The Effects of Q-Matrix Design on Classification Accuracy in the Log-Linear Cognitive Diagnosis Model

Educational and Psychological

Measurement

2015, Vol. 75(3) 491–511

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414539162

epm.sagepub.com

Matthew J. Madison¹ and Laine P. Bradshaw¹

Abstract

Diagnostic classification models are psychometric models that aim to classify examinees according to their mastery or non-mastery of specified latent characteristics. These models are well-suited for providing diagnostic feedback on educational assessments because of their practical efficiency and increased reliability when compared with other multidimensional measurement models. A priori specifications of which latent characteristics or *attributes* are measured by each item are a core element of the diagnostic assessment design. This item–attribute alignment, expressed in a Q-matrix, precedes and supports any inference resulting from the application of the diagnostic classification model. This study investigates the effects of Q-matrix design on classification accuracy for the log-linear cognitive diagnosis model. Results indicate that classification accuracy, reliability, and convergence rates improve when the Q-matrix contains isolated information from each measured attribute.

Keywords

diagnostic classification model, Q-matrix, cognitive diagnosis, log-linear cognitive diagnosis model, test design, diagnostic measurement

Diagnostic classification models (DCMs; e.g., Rupp, Templin, & Henson, 2010) are psychometric models that aim to classify examinees according to their mastery or non-mastery of specified latent characteristics or skills. In contrast, when using the

¹University of Georgia, Athens, GA, USA

Corresponding Author:

Matthew J. Madison, Department of Educational Psychology, University of Georgia, Aderhold Hall 125-A, Athens, GA 30602, USA.

Email: matthewm@uga.edu

Table 1. Example Q-Matrix.

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	0	0	0
2	0	1	0	1
3	1	0	1	0
4	1	1	0	1
5	1	1	0	1
6	1	1	0	1
7	1	1	0	1
8	1	0	0	0
9	1	1	0	1
10	1	1	0	1
11	1	1	0	1
12	1	0	0	0
13	1	0	0	0
14	1	1	0	1
15	1	1	0	1

most commonly applied educational measurement models, item response theory (IRT) models, the aim is to scale individuals on a continuum of general ability in a particular domain. The demand for individualized diagnostic information in education (Center for K-12 Assessment and Performance Management at ETS, 2014; Huff & Goodman, 2007), as well as other settings, has fueled the recent wave of research on DCMs. These models are well-suited for providing diagnostic feedback because of their practical efficiency and increased reliability when compared to other multidimensional measurement models (Templin & Bradshaw, 2013). DCMs have been applied in educational settings to classify examinees with respect to multiple latent characteristics (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014) and in psychological settings to diagnose mental disorders (e.g., Templin & Henson, 2006).

A core element of the design for a diagnostic assessment is a priori specifications of which latent characteristics or *attributes* are measured by each item. This item–attribute alignment is expressed in what is known as a *Q-matrix* (Tatsuoka, 1983). The Q-matrix is an item-by-attribute matrix of 0s and 1s indicating which attributes are measured on each item. If item *i* requires attribute *a*, then cell *ia* in the Q-matrix will be 1, and 0 otherwise. Table 1 displays a Q-matrix for a hypothetical assessment with 15 items and 4 attributes. Here, Item 1 measures Attribute 1, Item 2 measures both Attribute 2 and Attribute 4, Item 3 measures both Attribute 1 and Attribute 3, et cetera.

Q-matrix designs may vary by many features. We define a *Q-matrix design* as the deliberate arrangement of a set of test items according to the specific subset of attributes measured by each individual item. Basic features of Q-matrix designs include the number of items and the number of attributes measured on the assessment. Other features influence the *complexity* of the Q-matrix. Generally, the complexity of the Q-matrix increases as the number of nonzero entries in the Q-matrix increases. The complexity may differ by the number of items measuring each attribute, the number

of attributes measured within each item, and the number of attributes that are measured jointly with other attributes on the assessment.

The specification of the Q-matrix precedes and supports any inferences resulting from DCM analyses. In educational settings, the initial specification of a Q-matrix typically requires content-specific learning theory, as well as expert insights provided by teachers, researchers, and psychometricians. Verifying the item–attribute alignment expressed in the Q-matrix is a complex process that demands extensive qualitative research to investigate the specific mental processes that examinees use to respond to items. This verification is referred to as *content validity* (Boorsboom & Mellenberg, 2007) and is a critical component of the assessment development process because if the Q-matrix is not specified correctly, the inferences resulting from the application of the DCM will not be valid.

Because accurate specification of the Q-matrix is such a critical process, most prior research on the Q-matrix has examined the impact of misspecification. Simulation studies have shown that when the Q-matrix is not specified correctly, classification accuracy rates decrease by varying rates depending on the degree of misspecification (Rupp & Templin, 2008; Kunina-Habenicht, Rupp, & Wilhem, 2012). Rupp and Templin (2008) examined overfit (Q-matrix 0s incorrectly changed to 1s), underfit (1s incorrectly changed to 0s), balanced misfit Q-matrices (incorrectly exchanged 0s and 1s), and attribute-specific misfit (e.g., for every item measuring Attribute *a* and Attribute *b*, Q-matrix entry for Attribute *b* incorrectly changed to 0) for a constrained, noncompensatory DCM. They found that all four conditions resulted in item parameter bias and misclassification of examinees with profiles corresponding to the Q-matrix misspecifications. For example, when Attribute *a* was deleted from the true Q-matrix, many misclassifications occurred for attribute profiles involving Attribute *a*. Kunina-Habenicht et al. (2012) investigated misspecified Q-matrices with random permutations of 30% of the entries and incorrect dimensionality for a general diagnostic model. Similar to Rupp and Templin (2008), they found that classification accuracy decreases in both conditions.

Given that classification is the prime objective of DCMs, it is crucial that researchers and practitioners are aware of all the assessment design variables that can affect classification accuracy. While correctly specifying the Q-matrix is vital, another factor that may influence classification accuracy is the Q-matrix design. DeCarlo (2011) points out that in the deterministic-input, noisy-and-gate (DINA; e.g., Haertel, 1989; Junker & Sijtsma, 2001) model, and higher order models, the design of the Q-matrix can affect classification accuracy, *even if specified correctly*. DeCarlo showed through an analysis of the fraction subtraction data (Tatsuoka, 1990) and through an analytical argument, that when an attribute entered the DINA model only through interaction terms, posterior probabilities of attribute mastery were largely determined by prior probabilities. An attribute enters the DINA model only through interaction terms when the attribute is never measured in *isolation*, meaning that across the test, the attribute is never the only attribute being measured on an item. Following the terminology of McDonald (1999), items that measure an attribute in isolation are

referred to as *factorially simple items*. Chiu, Douglas, and Li (2009) analytically showed that the DINA model, as well as its disjunctive counterpart the deterministic-input, noisy-or-gate (DINO; Templin & Henson, 2006) model, require a Q-matrix design that specifies at least one factorially simple item for each unique attribute in order to identify all latent classes specified by the model. Given a diagnostic assessment of A attributes, there are 2^A unique groups or latent classes into which a diagnostic model aims to classify examinees: Each class is defined by a unique pattern of mastery across the set of A attributes. Essentially, attribute profiles can only be distinguished when the vector of expected item response probabilities across the assessment is unique for each hypothesized class.

Unlike the DINA model, a general DCM (e.g., the log-linear cognitive diagnosis model [LCDM]; Henson, Templin, & Willse, 2009) does not require attributes to be isolated for identification of all unique classes. A general DCM does not have extreme parameter constraints like the DINA model, but instead allows for attribute-specific main effects on individual items. With these main effects, the expected vector of item response probabilities across an assessment is unique for each hypothesized class as long as each attribute is measured by at least one item on the assessment. However, not all Q-matrix designs in a general diagnostic model provide equal information to classify examinees across all possible classes, and some designs may be particularly deficient in information to classify examinees, causing misclassifications.

We submit that although statistical identification is necessary to specify a diagnostic model, it is not sufficient for diagnostic accuracy. In other words, certain Q-matrix designs that yield an identified model as determined by distinct item response probability vectors across attribute profiles will remain problematic for a general DCM. For example, consider the Q-matrix in Table 1. This Q-matrix has several properties, any one of which may cause the Q-matrix to be problematic. Namely, (a) Attribute 1 is measured on almost all the items, (b) Attribute 3 is only measured once, (c) Attributes 2 and 4 are always measured together, which we will refer to as being measured in *conjunction*, and (d) Attributes 2, 3, and 4 are never measured in isolation. This article examines how these types of Q-matrix design characteristics affect DCM performance. The aforementioned studies quantify the impact of misspecification of the Q-matrix; however, research has not systematically quantified the degree to which the Q-matrix designs themselves influence classification accuracy. In this article, we assume that all Q-matrices are correctly specified; this assumption allows us to examine and quantify the extent to which problematic Q-matrix designs affect classification accuracy in a general DCM. We conclude the article by making recommendations for Q-matrix designs based on our findings.

Method

The Log-Linear Cognitive Diagnosis Model

The LCDM is a general DCM that flexibly models attribute effects and interactions at the item level. Other common DCMs (e.g., the DINA model) can be viewed as

special cases of the LCDM in which certain parameters are constrained a priori across all items or all attributes. Using the LCDM framework, these strict constraints do not need to be assumed from the onset of model specification. In contrast, the need for these types of constraints can be tested empirically to guide model specifications that best represent the relationships among items and attributes exhibited by the data (e.g., see Bradshaw et al., 2014).

The LCDM item response function (IRF) is similar to multidimensional IRT model specifications, with the distinguishing feature being that latent traits are not continuous, but rather, they are binary. These binary traits are referred to as attributes denoted by α , where $\alpha = 0$ indicates non-mastery and $\alpha = 1$ indicates mastery. For an assessment that measures A attributes, an examinee has one of 2^A unique patterns of attribute mastery. The LCDM item response function assumes item independence conditional on the examinee's attribute pattern. These patterns, or *attribute profiles*, are predetermined latent classes into which the LCDM probabilistically classifies examinees. To demonstrate the IRF of the LCDM, consider item i that measures two attributes, Attribute 3 (α_3) and Attribute 4 (α_4). The LCDM models the probability of a correct response for an examinee in class c as

$$P(X_{ic} = 1 | \alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1,(3)}(\alpha_3) + \lambda_{i,1,(4)}(\alpha_4) + \lambda_{i,2,(3,4)}(\alpha_3 * \alpha_4))}{1 + \exp(\lambda_{i,0} + \lambda_{i,1,(3)}(\alpha_3) + \lambda_{i,1,(4)}(\alpha_4) + \lambda_{i,2,(3,4)}(\alpha_3 * \alpha_4))}. \quad (1)$$

The person parameter in Equation (1) is the examinee's profile α_c , which is an A length vector denoting the individual attribute states within class c . Namely, $\alpha_c = [\alpha_1, \alpha_2, \dots, \alpha_A]$ where $\alpha_a = 1$ for every attribute a that examinees in class c have mastered and $\alpha_a = 0$ for every attribute a that examinees in class c have not mastered. The item parameters in Equation (1) include an intercept ($\lambda_{i,0}$), a simple main effect for Attribute 3 ($\lambda_{i,1,(3)}$), a simple main effect for Attribute 4 ($\lambda_{i,1,(4)}$), and an interaction between the two attributes ($\lambda_{i,2,(3*4)}$). Similar to a reference-coded analysis of variance (ANOVA) model, the non-masters of required attribute(s) are the reference group for each effect. For example, the simple main effect for Attribute 3 is only present in the IRF when an examinee's attribute profile indicates they have mastered Attribute 3 (i.e., when $\alpha_3 = 1$). Thus, likening the LCDM to ANOVA methods helps interpret the parameters. For example, the intercept is the log-odds of a correct response for examinees who have mastered neither Attribute 3 nor Attribute 4. The main effect for Attribute 3 is the increase in log-odds of a correct response for masters of Attribute 3 and non-masters of Attribute 4, and the main effect for Attribute 4 is the increase in log-odds of a correct response for masters of Attribute 4 and non-masters of Attribute 3. Finally, the interaction term is the change in log-odds for examinees who have mastered both Attributes 3 and 4.

Simulation Study

To systematically investigate the effects of Q-matrix design on classification accuracy, our simulation study manipulated two key factors: the design of the Q-matrix

Table 2. Q-Matrix Conditions.

Condition name	Number of attributes per item	Description
Version 1: FA measured ≈ 10 times total (more complex Q-matrices)		
ISO2 ₁₀	1.68	Most ideal: FA measured in isolation twice; paired with all other attributes
ISO1 ₁₀	1.73	FA measured in isolation once; paired with all other attributes
ISO0 ₁₀	1.77	FA not measured in isolation; paired with all other attributes
JOIN ₁₀	1.82	FA always measured with attribute 2; Attribute 2 not always measured with FA
CONJ ₁₀	1.73	Attribute 2 and FA are always measured together
*CONJ ₁₀	1.73	Attribute 2 and FA are always measured together; only main effects estimated
Version 2: FA measured 5 times total (less complex Q-matrices)		
ISO2 ₅	1.50	Most ideal: FA measured in isolation twice; paired with all other attributes
ISO1 ₅	1.55	FA measured in isolation once; paired with all other attributes.
ISO0 ₅	1.59	FA not measured in isolation; paired with all other attributes
JOIN ₅	1.59	FA is always measured with attribute 2; Attribute 2 is not always measured with FA
CONJ ₅	1.36	Attribute 2 and FA are always measured together
Version 3: FA measured nearly always		
ALL3	1.73	FA is measured on all but three items
ALL1	1.82	FA is measured on all but one item

Note. FA = focal attribute.

and the magnitude of the attribute effects. In total, we used 12 Q-matrix designs, which are outlined in Table 2 and described below. These designs were inspired in part by designs we have encountered in our collaborations with applied researchers who were interested in adapting their test designs for use with DCMs. Thus, we anticipate these designs will be relevant to practical diagnostic assessment development projects. For each of these Q-matrix designs, we fixed the number of items to 22 and the number of attributes to 4.

Our Q-matrix designs systematically differed with respect to Attribute 4, which we term the *focal attribute (FA)*. In the first Q-matrix design, *ISO2*, the assessment was expected to provide enough isolated information to accurately diagnose mastery states with respect to all four attributes. Specifically, in this initial condition, each attribute was measured in isolation at least twice, and each attribute was paired with every other attribute on one or more items. The next two Q-matrix designs were constructed to examine the extent to which less isolated information about the *FA* impacts classification accuracy. Specifically, *ISO1* only differed from *ISO2* in that the *FA* was measured in isolation once, and *ISO0* only differed from *ISO2* in that *FA* is never measured in isolation. For a fixed item parameter effect size, an item measuring the *FA* in isolation will provide more information for the model to make classifications with respect to *FA* than an item measuring the *FA* with one or more other attributes. We predicted that the loss of isolated information would lead to decreased classification accuracy of the *FA*. Because the other attributes' isolation is unaltered from *ISO2* to *ISO0*, we expected their accuracy to be relatively stable.

The next two Q-matrix designs explored the impact of the *FA* always being measured with another attribute, specifically with Attribute 2. In the *JOIN* condition, the *FA* is always measured with Attribute 2, but Attribute 2 is measured in isolation and with other attributes. In the *CONJ* condition, the *FA* and Attribute 2 are always measured in conjunction, meaning any item that measured *FA* also measured Attribute 2 and any item that measured Attribute 2 also measured *FA*. Because these attributes are partially and completely conjoined in the *JOIN* and *CONJ* designs, respectively, we hypothesized that the model would struggle to differentiate between examinees who possess one or neither of these attributes. As a result, we expected the classification accuracy for both the *FA* and Attribute 2 to decrease significantly under these designs. We also included an additional version of the *CONJ* condition, **CONJ*. The motivation for this additional condition is explained in subsequent sections.

The Q-matrix designs also systematically differed with respect to the complexity of the Q-matrix within each of these conditions. We had two specific Q-matrix versions of the five Q-matrix designs described above. The first version was more complex than the second, with approximately 1.75 attribute measured per item and approximately 10 items total measuring the *FA* across the assessment. The second version had approximately 1.5 attributes measured per item and 5 items total measuring the *FA* across the assessment. Although the *FA* is measured less frequently in the second version, we expected classification accuracy to be better because the Q-

matrix complexity is decreased, specifically with respect to the number of interaction terms estimated by the model.

The last two Q-matrix designs explored the extent to which an attribute being measured on nearly every item would impact attribute classification accuracy. In conditions *ALL3* and *ALL1*, the FA was measured on all but three items and one item, respectively. Conditions *ALL3* and *ALL1* are similar in design to the Q-matrix from the widely analyzed fraction subtraction data set (Tatsuoka, 1990), where Attribute 7 is measured on all but one item.

Given the base Q-matrix design (22 items $\times$ 4 attributes), there are more than six septillion possible Q-matrix designs that meet the minimum requirement that each attribute is measured by at least one item. Of these possible Q-matrix designs, only 0.02% isolate all four attributes at least once. In contrast, approximately 25% of these Q-matrix designs fail to isolate the FA. Probabilistically speaking, we are much more likely to find a Q-matrix that does not isolate an attribute than to find one that isolates all measured attributes. Therefore, it is worthwhile to examine the effects of attributes not being measured in isolation. Although we only explore a small portion of this total space of Q-matrix designs, our manipulation of the FA, while holding other attributes relatively constant, allows us to systematically examine the effects of two overarching issues in Q-matrix design: (1) isolation (or lack thereof) is explored in the *ISO* Q-matrix designs and (2) joint measurement of attributes is explored in the *JOIN*, *CONJ*, and *ALL* Q-matrix designs. Because of space considerations, these Q-matrix designs are not presented in this article; however, they are available on request from the first author.

In addition to manipulating the Q-matrix designs, we manipulated the item parameters to examine the extent to which the magnitude of attribute effects influenced classification accuracy and reliability for the different Q-matrix designs. For each Q-matrix design, effects were either “average” or “large.” The values of these effects were chosen to reflect a range of practical situations. Large parameters gave complete masters between a .75 and .90 probability of responding correctly. Complete masters are those examinees who have mastered all the required attributes on a particular item. Average sized parameters gave complete masters between a .60 and .75 probability of responding correctly. Similar simulation studies examining the Q-matrix have used parameter effects giving complete masters between a .75 and 1.0 probability of responding correctly (de la Torre, 2009; Kunina-Habenicht et al., 2012; Rupp & Templin, 2008). Because DCMs are still a relatively new area, there are few studies that have prospectively created tests designed to be modeled with a DCM; most studies have retrofitted data from assessments not designed for DCMs. For this reason, the magnitude of typical parameter effects on diagnostic tests is difficult to gauge. In one of the few diagnostic tests in education developed specifically to be modeled with a DCM, Bradshaw et al. (2014) observed parameters that would be considered average-sized parameters in our study.

We did not manipulate the number of examinees or the relationship among attributes in this study. We fixed the examinee sample size at 2,000. We expected this

Table 3. Simulation Conditions.

Characteristic	Value/interval
Number of items	22
Number of attributes	4
Number of examinees	2,000
Tetrachoric correlations among attributes	.70
Probability of correct response interval for complete non-masters ^a	(.10, .30)
Probability of correct response interval for complete masters ^b (average main effects)	(.60, .75)
Probability of correct response interval for complete masters (large main effects)	(.75, .90)

a. Complete non-masters are examinees who possess none of the required attributes on an item.

b. Complete masters are examinees who possess all the required attributes on an item.

size to be large enough to avoid confounding any results with a deficient sample size, but not too large to be practically unreasonable. An examinee sample size of 2,000 also falls in the range of the simulation studies aforementioned. Research studies examining the effect of attribute tetrachoric correlation in the context of DCMs have observed a slight positive relationship between correlation and model accuracy (Bradshaw & Templin, 2013; Cui, Gierl, & Chang, 2012; Henson & Douglas, 2005; Kunina-Habenicht et al., 2012). We would expect a similar result in this study. Therefore, rather than doubling the size of the study with an additional correlation condition, we chose to fix the correlation at .70. This size correlation falls in the range of the aforementioned studies and has been observed in educational settings (see Sinharay, Puhane, & Haberman, 2011), and is thus expected to be realistic.

In total, our study contained 24 conditions, which are summarized in Tables 2 and 3. The numerical subscript on the condition names refers to the number of times *FA* is measured (e.g., *ISO2₁₀*, *ISO2₅*). For each condition, we generated data using R Version 2.15.3 and estimated the LCDM with two-way interaction effects using marginal maximum likelihood in Mplus Version 7 (Muthén & Muthén, 1998-2012). We estimated 100 replications per condition, expecting this to be enough replications to provide stable summaries of results.

Results

The results of the simulation study will be discussed with a focus on the primary research inquiry: the effects of Q-matrix design on the accuracy and reliability of examinee classifications. For each condition, we examine the convergence rate, accuracy of classifications, and the reliability of classifications for each attribute. We also compare results for conditions with large and average parameter effects and for conditions with more versus less complex Q-matrices. We begin by reporting

Table 4. Convergence Rates by Condition.

Condition	Average effects	Large effects	Condition	Average effects	Large effects
<i>ISO2</i> ₁₀	.98	1	<i>ISO2</i> ₅	1	1
<i>ISO1</i> ₁₀	.98	1	<i>ISO1</i> ₅	.93	1
<i>ISO0</i> ₁₀	.87	1	<i>ISO0</i> ₅	.92	.98
<i>JOIN</i> ₁₀	.43	1	<i>JOIN</i> ₅	.73	.99
<i>CONJ</i> ₁₀	.03	.64	<i>CONJ</i> ₅	.51	.68
* <i>CONJ</i> ₁₀	.97	—			
<i>ALL3</i>	.95	.99			
<i>ALL1</i>	.54	.90			

convergence rates and then discuss other results only for the subset of replications that converged.

Convergence Rates

The convergence rate is the proportion of replications for which the estimation algorithm successfully converged on a solution. The convergence rates for all conditions are shown in Table 4. Convergence rates were higher when the specified Q-matrix contained items that isolated attributes. In conditions *ISO2*₁₀ and *ISO1*₁₀, where the *FA* is measured in isolation at least once, convergence was 98% and perfect for average and large parameters, respectively. In conditions *ISO0*₁₀, *JOIN*₁₀, and *CONJ*₁₀, where the *FA* is not isolated, convergence rates dropped significantly for conditions with average parameter effects. With large parameter effects and 10 items per attribute, only condition *CONJ*₁₀ resulted in convergence issues. The five-item conditions showed similar trends for these conditions, with large parameter effects overcoming convergence issues for all but the *CONJ*₅ condition.

Overall, the most problematic condition was *CONJ*₁₀, where the convergence rate dropped to 3% for average parameter effects. Similar trends were shown for *CONJ*₅ with a convergence rate of 51% under conditions with average parameter effects. The more severe drop in convergence for *CONJ*₁₀ in comparison with *CONJ*₅ shows the degree to which estimation complexity increased with the more complex Q-matrix designs. This result may not be anticipated as typically more information in a psychometric model aids in convergence. Because the *CONJ*₁₀ condition converged on only 3 of 100 replications for average parameters, for the subsequent sections, we replace this condition with condition **CONJ*₁₀, where we only estimated intercepts and main effects in the LCDM.

Conditions *ALL3* and *ALL1* had high convergence rates, > .9, except for condition *ALL1* with average parameters, where the convergence rate dropped to .54. As in the *CONJ* condition, this drop in convergence is a result of the model's trouble with estimating main effects for attributes that are not isolated.

Table 5. Marginal CCR by Condition.

Condition	Average effects				Large effects			
	α_1	α_2	α_3	FA	α_1	α_2	α_3	FA
ISO2 ₁₀	.899	.901	.905	.912	.948	.952	.956	.957
ISO1 ₁₀	.903	.902	.905	.893	.951	.951	.956	.938
ISO0 ₁₀	.903	.901	.911	.859	.951	.951	.960	.905
JOIN ₁₀	.890	.917	.877	.833	.937	.961	.930	.882
CONJ ₁₀	—	—	—	—	.970	.824	.981	.839
*CONJ ₁₀	.923	.698	.942	.644	—	—	—	—
ALL3	.836	.826	.846	.962	.892	.887	.903	.988
ALL1	.787	.820	.795	.964	.834	.884	.852	.989
ISO2 ₅	.905	.925	.929	.881	.954	.969	.973	.935
ISO1 ₅	.919	.914	.929	.863	.964	.962	.974	.912
ISO0 ₅	.917	.917	.930	.822	.963	.964	.973	.867
JOIN ₅	.924	.906	.931	.791	.968	.954	.974	.844
CONJ ₅	.948	.751	.958	.735	.986	.800	.990	.826

Note. Low accuracy rates (<.85) are given in boldface. CCR = correct classification rate; FA = focal attribute.

Accuracy of Classifications

Marginal Attribute Classification Accuracy. The effect of Q-matrix design on accuracy of classifications followed a similar trend to the convergence rates. In general, the model performed better with respect to classification with Q-matrix designs that had more isolated information. We report accuracy in terms of classification according to each of the four attributes individually, or marginally. Table 5 displays the correct classification rates (CCRs) of these marginal attribute classifications across Q-matrix and parameter effect conditions. Although conditions ISO2₁₀, ISO1₁₀, and ISO0₁₀ all had high accuracy rates (> 85% and 90% for average and large parameters, respectively), the accuracy rate of the FA steadily decreased as the number of times it is measured in isolation decreased. For example, we observed a 5% to 6% decrease in the classification accuracy of the FA from ISO2₁₀ to ISO0₁₀ for both average and large effect conditions.

Condition JOIN₁₀ was similar to ISO0₁₀ with the FA having a lower CCR than of all the other attributes that are not joined with another attribute at 83.3% and 88.2% for average and large parameters, respectively. In condition *CONJ₁₀, where the FA and Attribute 2 are conjoined, CCRs for Attribute 2 and the FA slipped to 69.8% and 64.4%, while Attribute 1 and 3 still maintain high CCRs at 92.3% and 94.2%, respectively, for average sized parameters. Even for large parameters, the CCRs for Attribute 2 and FA in the CONJ₁₀ condition were much lower than other attributes at 82.4% and 83.9%, respectively. In the five-item conditions, we observed a similar trend with the FA accuracy decreasing in the Q-matrix designs with non-isolated

Table 6. Classification Accuracy for Examinees With a Total Score of Zero.

Condition	Average effects				Large effects			
	α_1	α_2	α_3	FA	α_1	α_2	α_3	FA
ISO2 ₁₀	.996	.998	.999	.995	.999	1.000	1.000	.996
ISO1 ₁₀	.995	.996	.999	.984	.999	1.000	1.000	.992
ISO0 ₁₀	.994	.997	1.000	.958	.999	1.000	1.000	.976
JOIN ₁₀	.992	.997	.989	.969	.999	1.000	.999	.986
CONJ ₁₀	—	—	—	—	1.000	.983	1.000	.975
*CONJ ₁₀	.999	.952	1.000	.964	—	—	—	—
ALL3	.964	.960	.977	1.000	.989	.985	.992	1.000
ALL1	.938	.961	.962	1.000	.949	.984	.969	1.000

Note. FA = focal attribute.

attributes. Across all Q-matrix conditions, classification accuracy was higher for large parameter effects than average parameters effects.

With conditions *ALL3* and *ALL1*, where the *FA* is measured on the majority of the 22 items, classification accuracy for the *FA* is extremely high at approximately 96% and 99% for average and large parameters, respectively. However, this accuracy is achieved at the sacrifice of the accuracy of the other attributes. This is especially an issue for average parameters, where the accuracy of the other attributes falls as low as 78.7%.

Special Case of Marginal Attribute Accuracy. In DeCarlo’s (2011) analysis of the fraction subtraction data using the DINA model and higher order models, he found that examinees who answered all items incorrectly were misclassified as possessing most of the attributes. He credits these misclassifications to the fact that when attributes are not isolated in the DINA model, posterior probabilities of mastery of the non-isolated attributes are largely determined by prior probabilities. This is not the case with the LCDM; as previously discussed, even when attributes are not isolated, there is a main effect term for each measured attribute. To compare the performance of the LCDM to the DINA model with respect to this result, Table 6 displays the marginal attribute accuracy for examinees with a total score of zero for the more complex Q-matrix conditions. Although there is a decline in accuracy for the Q-matrix designs with non-isolated attributes, accuracy rates are still high with the lowest rate in the *ALL1* condition at 93.8%. These results indicate that the LCDM had little difficulty classifying examinees with a score of zero in these Q-matrix designs that may be problematic in other ways. Almost identical results were found for the less complex Q-matrix conditions, which are not displayed in Table 6.

Overall Profile Accuracy. Overall profile classifications are correct only if all marginal attribute classifications for a given profile are accurate. For this reason, overall profile

Figure 1. Correct classification rates for overall profile classifications.

Note. In (a), the more complex Q-matrix conditions measure the focal attribute (FA) approximately 10 times. In (b), the less complex Q-matrix conditions measure the focal attribute (FA) approximately 5 times. Average parameter effects give masters between .60 and .75 probability of responding correctly. Large parameter effects give masters between .75 and .90 probability of responding correctly.

Table 7. True and Estimated Classification Frequencies for the *CONJ₁₀ Condition.

		Estimated class				
		Class	[* ,0 ,* ,0]	[* ,1 ,* ,0]	[* ,0 ,* ,1]	[* ,1 ,* ,1]
True class	[* ,0 ,* ,0]	715	11	18	0	744
	[* ,1 ,* ,0]	193	26	32	1	252
	[* ,0 ,* ,1]	194	28	34	1	256
	[* ,1 ,* ,1]	69	372	271	37	748
		1,170	437	355	38	

Note. True and estimated class matches are displayed in the diagonal; the off-diagonal elements represent true and estimated class mismatches. Frequencies are rounded to the nearest integer and may not sum to the row and column totals displayed in the margins.

correct classification rates are lower than marginal correct classification rates. Figure 1 displays the expected a posteriori overall profile CCR for average and large parameter effects for each Q-matrix condition.

Because overall profile classifications are directly influenced by marginal attribute classifications, these figures illustrate the same trends seen in the previous section. That is, conditions *ISO2*, *ISO1*, and *ISO0* are much less problematic than conditions *JOIN*, *CONJ*, *ALL3*, and *ALL1*. In addition, the graphs in Figure 1 show that the magnitude of the parameters strongly influences classification accuracy, with large parameter effects partially compensating for the problematic Q-matrix designs.

Because the *CONJ conditions were particularly problematic, we further investigated where these misclassifications occurred. We collapsed the class membership into four classes that had a unique mastery pattern with respect to the two conjoined attributes, Attribute 2 and the *FA*. Namely, we classified examinees according to patterns [* ,0 ,* ,0], [* ,1 ,* ,0], [* ,0 ,* ,1], and [* ,1 ,* ,1] where the * entry in the attribute profile could equal either 0 or 1. Table 7 provides the true and estimated classifications according to these classes to illustrate which examinees were misclassified when two attributes were conjoined. We observed that examinees with exactly one of the conjoined attributes frequently were misclassified as having neither attribute; of the 252 and 256 examinees with profiles [* ,1 ,* ,0] and [* ,0 ,* ,1], on average, 193 and 194 were misclassified as having neither attribute, respectively. Examinees possessing both attributes were mostly misclassified as having only one of the attributes; of the 748 examinees with profile [* ,1 ,* ,1], on average, 372 and 271 were misclassified as having profile [* ,1 ,* ,0] and [* ,0 ,* ,1], respectively. Furthermore, more examinees with profile [* ,1 ,* ,1] were misclassified as having neither attribute ($n = 69$) than were accurately classified ($n = 37$).

Reliability of Classifications. Reliability is one of the most important characteristics of any assessment. For DCMs, reliability refers to the stability of examinee classifications on re-examination (Templin & Bradshaw, 2013). This is particularly important

Figure 2. Reliability of attribute classifications for the more complex Q-matrix designs. *Note.* The more complex Q-matrix conditions measure the focal attribute (FA) approximately 10 times. In (a), average parameter effects give masters between .60 and .75 probability of responding correctly. In (b), large parameter effects give masters between .75 and .90 probability of responding correctly. Att = Attribute.

Figure 3. Reliability of attribute classifications for the less complex Q-matrix designs.
Note. The less complex Q-matrix conditions measure the focal attribute (FA) approximately 5 times. In (a), average parameter effects gives masters between .60 and .75 probability of responding correctly. In (b), large parameter effects give masters between .75 and .90 probability of responding correctly. Att = Attribute.

for assessment developers who want to ensure that assessment results and inferences based on these results are not happenstance. Figures 2 and 3 display the reliability of attribute classifications for average and large parameters across Q-matrix conditions.

Generally, the reliability trends followed those observed in convergence rates and accuracy of classifications. Reliability did not fall to less than .80 for any condition for large parameter effects, so we focus our discussion on the average parameter effect conditions where reliability fell to less than .80. Consider first the conditions with 10 items per attribute (Figure 2a). For average parameters, although the reliability of the *FA* decreased from *ISO2*₁₀ to *JOIN*₁₀, reliability was still acceptable by measurement standards ($> .80$; Nunnally & Bernstein, 1994). Reliability under condition **CONJ*₁₀ indicated significant problems for average parameters where the reliability of Attribute 2 and *FA* dropped to .80 and .69, respectively. In conditions *ALL3* and *ALL1*, the reliability of the *FA* was nearly perfect. This result is because of the fact that the *FA* is measured on the majority of the 22 items. However, the other attributes' reliabilities fall to at or below the field standard of .80 in the *ALL3* and *ALL1* conditions. We observed a similar trend for five-item conditions (Figure 3), but to a lesser degree. As is typical for a reliability metric, decreasing the number of items measuring a latent trait decreases the reliability of the measurement of that trait, which explains the decrease in the reliability of the *FA* for the five-item conditions.

Discussion

Using simulation results, we demonstrated the performance of the LCDM using various Q-matrix designs with respect to classification accuracy and reliability. We simulated 12 Q-matrix designs that, statistically speaking, met the criteria for the model to be able to uniquely classify examinees according to distinct attribute profiles. We found that classification accuracy varied greatly across these designs, indicating that the Q-matrix design is an important feature of a diagnostic assessment design. Overall, when holding the number of times an attribute is measured constant, classification accuracy increased as number of items measuring the attribute in isolation increased. In contrast, classification accuracy suffered most when a pair of attributes was measured in conjunction.

Results from this study can be used as a guide for researchers or practitioners who seek to design diagnostic tests from a DCM framework. Based on our results, we suggest attempting to isolate each attribute at least once, preferably more than once (as in condition *ISO2*). While, unlike the DINA and DINO models, the LCDM is identified without this specification, including a few additional items that isolate attributes on a test may increase accuracy by a significant amount, making it worth additional effort to construct factorially simple items. However, practical instances may arise where the relationships among measured attributes do not allow for the isolation of certain attributes. In this case, we suggest ensuring that no attribute is completely conjoined with another attribute. That is, make sure that each attribute is measured with other attributes, even if it cannot be measured alone (as in condition *ISO0*). If

Figure 4. Q-matrix design flowchart.

two attributes are truly attached and items cannot be written to measure either attribute without the other, we suggest combining the two attributes to form one composite attribute. In this case, marginal attribute information will not be attained, but the classification accuracy of the composite attribute will likely be better than the marginal classification accuracy of the conjoined attributes.

Various practical considerations often make Q-matrix design decisions nontrivial. Figure 4 illustrates a flow chart that can be used to guide test developers in designing sound Q-matrix designs prior to administering a diagnostic assessment. Grounded not only this research but also in other DCM literature, this chart concisely presents a scheme for recognizing problematic designs and a guide for how to proceed when these designs are found. Test developers should progress through the flow chart for each individual attribute one at a time, beginning the chart over again until the result is to proceed or to proceed with certain cautions. While there are many factors that affect DCM performance, iterating through this scheme for each measured attribute will help lead to sound Q-matrix design.

Our results also indicate that the quality of items on a diagnostic assessment strongly influence the ability to make valid inferences about the diagnostic state of examinees. In addition to attending to characteristics of the Q-matrix design, items need to be constructed to clearly discriminate between non-masters and masters of the attributes measured by the items. Large parameter effects mitigated the effects of the problematic Q-matrix design to a degree, emphasizing the critical role that the quality of the items plays in a diagnostic assessment. Results also indicated that the quality of items cannot be replaced by the increasing the quantity of items measuring a given attribute on an assessment, when the assessment length is held constant. In fact, if the overall number of items on an assessment remains constant and the number of items measuring a given attribute increases, the increase in complexity of the Q-matrix may detrimentally impact estimation capabilities resulting in reduced effectiveness of the model to classify examinees accurately.

There are a number of other factors not manipulated in this study that could potentially affect classification accuracy. First and foremost, the accuracy of the Q-matrix specifications was not a factor in our study that will impact classification accuracy; although there may be interplay among types of Q-matrix misspecifications and designs, Q-matrices that contain incorrectly specified entries are expected to decrease classification accuracy, regardless of the Q-matrix design. In addition, examinee sample size, the length of assessment, the number of attributes, the correlations among attributes, and the interaction among these factors all affect classification accuracy. While the results of this study show clear trends that offer insights for Q-matrix designs that best support DCM-based inferences, these results may not generalize to every assessment design situation.

Because of the many factors that may be manipulated in a diagnostic assessment design, we encourage researchers interested in using a DCM to conduct a similar simulation study under their assessment's specific conditions to obtain predicted accuracy rates. This type of prior analysis under hypothesized or theorized conditions would be

analogous to an a priori or prospective power analysis in more traditional statistical models (ANOVA, regression, etc.). This will allow researchers to adjust their study design before collecting and analyzing data. In practical situations, we will never know the true profile for examinees, and hence we can never obtain a measure of model accuracy. However, this type of prior analysis will give researchers and consumers a sense of confidence in the inferences resulting from the DCM. Hopefully, this practice will help researchers anticipate problems in their Q-matrix designs, which, if unknown to researchers, may result in making incorrect inferences about examinees' abilities.

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., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational number: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, 33(1), 2-14.
- Bradshaw, L., & Templin, J. (2013). Combining scaling and classification: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*. Advance online publication. doi:10.1007/s11336-013-9350-4
- Center for K-12 Assessment and Performance Management at ETS. (2014, March). *Coming together to raise achievement: New assessments for the common core state standards*. Retrieved from <http://www.k12center.org>
- 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.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115-130.
- 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.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 333-352.
- Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, 29, 262-277.

- 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.
- Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic 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.
- 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.
- Kunina-Habenicht, O., Rupp, A. A., & Wilhem, 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.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Nunnally, J. C., & Bernstein, I. R. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.
- 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.
- Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Sinharay, S., Puhon, 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. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum.
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, 30, 251-275.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.