PSYCHOMETRIKA
2013
DOI: 10.1007/s11336-013-9362-0

HIERARCHICAL DIAGNOSTIC CLASSIFICATION MODELS: A FAMILY OF MODELS FOR ESTIMATING AND TESTING ATTRIBUTE HIERARCHIES

JONATHAN TEMPLIN

UNIVERSITY OF KANSAS

LAINE BRADSHAW

UNIVERSITY OF GEORGIA

Although latent attributes that follow a hierarchical structure are anticipated in many areas of educational and psychological assessment, current psychometric models are limited in their capacity to objectively evaluate the presence of such attribute hierarchies. This paper introduces the Hierarchical Diagnostic Classification Model (HDCM), which adapts the Log-linear Cognitive Diagnosis Model to cases where attribute hierarchies are present. The utility of the HDCM is demonstrated through simulation and by an empirical example. Simulation study results show the HDCM is efficiently estimated and can accurately test for the presence of an attribute hierarchy statistically, a feature not possible when using more commonly used DCMs. Empirically, the HDCM is used to test for the presence of a suspected attribute hierarchy in a test of English grammar, confirming the data is more adequately represented by hierarchical attribute structure when compared to a crossed, or nonhierarchical structure.

Key words: diagnostic classification models, cognitive diagnosis, attribute hierarchies, LCDM, latent class models.

Diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010) have become an active area of psychometric research over the past decade. DCMs are psychometric models that characterize examinee item responses with categorical latent variables, or *attributes* that represent the knowledge state of an examinee. Most DCMs use binary attributes, considering examinees to either possess or not possess each attribute. In an educational testing context, possessing an attribute is referred to as mastery of an attribute, while lacking an attribute is referred to as nonmastery. In many cases, attributes may be hierarchical, such that the mastery of one attribute is a precursor to the mastery of another. In such cases, traditional non-hierarchical DCMs will not be as appropriate or useful as models that can explicitly account for an attribute hierarchy. Accordingly, this paper forges a link between two diagnostic modeling paradigms: the Attribute Hierarchy Method (AHM; Leighton, Gierl, & Hunka, 2004) and the Log-linear Cognitive Diagnosis Model (LCDM; Henson, Templin, & Willse, 2009). We introduce an adapted form of the LCDM (the Hierarchical Diagnostic Classification Model, or HDCM) that can be used to form a statistical hypothesis test to investigate the presence of an attribute hierarchy, allowing researchers to gain understanding of the underlying structure of attributes measured by a test.

For consistency throughout the paper, we refer to test takers as *examinees*, the questions of a test as *items*, and the categorical latent variables measured by the test as *attributes* that are either mastered or not mastered by examinees. We first discuss DCMs and the LCDM, a general latent class-based model used for diagnostic modeling, in the context of a test of English grammar rules where an attribute hierarchy may exist. We then discuss the nature of attribute hierarchies and the diagnostic models currently used to detect them. We then introduce the HDCM, followed

Requests for reprints should be sent to Jonathan Templin, Department of Psychology and Research in Education, University of Kansas, 1122 West Campus Rd., Joseph R. Pearson Hall, Room 621, Lawrence, KS 66045, USA. E-mail: jtemplin@ku.edu

Published online: 30 January 2014

by a simulation study to investigate its properties. We conclude the paper by returning to the empirical data analysis to show how the HDCM can detect attribute hierarchies, providing a general discussion of how the HDCM fits into the diagnostic modeling taxonomy.

1. Diagnostic Classification Models

DCMs, also known as *cognitive diagnosis models* (e.g., Leighton & Gierl, 2007), are psychometric models that characterize the relationship of observed responses to a set of latent categorical *attributes*. For an examinee e, the attributes ($\boldsymbol{\alpha}_e = [\alpha_{e1}, \alpha_{e2}, \ldots, \alpha_{eA}]$) are binary indicators of the mastery or nonmastery of a set of A attributes representing the multiple dimensions thought to underlie an examinee's responses to the items of a test. To indicate the attributes measured by each item, an item-by-attribute *Q-matrix* is constructed, with $\boldsymbol{q}_i = [q_{i1}, q_{i2}, \ldots, q_{iA}]$ for an item i measuring A possible attributes. Q-matrix indicators are binary—either the item measures an attribute ($q_{ia} = 1$) or the item does not ($q_{ia} = 0$).

General diagnostic models have been developed based on log-linear models with latent classes, with the Log-linear Cognitive Diagnosis Model (LCDM; Henson et al., 2009) providing a general approach for diagnostic modeling based on an extension of the General Diagnostic Model (GDM; von Davier, 2005). Other general diagnostic models exist, the most recent being the Generalized Deterministic Inputs Noisy and Gate model (G-DINA; de la Torre, 2011). The LCDM and the G-DINA are reparameterizations of each other, providing the same model log-likelihood under marginal maximum likelihood estimation. The LCDM allows for both conjunctive (non-compensatory) and disjunctive (compensatory) links between attributes at the item level, subsuming most commonly used latent class-based DCMs, models such as the Deterministic Inputs Noisy And Gate model (or DINA; Haertel, 1989; Junker & Sijtsma, 2001), the Noisy Inputs Deterministic And Gate model (or NIDA; Maris, 1999), the Reduced Reparameterized Unified Model (or RUM; Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007), the Deterministic Inputs Noisy Or Gate model (or DINO; Templin & Henson, 2006), the Noisy Inputs Deterministic Or Gate model (or NIDO; Templin, 2006), and the Compensatory Reparameterized Unified Model (or C-RUM; Hartz, 2002).

To explain the LCDM, consider an item written to measure two attributes, resulting in Q-matrix entries of $q_{i1} = 1$ and $q_{i2} = 1$. Conditional on an examinee e's attribute profile for these two attributes, $\boldsymbol{\alpha}_e = [\alpha_{e1}, \alpha_{e2}]$, the LCDM item response function for the item is

$$P(X_{ei} = 1 \mid \boldsymbol{\alpha}_e) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(2)}\alpha_{e2} + \lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2})}{1 + \exp(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(2)}\alpha_{e2} + \lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2})}. \tag{1}$$

When attributes are coded as binary latent variables, the LCDM item parameters are analogous to the differing levels of effects found in a reference (or dummy) coded analysis of variance (ANOVA) model, but with a logistic link for dichotomous data. The attribute pattern, $\boldsymbol{\alpha}_e = [\alpha_{e1}, \alpha_{e2}]$, uses reference cell (or dummy) coding to indicate whether examinee e is or is not a master of attribute a ($\alpha_{ea} = 1$ or $\alpha_{ea} = 0$, respectively). The intercept for the item ($\lambda_{i,0}$) represents the log-odds of a correct response for an examinee who is not a master of either attribute. The main effects ($\lambda_{i,1,(1)}$ and $\lambda_{i,1,(2)}$) indicate the increase in log-odds of a correct response given mastery of each attribute, individually. The two-way interaction between the two attributes ($\lambda_{i,2,(1,2)}$) allows the log-odds of a correct response to change given mastery of both attributes by the examinee. The first subscript for the λ parameters in Equation (1) indicates the item i; the second subscript indicates the parameter type (i.e., 0 for intercept, 1 for main effect, 2 for two-way interaction, etc.); the parenthetical subscripts indicate the specific attributes to which the main effect or interaction refers. The LCDM follows a factorial ANOVA model in which the attributes are crossed factors, assuming all combinations of attributes are possible.

The LCDM can be specified to model up to A attributes measured by an item; however, considerations of model identification, computation resources, and algorithm execution time may limit the number of attributes that can be measured per item. The general form of the LCDM is given by

$$P(X_{ei} = 1 \mid \boldsymbol{\alpha}_e = \boldsymbol{\alpha}_c) = \frac{\exp(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \mathbf{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i))}{1 + \exp(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \mathbf{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i))}, \tag{2}$$

where $\boldsymbol{q}_i$ is the vector of Q-matrix entries for item i. The intercept parameter, $\lambda_{i,0}$, retains the same meaning as described previously. The vector $\boldsymbol{\lambda}_i$ (size $(2^A - 1) \times 1$) contains the LCDM parameters for item i, and the function $\boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_e)$ is a vector-valued function of size $(2^A - 1) \times 1$ that contains indicators as to whether or not a parameter is present for an item, as given by the crossed factors ANOVA-like response function of attributes:

$$\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i) = \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ea} q_{ia} + \sum_{a=1}^{A-1} \sum_{b>a} \lambda_{i,2,(a,b)} \alpha_{ea} \alpha_{eb} q_{ia} q_{ib} + \cdots. \tag{3}$$

For an item i, (3) includes all main effects ($\lambda_{i,1,(a)}$ for an attribute a) and interactions between attributes (e.g., $\lambda_{i,2,(a,b)}$ is a two-way interaction for an attribute a and an attribute b). Therefore, for a Q-matrix with A total attributes, the first A elements of $\boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_e)$ are indicators for the A main effect parameters; each main effect indicator is the product of attribute α_{ea} and its corresponding Q-matrix entry, q_{ia}. The second set of elements of $\boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_e)$ are indicators of all $\binom{A}{2}$ possible two-way interactions, which are present when items measure and examinees master a specified pair of attributes. This set of elements results from the multiplication of two α (attribute) values and two entries in the Q-matrix (e.g., the two-way interaction indicator between Attributes 1 and 2 is represented by the product $\alpha_{e1} \times \alpha_{e2} \times q_{i1} \times q_{i2}$). The remaining linear combinations of $\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_e)$ are defined as all possible three-way interactions (for items measuring three attributes) up to a possible final A-way interaction. Monotonicity constraints are placed on the elements of $\boldsymbol{\lambda}_i$ so that the probability of a positive response increases as an examinee masters additional Q-matrix indicated attributes.

In DCMs, the attributes are assumed to follow a discrete distribution, the parameters of which are commonly called the structural portion of the model. The LCDM is a constrained version of a more general latent class model (e.g., Lazarsfeld & Henry, 1968), in which the full model contains the attribute distribution information in so-called "base-rate" parameters (π_c) that represent the probability a given examinee from a population has a given attribute pattern c ($c = 1, \ldots, 2^A$). When coupled with the item response (or measurement) model, the marginal LCDM likelihood function for binary items with binary attributes for a single examinee becomes

$$P(\boldsymbol{X}_e) = \sum_{c=1}^{2A} \pi_c \prod_{i=1}^{I} P(X_{ei} = 1 \mid \boldsymbol{\alpha}_e)^{X_{ei}} \big(1 - P(X_{ei} = 1 \mid \boldsymbol{\alpha}_e)\big)^{1 - X_{ei}}. \tag{4}$$

Because attributes in DCMs are categorical (discrete) latent variables, π_c represents the probability an examinee has a given attribute profile c. Further, the $2^A \pi_c$ terms form the parameters of the joint distribution of the attributes (as opposed to the mean vector and covariance matrix for multivariate normal distributions). Embedded within this distribution are estimates of all pair-wise associations between attributes and, for each attribute, the marginal frequency with which the attribute is mastered in the population of examinees from which a sample is drawn (see Chapter 8 of Rupp et al., 2010). The Multivariate Bernoulli Distribution (or MVB; e.g., Maydeu-Olivares & Joe, 2005) is a general categorical distribution that maps onto two-category latent attributes. Under the general LCDM reflecting an ANOVA-like model with fully crossed factors, the MVB distribution has $2^A - 1$ estimated π_c parameters. The parameters of the MVB

distribution are often modeled using a log-linear parameterization to reduce their number (i.e., Henson & Templin, 2005; Rupp et al., 2010; von Davier & Yamamoto, 2004; Xu & von Davier, 2008).

2. Attribute Hierarchies

In education, the process of teaching generally proceeds sequentially, with each step building upon the last. That is, the curriculum is designed to follow a hierarchical structure that reflects theories about students' learning, where students are taught a given skill as a pre-requisite for the next lesson where that skill will be required. The resulting hierarchy of attributes representing students' knowledge structures may be evident in their responses to test items measuring those attributes and consequently, should be detectable with a statistical model.

Earlier diagnostic classification models anticipated such hierarchies. The Rule Space Model (or RSM; Tatsuoka, 1983, 1990, 1993, 1995, 2009) incorporates an adjacency matrix in which hierarchical attribute structures can be specified and then implemented (although hierarchies are not needed to use the RSM). When compared to the LCDM, the RSM uses a much different approach for classification. Given a set of attributes measured by a test and a set of test items that map onto the attributes, the RSM creates a set of "ideal" or "expected" response patterns for examinees, assuming a conjunctive (all-or-none) relationship between attributes measured by an item. Examinee classification then occurs via a set of clustering procedures mapping the estimate of an examinee's continuous ability from an item response model to that which is provided by the nearest "ideal" response pattern. As such, the RSM presents more of a statistical pattern recognition approach to diagnosis, which makes inferences at the item level difficult due to the aggregate nature of the process (Rupp et al., 2010).

The Attribute Hierarchy Method (or AHM; Gierl, Cui, & Hunka, 2007a; Gierl, Leighton, & Hunka, 2007b; Leighton et al., 2004) expanded upon the RSM to model hierarchical attribute structures that are explicitly defined a priori. The AHM is a probabilistic approach for classifying examinees that requires a formal representation of how the measured attributes of a test are related. An AHM analysis starts with the construction of a matrix of possible attribute profiles given the structure of the attribute hierarchies, followed by a mapping of the test onto the possible attribute profiles. The AHM then uses neural networks to map examinees to attribute profiles based on their item responses. To check model fit, the AHM uses various statistical tests (Cui & Leighton, 2009; Leighton, Cui, & Corr, 2009) and an index of classification reliability (Gierl, Cui, & Zhou, 2009) to determine if the selected attribute profiles match the observed data. As with the RSM, the AHM is more of a pattern-recognition approach to measurement than a statistical measurement model. One of the open questions about the AHM is whether the observed results fit the data as model fit tests provide summary measures but lack inferential statistics to falsify hypotheses about attribute hierarchies.

Accordingly, the purpose of this paper is to bridge the LCDM and latent class-based DCMs with the AHM and RSM in order to (1) provide a statistical test for the presence of attribute hierarchies, and (2) specify a DCM that can model attribute hierarchies within the LCDM framework. That is, just as the AHM expanded upon the RSM to examine attribute hierarchies defined a priori in a confirmatory model, the methods developed in this paper will provide researchers and analysts with a method to empirically evaluate potential hierarchies in the structure of measured latent attributes, as described next.

3. Attribute Hierarchies and Fully Parameterized DCMs

Numerous empirical studies have been conducted using fully parameterized, non-hierarchical latent class-based DCMs to evaluate tests and to estimate examinee attribute profiles. A nonhier-

archical DCM estimates all possible (i.e., 2^A) patterns of attribute mastery, reflecting the theory that no hierarchical relationships among attributes (i.e., no attribute hierarchies) exist or that existing hierarchies within the population would be adequately reflected in model estimation and classification. A hierarchical relationship exists among a pair of attributes when the mastery of an attribute a is required prior to mastery of another attribute b. In other words, for an examinee e, $\alpha_{ea} = 1$ is a necessary (but not sufficient) condition for $\alpha_{eb} = 1$, meaning any attribute profile where elements $\alpha_{ea} = 0$ and $\alpha_{eb} = 1$ would have zero class membership if estimated as a class in a nonhierarchical DCM. In the set of A attributes, as few as one pair or as many as $\binom{A}{2}$ pairs of attributes may have this type of hierarchical relationship, which we refer to as an attribute hierarchy throughout this paper.

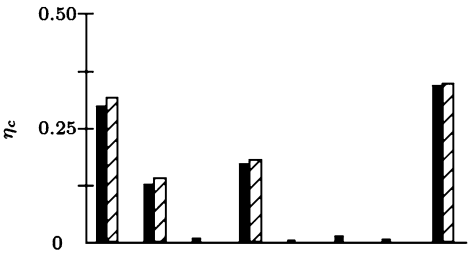
Although nonhierarchical DCMs can provide evidence that an attribute hierarchy is present (e.g., Henson & Templin, 2007), statistical hypothesis tests for the presence of an attribute hierarchy have yet not been developed. This is a problem because if such hypothesized attribute hierarchies are present, the nonhierarchical DCM over-fits the data in two ways: (1) the structural model allows for classes to exist that are not present due to attribute hierarchies, and (2) the measurement model specifies some item parameters that are redundant due to attribute hierarchies (i.e., the model is empirically under-identified). Ultimately, we will form a model where these redundant parameters are fixed to zero, with constraints guided by theoretical considerations of the attribute hierarchy structure. We will expand on these points in the next section when we present the HDCM but for now, we will demonstrate how attribute hierarchies may be manifested when using nonhierarchical DCMs.

4. A Numerical Example: An Application of the LCDM to the Examination for the Certificate of Proficiency in English (ECPE)

As an example, we cite an LCDM analysis of the Examination for the Certificate of Proficiency in English (ECPE), a test developed and scored by the English Language Institute of the University of Michigan. Initially, the ECPE was scored with unidimensional IRT, but more recently the test has been adapted for the Rule Space Model (RSM) in order to examine the individual cognitive skills required for obtaining a complete understanding of English language grammar. Like IRT, the RSM assumes a test is measuring an underlying unidimensional trait, and thus provides an examinee a single score representing a continuous ability. Subsequent to scaling examinee ability, RSM uses a statistical pattern recognition approach to deduce mastery of more fine-grained skills. To investigate the multidimensional cognitive attributes believed to be underlying a test, content experts partnered with psychometricians to conduct DCM analyses of the EPCE (Henson & Templin, 2007; Templin & Hoffman, 2013; Templin, Rupp, Henson, Jang, & Ahmed, 2008; and Buck & Tatsuoka, 1998). The ECPE purports to measure three attributes: knowledge of (1) morphosyntactic rules, (2) cohesive rules, and (3) lexical rules (Buck & Tatsuoka, 1998).

Templin and Hoffman (2013) analyzed a sample of 2,922 examinees who took the ECPE with the nonhierarchical LCDM. This was a didactic paper that showed how to estimate the LCDM using *Mplus* (Muthén & Muthén, 1998–2013). The ECPE data set was used as an example to facilitate demonstrations for setting model specifications and understanding output. Upon further investigation not provided in Templin and Hoffman (2013), results indicated that an attribute hierarchy appeared to be present in the ECPE data.

Evidence for the hierarchy can be seen in both examinee classifications and item parameter estimates reported in Templin and Hoffman (2013). Figure 1 shows the estimates for the structural model with the solid bars, which represent the probabilities an examinee at large has



Morphosyntactic	0	0	0	0	1	1	1	1
Cohesive	0	0	1	1	0	0	1	1
Lexical	0	1	0	1	0	1	0	1
Attribute Profile c	1	2	3	4	5	6	7	8
$\hat{\eta}_c$ - LCDM (solid)	.301	.129	.012	.175	.009	.018	.011	.346
$\hat{\eta}_c$ - HDCM (dashed)	.320	.144	-	.184	-	-	-	.351

FIGURE 1.
Structural parameter estimates from the full LCDM (no attribute hierarchy assumed) and the HDCM (linear attribute hierarchy assumed).

a given attribute profile (π_c in (4)). The eight profiles represent all possible combinations of attribute mastery/nonmastery for the three attributes measured by the ECPE. Of these profiles, four (Profile 3, $\boldsymbol{\alpha}_3 = [010]$; Profile 5, $\boldsymbol{\alpha}_5 = [100]$; Profile 6, $\boldsymbol{\alpha}_6 = [101]$; and Profile 7, $\boldsymbol{\alpha}_7 = [110]$) had very small proportions, meaning few examinees were estimated to have each of these profiles (a combined 5 % of the sample). The remaining 95 % of the sample was estimated to have one of the other four profiles. Specifically, 30.1 % of the sample had not mastered any of the three attributes (Profile 1, $\boldsymbol{\alpha}_1 = [000]$), 12.9 % of the sample had only mastered lexical rules (Profile 2, $\boldsymbol{\alpha}_2 = [001]$), 17.5 % of the sample had mastered both cohesive and lexical rules (Profile 4, $\boldsymbol{\alpha}_4 = [011]$), and 34.6 % of the sample had mastered all three attributes: morphosyntactic, cohesive, and lexical rules (Profile 8, $\boldsymbol{\alpha}_8 = [111]$). Substantial membership in only these four profiles suggests an attribute hierarchy is present: Examinees must master lexical rules before mastering cohesive rules, and must master cohesive rules before mastering morphosyntactic rules. Gierl et al. (2007a, 2007b) call this structure a linear hierarchy, where mastery of each attribute follows a linear progression.

Examining the LCDM item parameter estimates provides further insight about the potential attribute hierarchy. Table 1 provides the parameter estimates for all 28 ECPE items. For items measuring more than one attribute (i.e., Items 1, 3, 7, 11, 12, 16, 20, and 21), the item parameter estimates exhibit behavior discordant with the estimated nonhierarchical LCDM. For instance, main effects for morphosyntactic rules on Items 1 and 12 have values of zero with no standard errors. This result occurred because the estimation software fixed these estimates to the boundary value of the parameters, as main effects cannot be negative under the monotonicity constraints of the LCDM. For Items 3, 11, 16, 20, and 21, the estimated main effects for morphosyntactic rules have large standard errors. If the linear hierarchy is present where morphosyntactic rules can only be mastered after the other rules, the large standard errors would be expected because the main effects represent the increase in the log-odds of a correct response for mastering morphosyntactic rules *in the absence of mastery of the other skills measured by the item* (i.e., a simple effect). That is, if this hierarchy is present, then there would be no one in the sample for whom these main effects would apply. Consequently, these main effects are empirically under-identified because there is no information with which they could be estimated precisely.

The interaction parameters also exhibit misfitting behavior. For Items 7 and 16, the estimated interaction between morphosyntactic rules and lexical rules is equal to (Item 7) or nearly equal to

TABLE 1.
ECPE Q-matrix and LCDM item parameter estimates and standard errors.

i	$\lambda_{i,0}$	$\lambda_{i,1,(1)}$	$\lambda_{i,1,(2)}$	$\lambda_{i,1,(3)}$	$\lambda_{i,2,(1,2)}$	$\lambda_{i,2,(1,3)}$	$\lambda_{i,2,(2,3)}$
1	0.86 (0.07)	0.00 (0.00)	0.60 (0.22)		1.22 (0.27)		
2	1.04 (0.07)		1.25 (0.15)				
3	−0.34 (0.08)	0.75 (0.80)		0.35 (0.13)		0.54 (0.81)	
4	−0.14 (0.08)			1.69 (0.11)			
5	1.08 (0.08)			2.02 (0.16)			
6	0.87 (0.08)			1.69 (.11)			
7	−.11 (.08)	2.86 (0.19)		0.95 (0.14)		−0.95 (0.14)	
8	1.48 (0.08)		1.92 (0.24)				
9	0.12 (0.07)			1.20 (0.10)			
10	0.06 (0.05)	2.05 (0.15)					
11	−0.04 (0.08)	0.82 (0.77)		0.96 (0.14)		0.78 (0.82)	
12	−1.77 (0.11)	0.00 (0.00)		1.29 (0.16)		1.52 (0.14)	
13	0.66 (0.05)	1.63 (0.15)					
14	0.18 (0.05)	1.37 (0.12)					
15	1.00 (0.08)			2.11 (0.16)			
16	−0.10 (0.09)	2.34 (2.08)		0.89 (0.13)		−0.86 (2.10)	
17	1.35 (0.10)		0.77 (1.52)	0.60 (0.30)			0.08 (1.62)
18	0.93 (0.08)			1.39 (0.13)			
19	−0.20 (0.08)			1.85 (0.11)			
20	−1.39 (0.11)	0.24 (0.93)		0.91 (0.15)		1.41 (0.96)	
21	0.16 (0.08)	1.05 (0.76)		1.13 (0.14)		0.04 (0.80)	
22	−0.87 (0.09)			2.25 (0.11)			
23	0.66 (0.07)		2.07 (0.18)				
24	−0.67 (0.07)		1.52 (0.11)				
25	0.09 (0.05)	1.14 (0.11)					
26	0.16 (0.07)			1.12 (0.10)			
27	−0.89 (0.06)	1.71 (0.11)					
28	0.57 (0.08)			1.75 (0.12)			

Note: A column for $\lambda_{i,3,(1,2,3)}$ would exist if the ECPE had any items that measured all three attributes. Attribute 1—Morphosyntactic rules; Attribute 2—Cohesive rules; Attribute 3—Lexical rules.

(Item 16) its lower bound under the monotonicity constraints of the LCDM, as the smallest value an interaction can be is negative one times the value of the largest main effect. For Items 3, 11, 16, 17, 20, and 21, the interactions have large standard errors, resulting in Wald-test p-values that do not indicate statistical significance. These interaction terms each appear in items where at least one attribute being measured by the item has prerequisite attributes in the suspected hierarchy. Thus, if the hierarchy is present, we would expect interactions to be incorrectly estimated because the model is not correctly being specified.

In sum, item parameters for items measuring more than one attribute show a pattern of results that indicates an issue with estimation due to some type of model-data misfit. Although the estimation issues seem erratic and difficult to understand initially, the suspected linear attribute hierarchy could systematically produce these observed trends in item parameters for the nonhierarchical LCDM and explain the parameter under-identification. Given that an estimated 95 % of the sample has one of the four patterns in the linear hierarchy, the question remains: Does the remaining 5 % of the sample indicate a departure from a required attribute hierarchy (i.e., that the hierarchy is prevalent in the population but attributes can be mastered by some examinees in alternative sequences), or is it statistical, measurement, or sampling error? To answer this ques-

tion for these data specifically and to create a model to evaluate attribute hierarchies in the DCM framework more generally, we developed the HDCM.

5. The Hierarchical Diagnostic Classification Model

As previously described, the LCDM is a model analog to a factorial ANOVA model where attributes are fully crossed "design" factors. As such, the LCDM assumes that every possible combination of attributes exists in the population of examinees measured. Continuing with the terminology of the ANOVA model, under an attribute hierarchy specified by the HDCM, attributes no longer represent fully crossed factors but instead represent *nested* factors. In the ECPE example, the data and results suggest a linear attribute hierarchy: Examinees must master Attribute 3 (lexical rules) before mastering Attribute 2 (cohesive rules) before mastering Attribute 1 (morphosyntactic rules). Using ANOVA terminology, we would say that Attribute 2 was *nested* within Attribute 3 (denoted $\alpha_{2(3)}$) and that Attribute 1 was *nested* within Attribute 2 (and consequently, Attribute 3, denoted $\alpha_{1(2,3)}$, which we shorten to $\alpha_{1(2)}$ as the further nesting structure is already implied). If attributes truly are hierarchical, the number of possible attribute profiles is reduced from $2^3 = 8$ under the fully crossed attribute LCDM to 4 (Profiles 1, 2, 4, and 8) in the nested attribute HDCM, as the remaining four profiles cannot exist.

The fully crossed LCDM has model parameters that are redundant in the presence of an attribute hierarchy. For example, in the ECPE fully crossed LCDM analysis, no examinees mastered Attribute 1 (morphosyntactic rules) who had not already mastered Attribute 3 (lexical rules). However, the fully crossed LCDM assumes such examinees exist and estimates the corresponding item parameter for the main effect of Attribute 1: the simple effect of mastering morphosyntactic rules when not mastering lexical rules. In the case of the fully crossed LCDM for an item measuring two attributes, the assumption is that four parameters (an intercept, two main effects, and one two-way interaction) are estimable. If the attributes measured by the item are nested, however, only three cells can be observed, meaning only three parameters can be estimated (an intercept, a main effect of the nonnested attribute, and a two-way interaction of the nested attribute within the nonnested attribute).

6. The Numerical Example Continued: Issues with the Full LCDM Estimation and Potential Remedies

The ECPE item parameter results suggest that such an overfitting of the data is indeed happening. Consider the estimates of Item 7 in Table 1, measuring Attributes 1 (morphosyntactic rules) and 3 (lexical rules), with estimated values of $\hat{\lambda}_{7,0} = -0.11$ for the intercept, $\hat{\lambda}_{7,1(1)} = 2.86$ for the main effect of morphosyntactic rules, $\hat{\lambda}_{7,1(3)} = 0.95$ for the main effect of lexical rules, and $\hat{\lambda}_{7,2(1,3)} = -0.95$ for their interaction (which is at the lowest possible value under the monotonicity assumption of the LCDM). If morphosyntactic and lexical rules were crossed as assumed by the LCDM and were not hierarchically related, we could envision a predicted response for the four possible mastery states of these attributes: (1) neither attribute mastered; (2) lexical rules mastered and morphosyntactic rules not mastered; (3) morphosyntactic rules mastered and lexical rules not mastered; and (4) both attributes mastered. Using the estimates and the item response function for the fully crossed LCDM in Equation (1), the log-odds of the probability of a correct response for each is (1) −0.11, (2) 0.84, (3) 2.75, (4) 2.75. Of the four unique cells, only three values occur. Although these estimates could happen without a hierarchy present, such results occur for nearly every ECPE item measuring more than one attribute.

7. HDCM Parameterization

To avoid the over-parameterization of the LCDM when applied to an attribute hierarchy, the HDCM changes this parameterization to reflect the nested structure of the attributes. Under the LCDM $\boldsymbol{\alpha}_c$ represents the attributes in profile c, where $c = 1, \ldots, 2^A$. Under the HDCM, $\boldsymbol{\alpha}_c^*$ represents the attributes in profile c, but with the number of profiles reduced by eliminating the attribute profiles that could not be possible under the attribute hierarchy. A linear hierarchy, such as that thought to be present in the ECPE data, greatly reduces the number of attribute profiles from 2^A to $A + 1$. For linear hierarchies (where attribute c is nested within attribute b, which is nested in attribute a, and so on), the matrix product in Equation (2) becomes $\boldsymbol{\lambda}_i^T \mathbf{h}(\boldsymbol{\alpha}_e^*, \boldsymbol{q}_i)$, resulting in a sum across the reduced set of attribute profiles (for an examinee e with $\boldsymbol{\alpha}_c^* = \boldsymbol{\alpha}_e^*$):

$$\begin{aligned}\boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_e^*, \boldsymbol{q}_i\right) = {} & \lambda_{i,1,(a)}\alpha_{ea}q_{ia} + \lambda_{i,2(b(a))}\alpha_{ea}\alpha_{eb}q_{ia}q_{ib} \\ & + \lambda_{i,3(c(b,a))}\alpha_{ea}\alpha_{eb}\alpha_{ec}q_{ia}q_{ib}q_{ic} + \cdots. \end{aligned} \tag{5}$$

As a result, the set of parameters for an item now reflects the nested structure of the attributes in the pattern $\boldsymbol{\alpha}_c^*$. Therefore, for an item measuring two attributes a and b, with attribute b nested within attribute a, the HDCM item response function for an examinee e on an item i is

$$P\left(X_{ei} = 1 \mid \boldsymbol{\alpha}_e^*\right) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1,(a)}\alpha_{ea} + \lambda_{i,2,(b(a))}\alpha_{ea}\alpha_{eb})}{1 + \exp(\lambda_{i,0} + \lambda_{i,1,(a)}\alpha_{ea} + \lambda_{i,2,(b(a))}\alpha_{ea}\alpha_{eb})}. \tag{6}$$

The item response function includes three parameters: an intercept ($\lambda_{i,0}$), a main effect for attribute a($\lambda_{i,1,(a)}$), and an interaction for attribute b nested within attribute a($\lambda_{i,2,(b(a))}$). The interaction parameters of the HDCM must be positive to ensure monotonicity of the response probability for additional mastered attributes.

Similarly, the structural model for the HDCM is reduced. An attribute hierarchy suggests that certain attribute profiles are not possible and, therefore, no examinee's mastery profile should reflect such patterns. Therefore, no longer are there 2^A base-rate (π_c) parameters in the structural model in (4). The new structural model contains the reduced set of base-rate parameters. The reduction in the complexity of the structural model results in far fewer parameters under the HDCM when compared to the LCDM.

8. Theoretical Considerations: Development of a Hypothesis Test for Attribute Hierarchies

The HDCM is a model nested within the full LCDM in that the reduction of parameters can be obtained by setting some parameters of the full LCDM to zero. As such, it is possible to construct a nested model comparison hypothesis test through use of -2 times the difference in model log-likelihood values when both the full LCDM and nested HDCM have been estimated (sometimes called a deviance or likelihood ratio test). Typical deviance tests are compared to a Chi-Square distribution with the difference in model parameters as the degrees of freedom. Under the HDCM, however, some LCDM main effect parameters and some LCDM structural parameters are fixed to their lower boundary of zero, violating the regularity conditions for use of the deviance test (e.g., Stoel, Garre, Dolan, & van den Wittenboer, 2006). Instead, the true distribution represents a mixture of Chi-Square distributions and can be derived analytically for simpler models or through simulation for models with a large number of bounded parameters (Shapiro, 1985). For the HDCM, the number of bounded parameters depends on the nature of the attribute hierarchy. In most cases, the p-value for the hypothesis test of the full LCDM versus the nested HDCM will have to be derived through simulation, as the information matrix of LCDM parameters will have off-diagonal elements that cannot easily be found analytically (Verbeke & Molenberghs, 2000). However, when such boundary constraint conditions are violated

yet ignored and the incorrect Chi-Square distribution is used rather than the correct Mixture Chi-Square distribution, the result is typically an overly conservative hypothesis test. Using a simulation study and empirical data analysis, we will demonstrate the distribution of the test statistic under the null hypothesis that an attribute hierarchy exists, and further, we will use a simulation to derive the p-value for this test in our analysis of the ECPE data with the HDCM.

The ability of the HCDM to detect attribute hierarchies results from the linear-model structure of the LCDM, which itself subsumes many other latent class-based DCMs. Because each of the models subsumed by the LCDM does not parameterize the LCDM's crossed-effects ANOVA-style structure for item parameters, it is difficult, if not impossible, to use such sub-models to detect an attribute hierarchy. In our simulation study, we will demonstrate this deficiency in two of the most constrained versions of the LCDM. The DINA model constrains item parameters to reflect *completely non-compensatory* attributes on every item, and the DINO model constrains items parameters to reflect *completely compensatory* attributes on every item. Likely because of their simplicity, the DINA and DINO models are two of the most commonly-used DCMs. Other methods for modeling attribute hierarchies (i.e., RSM and AHM) do not utilize statistical measurement models for attributes and as such do not yield parameters for which statistical significance can be evaluated. Thus, our simulation study only contains models within the latent class DCM framework that allow for the parameterization of attribute mastery.

We next present a simulation study investigating the theoretical properties of the HDCM and the nested model hypothesis test for attribute hierarchies. Following that, we present an analysis of the ECPE responses using the HDCM to demonstrate how to detect and test for attribute hierarchies and how to interpret the HDCM parameters.

9. Simulation Study

We conducted a simulation study to investigate the theoretical properties of the HDCM in mapping a set of item responses onto a set of hierarchical attributes. We designed the study (1) to determine how the parameters of the HDCM were recovered under ideal conditions, (2) to develop the sampling distribution for the deviance statistic for comparing the HDCM to the full LCDM, and (3) to see how well the DINA and DINO models can detect attribute hierarchies.

9.1. Method

All conditions of the simulation study included 3000 examinees, 30 items, and 3 attributes, with 500 replications. Data were generated from three different models: the full LCDM (with a total of $2^3 = 8$ attribute profiles, referred to as the LCDM), the HDCM with one nested attribute (with a total of 6 attribute profiles, referred to as the 6-profile HDCM), and the HDCM with two nested attributes (with a total of 4 attribute profiles, referred to as the 4-profile HDCM). For each model, we used a balanced Q-matrix with each item measuring either one or two attributes, resulting in five items measuring each type of attribute combination. To remove a confound potentially caused by the magnitudes of item parameters, the main effects were all set to a value of 2, the interactions were all set to a value of 1, and the intercept was set to $-0.5 \cdot \boldsymbol{\lambda}_i^T \mathbf{h}(\boldsymbol{\alpha}_{2^A}, \boldsymbol{q}_i)$ where $\boldsymbol{\alpha}_{2^A} = \{1, 1, \ldots, 1\}$. The intercept value was chosen to give symmetry around a log-odds of zero for masters of all and masters of no attributes in order to make attribute misclassifications roughly equal. Each simulated data set was then analyzed by nine different models: the LCDM, the 6-profile HDCM, the 4-profile HDCM, an 8-, 6-, and 4-profile DINA model, and an 8-, 6-, and 4-profile DINO model. All analyses were estimated using marginal maximum likelihood in *Mplus* 6.1 (Muthén & Muthén, 1998–2013).

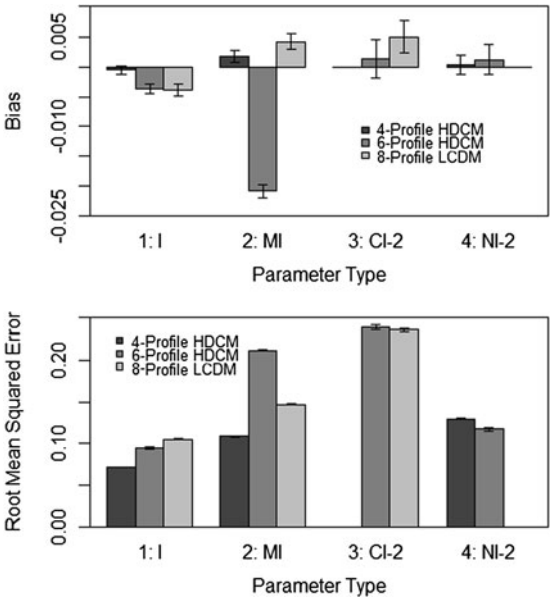


FIGURE 2.
Simulation study item parameter results (I: intercept; ME: main effect; CI-x: x-way crossed interaction; NI-x: x-way nested interaction).

9.2. Results

We present the results of the simulation study in three sections corresponding to the three goals outlined previously: (1) a parameter recovery study, (2) a hypothesis test error and power study, and (3) an alternate DCM efficacy study.

9.2.1. Parameter Recovery Study To understand the estimation accuracy of the Mplus marginal maximum likelihood estimator, a parameter recovery study was conducted. In the three cases where the estimation model matched the data generation model (LCDM, the 6-profile HDCM, and the 4-profile HDCM), model parameter recovery was evaluated. Three types of parameters were evaluated: item parameters (intercepts, main effects, crossed interactions, and nested interactions), structural model parameters (also intercepts, main effects, crossed interactions, and nested interactions), and classification accuracy of examinee attribute profiles.

Figure 2 displays the LCDM item parameter bias and root mean squared error (RMSE) results, aggregated across all 500 simulation replications. In general, as the number of attribute profiles decreased (as the number of hierarchical/nested attributes increased) RMSE values and bias also decreased, indicating an increase in estimation accuracy. The full LCDM had the worst accuracy and bias, which was expected considering it had the highest number of parameters of all models while the number of simulated items and examinees was held constant. The lone exception was the bias/accuracy of the main effect parameters for the 6-profile HDCM, which had one hierarchical/nested attribute. In this case, the parameter had a negative bias whereas in the 4-profile HDCM and the full LCDM this parameter had a slight positive bias. We speculate this may have been due to a mix of the types of interactions present in this model, with five items having nested-model interactions and ten items having crossed-model interactions. Even with this result, bias was minor, ranging between -0.02 and 0.005 across all parameters. Except for the 6-profile HDCM main effect result, the trend in our results mirrored those reported for similar parameters in Bradshaw and Templin (2013) and Choi (2010).

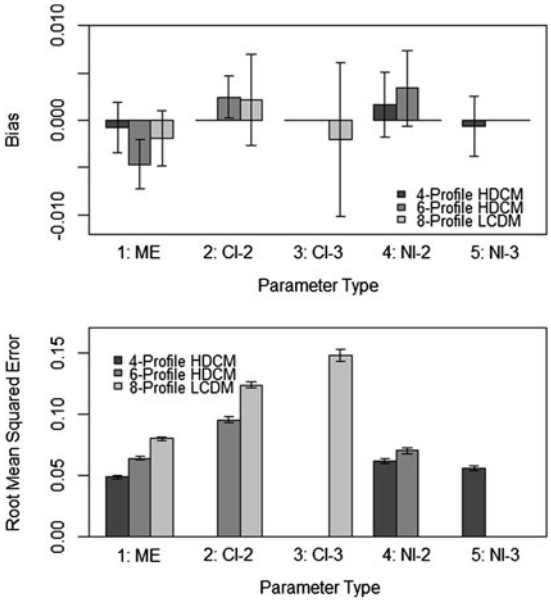


FIGURE 3.
Simulation study structural parameter results (ME: main effect; CI-x: x-way crossed interaction; NI-x: x-way nested interaction).

The LCDM structural parameters, representing the probability an examinee has a given attribute pattern, were modeled using a fully saturated log-linear model (e.g., Henson & Templin, 2005; Rupp et al., 2010; von Davier & Yamamoto, 2004; Xu & von Davier, 2008) to allow for an unbounded estimate of the parameters. Figure 3 displays the bias and RMSE of the structural parameters, shown by parameter type (main effect, crossed interaction, or nested interaction). The structural parameter results mirrored the results for the item parameters, in that the bias and RMSE were worse for models with higher numbers of parameters. The 6-profile HDCM condition exhibited the most variability in the parameters. The bias was still negligible, ranging from -0.05 to 0.05 across all parameter types and models.

Table 2 displays two measures of classification accuracy, the attribute profile classification rate and marginal attribute classification rate, for each model (4-profile, 6-profile, and 8-profile). The attribute profile classification rate is the proportion of examinees that were correctly classified as having a given attribute profile (i.e., all attributes were correctly classified). The marginal attribute classification rate is the proportion of times a single attribute was classified correctly, across all examinees and attributes. Along the diagonal, when the estimation model was the same as the generation model, for each of the hierarchy structures, attribute profile classification rate was between 0.92 and 0.93 with little variability. A similar result was found for the marginal attribute correct classification rate, with the average rate being near 0.97 for all models. In all, when the estimation model matched the true model, classification rates are consistent with those found in the DCM literature (i.e., Henson et al., 2009).

Perhaps more revealing were the classification results for conditions where the estimated model did not match the model generating the data. Specifically, in cases where the estimated structural model had *more* profiles than the true structural model, classification was roughly equal to that found when the estimated and true models matched. In other words, additional (empty) profiles did not affect classification of examinees. When the estimated structure model had *fewer* profiles than the true structural model, classification rates plummeted (as much as 30 % for the whole pattern classification rate and 12 % for the average attribute classification rate). As we

TABLE 2.
Simulation study whole pattern and marginal attribute correct classification rates.

True structure	LCDM/HDCM Estimated structure		
	4-Profile (HDCM)	6-Profile (HDCM)	8-Profile (LCDM)
4-Profile (HDCM)	0.922(0.005)/0.973(0.002)	0.921(0.005)/0.972(0.002)	0.920(0.005)/0.972(0.001)
6-Profile (HDCM)	0.730(0.009)/0.901(0.003)	0.929(0.005)/0.974(0.002)	0.928(0.005)/0.973(0.002)
8-Profile (LCDM)	0.627(0.009)/0.853(0.004)	0.776(0.008)/0.918(0.003)	0.924(0.005)/0.972(0.002)
True structure	DINA Estimated structure		
	4-Profile	6-Profile	8-Profile
4-Profile	0.735(0.011)/0.908(0.005)	0.688(0.014)/0.891(0.006)	0.687(0.014)/0.891(0.006)
6-Profile	0.646(0.010)/0.876(0.004)	0.805(0.009)/0.933(0.003)	0.800(0.010)/0.932(0.003)
8-Profile	0.587(0.001)/0.843(0.004)	0.704(0.011)/0.895(0.004)	0.825(0.008)/0.942(0.003)
True structure	DINO Estimated structure		
	4-Profile	6-Profile	8-Profile
4-Profile	0.888(0.007)/0.962(0.002)	0.888(0.007)/0.962(0.002)	0.885(0.016)/0.961(0.005)
6-Profile	0.611(0.012)/0.863(0.005)	0.685(0.017)/0.893(0.006)	0.661(0.019)/0.887(0.008)
8-Profile	0.518(0.009)/0.830(0.003)	0.547(0.021)/0.848(0.017)	0.574(0.024)/0.863(0.018)

Note: Standard errors are given in parentheses.

expected, examinees having profiles that were no longer in the estimation routine were the cause of this misclassification.

9.2.2. Hypothesis Test for Attribute Hierarchies To investigate the behavior of the deviance statistic, we constructed three different model comparisons for a Type-I error study: the 6-profile HDCM versus the full LCDM (T6-8); the 4-profile HDCM versus the 6-profile HDCM (T4-6); and the 4-profile HDCM versus the full LCDM (T4-8). A Type-I error was committed when a reduced attribute hierarchy was incorrectly rejected in favor of a model with more parameters.

For each simulated data set, the deviance statistic was calculated to compare the null model (the nested model with fewer attribute profiles due to an attribute hierarchy) to the alternative model with more attribute profiles. The distribution of this deviance statistic is a mixture of Chi-Square distributions that is not directly derivable due to dependencies between parameters in the information matrix. Therefore, this portion of our simulation study is to describe the behavior of the deviance test statistic under the null hypothesis and to demonstrate its Type-I error rate when naïvely testing it against a conventional Chi-Square distribution with number of degrees of freedom equal to the difference in the number of freed parameters between the alternative and null models.

At varying levels of significance, Table 3 displays the Type-I error rates for the naïve deviance test, the estimated critical values from the simulations, and the corresponding critical values from the naïve tests. At each significance level, the naïve hypothesis test was conservative, with Type-I error rates well below each level of significance. This indicates that if a p-value less than a stated level of significance was found, then one could have a high degree of confidence in the decision. Similarly, the critical values estimated from the distribution all are well under that given by the naïve test, indicating the distribution of the deviance statistic is not a conventional Chi-Square. As such, we recommend using the simulation-based approach to find-

TABLE 3.
Simulation study HDCM hypothesis test Type-I error rates, critical values, and relative fit rates.

Type-I error rate at:	Model comparison		
	T6-8 ($df = 7$; 447R)	T4-6 ($df = 12$; 330R)	T4-8 ($df = 19$; 330R)
0.01	0.002	0.000	0.000
0.05	0.009	0.003	0.000
0.10	0.020	0.006	0.000
0.20	0.034	0.030	0.019
0.50	0.159	0.188	0.136
Estimated critical value at:	T6-8 ($df = 7$; 447R)	T4-6 ($df = 12$; 330R)	T4-8 ($df = 19$; 330R)
0.01	13.774 (18.475)	17.592 (26.217)	24.546 (36.191)
0.05	9.110 (14.067)	14.732 (21.026)	21.716 (30.144)
0.10	7.622 (12.017)	13.050 (18.549)	19.463 (27.204)
0.50	3.142 (6.346)	7.479 (11.340)	11.923 (18.338)
Information criteria error rate	T6-8 ($df = 7$; 447R)	T4-6 ($df = 12$; 330R)	T4-8 ($df = 19$; 330R)
AIC	0.000	0.009	0.000
BIC	0.000	0.000	0.000
Adjusted BIC	0.000	0.000	0.000

Note: The number of converged replications are listed with the letter R. Parentheses display critical value from naïve Chi-Square test with degrees of freedom equal to difference in freed parameters of the model. Adjusted BIC is the Sample-Size Adjusted BIC statistic as provided by *Mplus* (Muthén & Muthén, 1998–2013).

ing the p-value, as outlined in the subsequent empirical data analysis results section. Table 3 also lists the error rates for the three information criteria reported by *Mplus*, the AIC, BIC, and Sample Size Adjusted BIC. For each criterion and in nearly every model, the information criteria selected the correct model. Although we caution against the use of the information criteria when a hypothesis test is available, we provide these results for a matter of comparison with the results of other DCMs and for our empirical data analysis of the ECPE.

Finally, to investigate the power of the test to detect the alternative, fully crossed LCDM, we calculated deviance statistics from model comparisons when the data were generated by the LCDM. In all conditions, for all levels of significance, the power of the test was 100 %. We report this result in order to provide some insight into the power of the test under the naïve hypothesis test. However, we expect the power of this hypothesis test to be impacted by the sample size and number of items and attributes measured by the test.

9.2.3. Detection of Attribute Hierarchies and Classification Rates Under Various DCMs
To compare the ability of the LCDM/HDCM family to detect attribute hierarchies to that of other DCMs, we replicated our Type-I error study with the DINA and DINO models. As mentioned above, the DINA and DINO models are subsumed by the LCDM with a simplified structure. In both models, each item has two parameters, a slipping and a guessing parameter, no matter how many attributes are measured. Specifically, the DINA model is parameterized as

$$P(X_{ei} = 1 \mid \boldsymbol{\alpha}_e) = (1 - s_i)^{\eta_i} g_i^{1-\eta_i}. \tag{7}$$

The guessing parameter g_i represents the probability of answering an item i incorrectly conditional on examinee e having not mastered all attributes measured by the item ($\eta_i = \prod_{a=1}^{A} \alpha_{ea}^{q_{ia}} = 0$). The slipping parameter s_i represents the probability of answering an item i *incorrectly* con-

ditional on examinee e having mastered all q-matrix attributes for the item ($\eta_i = \prod_{a=1}^{A} \alpha_{ea}^{q_{ia}} = 1$). The LCDM subsumes the DINA by specifying the guessing parameter as

$$g_i = \frac{\exp(\lambda_{i,0})}{1 + \exp(\lambda_{i,0})}. \tag{8}$$

Likewise, the LCDM formulation for the slipping parameter comes from estimating the highest-level interaction between all attributes where no lower order parameters are present, yielding

$$1 - s_i = \frac{\exp(\lambda_{i,0} + \lambda_{i,A,(1,2,\ldots)} \prod_{a=1}^{A} \alpha_{ea}^{q_{ia}})}{1 + \exp(\lambda_{i,0} + \lambda_{i,A,(1,2,\ldots)} \prod_{a=1}^{A} \alpha_{ea}^{q_{ia}})}. \tag{9}$$

Similar to the DINA model, the DINO model expresses the probability an examinee answers an item correctly as a function of the same two slipping and guessing parameters, only with a different function differentiating the types of examinees to which the parameters apply:

$$P(X_{ei} = 1 \mid \boldsymbol{\alpha}_e) = (1 - s_i)^{\eta_i} g_i^{1-\eta_i}. \tag{10}$$

Unlike the DINA model, under the DINO model, the slipping parameter applies when an examinee has mastered at least one of the attributes measured by the item ($\eta_i = 1 - \prod_{a=1}^{A} (1 - \alpha_{ea})^{q_{ia}} = 1$). If the examinee has not mastered any of the measured attributes, then the guessing parameter applies ($\eta_i = 1 - \prod_{a=1}^{A} (1 - \alpha_{ea})^{q_{ia}} = 0$). The LCDM subsumes the DINO model in a similar manner to the DINA in that only two parameters are estimated: the intercept and a common slope. The difference is that for every attribute measured by the item, the item parameters must counteract so that there are only two possible probabilities expressed by the model. This counteracting requires less-straightforward constraints of the LCDM to yield the DINO model. These constraints are detailed elsewhere (Henson et al., 2009), but ultimately the slipping parameter for the DINO model is defined by

$$1 - s_i = \frac{\exp(\lambda_{i,0} + \lambda_{i,1}(1 - \prod_{a=1}^{A} (1 - \alpha_{ea})^{q_{ia}}))}{1 + \exp(\lambda_{i,0} + \lambda_{i,1}(1 - \prod_{a=1}^{A} (1 - \alpha_{ea})^{q_{ia}}))}. \tag{11}$$

Because each item has only two parameters, the difference in parameters for the nested and full versions of the DINA and DINO models is the difference in the number of attribute profiles measured. As with the LCDM/HDCM Type-I error study, for each simulated data set, for both the DINA and DINO models, the null model (i.e., the nested model with fewer attribute profiles because of the hierarchy) was used as the baseline where three comparisons were made: the 6-profile versus the full 8-profile (T6-8); the 4-profile versus the 6-profile (T4-6); and the 4-profile versus the full 8-profile (T4-8).

We examined the Type-I error rates using the naïve hypothesis test along with the error rate using the three information criteria. Table 4 shows the Type-I error rates and information criteria error rates for all three comparisons (T6-8, T4-6, and T4-8). For the DINA model, the results showed inflated Type-I error rates for each comparison, with 100 % Type-I error rates for the T4-6 and T4-8 conditions. Further, in the T4-6 and T4-8 conditions, the information criteria selected the wrong model 100 % of the time, and they selected the wrong model more than 50 % of the time in the T6-8 condition. Simply put, *the DINA model cannot detect attribute hierarchies* as the measurement portion of the model is too rigid to show missing parameters.

A similar, yet less drastic set of results was found for the DINO model, as the pattern of parameters in the DINO model includes fixed values for main effects and interactions that may dampen the impact of the hierarchy and explain the deviation from the DINA model results. In the T6-8 condition, 100 % Type-I error rates for all levels of significance and information criteria

TABLE 4.
Simulation study alternate DCM hierarchy hypothesis test and information criteria.

Type-I error rate at:	DINA model Hypothesis test		
	T6-8 ($df = 2$; 393R)	T4-6 ($df = 2$; 455R)	T4-8 ($df = 4$ 455R)
0.01	0.761	1.000	1.000
0.05	0.863	1.000	1.000
0.10	0.911	1.000	1.000
0.20	0.933	1.000	1.000
0.50	0.980	1.000	1.000
	Information criteria model selection error rates		
AIC error rate	0.924	1.000	1.000
BIC error rate	0.492	1.000	1.000
SSABIC error rate	0.749	1.000	1.000
Type-I error rate at:	DINO model Hypothesis test		
	T6-8 ($df = 2$; 482R)	T4-6 ($df = 12$; 330R)	T4-8 ($df = 4$; 455R)
0.01	1.000	0.003	0.055
0.05	1.000	0.024	0.116
0.10	1.000	0.038	0.163
0.20	1.000	0.113	0.262
0.50	1.000	0.323	0.468
	Information criteria model selection error rates		
AIC error rate	1.000	0.241	0.187
BIC error rate	1.000	0.000	0.000
SSABIC error rate	1.000	0.018	0.020

error rates were found. Under the T4-6 and T4-8 conditions, however, the DINO model found more appropriate error rates that were under each level of significance and found well-performing information criteria (the BIC had a 0 % error rate). We speculate this improvement in attribute hierarchy detection for these conditions is due to the presence of a more limiting hierarchy in the true models.

Table 2 also lists the attribute profile and marginal attribute correct classification rates for the DINA and DINO models. When compared with the LCDM/HDCM, the DINA and DINO models had markedly reduced classification accuracy, by as much as 35 % for the 8-profile DINO. However, when the differing structural models were compared within the DINA or DINO, the same general trend found in the LCDM/HDCM was observed: An over-specification of the attribute structural model (with more profiles) has less of a negative effect on classification accuracy compared to an under-specification of the attribute structural model (with fewer profiles). The restricted parameter structures of the alternate DCMs cause difficulty in detection of attribute hierarchies and leads to diminished classification accuracy.

10. The Numerical Example Continued: Testing for Attribute Hierarchies in the ECPE

To evaluate the HDCM in an empirical setting, we return to the analysis of the 2,922 examinees who took the ECPE. As described previously, the ECPE is thought to have measured three grammar attributes (1: morphosyntactic rules, 2: cohesive rules, 3: lexical rules). The full LCDM results of the ECPE data suggested the presence of an attribute hierarchy, with approximately 95 % of the sample following a structure in which the morphosyntactic attribute was

TABLE 5.
ECPE HDCM (full linear hierarchy) item parameter estimates and standard errors.

i	$\lambda_{i,0}$	$\lambda_{i,1,(1)}$	$\lambda_{i,1,(2)}$	$\lambda_{i,1,(3)}$	$\lambda_{i,2,(1(2))}$	$\lambda_{i,2,(1(3))}$	$\lambda_{i,2,(2(3))}$
1	0.87 (0.07)		0.61 (0.23)		1.04 (0.27)		
2	1.08 (0.07)		1.16 (0.14)				
3	−0.32 (0.08)			0.35 (0.13)		1.30 (0.14)	
4	−0.16 (0.08)			1.69 (0.11)			
5	1.06 (0.09)			2.00 (0.16)			
6	0.85 (0.09)			1.66 (0.13)			
7	−0.04 (0.09)			0.92 (0.14)		1.95 (0.23)	
8	1.53 (0.08)		1.77 (0.22)				
9	0.11 (0.08)			1.19 (0.10)			
10	0.11 (0.06)	2.06 (0.16)					
11	−0.02 (0.08)			0.96 (0.14)		1.47 (0.14)	
12	−1.79 (0.12)			1.34 (0.17)		1.52 (0.14)	
13	0.70 (0.06)	1.65 (0.16)					
14	0.21 (0.05)	1.41 (0.12)					
15	0.96 (0.10)			2.13 (0.16)			
16	−0.05 (0.09)			0.86 (0.14)		1.51 (0.20)	
17	1.39 (0.10)			0.45 (0.28)			0.97 (0.34)
18	0.90 (0.09)			1.40 (0.13)			
19	−0.22 (0.09)			1.85 (0.11)			
20	−1.39 (0.10)			0.95 (0.15)		1.63 (0.15)	
21	0.19 (0.09)			1.12 (0.14)		1.11 (0.20)	
22	−0.90 (0.10)			2.24 (0.12)			
23	0.71 (0.08)		1.96 (0.17)				
24	−0.62 (0.07)		1.44 (0.11)				
25	0.11 (0.05)	1.18 (0.12)					
26	0.15 (0.08)			1.12 (0.10)			
27	−0.82 (0.06)	1.69 (0.11)					
28	0.54 (0.09)			1.75 (0.12)			

Note: Attribute 1: Morphosyntactic rules; Attribute 2: Cohesive rules; Attribute 3: Lexical rules.

nested within the cohesive attribute, which was nested within the lexical attribute. The data were analyzed using the HDCM with four profiles representing the suspected linear attribute hierarchy. To describe the HDCM, we first present estimates of the model parameters and then construct the hypothesis test for the attribute hierarchy using simulation methods for the p-value.

Table 5 lists the item parameter estimates for the nested-attribute HDCM, and Table 1 lists the item parameters estimates for the full LCDM. In comparison to the LCDM, the HDCM intercepts were nearly identical, with a Pearson correlation of 0.999. Unlike the LCDM, however, the estimates of the HDCM item parameters, both in terms of main effects and nested interactions, exhibited more stable behavior, such that no parameters were near their boundary values and the largest standard error was 0.34 (as compared to 2.10 in the fully crossed LCDM). For example, consider Item 12, measuring morphosyntactic and lexical rules. Under the fully crossed LCDM, the intercept was −1.77 (0.11), the main effect of morphosyntactic rules was 0.00 (0.00; fixed by *Mplus* at the boundary), the main effect of lexical rules was 1.29 (0.16), and the interaction between the two attributes was 1.52 (0.14). Under the HDCM, the intercept was −1.79 (0.12), approximately the same as the LCDM. The HDCM does not have a main effect of morphosyntactic rules as this attribute is nested within the lexical rules attribute. The main effect of lexical rules under the HDCM was 1.34 (0.17), only slightly higher than the LCDM estimate. Under the HDCM, the nested interaction of morphosyntactic rules and lexical rules was 1.52 (0.14),

TABLE 6.
Comparison of classifications between HDCM and LCDM for ECPE data.

HDCM	LCDM classification (most likely profile)							
	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_3$	$\boldsymbol{\alpha}_4$	$\boldsymbol{\alpha}_5$	$\boldsymbol{\alpha}_6$	$\boldsymbol{\alpha}_7$	$\boldsymbol{\alpha}_8$
$\boldsymbol{\alpha}_1^*$	960	0	0	2	8	0	7	1
$\boldsymbol{\alpha}_2^*$	5	296	0	0	0	7	0	29
$\boldsymbol{\alpha}_3$	0	0	0	0	0	0	0	0
$\boldsymbol{\alpha}_4^*$	2	0	0	463	0	0	2	83
$\boldsymbol{\alpha}_5$	0	0	0	0	0	0	0	0
$\boldsymbol{\alpha}_6$	0	0	0	0	0	0	0	0
$\boldsymbol{\alpha}_7$	0	0	0	0	0	0	0	0
$\boldsymbol{\alpha}_8^*$	3	7	0	27	0	2	3	1015

Note: The HDCM profiles denoted with an * are estimated as part of the attribute hierarchy.

which was identical to the crossed model interaction under the LCDM. The results indicate that the LCDM attempts to use four parameters to fit three values (the three average log-odds for the combinations of attributes mastered by examinees), whereas the HDCM uses the appropriate number of parameters due to the nested-model structure.

Figure 1 lists the estimated structural model parameters of the HDCM and the fully crossed LCDM. Of note are the four structural model parameters that are fixed to zero in the HDCM and estimated in the LCDM, with a sum of 0.05. The HDCM distributed the probability mass of these parameters more toward the first and second profiles (mastery of no attributes and mastery of lexical only), with both showing an increase in the probability of an examinee having each profile of 0.019 and 0.015, respectively. The remaining probability mass was distributed almost equally between the other two profiles. Related to the structural model results, as a comparison of the examinee classification estimates of the two models, Table 6 displays the cross-classification count for the most likely profiles assigned to examinees by the HDCM and the LCDM. Agreement between the two models was very high, providing matching classifications for 93.6 % of examinees (Cohen's kappa of 0.909). The two models provide similar classifications of examinees with only a few cases where classifications differed substantially.

Finally, to formulate the hypothesis test of whether an attribute hierarchy existed, we constructed the deviance test statistic and used both the naïve distribution and a simulation to obtain the correct p-value. Specifying the 4-profile HDCM as the null model that was nested within the 8-profile LCDM, we first calculated the deviance test statistic as -2 times the difference in model log-likelihood values, which was 23.06. The naïve hypothesis test compares this value to a Chi-Square distribution with degrees of freedom equal to the difference in number of estimated parameters in the two models, which was 13, resulting in a p-value of 0.039. From the simulation study, we know this p-value is very conservative, such that if we had chosen 0.05 as our level of significance, we would have a high degree of confidence in rejecting the null hypothesis that the attribute hierarchy was present. However, given the p-value is between 0.05 and 0.01, researchers may either reject *or* fail to reject this test. Therefore, to demonstrate the appropriate way to construct the test statistic, we conducted a simulation approach to determine a more appropriate p-value. Using the HDCM model parameter estimates, we simulated 100 data sets and fit both the HDCM and the LCDM to each. The number of simulated data sets was chosen due to the time each analysis takes to complete, particularly for the overspecified alternative model, the LCDM, which could take over a day to finish. For each of the 76 cases where both models converged, we calculated the deviance statistic and used those deviances to

construct our p-value. The average deviance statistic from the simulation was 7.74, with a maximum of 52.62 and a minimum of 0.02. Our observed deviance statistic of 23.06 was higher than all but one case (the maximum), meaning our estimated p-value for the hypothesis test was $1/76 = 0.013$.

Based on the results of the hypothesis tests, researchers selecting a 0.01 level of significance would fail to reject the test and conclude an attribute hierarchy was present in these data whereas researchers selecting a 0.05 level of significance would reject the test and conclude an attribute hierarchy was not present in these data. As part of our analysis, the HDCM had lower AIC (85,638.63 to 85,641.43), BIC (86,045.08 to 86,125.81), and sample-size adjusted BIC (85,829.21 to 85,868.44). Coupled with the performance of the information criteria in our simulation study showing virtually no errors when choosing the correct model under the HDCM/LCDM comparison, we conclude that we have an attribute hierarchy present in our data.

11. Comparisons of Linear Attribute Hierarchies with Other Models

The results of the analysis with the HDCM indicated a linear attribute hierarchy was present in these data. Linear attribute hierarchies share many similarities with unidimensional latent variable models, such as multicategory, unidimensional DCMs; ordered latent class models; and the two-parameter logistic model. Although we primarily present the ECPE results for illustrative purposes to demonstrate how to test for a suspected attribute hierarchy, in practice if such linear attribute hierarchies are detected in data, unidimensional models may be a better fit and should be investigated accordingly. To demonstrate, we compare the HDCM results with the previously mentioned unidimensional models in order to investigate and avoid overfitting the data with multiple dimensions.

We compared the HDCM to six alternative unidimensional models. The first four models are constrained versions of located latent class models (e.g., Lazarsfield & Henry, 1968; Lindsay, Clogg, & Grego, 1991), referred to here as unidimensional DCMs (UDCMs) to put them into the context of our analysis. These UDCMs differ by the number of categories the discrete attribute is assumed to have, with categories ranging from two to five. For example, the 5-category DCM is still defined by (2) with locations specified by $\alpha_c \in \{0, 1, 2, 3, 4\}$. These multicategory UDCMs help answer the question: Are there three distinct, hierarchical attributes or is there one attribute with a number of distinct levels? The maximum number of levels an attribute may have approaches 2^I, as is equal to the number of unique response patterns and examinee estimates modeled in the 2-PL IRT model. We included the 2-PL model as an alternative because UDCMs can roughly be understood as approximations to the 2-PL. Last, we included an ordered latent class model (e.g., Croon, 1990) to align with the suspected linear attribute hierarchy that yielded four classes. In comparison to the UDCMs specified above, this model is less constrained. Conditional item response probabilities in the ordered latent class model are constrained to increase monotonically across as the attribute categories, but the item response probabilities are not parameterized as a function of attribute category as in the UDCMs.

Table 7 presents model fit results comparing the HDCM to the alternative unidimensional models. All three information suggested the best fitting model for the ECPE data is the UDCM windicesith a 5-category attribute. These results are provided to emphasize that the HDCM is an intermediate step in the model-fitting process: The LCDM should be tested against the HDCM in the case of a suspected attribute hierarchy, and the HDCM should be tested against simpler models that are reasonable alternatives, as is generally recommended in the psychometric practice of parsimonious model fitting.

TABLE 7.
Model fit statistics comparing HDCM to alternative models.

Model	Number of parameters	AIC	BIC	SSA BIC
HDCM	68	85,638.63	86,045.27	85,829.21
UDCM 2 category attribute	57	86,059.48	86,400.34	86,219.23
UDCM 3 category attribute	58	85,354.26	85,701.1	85,516.82
UDCM 4 category attribute	59	85,204.00	85,556.82	85,369.35
UDCM 5 category attribute	60	85,163.48	85,522.28	85,331.64
Ordered latent class 4 class model	115	85,252.89	85,940.59	85,575.19
2PL	56	85,205.33	85,540.21	85,362.28

Note: UDCM stands for unidimensional diagnostic classification model. 2PL stands for the two-parameter logistic item response model.

12. Discussion

The presence of attribute hierarchies may be a commonly occurring phenomenon in psychometrics. Many theories, in both education and the social sciences, presume a hierarchical structure for attributes mastered by examinees. Diagnostic classification models (DCMs) are statistical models that classify examinees based on a set of discrete attributes. Up to this point, latent class-based DCMs, such as the LCDM and the models it subsumes, have assumed that attributes do not have a hierarchical structure, meaning all patterns of attributes are assumed to be present in a population of interest. The Hierarchical Diagnostic Classification Model (HDCM) was introduced in this paper to adapt the Log-linear Cognitive Diagnosis Model (LCDM) to the case where attribute hierarchies are present. The HDCM provides a link between the latent class-based DCMs and another frequently used method for cognitive diagnosis, the attribute hierarchy method. Because the HDCM can parameterize attribute hierarchies and—coupled with the statistical test vetted herein—determine the presence of attribute hierarchies, it provides a model within the DCM framework that can model phenomena with applicability to the AHM.

The simulation studies provide evidence that the HDCM can be used to detect hierarchical attribute structures. The HDCM is nested within the LCDM, allowing for the use of a simulation-constructed p-value to provide a hypothesis test where the null hypothesis is that an attribute hierarchy is present. This hypothesis test, also supplemented with the ability of information criteria to aid in hierarchy detection, provided an empirical method to assess the presence or absence of an attribute hierarchy. The HDCM can be estimated using existing commercial psychometrics packages (such as *Mplus*) and was shown to have more accurate estimation of its model item and structural parameters when compared to the LCDM, a result that is due to the exclusion of the redundant parameters of the LCDM under an attribute hierarchy.

Additionally, the simulation study compared the performance of two commonly-used DCMs, the DINA and DINO models, under the presence of an attribute hierarchy. These models showed a greatly reduced ability to detect any hierarchy present (with the DINA model having little to no success), even though the hierarchy presence resulted in a reduction in classification accuracy. The results of this study strongly caution against specifying a DINA and DINO model before the test for the presence of an attribute hierarchy has been conducted. We recommend estimating the general LCDM first and then reducing the model based on item parameters that are found to not be significant. This top-down approach will allow for the detection of an attribute hierarchy, in addition to identifying when constraints on the LCDM to yield item-based DINA- or DINO- equivalent specifications are appropriate (Henson et al., 2009). Often constraints specified in the DINA and DINO models are superficially acknowledged, but justified with neither

cognitive theory, nor empirical evidence; results from this study further substantiate the need to justify these strict model assumptions.

Attribute hierarchies, if present, are important structural features of DCMs that provide actionable information about the nature of attributes to researchers and end-users of DCMs. For researchers, the HDCM provides a means to statistically falsify theories about knowledge acquisition. In practice, if an attribute hierarchy is present, remediation plans can focus on getting examinees to study the nonmastered attributes in sequential order, matching the structure of knowledge indicated from the DCM results. In clinical settings, treatment plans can take advantage of an attribute hierarchy, treating the behavioral or cognitive characteristics that are most likely the antecedents of a disorder. If an attribute hierarchy is not detected by the DCM, these potential benefits of the hierarchical structure may be lost. The HDCM was shown to provide the ability to effectively detect attribute hierarchies. We note that any hierarchy, if detected from cross-sectional data, does not necessarily indicate a longitudinal process and that subsequent investigation should include a longitudinal analysis design.

The analysis of the ECPE with the LCDM/HDCM demonstrated how the HDCM can be used to evaluate and test for the existence of an attribute hierarchy. Specifically, under alternate approaches such as the Rule Space Model or the Attribute Hierarchy Method, the hypothesis that an attribute hierarchy exists cannot be tested using statistical hypothesis tests. Instead, goodness of fit indices summarize model fit, with no ability to provide inference at a population level, nor do they account for a balance of model complexity with model-data fit. The HDCM provides a mechanism to statistically test and reject the null hypothesis that a hierarchy is present. If evidence from the HDCM suggests there is a hierarchy present, one could envision using the AHM or RSM in analyses with a degree of confidence about the attribute structure. Further comparisons of the HDCM and AHM are needed and are left to future research.

The ECPE data presented a special case in which a linear attribute hierarchy better represented the measured construct than the fully crossed attribute profiles; however, a single multicategory attribute better represented the construct than the linear attribute hierarchy. Although unidimensional models are an alternative to test against linear attribute hierarchies, we emphasize that the ECPE data showed better fit with the unidimensional models because the suspected linear attribute hierarchy may not have been truly a linear attribute hierarchy, not because unidimensional models always fit better in the presence of a linear attribute hierarchy. To emphasize this point, we reanalyzed our simulated data that contained linear attribute hierarchies (i.e., data generated by the 4-profile HDCM) with the 5-category UDCM and the familiar 2PL IRT model. For all 500 replications, according to all three information criteria, the 4-profile HDCM was the best-fitting model. In theory, if learning progressions mapped onto an attribute hierarchy, and a test was constructed using the HDCM as the standard for screening items, then the HDCM would be the preferred model. In fact, in all cases where a linear attribute hierarchy is truly underlying the item responses, the HDCM would be preferred.

Although we provide simulated and empirical results for comparing the HDCM to more familiar IRT models in order to acknowledge statistical model-data fit considerations, we want to emphasize that the selection of a psychometric model should be determined with respect to the purpose of the test. The process of determining the best fitting model is a balance between the statistical evidence held by the items of a test and the purpose for why the test was constructed in the first place. Often in large scale assessment, miss-fitting test items are discarded so that the remaining items of the test fit the model and purpose. In our study, we did not have the benefit of the use of a large item bank for purposes of demonstration. Rather, we feel the HDCM is aligned with the curious results of Templin and Hoffman (2013). Although UDCMs or the HDCM with a linear attribute hierarchy may be seen as an approximation to a 2-PL model, the models serve different purposes: DCMs classify examinees and IRT models scale examinees. Frequently in educational testing, IRT models are used to scale examinees only to subsequently

classify them into groups with labels like "proficient" or "not proficient" by using standard setting procedures. DCMs present an alternate measurement paradigm to directly classify examinees when classification according to one or more latent variables is the purpose of the test. In addition to tests for classification, DCMs are well-purposed for tests seeking to measure more than one latent variable in general. Multidimensional IRT models require unreasonably long tests to scale multiple latent variables with a level of reliability sufficient for their subsequent reporting and use (Templin & Bradshaw, 2013), making DCMs a practical choice to suit test purposes requiring multidimensional scoring.

As with the ECPE data, DCMs are frequently applied to existing tests that were created and calibrated using models that assume an underlying continuous dimension. A common issue in such retrofitting of DCMs is dimensionality: specifying more attributes than can be measured by the test. In these situations, attributes lose their dimensionality, degenerating to having only a few attribute profiles present in the data, as shown by the structural model parameters. Such cases *may* indicate a hierarchy is present, or that a more appropriate model for the data may be a unidimensional latent variable model. The HDCM provides a mechanism to test for the presence of such hierarchies, allowing for a more critical examination of a test's dimensionality and an assessment of the appropriateness of the use of a DCM on the test, in general.

In sum, the HDCM is a new psychometric model that theoretically can be used to detect attribute hierarchies and model attributes using a reduced set of parameters from the fully crossed LCDM. We believe the HDCM expands the methodological toolbox available to researchers, practitioners, and psychometricians by adding modeling options that more accurately reflect the reality of current educational and social science theories. Furthermore, as general DCM test construction methods become better understood with practice, we anticipate better-fitting applications of the HDCM from tests designed a priori for the classification of examinees according to theorized attribute hierarchies.

Acknowledgements

This research was supported by the National Science Foundation under grants DRL-0822064, SES-0750859, and SES-1030337. The opinions expressed are those of the authors and do not necessarily reflect the views of NSF.

References

Bradshaw, L.P., & Templin, J. (2013). Combining scaling and classification: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*. doi:10.1007/S11336-013-9350-4.

Buck, G., & Tatsuoka, K.K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing*, *15*, 119–157.

Choi, H.-J. (2010). *A model that combines diagnostic classification assessment with mixture item response theory models*. Unpublished doctoral dissertation, University of Georgia, Athens, Georgia.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical & Statistical Psychology*, *43*, 171–192.

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

Gierl, M. J., Cui, Y., & Hunka, S. (2007a). *Using connectionist models to evaluate examinees' response patterns on tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gierl, M.J., Leighton, J.P., & Hunka, S.M. (2007b). Using the attribute hierarchy method to make diagnostic inferences about respondents' cognitive skills. In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: theory and applications* (pp. 242–274). Cambridge: Cambridge University Press.

Gierl, M.J., Cui, Y., & Zhou, J. (2009). Reliability of attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 293–313.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Henson, R., & Templin, J. (2005). *Hierarchical log-linear modeling of the joint skill distribution*. Unpublished manuscript.

Henson, R., & Templin, J. (2007). *Large-scale language assessment using cognitive diagnosis models*. Paper presented at the annual meeting of the National Council for Measurement in Education in Chicago, Illinois.

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.

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.

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Leighton, J.P. & Gierl, M.J. (Eds.) (2007). *Cognitive diagnostic assessment for education: theory and applications*. Cambridge: Cambridge University Press.

Leighton, J.P., Gierl, M.J., & Hunka, S.M. (2004). The attribute hierarchy model for cognitive assessment: a variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.

Leighton, J.P., Cui, Y., & Corr, M.K. (2009). Testing expert-based and student-based cognitive models: an application of the attribute hierarchy method and hierarchy consistency index. *Applied Measurement in Education*, *22*, 229–254.

Lindsay, B., Clogg, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*, 96–107.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 197–212.

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.

Muthén, L.K., & Muthén, B.O. (2013). *Mplus user's guide (Version 6.1) [Computer software and manual]*. Los Angeles: Muthén & Muthén.

Roussos, L., DiBello, L., Stout, W., Hartz, S., Henson, R., & Templin, J. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education* (pp. 275–318). New York: Cambridge University Press.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: theory, methods, and applications*. New York: Guilford.

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, *72*, 133–144.

Stoel, R.D., Garre, F.G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*, 439–455.

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. Fredrickson, R.L. Glaser, A.M. Lesgold, & M.G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453–488). Hillsdale: Erlbaum.

Tatsuoka, K.K. (1993). Item construction and psychometric models appropriate for constructed responses. In R.E. Bennett & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 107–133). Hillsdale: Erlbaum.

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: a statistical pattern recognition and classification approach. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale: Erlbaum.

Tatsuoka, K.K. (2009). *Cognitive assessment: an introduction to the rule space method*. New York: Routledge.

Templin, J. (2006). *CDM user's guide*. Unpublished manuscript.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251–275.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50.

Templin, J., Rupp, A., Henson, R., Jang, E., & Ahmed, M. (2008). *Nominal response diagnostic models*. Paper presented at the annual meeting of the National Council on Measurement in Education in New York, NY.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16).

von Davier, M., & Yamamoto, K. (2004). *A class of models for cognitive diagnosis*. Paper presented at the 4th Spearman Conference in Philadelphia, PA.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (RR-08-27). Princeton: Educational Testing Service.

Manuscript Received: 8 APR 2011