Misspecification of Attribute Structure in Diagnostic Measurement

Educational and Psychological

Measurement

2018, Vol. 78(4) 605–634

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164417702458

journals.sagepub.com/home/epm

Ren Liu¹

Abstract

Attribute structure is an explicit way of presenting the relationship between attributes in diagnostic measurement. The specification of attribute structures directly affects the classification accuracy resulted from psychometric modeling. This study provides a conceptual framework for understanding misspecifications of attribute structures. Under the framework, each attribute structure can be represented through an external shape and an internal organization. A simulation study and an application example were used to investigate how misspecification of external shapes and internal organizations affects model fit and item fit assessments, and respondent classification. The proposed framework and simulation results aim to support using attribute structures to (a) develop better diagnostic assessments and (b) inform theories of constructs.

Keywords

diagnostic measurement, diagnostic classification modeling, attribute structure, misspecification, model fit, classification accuracy

Diagnostic measurement is a newer psychometric framework that has received increasing attention in recent years. It differs from traditional frameworks by providing multidimensional classifications to respondents through designing items that measure multiple discrete latent traits (or *attributes*) and analyzing item responses using *diagnostic classification models* (DCMs; Rupp, Templin, & Henson, 2010).

In traditional multidimensional item response theory models, attributes are often hypothesized as correlated (Reckase, 2009). In DCMs, attributes by default are

¹University of Florida, Gainesville, FL, USA

Corresponding Author:

Ren Liu, School of Human Development and Organizational Studies in Education, University of Florida, Gainesville, FL 32611, USA.

Email: liurenking@ufl.edu

correlated, but can be hypothesized as sequentially ordered (Templin & Bradshaw, 2014). Correlated attributes imply that possessing one attribute does not require a possession of another, whereas sequentially ordered attributes imply that the possession of an attribute is a prerequisite of the possession of another attribute. This type of dependent relationship between attributes, known as *attribute hierarchy* (Leighton, Gierl, & Hunka, 2004) or *attribute structure* (Liu & Huggins-Manley, 2016), is common in both education and psychology. In education, student learning follows a certain curriculum and one skill can only be mastered after the mastery of another skill (Darling-Hammond et al., 2015; Entwistle & Ramsden, 2015). In psychology, certain behaviors are developed sequentially over age or time (Bongers, Koot, Van Der Ende, & Verhulst, 2004; Broidy et al., 2003). Attribute structures are thus reflective of the presence or absence of certain skills/behaviors given the presence or absence of other skill/behaviors (Leighton et al., 2004). If an attribute structure is misspecified, respondents will likely be misclassified.

Taking an alternative perspective to study attribute structure, the relationship between attributes directly informs possible attribute combinations when DCMs are fit to item responses. As confirmatory latent class models, any DCM is composed of a structural component and a measurement component. Attribute structure is a direct reflection of the structural component of DCMs. Previous studies on misspecifications focus on the measurement component of DCMs. Those studies have examined the relationships between items and attributes, concluding that misspecifications of attribute–item relationships have a big impact on respondent classifications (e.g., Im & Corter, 2011; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp & Templin, 2008). However, researchers have not examined the misspecification of the structural component of DCMs. This study aims to investigate the effects of attribute structure misspecification on model fit and item fit assessments, and respondent classification.

In the next section, a framework of exploring different types of attribute structure is provided. Then, a framework for investigating misspecifications is proposed, followed by a description of a hierarchical DCM for modeling attribute structures. Next, a simulation study and an applied example are provided. This article concludes with a discussion of the findings, limitations, and practical considerations.

Attribute Structure

Attribute structure is an explicit way of presenting the relationship between attributes. Based on previous studies (e.g., de la Torre, Hong, & Deng, 2010; Leighton et al., 2004; Liu & Huggins-Manley, 2016; Templin & Bradshaw, 2014), each attribute structure can be portrayed through both a numerical and graphical presentation, named as its *internal organization* and *external shape*, respectively.

The internal organization of an attribute structure is the hypothesized permissible and impermissible attribute profiles (e.g., Templin & Bradshaw, 2014). For an assessment that measures A binary attributes, a respondent is classified into one of the 2^A

Figure 1. Types of attribute structures.

latent classes. These latent classes, often called *attribute profiles*, make up the structural component of DCMs. This article will limit discussions to dichotomous attributes although polytomous attributes are also possible (e.g., Chen & de la Torre, 2013; von Davier, 2005). Suppose there are two dichotomous attributes: *a* and *b*, there are four possible attribute profiles: {0, 0}, {1, 0}, {0, 1}, and {1, 1}. When attributes are not hierarchically related, respondents are assigned with one of the four profiles. If an attribute structure is considered, for example, attribute *a* is a prerequisite to attribute *b*, one of the four profiles is *impermissible*: {0, 1}, so respondents are assigned with one of the remaining three *permissible* profiles. In any two-attribute relationship, the *parent* attribute is the prerequisite attribute, and the other one is thus the *child* attribute (e.g., Leighton et al., 2004). In essence, specifying an attribute structure in DCMs informs the modeling of the hypothesized permissible and impermissible attribute profiles.

The external shape of an attribute structure is a visual representation of attribute relationships for both measurement and nonmeasurement professionals. Leighton et al. (2004) first proposed four shapes (or types) of structures. Liu and Huggins-Manley (2016) refined the Leighton taxonomy by disentangling overlapping structures and introducing new structures. In Liu and Huggins-Manley’s framework, attributes can form a *single structure* or a *complex structure*. Forming a single structure necessitates at least two attributes whereas a complex structure demands at least three.

An example of external shapes and internal organizations of different types of attribute structures is presented in Figure 1. For external shapes, attributes at beginning of arrows are prerequisite (i.e., parent) attributes for the ones (i.e., child attributes) at the end. The internal organizations are represented by mosaic plots, where black boxes signify permissible attribute profiles in the structure, with gray boxes indicating impermissible ones. The internal organization for the example complex structure is not presented because the space is limited for laying out $2^5 = 32$ attribute

profiles. For single structures, attributes can form a *linear*, *pyramid*, or *invert pyramid* structure. In a *linear* structure, all attributes are linearly ordered in one line. This linear structure shows, for example, respondents who have mastered the child attribute (α_2) should have mastered the preceding parent attribute (α_1). In a *pyramid* structure, multiple parent attributes yield a child attribute, and one parent can produce at most one child. Respondents who have mastered the child (α_3) should have mastered both parents (α_1 and α_2). In an *inverted pyramid* structure, one parent generates multiple child attributes, and a child does not have more than one parent. In this case, respondents who have mastered either of the higher level attributes (α_2 or α_3) should have mastered the parent attribute (α_1). Attributes can also form a complex structure, which is a combination of two or more simple structures. Those complex structures are referred to as *diamond* structures in Liu and Huggins-Manley's framework. In Figure 1, the diamond structure is formed from the combination of a linear (between α_1 and α_2), an invert pyramid (among α_2 , α_3 , and α_4), and a pyramid (among α_3 , α_4 , and α_5) structure.

Please note that the external shapes and internal organizations of an attribute structure are not distinct entities, instead, they are different ways of representing and understanding how attributes are related. Based on experience, it is usually easier to begin with the external shape when communicating the idea of attribute structure with content experts.

When developing a diagnostic assessment, an attribute structure can be established by using either a theory- or data-driven approach (Liu, Huggins-Manley, & Bradshaw, 2016; Templin & Bradshaw, 2014). By using a theory-driven approach, attribute structures are defined prior to item development based on strong theories of the construct. The benefit of this approach is that it decreases burden in model estimation by reducing the number of possible attribute profiles and the number of parameters. This is especially helpful when the number of attributes is large. Alternatively, attribute structures can be ignored at the design stage but detected at the modeling stage by comparing the fit of two psychometric models, where one takes an attribute structure into account and the other does not. This data-driven approach has the benefit of providing statistical evidence to support an attribute structure. Either way, a correct specification of attribute structure is essential to correct respondent classifications. Correct classification of respondents is typically of primary interest in diagnostic measurement because respondent classifications are the outcome "scores" for respondents.

Although the discussion and application of attribute structure have existed for years (e.g., Gierl, 2007; Gierl, Alves, & Majeau, 2010; Gierl, Leighton, & Hunka, 2000; Leighton et al., 2004), there is limited research on evaluating the effects of attribute structure on diagnostic classification accuracy except for Templin, Henson, Templin, and Roussos (2008), Templin and Bradshaw (2014), Liu and Huggins-Manley (2016) and Liu et al. (2016). The first study demonstrated that incorrect specifications of the structural component of DCMs have a negative impact on parameter and respondent estimation accuracy. The second study proposed a family of

hierarchical DCMs that could be used to estimate and validate the presence of attribute structures. They also found that incorporating fewer attribute profiles than the true profiles significantly decreases classification accuracy, whereas incorporating attribute profiles that are not present in the true profiles does not have a strong influence. The third study applied the hierarchical model proposed in the second study and found that a wide range of factors such as the number of attributes, the number of total levels of an attribute structure, the shape of an attribute structure, and the relative number of attributes in each level, have a significant effect on respondent classification accuracy. For example, they found that (a) child attributes are more likely to be correctly classified than parent attributes and (b) the pyramid structure produces lower accuracy when other factors are held constant. This study investigated correct specifications of attribute structure. The last study found that using attribute structures to design item–attribute relationships can increase classification accuracy. For example, designing items that measure two directly related attributes in the external shape can achieve acceptable classification accuracy in a limited test length. The current study builds on previous studies and investigates the effects of attribute structure misspecifications. The “Discussion” section will elaborate how findings from the current study confirm and extend the previously reported results.

Misspecifications of Attribute Structure

For diagnostic measurement, misspecifications of attribute structure will provide incorrect mastery classifications for respondents. This detrimental impact cannot be remedied by well-designed items or correct item–attribute relationships (e.g., Kunina-Habenicht et al., 2012; Rupp & Templin, 2008). In light of this, it is critical to provide evidence that the hypothesized attribute structures are supported by both prior theories of the construct and statistical evidence from the data (Bradshaw, Izsák, Templin, & Jacobson, 2014; Templin & Bradshaw, 2014). During item development, it is common to first specify the external shape of the structure, on which the internal organization is obtained thereafter. The same approach applies to discussing misspecifications of attribute structures, which means that attribute structures are first misspecified from the external shape, and then are investigated from both their external shapes and internal organizations.

For any two attributes, there are three ways to misspecify the relationship between them: (a) breaking a connection, (b) establishing a connection, and (c) switching the parent–child sequence. Any one of these three ways is expected to alter both the external and internal facets of attribute structures. To systematically understand the impact of misspecifications, several types of misspecifications are introduced below and presented in Figure 2 from perspectives of both external and internal facets. Please note that the structures used in the two perspectives in Figure 2 are hypothetical and unrelated to each other.

The misspecification between two attributes can be generalized to an attribute structure. Misspecifications of external shapes may happen in the form of switching

Figure 2. Example attribute structure misspecifications.

attributes in the chain or breaking/establishing connections; thus, they are consequently referred to as *sequence misspecification* and *design misspecification*.

In a sequence misspecification, the order of some attributes is misspecified but the shape of the structure and the number of chains remain unchanged. For attribute structures that have at least three levels, the position-switch between the higher level and midlevel attributes is referred to as *higher level misspecification* whereas the position-switch between lower level and midlevel attributes is referred to as *lower level misspecification*. In Figure 2, the position-switch between α_2 and α_3 is a higher level misspecification, and between α_1 and α_2 a lower level. Given the findings in Liu and Huggins-Manley (2016), misspecifications at different levels are hypothesized to affect classification accuracy. In a design misspecification, the shape of structure and/or the number of chains may change. In Figure 2, the disconnection between α_1 and α_2 changes the external shape from linear to pyramid.

Misspecifications of external shapes also affect the internal organization of attribute structures. Deviation from the true attribute profiles can produce an *overfitting*, *underfitting*, or *misfitting* structure. An overfitting arises when the misspecified structure includes additional profiles other than all the profiles in the true structure, an underfitting occurs when the true structure includes additional profiles other than all the profiles in the misspecified structure, and a misfitting ensues when there are profiles in one structure that do not exist in the other. In Figure 2, the example overfitting structure includes all attribute profiles in the true structure plus $\{0, 1, 0\}$ that does not exist in the true structure. The example underfitting structure is nested within the true structure because the underfitting structure does not include $\{1, 1, 0\}$, which exists in

the true structure. The example misfitting structure differs from the true structure in patterns $\{1, 0, 0\}$, $\{0, 1, 0\}$, $\{1, 1, 0\}$, and $\{0, 0, 1\}$, where these profiles either exist in the true or misfitting structure.

The Hierarchical Diagnostic Classification Model

There are three promising approaches for modeling attribute structures: the hierarchical diagnostic classification model (HDCM; Templin & Bradshaw, 2014), the attribute hierarchy method (Leighton et al., 2004), and the Bayesian inference networks (e.g., Almond, Mislevy, Steinberg, Williamson, & Yan, 2015; Wu, 2013). Among the three, the HDCM offers the flexibility of directly modeling attribute structures through the presence and/or absence of attribute effects (Templin & Bradshaw, 2014).

The HDCM is an extension of its saturated model: the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009). Suppose item i measures attribute a and attribute b , the LCDM expresses the probability of a respondent r correctly answering item i given the respondent's attribute profile α_r as

$$\text{logit}[P(y_{ri} = 1 | \alpha_r)] = \lambda_{i,0} + \lambda_{i,1,(a)}(\alpha_{ra}) + \lambda_{i,1,(b)}(\alpha_{rb}) + \lambda_{i,2,(a*b)}(\alpha_{ra})(\alpha_{rb}), \quad (1)$$

where $\lambda_{i,0}$ is the intercept, $\lambda_{i,1,(a)}$ is the main effect for attribute a , $\lambda_{i,1,(b)}$ is the main effect for attribute b , and $\lambda_{i,2,(a,b)}$ is the interaction effect between attribute a and b , representing the change when both attributes are mastered (i.e., when $\alpha_{ea} = 1$ and $\alpha_{eb} = 1$).

If mastering attribute a is a prerequisite for mastering attribute b , the HDCM can be used to model such an attribute structure through directly removing the main effect of the child attribute and refraining respondents from being classified into the impermissible attribute patterns. Given that attribute a is the parent of attribute b , the item response function for the HDCM is

$$\text{logit}[P(y_{ri} = 1 | \alpha_r)] = \lambda_{i,0} + \lambda_{i,1,(a)}(\alpha_{ra}) + \lambda_{i,2,(a*b)}(\alpha_{ra})(\alpha_{rb}), \quad (2)$$

which does not have a main effect ($\lambda_{i,1,(b)}$) for the child attribute. $\lambda_{i,1,(b)}$ denotes the change of the logit of the probability of a correct response for respondents who have mastered attribute b but not attribute a , and this is not permissible in the attribute structure modeled by the HDCM. The HDCM can be extended to a generalized form by using multiple link functions, similar to de la Torre (2011). The generalized HDCM can be written as

$$g[P(\alpha_r)] = \lambda_{i0} + \lambda_{i1(a)}(\alpha_{ra}) + \lambda_{i2(a*b)}(\alpha_{ra})(\alpha_{rb}), \quad (3)$$

where $g[P(\alpha_r)]$ is $P(\alpha_r)$, $\log[P(\alpha_r)]$, and $\text{logit}[P(\alpha_r)]$ when identity, log and logit link is used, respectively. The HDCM with the logit link was used for both data generation and estimation in the simulation study and the applied example.

Table 1. Simulation Conditions.

Condition	True	Sequence misspecification			Design misspecification		
1	L3-1	L3-2	L3-3	P3-1	IP3-1		
2	P3-1	P3-2		L3-1	IP3-1		
3	IP3-1	IP3-2		L3-1	P3-1		
4	L5-1	L5-2	L5-3	P5-1	P5-3	IP5-1	IP5-3
5	P5-1	P5-2		P5-3	L5-1	IP5-1	IP5-3
6	P5-3	P5-4	P5-5	P5-1	L5-1	IP5-1	IP5-3
7	IP5-1	IP5-2		IP5-3	L5-1	P5-1	P5-3
8	IP5-3	IP5-4	IP5-5	IP5-1	L5-1	P5-1	P5-3
9	D5-1.1	D5-1.2	D5-1.3	D5-2.1	D5-3.1	D5-4.1	D5-4.1
10	D5-2.1	D5-2.2	D5-2.3	D5-2.4	D5-1.1	D5-3.1	
11	D5-3.1	D5-3.2	D5-3.3	D5-1.1	D5-2.1	D5-4.1	
12	D5-4.1	D5-4.2	D5-4.3	D5-1.1	D5-2.1	D5-3.1	

Simulation Study

Simulation Design

Based on studies that were reviewed earlier, the current study manipulated three key factors that might have a strong influence on the effects of attribute structure misspecification. These three factors are (a) the number of attributes, (b) the type of attribute structures, and (c) the type of misspecifications. In total, 12 conditions were generated, which were outlined in Table 1 and described below.

For each condition, item responses were generated under the true structure, and multiple misspecified structures were modeled to fit the responses. External shapes of the corresponding attribute structures are shown in Figure 3. The names of the attribute structures begin with the type of structures (e.g., L for linear, P for pyramid, IP for invert pyramid, and D for diamond), followed by the number of attributes (e.g., 3 or 5). Both sequence and design misspecifications were considered in each condition.

In Conditions 1 to 3, three attributes were used. The true structures in these three conditions were linear, pyramid, and invert pyramid, respectively. When the true structure was L3-1, there were three possible sequence misspecifications: exchanging α_1 and α_2 , α_2 and α_3 , or α_1 and α_3 . The first two misspecifications were considered in this simulation and the last one was not because adjacent attributes (i.e., attributes that are directly, instead of indirectly connected in a chain) may be more likely to be placed in wrong positions than nonadjacent attributes in practice. The scope of this study is limited to misspecifications of adjacent attributes. Both exchanging α_2 and α_3 (in L3-2) and exchanging α_1 and α_2 (in L3-3) were used in the simulation study because the former represents a higher level structure misspecification and the latter a lower level. It was conjectured that misspecifications at various levels in the attribute structures may have different impacts on parameter estimates and classification accuracy. Then, for design misspecifications, L3-1 was misspecified into P3-1 and

Figure 3. External shapes of attribute structures in the simulation.

IP3-1, corresponding to pyramid and invert pyramid structures, respectively. For P3-1, α_1 and α_2 were parent attributes and α_3 was the child attribute. It was possible that α_2 and α_3 were parent attributes and α_1 was the child attribute, and the classification was expected to be worse. This was considered when the true structure was P3-1.

In Conditions 4 to 12, five attributes were used. Simple structures were defined in Conditions 4 to 8, and complex structures were specified in Conditions 9 to 12. When the true structure was pyramid or invert pyramid, both two- and three-level structures were specified because it was conjectured that misspecifications may have different impacts on structures with different numbers of levels. In this study, misspecifications that are statistically equivalent were not included. For example, when P5-1 was misspecified into P5-2, the positions of α_1 and α_5 were switched. It is possible to switch the positions of α_2 and α_5 , but the model estimation will be identical with P5-2.

In diamond (i.e., complex) structures, two, three, and four levels were considered. The four true structures in Conditions 9 to 12 also aimed to represent (a) mixed three simple structures (including linear, pyramid, and invert pyramid in D5-1.1), (b) connected three simple structures (including linear, pyramid, and invert pyramid in D5-2.1), (c) overall pyramid shape structure (including both pyramid and invert pyramid in D5-3.1), and (d) overall invert pyramid shape structure (including both pyramid and invert pyramid in D5-4.1). Although the simulation study considered four true

Table 2. Q-matrices Used in the Three-Attribute and Five-Attribute Conditions.

	Three attributes			Five attributes				
	α_1	α_2	α_3	α_1	α_2	α_3	α_4	α_5
Item 1	1	0	0	1	0	0	0	0
Item 2	0	1	0	0	1	0	0	0
Item 3	0	0	1	0	0	1	0	0
Item 4	1	1	0	0	0	0	1	0
Item 5	0	1	1	0	0	0	0	1
Item 6	1	0	1	1	1	0	0	0
Item 7	1	0	0	0	1	1	0	0
Item 8	0	1	0	0	0	1	1	0
Item 9	0	0	1	0	0	0	1	1
Item 10	1	1	0	1	0	1	0	0
Item 11	0	1	1	1	0	0	1	0
Item 12	1	0	1	1	0	0	0	1
Item 13	1	0	0	0	1	0	1	0
Item 14	0	1	0	0	1	0	0	1
Item 15	0	0	1	0	0	1	0	1
Item 16	1	1	0	1	0	0	0	0
Item 17	0	1	1	1	0	0	1	0
Item 18	1	0	1	1	0	0	0	1
Item 19	1	1	0	0	0	0	1	0
Item 20	0	0	1	0	1	0	0	1
Item 21	1	0	1	0	0	1	0	0
Item 22	0	1	1	1	1	1	0	0
Item 23	1	1	0	0	1	1	1	0
Item 24	1	1	1	0	0	1	1	1
Item 25	1	1	1	0	1	0	0	1

diamond structures and at least five misspecifications were considered for each true structure, this was by no means exhaustive. It is the hope that some patterns of the effects of structure misspecifications can be revealed through the examples used in the simulation study.

In addition to specifying the relationship among attributes, the relationship between items and attributes also needs to be hypothesized and specified prior to the simulation. The information that which items measure which attributes is contained in an item by attribute matrix known as a Q-matrix (Tatsuoka, 1983). In a binary Q-matrix, an entry, q_{ia} , equals to 1 if item i measures attribute a , and 0 otherwise. The two Q-matrices used in three- and five-attribute conditions are presented in Table 2. The number of items was fixed at 25, and the Q-matrix design was sought to be balanced so that each single attribute and combination of attributes was measured the same number of times (Ayers, Nugent, & Dean, 2009). In the three-attribute Q-matrix, each attribute was measured 14 times. In the five-attribute Q-matrix, each attribute was measured 9 times. The item pool was simulated from $U(0.10, 0.30)$ for

$P(1|\alpha_c = 0)$ and from $U(0.70, 0.90)$ for $P(1|\alpha_c = 1)$. Respondent profiles were generated under a multinomial distribution and the proportion of respondents in each profile was uniformly distributed. The proportion of respondents in each permissible attribute profile in the true model was uniformly distributed. The tetrachoric correlations among attributes were fixed at .7. For each model in every condition, 300 data sets were generated using 2,000 respondents, and each data set was analyzed using marginal maximum likelihood in R (R Core Team, 2016).

Outcome Measures

Multiple measures of model fit, item fit, respondent classification accuracy, and parameter estimates were computed. This section discusses and selects those that are informative to report in the results section.

For model fit, seven indices were computed. The first three were relative fit indices including $-2 \log$ -likelihood ($-2LL$), Akaike's information criterion (AIC; Akaike, 1987), and Bayesian information criterion (BIC; Schwarz, 1978). These three relative fit indices were computed using the contingency table of the expected and observed item responses (i.e., using full-information), and they were useful to detect misspecified DCMs when the saturated model is involved (Chen, de la Torre, & Zhang, 2013). In the current study, the comparisons were conducted among models with different attribute structures. These models were all nested models, and the saturated model was not used for comparisons. However, we still report AIC and BIC in the "Results" section given that researchers are interested in the performance of relative fit indices.

The other four were absolute fit indices including (a) the standardized root mean square residuals (SRMSR; Maydeu-Olivares & Joe, 2014), (b) the mean of absolute deviations between observed and expected correlations (DiBello, Roussos, & Stout, 2007), (c) the mean of absolute values of Q_3 statistic (Yen, 1984), and (d) absolute deviations of residual covariances (McDonald & Mok, 1995). These four indices were computed using limited-information methods and had many advantages over the first three indices that were computed using the full-information methods (Maydeu-Olivares & Joe, 2005; Templin, 2007). In the current study, every index showed the same trend and could differentiate both models with the true and misspecified attribute structures. In the "Results" section, the SRMSR is chosen to be reported because its interpretation is straightforward, intuitive, and not affected by the number of items (Maydeu-Olivares, 2013). The SRMSR has the following structure:

$$SRMSR = \sqrt{\frac{2}{I(I-1)} \sum_{i < j} (r_{ij} - \hat{r}_{ij})^2}, \quad (4)$$

where r_{ij} and $\hat{r}_{ij}$ are the observed correlations and the model predicted correlations between item pairs (i, j) .

Item fit was evaluated by the root mean square error of approximation (RMSEA; von Davier, 2005; cited in Kunina-Habenicht, Rupp, & Wilhelm, 2009), and was computed as

$$RMSEA_i = \sqrt{\sum_v \sum_c \pi(\alpha_c) \left(P_i(\alpha_c) - \frac{n_{icv}}{N_{ic}} \right)^2}, \quad (5)$$

where $\alpha_c = (\alpha_1, \dots, \alpha_A)$ is the attribute vector for latent class c , v is the item category, $\pi(\alpha_c)$ is the estimated class probability of α_c , P_i is the item response function for item i , n_{icv} is the expected number of respondents in latent class c on item i in category v , and N_{ic} is the expected number of respondents in latent class c on item i . Although the RMSEA in von Davier (2005) was computed differently than the “RMSEA” often used in structural equation modeling, it has already been unfortunately named and remains to be used as an item-level fit index in diagnostic measurement studies.

The accuracy of respondent classifications was evaluated by attribute-wise classification accuracy (ACA) and profile classification accuracy (PCA). The reliability of classification was computed using tetrachoric correlation coefficient (Templin & Bradshaw, 2014). However, reliability results were not reported because they showed a similar trend with results obtained from the accuracy indices. Such similarity in the trend of classification accuracy and reliability was also observed in previous studies (e.g., Liu et al., 2016; Madison & Bradshaw, 2015). Respondents’ class distributions were also examined to see how respondents were incorrectly classified under misspecified structures.

The impact of misspecifications on item parameter estimation was also assessed by computing the *bias* and the *root mean square error*. However, different attribute patterns produced different number of parameters that were associated with different attributes. Although the estimates of item parameters in misspecified attribute structures were expected to diverge from the true structure, the impact of misspecifications on parameter estimates was not reported because it might be challenging, at this point, to report those divergences in a systematic way.

Results

To understand the impact of structure misspecifications, the internal organizations of true and misspecified structures in each condition are presented in Figures 4 and 5 using 12 mosaic plots. These plots aim to help readers visualize the differences in permissible and impermissible attribute patterns among different structures in the 12 conditions. The words above each plot are condition numbers and the corresponding true structures. Each plot resembles a matrix with rows as attribute patterns and columns as attribute structures. The first column is always the true structure, and the column sequence is the same as specified in Table 1. Black boxes signify permissible attribute patterns, and gray boxes denote impermissible patterns. Results of relative

Figure 4. Internal organizations of attribute structures in Conditions 1 to 7.

Figure 5. Internal organizations of attribute structures in Conditions 8 to 12.

model fit indices (AIC and BIC), absolute model fit index (SRMSR), and item fit index (RMSEA) are presented in Figures 6, 7, and 8, respectively. The box plots in Figure 8 contain information about the mean RMSEA of each item across the 300 data sets, with the white dots representing the mean RMSEA across 25 items. Results of PCAs and ACAs are shown in Figure 9. Each of the 12 boxes is split into two parts: The top halves show the PCAs and mean ACAs with the names of true and misspecified structures on the x -axis, whereas the bottom halves are ACAs across different structures where each attribute is on the x -axis and each line represents a structure.

Results of the simulation study are presented in two sections. The first section summarizes the macro impact of misspecifications, which includes model fit and item fit assessments and PCAs. The second section presents the impact of misspecifications at the attribute level, which includes the ACAs.

Impact of Structure Misspecification on Model Fit and Item Fit Assessments and Respondent Profile Classification Accuracy. Absolute fit indices are commonly compared with some rule-of-thumb criteria to determine the level of acceptable fit. For absolute model fit, SRMSR smaller than .05 indicates a substantively negligible amount of misfitting, thus, is considered good fit (Maydeu-Olivares, 2013). For AIC and BIC, smaller values indicate better relative fit. AIC, BIC, and SRMSR show the same trend across conditions. For item fit, RMSEA smaller than .05 may be considered a good fit (Kunina-Habenicht et al., 2009). In this study, it is of more interest to compare the model fit, item fit, and PCAs across true and misspecified structures within each condition. Overall, although the true model always had the best model fit and item fit results, the differences between the true and misspecified models in each condition were partly very small for the RMSEA, and sometimes for the SRMSR. The SRMSR may be sufficient as a relative index in most conditions, but the RMSEA values between true and misspecified structures were very similar. This will be discussed further as a limitation of the study in the “Discussion” section.

Three attributes and simple structures (Conditions 1-3). The following three patterns from the external shapes were observed. First, across the three conditions, misspecifications had a larger impact when the true structure is invert pyramid (in Condition 3), as evidenced by poor model fit and item fit in the misspecified structures. Second, structures with sequence misspecifications had worse model fit, item fit, and lower PCAs than structures with design misspecifications in these three-attribute conditions. Take Condition 2 as an example. L3-1 and IP3-1, which were both design misspecifications, had better model fit, item fit, and higher PCAs than P3-2 which was a sequence misspecification. Third, misspecifications at a higher level in the structure produced worse model fit, item fit, and lower PCAs than lower level misspecifications. For example, the external shapes of L3-2 and L3-3 differed in terms of the level of misspecifications. The sequence of higher level attributes (α_2 and α_3) was misspecified in L3-2, whereas that of lower level attributes (α_1 and α_2) was

Figure 6. AIC and BIC for relative model fit.

Note. This is the only figure in which the range of y-axis differs in each condition. The purpose is to show differences in each condition. AIC = Akaike information criterion; BIC = Bayesian information criterion.

Figure 7. Standardized root mean square residual (SRMSR) for absolute model fit.

Figure 8. Root mean square error of approximation (RMSEA) for item fit.

Figure 9. Profile and attribute-wise classification accuracy.

misspecified in L3-3. For model fit and PCA, L3-2 had the worst model fit (.086) and lowest PCA (.655) among all misspecified attribute structures. At the item level, the mean of the RMSEA distribution in L3-2 (.030) was the largest among all the misspecified structures.

Regarding internal organizations, results show that misfitting structures had worse model fit and item fit than overfitting structures. Take Condition 1 as an example. Overfitting structures such as P3-1 and IP3-1 had better model fit and PCAs than misfitting structures such as L3-2 and L3-3.

Five attributes and simple structures (Conditions 4-8). There are three major findings regarding external shapes in these conditions. First, similar to the conditions with three attributes, misspecifications had the largest impact on the invert pyramid structure (in Conditions 7 and 8). Second, structures with design misspecifications had worse model fit, item fit, and lower PCAs than structures with sequence misspecifications in these conditions. Take Condition 7 as an example. Sequence misspecifications such as IP5-2 and IP5-3 had better model fit, item fit, and higher PCAs than design misspecifications such as L5-1, P5-1, and P5-3. This finding is different from that in the three-attribute conditions. Third, misspecifications at a higher level in the structure produced worse model fit, item fit, and lower PCAs than lower level misspecifications. For example, L5-2, P5-4, and IP5-4 had better model fit, item fit, and higher PCAs than L5-3, P5-5, and IP5-5, respectively in Conditions 4, 6, and 8. This finding is aligned with the finding in the three-attribute conditions.

In terms of internal organizations, one noticeable discrepancy between the five-attribute conditions and the three-attribute conditions is that the number of attribute patterns increased from $2^3 = 8$ to $2^5 = 32$. As shown in the mosaic plots, linear structures always had the least amount of permissible attribute patterns and they can be viewed as the most stringent requirement on attribute relationships among all simple structures. There are three major findings about the internal organizations. First, overfitting structures had better fit and higher PCAs than underfitting and misfitting structures. For example, in Condition 6, P5-1 (an overfitting structure) showed better model fit and higher PCA than L5-1 (an underfitting structure). Second, the internal organizations of the misspecifications that deviated less from the corresponding true structures have better model fit and item fit than those that deviated more. For example, in Condition 5, P5-3 deviated the least from P5-1 so it had the best model fit, item fit, and higher PCA among all misspecified structures. Third, the degree of deviations from true structures also had a larger impact than the type of deviations (i.e., overfitting, underfitting or misfitting). For example, in Condition 8, the attribute patterns in IP5-4 and IP5-1 deviated less than other misspecified structures from the true structures, and they showed better model fit, item fit, and higher PCA, although IP5-4 was a misfitting structure while IP5-1 was an overfitting structure.

Five attributes and simple structures (Conditions 9-12). There are five major findings regarding the external shapes in these conditions. First, across the four conditions, the misspecified structures produced relatively better model fit, item fit, and higher PCAs in Condition 10 where the true structure was D5-2.1. D5-2.1 also had the least

number of attribute patterns among the four true structures in Conditions 9 to 12, because the linear part was more independent in the external shape than that in D5-1.1. Second, misspecified structures produced relatively worse model fit, item fit, and lower PCAs in Condition 12 where the true structure was D5-4.1. D5-4.1 had the shape of an invert pyramid overall. Third, structures with design misspecifications had noticeable worse model fit and item fit than those with sequence misspecifications in Conditions 11 (pyramid shape) and 12 (invert pyramid shape). Fourth, higher level misspecifications produced worse model fit than lower level ones. For example, D5-2.2 had better model fit and higher PCA than D5-2.3 which was better and higher than D5-2.4. Fifth, structures with design misspecifications had worse model fit, item fit, and lower PCAs than sequence misspecifications. This is more noticeable in Conditions 11 and 12 in terms of model fit and item fit, and in Conditions 9 and 10 in terms of PCAs.

Regarding internal organizations, results show that underfitting structures produced worse model fit and lower PCAs than misfitting structures when underfitting and misfitting structures had similar degrees of deviation from the true structures. For example, D5-2.1 was an underfitting structure in both Conditions 11 and 12, and it produced the worst fit and lowest PCAs across different misspecifications in those two conditions.

Impact of Structure Misspecification on Attribute Classification Accuracy. The presentation of ACA results is broken down from the perspectives of external shape and internal organization. The patterns are consistent across all simulation conditions. For external shape, results are presented spanning the three situations where relationships between two attributes are misspecified: (a) a connection is broken, (b) a connection is established, and (c) their sequence is switched.

First, breaking the connection between two attributes was less influential on ACAs than connecting or switching them. Take Condition 1 as an example. Although α_1 and α_2 , and α_2 and α_3 were disconnected in P3-1 and IP3-1, respectively, the ACAs for all three attributes in both conditions were close to the ACAs in the true structure. Similarly, in Condition 10, D5-1.1, D5-3.1, and D5-4.1 each broke some connections between different pairs of attributes, but ACAs in those misspecified structures were very close to the corresponding true structures.

Second, connecting two attributes had a substantial impact on the child (i.e., higher level) attribute. Take Condition 2 as an example. α_1 and α_2 were connected in L3-1 and IP3-1 whereas α_2 had very low ACA. Similar situations happen for α_3 in L3-1 and P3-1 in Condition 3; α_2 , α_3 , α_4 , and α_5 in L5-1, IP5-1, and IP5-3 in Condition 5; α_4 in L5-1, IP5-1, and IP5-3 in Condition 6; α_2 in L5-1 and P5-1 in Condition 7.

Third, switching the sequence of two attributes was more influential on the parent (i.e., lower level) attribute. For example, in Condition 1, in L3-2, α_2 and α_3 are switched, and the ACA for α_2 is significantly lower. Similar situations happen for α_1 in L3-3 in Condition 1, α_1 in IP3-2 in Condition 3, α_1 in P5-2 in Condition 5, α_1 in P5-4 in Condition 6, α_2 in IP5-2 and IP5-3 in Condition 7, α_1 in IP5-4 in Condition

Table 3. Respondent Class Distributions in Conditions 1 to 3.

	Condition 1					Condition 2				Condition 3			
	L3-1	L3-2	L3-3	P3-1	IP3-1	P3-1	P3-2	L3-1	IP3-1	IP3-1	IP3-2	L3-1	P3-1
A000	0.260	0.486	0.490	0.258	0.260	0.187	0.577	0.396	0.396	0.189	0.561	0.183	0.182
A100	0.235	0.255		0.233	0.235	0.194		0.204	0.204	0.210		0.397	0.394
A010			<u>0.004</u>	<u>0.002</u>		0.212	0.218				<u>0.002</u>		<u>0.001</u>
A110	0.248		0.249	0.249	0.248	0.204		0.197	0.197	0.201	0.203	0.203	0.204
A001							<u>0.001</u>						
A101		<u>0.001</u>			<u>0.001</u>				<u>0.001</u>	0.184			
A011											<u>0.011</u>		
A111	0.257	0.257	0.257	0.257	0.258	0.203	0.205	0.203	0.203	0.217	0.223	0.217	0.219

Note. Boldfaced values are excessively higher than the true value. Underlined values are close to 0. Gray-shaded cells are impermissible profiles.

8, α_1 in D5-1.2 and α_2 in D5-1.3 in Condition 9, α_2 in D5-2.3 and α_3 in D5-2.4 in Condition 10, α_1 in D5-3.2 and α_2 in D5-3.3 in Condition 11, α_1 in D5-4.2 and α_1 in D5-4.3 in Condition 12.

For internal organizations, attribute structures that were misspecified into a linear or pyramid shape may produce low ACAs for higher level attributes because of restrictions on permissible attribute patterns. For example, in Conditions 7 and 8, mosaic plots for the two conditions show that mastering α_5 was only permissible in 1 of 32 patterns ($\{1, 1, 1, 1, 1\}$) in L5-1, P5-1, and P5-3. Results show that ACAs for α_5 were low when the structure was misspecified into L5-1, P5-1, and P5-3.

For respondent class distributions, Tables 3 and 4 show results from Conditions 1 to 3 and Condition 4, respectively. Because of space limitations, results from Conditions 5 to 12 can be requested from the author. There are two key findings across both simple and complex structures. First, although additional profiles that did not exist in the true structure were created in misspecified structures, few respondents were classified into those “should-not-exist” profiles. In both tables, underlined values are proportions of respondents in the “should-not-exist” profiles, and those values are very close to 0. Second, respondents whose true attribute profiles did not exist in the misspecified structures were classified to “lower” categories. For example, in Condition 1, $\{1, 0, 0\}$ was not permissible in L3-3, and respondents with $\{1, 0, 0\}$ were classified into $\{0, 0, 0\}$. In Condition 3, $\{1, 0, 1\}$ was not permissible in L3-1 or P3-1, and respondents with $\{1, 0, 1\}$ were classified into $\{1, 0, 0\}$.

Real-Data Example

Templin and Hoffman (2013) and Templin and Bradshaw (2014) analyzed a dichotomously scored data set consisting 2,922 respondents and 28 items from the

Table 4. Respondent Class Distributions in Condition 4.

	Condition 4						
A00000	0.173	0.280	0.173	0.159	0.160	0.164	0.166
A10000	0.167		0.170	0.156	0.160	0.168	0.171
A01000		<u>0.050</u>		<u>0.015</u>	<u>0.013</u>		
A11000	0.169	<u>0.178</u>	0.168	0.170	<u>0.172</u>	0.169	0.166
A00100				<u>0.003</u>			
A10100				<u>0.002</u>		<u>0.001</u>	<u>0.001</u>
A01100				<u>0.000</u>			
A11100	0.165	0.165	0.191	<u>0.163</u>	0.164	0.168	0.171
A00010				<u>0.003</u>	<u>0.003</u>		
A10010				<u>0.001</u>	<u>0.001</u>	<u>0.002</u>	
A01010				<u>0.001</u>	<u>0.001</u>		
A11010				<u>0.000</u>	<u>0.000</u>	0.000	<u>0.001</u>
A00110				<u>0.001</u>			
A10110				<u>0.000</u>		<u>0.000</u>	
A01110				<u>0.000</u>			
A11110	0.156	0.156		<u>0.155</u>	0.156	0.155	0.152
A00001							
A10001						<u>0.002</u>	
A01001							
A11001						<u>0.000</u>	<u>0.001</u>
A00101							
A10101						<u>0.000</u>	
A01101							
A11101			<u>0.001</u>			<u>0.001</u>	<u>0.001</u>
A00011							
A10011						<u>0.000</u>	
A01011							
A11011						<u>0.000</u>	<u>0.000</u>
A00111							
A10111						<u>0.000</u>	
A01111							
A11111	0.170	0.170	0.297	0.170	0.171	0.169	0.169

Note. Boldfaced values are excessively higher than the true values. Underlined values are close to 0. Gray-shaded cells are impermissible profiles.

Examination for the Certificate of Proficiency in English. This data set was obtained from the “CDM” R package (Robitzsch, Kiefer, George, & Uenlue, 2017). In Templin and Bradshaw (2014), three attributes with a linear hierarchy was modeled as lexical rules (α_1) $\rightarrow$ cohesive rules (α_2) $\rightarrow$ morphosyntactic rules (α_3). The simulation Condition 1 was thus replicated with the Examination for the Certificate of Proficiency in English data set. Attribute structures and results are shown in Figure 10.

Figure 10. Results from the real-data example.

To clearly differentiate the same attribute structures used in both real-data example and simulation study, a “D” was added to all the structures in the real-data example. For example, L3-1 in the simulation study was denoted as DL3-1 in the real-data example. Results for the absolute model fit index SRMSR, relative model fit index AIC and BIC, and item fit index RMSEA show the same trend with the simulation study: (a) DL3-1 as the true model always had the best model fit and item fit, (b) the differences of fit results between the true and misspecified models were very small. This confirms that these fit indices are not sensitive to misspecification of attribute structures.

Among the misspecifications, DL3-2 and DL3-3 as sequence misspecifications and misfitting structures had larger SRMSR, AIC, and BIC than DP3-1 and DIP3-1 as design misspecifications and overfitting structures. Between DL3-2 as a higher level misspecification and DL3-3 as a lower level misspecification, DL3-2 had worse model fit and item fit, aligning with the simulation results. For respondent class distributions, few respondents were classified into those “should-not-exist” profiles, which agrees with the simulation results. However, respondents whose true attribute profiles did not exist in the misspecified structures were not consistently classified to “lower” categories. For example, it seems that respondents with $\{1, 0, 0\}$ were classified into $\{1, 1, 0\}$. However, given that the true class membership is unknown, it is unsure that whether the class membership under DL3-1 is the “true.”

In sum, the real-data analysis confirms all findings in the simulation study, except for the true class membership that cannot be tested.

Discussion

Attribute structures are commonly observed in education and psychology because students, patients, or respondents often possess a progressive path for certain skills, behaviors, or reasoning (Templin & Bradshaw, 2014). Correct specifications of attribute structures are crucial in diagnostic measurement and are expected to be time-consuming because it involves multiple stages of diagnostic measurement from assessment design to psychometric modeling (Leighton et al., 2004). Up to this point, assessment developers have incorporated attribute structures in assessment designs (e.g., Bradshaw et al., 2014), yet no studies have examined the effects of structure misspecifications on respondent classifications.

This study provides a conceptual framework to understand attribute structure misspecifications. Under the framework, both external shapes and internal organizations should be considered when specifying attribute structures. Several findings in the current study are in line with previously reported results. For example, this study showed that misspecifications of attribute structure had a large impact on respondent classification accuracy, which has been discussed in Templin et al. (2008). Also, both the current study and Templin and Bradshaw (2014) showed that incorporating additional attribute profiles than the true profiles (i.e., overfitting) has very small impact on classification accuracy. Built on previous studies, the current simulation and real-

data study contribute to a systematic understanding of the effects of attribute structure misspecifications on model fit, item fit, and respondent classification accuracy. Key results are summarized as follows. For external structures (a) misspecifications had a larger impact on the invert pyramid structure; (b) sequence misspecifications had worse model fit, item fit, and lower PCAs than design misspecifications in conditions with three attributes, but the other way around in conditions with five attributes; (c) misspecifications at higher levels produced worse model fit, item fit, and lower PCAs than lower level misspecifications. For internal organizations (a) overfitting structures had better fit and higher PCAs than underfitting structures, which was better than misfitting structures; (b) less deviation from the true structure had better model fit, item fit, and PCAs; (c) the degree of deviation had a larger impact than the type of the deviation; and (d) linear structures posed most stringent requirement on attribute relationships among all simple structures because it had the least amount of permissible attribute patterns. For biattribute relationships (a) breaking the connection between two attributes was less influential on ACAs than connecting or switching, (b) connecting two attributes had a substantial effect on the child attribute, and (c) switching the sequence of two attributes had a big impact on the parent attribute.

Limitations and Future Research

This study has shown that much can be learned from investigating attribute structure misspecifications from multiple perspectives. However, the current study is limited in at least three ways. First, as mentioned in the simulation design, the impact of misspecifications on parameter estimation bias was not reported because item parameters in misspecified models could have different meaning and comparing them with ones in the true model may be challenging. However, one could explore overfitting or underfitting structures and investigate how misspecifications affect parameter estimation bias. Second, the RMSEA was not a sensitive indicator for item fit, and most RMSEA values in all conditions were below the .05 benchmark. This similar issue was also presented by Kunina-Habenicht et al. (2012). This index may not be meaningful for practical applications when the true structure and/or model is unknown. A more sensitive index may be needed to better evaluate item fit in diagnostic measurement. Third, there are other promising indices of model fit that can be used and the current study is by no means exhaustive (e.g., Chen et al., 2013). For example, model fit can also be evaluated through the M_r test statistics (Maydeu-Olivares & Joe, 2005) or marginal and bivariate item fit statistics (Rupp et al., 2010). Those indices may be also sensitive to misfitting structures (J. Templin, personal communication, 2016).

Future research could also extend the current investigations in important ways. First, researchers could narrow down to investigate variants in specific types of attribute structures, and in particular the complex structure, as a combination of two or more simple structures. Second, misspecifications in current study are limited to adjacent attributes. It would be interesting to explore the effect of misspecifications

Figure 11. The attribute structure trinity: How it interacts with model fitting and theories of a construct.

among nonadjacent attributes as well. Third, investigating the interplay among attribute structure misspecifications and Q-matrix misspecifications may also be helpful for assessment developers. Fourth, it would be beneficial to conduct a similar study that begins with including all possible attribute patterns, but drops one attribute pattern at a time to obtain a more nuanced understanding of the effect of structure misspecifications.

Practical Considerations

Attribute structures are developmental patterns of respondents that can be used for intervention, curriculum design, theory refinement, and policy making. A correct and useful attribute structure should be involved in an iterative process associated with theories of a construct and model fitting. Figure 11 shows a conceptual understanding of the attribute structure trinity. Beginning with theories of the construct that guide assessment design, assessment developers can either (a) hypothesize that attributes ontologically form a structure, and directly include the attribute structure into the modeling process or (b) ignore possible attribute structures before modeling, and detect an attribute structure epistemologically. Either way, a saturated LCDM is always recommended to fit to the data, where the number of respondents in each attribute pattern may or may not support or inform which patterns are possible and which attributes may be dependent on each other (Templin & Bradshaw, 2014). As a feedback loop, results obtained from the analyses and the presence/absence of attribute structures can be cycled back to inform theories of a construct.

Beyond traditional assessments, some state-of-the-art programs and grants in the measurement field are also investigating ways to capture student learning progressions. For example, the Dynamic Learning Map project at the University of Kansas seeks to illustrate knowledge learning sequence in academic domains (Clark, Kingston, Templin, & Pardos, 2014). A newer NSF grant by Fellouris, Chang,

Douglas, and Culpepper (2016) also intends to identify student learning sequence and develop algorithms for item selection in cognitive diagnostic computerized adaptive testing. The present study has demonstrated that misspecifications of attribute structures are detrimental to classification results from which the inferences about respondents are drawn. It is the hope that the proposed framework and simulation results can not only help practitioners develop better diagnostic assessments but also open doors for future research on modeling the development of learning and behavior.

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

- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317-332.
- Almond, R., Mislevy, R. J., Steinberg, L. S., Williamson, D. M., & Yan, D. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.
- Ayers, E., Nugent, R., & Dean, N. (2009, July 1-3). *A comparison of student skill knowledge estimates*. Paper presented at the Second International Conference on Educational Data Mining (EDM), Cordoba, Spain.
- Bongers, I. L., Koot, H. M., Van Der Ende, J., & Verhulst, F. C. (2004). Developmental trajectories of externalizing behaviors in childhood and adolescence. *Child Development*, 75, 1523-1537.
- Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, 33(1), 2-14.
- Broidy, L. M., Nagin, D. S., Tremblay, R. E., Bates, J. E., Brame, B., Dodge, K. A., & Lynam, D. R. (2003). Developmental trajectories of childhood disruptive behaviors and adolescent delinquency: A six-site, cross-national study. *Developmental Psychology*, 39, 222.
- Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, 37, 419-437.
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50, 123-140.
- Clark, A., Kingston, N., Templin, J., & Pardos, Z. (2014). *Summary of results from the fall 2013 pilot administration of the Dynamic Learning Maps™ Alternate Assessment System* (Technical Report No. 14-01). Lawrence: University of Kansas Center for Educational Testing and Evaluation.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, 47, 227-249.

- Darling-Hammond, L., Barron, B., Pearson, P. D., Schoenfeld, A. H., Stage, E. K., Zimmerman, T. D., ... Tilson, J. L. (2015). *Powerful learning: What we know about teaching for understanding*. New York, NY: Wiley.
- DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 979-1030). Amsterdam, Netherlands: Elsevier.
- Entwistle, N., & Ramsden, P. (2015). *Understanding student learning (Routledge revivals)*. New York, NY: Routledge.
- Fellouris, G., Chang, H., Douglas, J., & Culpepper, S. (2016). *NSF award search: Award No.1632023—Modeling and detection of learning in cognitive diagnosis*. Retrieved from https://www.nsf.gov/awardsearch/showAward?AWD_ID=1632023
- Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement*, 44, 325-340.
- Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, 10, 318-341.
- Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2000). An NCME instructional module on exploring the logic of Tatsuoaka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice*, 19(3), 34-44.
- 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.
- Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement*, 71, 712-731.
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, 35, 64-70.
- 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* (pp. 243-254). New York, NY: Springer.
- Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2016). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement*. doi:10.1177/0013164416645636
- Madison, M., & Bradshaw, L. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, 75, 491-511.
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measurement: Interdisciplinary Research and Perspectives*, 11, 71-137.

- Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, 100, 1009-1020.
- Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, 49, 305-328.
- McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30, 23-40.
- R Core Team. (2016). R (Version 3.3) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.
- Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2017). *CDM: Cognitive diagnosis modeling* (R package Version 5.4.0) [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=CDM/>
- Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68, 78-96.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Templin, J. (2007, October). *Assessing cognitive diagnosis model fit using limited information methods*. Paper presented at the International Conference on Advances in Interdisciplinary Statistics and Combinatorics in Greensboro, NC.
- Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79, 317-339.
- Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, 32, 559-574.
- Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, 32, 37-50.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16). Princeton, NJ: ETS.
- Wu, H. (2013). *A comparison of general diagnostic models (GDM) and Bayesian networks using a middle school mathematics test* (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.