the development of diagnostic tests.

This study extended previous research in several ways. First, we have applied the hierarchical structure among attributes a priori in the design of the Q-matrix and explored the impact of Q-matrix designs on classification results when an attribute hierarchy is present. We observed that the classification results for the adjacent approach are similar to or slightly better than those for the independent approach when the designs are placed onto tests of similar length. When test length is limited or the quality of items is low, the adjacent approach produces better classification results than the other two approaches. The adjacent approach performs well partly because (a) it utilizes the information gained from the hierarchical structure through eliminating impossible attribute mastery patterns and decreasing the number of parameters in the structural model and (b) it limits the number of attributes one item can measure to two. Second, this study shows that classification accuracy lost from not

measuring an attribute in isolation can mostly be regained by increasing the number of times each attribute is measured. If the number of times each attribute is measured is fixed, measuring attributes in isolation (i.e., using the independent approach) has higher classification accuracy than other item-loading approaches. However, this would be an unfair comparison because when we design diagnostic tests, it is the number of test items that often limits the design. Given a short test length, such as 10 items in Figure 2, the adjacent approach approximately doubles the number of times each attribute is measured, which improves classification accuracy and reliability.

Our simulation demonstrated that the quality of items has a major influence on classification accuracy, which has been found in several previous studies (Kunina-Habenicht et al., 2012; Madison & Bradshaw, 2015; Rupp & Templin, 2008). Higher-quality items were consistently associated with higher classification and reliability indices regardless of other study factors, indicating the need for much time and effort during the item development stage. Among the three approaches, the reachable approach is strongly affected by the quality of the item bank because with poor items the higher-order interactions modeled in the reachable approach are difficult to estimate.

The results from our study suggest that the adjacent approach shows promise when the test length is limited and/or item quality is low. The simulation design in our article shows an either/or scenario (i.e., either *all* independent structure or *all* adjacent) in isolating the effects of the three design approaches. In practice, we recommend that test developers consider the adjacent approach while incorporating some aspects of the independent approach by measuring each attribute in isolation at least once. Developers are able to gain efficiency with complex items and accuracy with isolated items. If one can isolate each attribute once, classification accuracy increases. Of the design approaches we studied, the reachable approach is the least recommended.

We present a flowchart in Figure 5 to help test developers design diagnostic tests based on these results. In the figure, we suggest that practitioners increase the test length or combine attributes if each attribute cannot be measured at least five times under different approaches. Test developers are recommended to follow the flow chart, which provides a basis for a sound design of the Q-matrix and further supports any inferences based on the diagnostic tests designed. Researchers are advised to add future research findings to this flowchart to continue improving it over the coming years.

There are other factors that need to be considered during the design of diagnostic tests that are not presented in the flowchart because they are more overarching and need iterative validation. For example, the misspecification of the Q-matrix has been heavily discussed in the DCM literature, and it is expected to have a negative impact on classification accuracy (Rupp & Templin, 2008). We refer readers to Chiu (2013) and de la Torre and Chiu (2016) for Q-matrix refinement methods that could be used to identify and correct misspecified entries in the Q-matrix. Also, specifying the attribute structure is another of the major steps in developing a diagnostic test. It can be subjective and requires content knowledge. Test developers should be aware of these factors when designing Q-matrices in diagnostic tests.

Figure 5. Q-matrix design flowchart.

Although the current study introduced three approaches to designing diagnostic tests and tested their classification accuracy and reliability under several conditions, it has limitations in several aspects. First, the specification of attribute structures and Q-matrices was assumed to be correct in the study. Future research should investigate the interplay among three factors in hierarchical diagnostic classification modeling: the misspecification of attribute structures, the three design approaches, and the misspecification of Q-matrices. Second, the independent approach was not examined under different test lengths because reduced test length would affect the number of times each attribute was measured. Achieving a balance between the minimum-measurement occurrences for each attribute and test length across the three approaches to achieve a certain level of classification accuracy is worth investigating. Third, we recommend that researchers look into the impact of the three approaches on specific attributes that have different locations in the hierarchical structure, using the newer framework of attribute structures introduced in Liu and Huggins-Manley (2016). Fourth, it would be interesting to investigate how alternative approaches to examinee classification (e.g., Chiu & Douglas, 2013) perform with different designs of the Q-matrix. Fifth, we advise researchers to investigate if certain Q-matrix designs are better able to detect certain types of hierarchy than others. This, too, would be a helpful research topic for providing useful information to test developers.

Learning is often a sequential process, and yet modeling hierarchies in data is a complex process. Specifically, it is difficult to obtain reliable information about latent traits when those traits are not measured in isolation, yet a hierarchical learning sequence and tests designed to measure that learning sequence pose issues for isolating attribute measurement. The adjacent approach proposed in the study utilizes the information gained from the hierarchical structure and allows more times of measurement for each attribute. Although the recommendation of the adjacent approach may be somewhat counterintuitive to measurement researchers, it opens doors for future research with hierarchical attributes, thereby further illuminating their role and usefulness as a way to describe the relationship among latent traits.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Boorsboom, D., & Mellenberg, G. D. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). Cambridge, England: Cambridge University Press.

- Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, 79, 403-425.
- Chiu, C. Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37, 598-618.
- Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30, 225-250.
- Chiu, C.-Y., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74, 633-665.
- Cui, Y., Gierl, M. J., & Chang, H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, 49, 19-38.
- Dahlgren, M. A., Hult, H., Dahlgren, L. O., af Segerstad, H. H., & Johansson, K. (2006). From senior student to novice worker: Learning trajectories in political science, psychology and mechanical engineering. *Studies in Higher Education*, 31, 569-586.
- DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, 35, 8-26.
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343-362.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115-130.
- de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81, 253-273.
- Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about respondents' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242-274). Cambridge, England: Cambridge University Press.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301-321.
- Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika*, 74, 191-210.
- Hu, J., Miller, M. D., Huggins-Manley, A. C., & Gao, M. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, 16, 119-141. doi: 10.1080/15305058.2015.1133627
- Jimoyiannis, A., & Komis, V. (2001). Computer simulations in physics teaching and learning: A case study on students' understanding of trajectory motion. *Computers & Education*, 36, 183-204.
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, 49, 59-81.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoaka's rule-space approach. *Journal of Educational Measurement*, 41, 205-237.
- Liu, R., & Huggins-Manley, A. C. (2016). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research*. New York, NY: Springer.

- Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75*, 491-511.
- R Core Team. (2015). R (Version 3.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.
- Rupp, A. A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement, 68*, 78-98.
- Simon, M. A., & Tzur, R. (2004). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning, 6*, 91-104.
- Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement, 67*, 239-257.
- Sinharay, S., Puhane, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29-40.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.
- Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, NY: Routledge.
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251-275.
- Templin, J., & Bradshaw, L. P. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317-339.
- Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.
- U.S. Department of Education. (2014). *Secretary's final supplemental priorities and definitions for discretionary grant programs*. Retrieved from: <https://www.federalregister.gov/articles/2014/12/10/2014-28911/secretarys-final-supplemental-priorities-and-definitions-for-discretionary-grant-programs#h-28>
- Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*, 625-649.