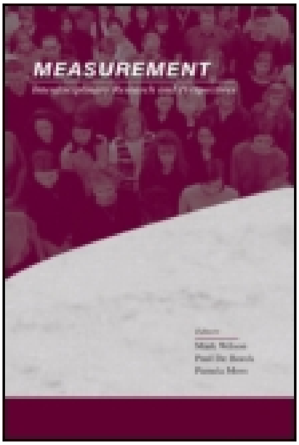


Measurement: Interdisciplinary Research and Perspectives

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmes20

Unique Characteristics of Diagnostic Classification Models: A Comprehensive Review of the Current State-of-the-Art

André A. Rupp [a] & Jonathan L. Templin [b]

[a] University of Maryland

[b] University of Georgia

Version of record first published: 17 Feb 2011.

To cite this article: André A. Rupp & Jonathan L. Templin (2008): Unique Characteristics of Diagnostic Classification Models: A Comprehensive Review of the Current State-of-the-Art, Measurement: Interdisciplinary Research and Perspectives, 6:4, 219-262

To link to this article: http://dx.doi.org/10.1080/15366360802490866

Measurement, 6: 219–262, 2008
Copyright © Taylor & Francis Group, LLC
ISSN 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366360802490866

Psychology Press
Taylor & Francis Group

FOCUS ARTICLE

Unique Characteristics of Diagnostic Classification Models: A Comprehensive Review of the Current State-of-the-Art

André A. Rupp
University of Maryland

Jonathan L. Templin
University of Georgia

Diagnostic classification models (DCM) are frequently promoted by psychometricians as important modelling alternatives for analyzing response data in situations where multivariate classifications of respondents are made on the basis of multiple postulated latent skills. In this review paper, a definitional boundary of the space of DCM is developed, core DCM within this space are reviewed, and their defining features are compared and contrasted with those of other latent variable models. The models to which DCM are compared include unrestricted latent class models, multidimensional factor analysis models, and multidimensional item response theory models. Attention is paid to both statistical considerations of model structure, as well as substantive considerations of model use.

Key words: review, diagnostic classification models, item response theory, factor analysis, multidimensional models, latent class models, latent variable models

Correspondence should be addressed to André A. Rupp, Department of Measurement, Statistics, and Evaluation, University of Maryland, 1230 Benjamin Building, College Park, MD 20742. E-mail: ruppandr@umd.edu

INTRODUCTION

Over the last 20 years there has been a renewed and widened psychometric interest in statistical models with latent variables that provide *multidimensional classifications of respondents* for the purpose of a fine-grained *diagnosis* (Winter 2007 special issue of the *Journal of Educational Measurement*). These models will be referred to as *diagnostic classification models* (DCM) in this review paper (Rupp, Templin, & Henson, 2010). The objective of the following exposition is to raise awareness about the unique characteristics of DCM vis-à-vis popular scaling alternatives for contexts that call for the analysis of data from *diagnostic assessments* in a certain discipline. It also serves to address the resulting advantages and disadvantages of DCM by focusing on statistical as well as substantive considerations. The objective is, thus, to provide an overview of these models in an accessible manner for a wide audience, while sacrificing neither expositional depth nor clarity.

To achieve this purpose, the paper is organized into five sections as follows. In the first section, a brief context for the utilization of DCM in educational and psychological assessment is presented. In the second section, current alternative labels for DCM are reviewed and a consensual definitional boundary of the space of DCM is developed. In the third section, DCM are then compared and contrasted with other well-known psychometric models such as *latent class models* (Haagenars & McCutcheon, 2002), *factor analysis* (FA) models (McDonald, 1999) and various *item response theory* (IRT) models (Embretson & Reise, 2000) by drawing on nine statistical and substantive characteristics grounded in the definition proposed in the second section. In the fourth section, a taxonomy of the core DCM that are currently advocated and used in the methodological literature is presented in which the models are organized on the basis of three key statistical properties. In the fifth section, challenges in estimating DCM and in assessing goodness-of-fit at the model, item, and respondent level are discussed in detail. The paper closes with a section that lays out uncharted areas for future research.

SECTION 1: CONTEXT OF USE

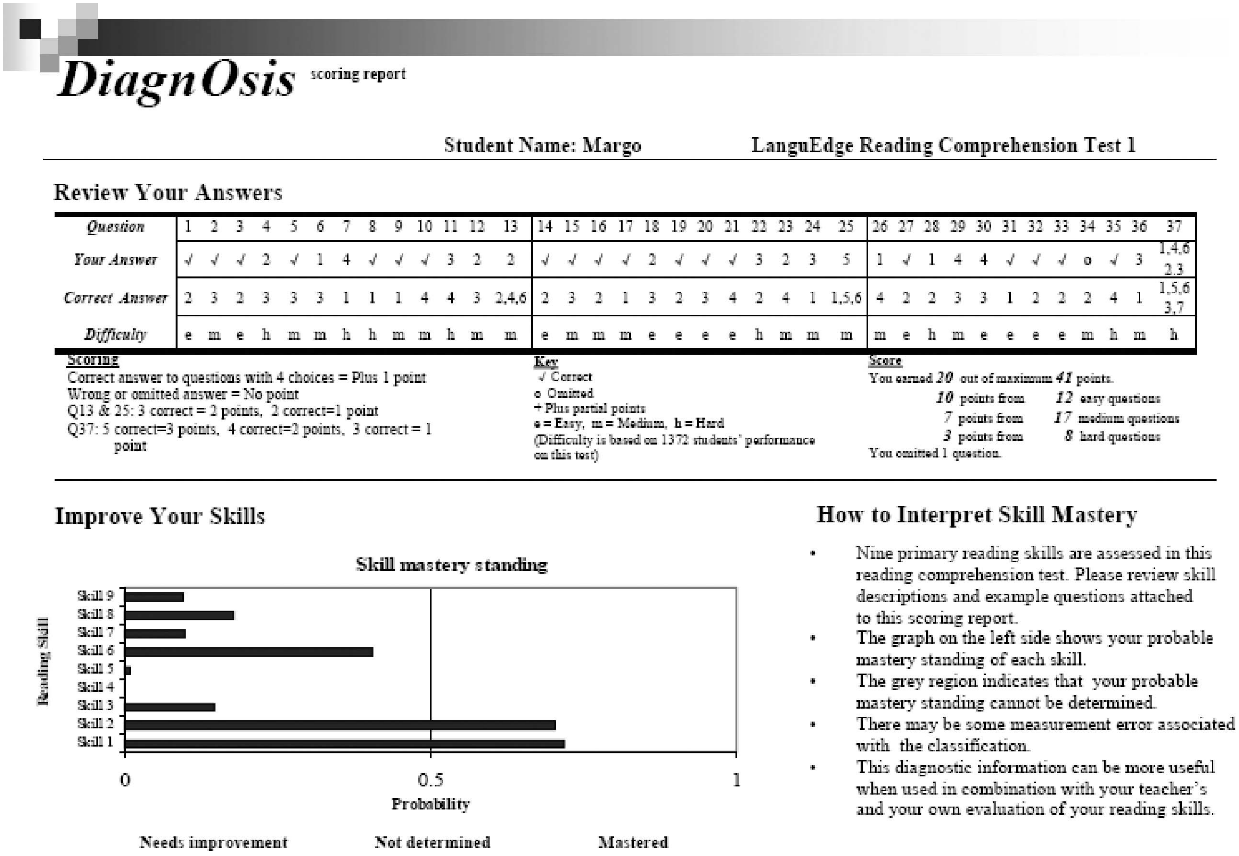
On one end of the application spectrum, DCM can be used to test rather precise hypotheses about the nature of the response processes that respondents engage in when they react to assessment or questionnaire items. This context is particularly representative of current trends in *cognitively diagnostic educational assessment* (Leighton & Gierl, 2007; Nichols, Chipman, & Brennan, 1995). If the data-collection design and the substantive response theory are developed to a sufficient degree, detailed empirical information about the mental components that are involved in the response processes and the manner in which these components interact can be obtained in this case. In an educational assessment context, it is commonly believed that an identification of these mental components helps to identify remedial pathways toward mastery on all components that are instructionally relevant and educationally meaningful to the respondents (diBello, Roussos, & Stout, 2007).

To illustrate this idea, consider a context where experts in didactics, teachers, and measurement specialists have collaborated to design a formative *diagnostic assessment* of reading comprehension. The purpose of this assessment is to provide detailed, fine-grained feedback to learners about their mastery of the necessary component skills of reading (e.g., determining word meaning out of context, comprehending negatively stated information) and to illustrate to the learners the pathways that they can take to remedy those skills that they have not yet sufficiently mastered. An excellent example of a feedback mechanism for such an assessment that is given to learners, their teachers, and their parents is the *diagnostic report card* presented in Figure 1 that is taken from the dissertation by Eunice Jang (2005).

The section of the diagnostic report card that is of most interest for the purpose of this review paper is the skills profile in the lower left-hand corner, which shows, for an individual learner, the estimated probabilities for having mastered each of the nine skills measured by the reading assessment. The card in Figure 1 thus shows that Margo has most likely already mastered skills 1 and 2, is most likely approaching mastery of skill 6, and has almost certainly not yet mastered the remaining six skills. Note that all of these inferences assume that a probability exceeding .5 reliably indicates mastery of a skill. This nine-dimensional skill profile was estimated using a rather complex DCM, and it is the statistical and substantive properties of the DCM behind such skill profiles that are the focus of the following narrative.

At this early juncture it is worth reflecting on what is typically meant by the term “cognitive response processes” when it is used in educational assessment contexts that are concerned with applying DCM to data from diagnostic assessments. As discussed, for example, by Rupp (2007) and Mislevy (2007), the connotations of the word “cognition” in educational assessment differ from the core meaning of the word within the discipline of cognitive psychology. Specifically, when experts in educational measurement propose new DCM and apply them to data from a diagnostic assessment to show that they are estimable, they typically draw on theories of response processing that are grounded distinctly in *applied* cognitive psychology.

This is partly a result of differing disciplinary traditions and objectives. Cognitive psychology is a diverse field that includes research in perception and attention, language and communication, the development of expertise, situated and sociocultural psychology, and neurological bases of cognition (Mislevy, 2008).



DiagnOsis scoring report

Student Name: Margo LanguEdge Reading Comprehension Test 1

Review Your Answers

Question	1	2	3	4	5	6	7	8	9	10	11	12	13
Your Answer	√	√	√	2	√	1	4	√	√	√	3	2	2
Correct Answer	2	3	2	3	3	3	1	1	1	4	4	3	2,4,6
Difficulty	e	m	e	h	m	m	h	h	m	m	h	m	m

Question	14	15	16	17	18	19	20	21	22	23	24	25
Your Answer	√	√	√	√	2	√	√	√	3	2	3	5
Correct Answer	2	3	2	1	3	2	3	4	2	4	1	1,5,6
Difficulty	e	m	m	m	e	e	e	e	h	m	m	m

Question	26	27	28	29	30	31	32	33	34	35	36	37
Your Answer	1	√	1	4	4	√	√	√	o	√	3	1,4,6 2,3
Correct Answer	4	2	2	3	3	1	2	2	2	4	1	1,5,6 3,7
Difficulty	m	e	h	m	e	e	e	e	m	h	m	h

Scoring
Correct answer to questions with 4 choices = Plus 1 point
Wrong or omitted answer = No point
Q13 & 25: 3 correct = 2 points, 2 correct=1 point
Q37: 5 correct=3 points, 4 correct=2 points, 3 correct = 1 point

Key
√ Correct
o Omitted
+ Plus partial points
e = Easy, m = Medium, h = Hard
(Difficulty is based on 1372 students' performance on this test)

Score
You earned ***20*** out of maximum ***41*** points.

10 points from	***12***	easy questions
7 points from	***17***	medium questions
3 points from	***8***	hard questions

You omitted 1 question.

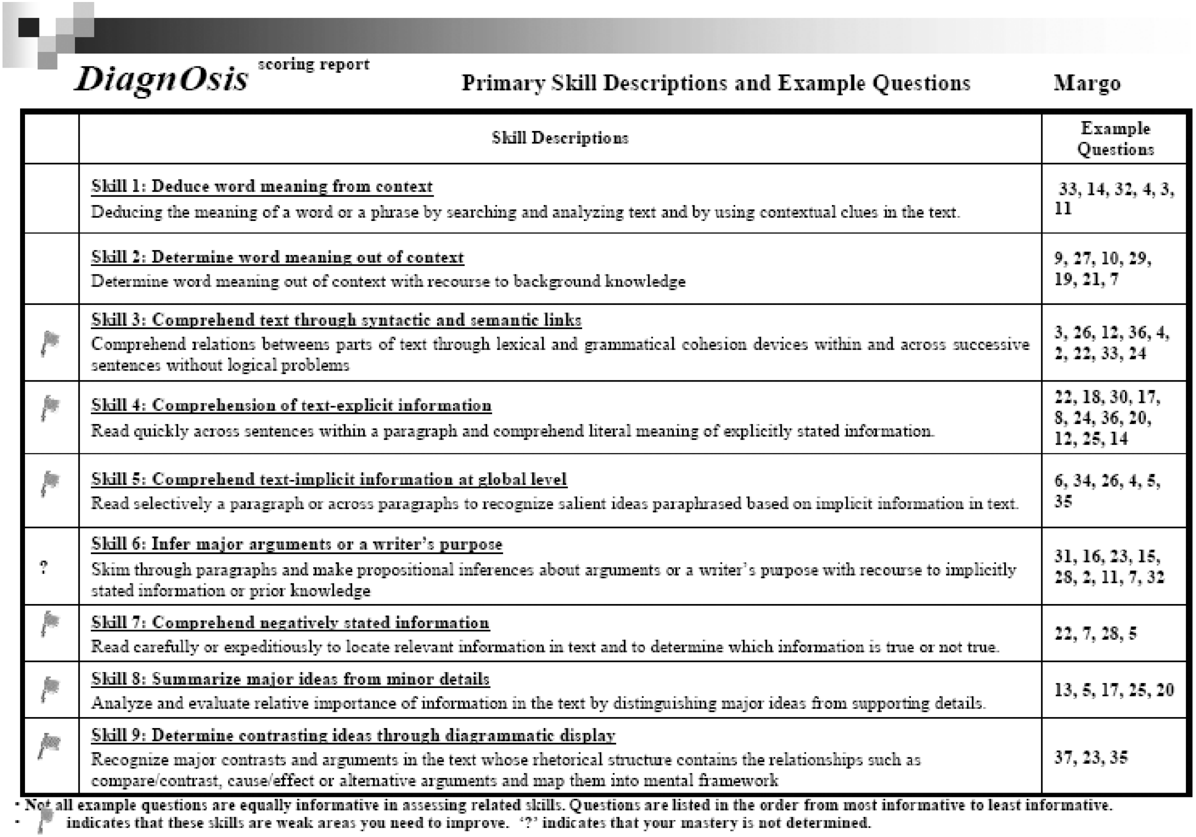
Improve Your Skills

Skill mastery standing

How to Interpret Skill Mastery

- Nine primary reading skills are assessed in this reading comprehension test. Please review skill descriptions and example questions attached to this scoring report.
- The graph on the left side shows your probable mastery standing of each skill.
- The grey region indicates that your probable mastery standing cannot be determined.
- There may be some measurement error associated with the classification.
- This diagnostic information can be more useful when used in combination with your teacher's and your own evaluation of your reading skills.

FIGURE 1 Sample report card for a reading assessment diagnosing proficiency in nine component skills (from Jang, 2005).



DiagnOsis scoring report Primary Skill Descriptions and Example Questions Margo

	Skill Descriptions	Example Questions
	Skill 1: Deduce word meaning from context Deducing the meaning of a word or a phrase by searching and analyzing text and by using contextual clues in the text.	33, 14, 32, 4, 3, 11
	Skill 2: Determine word meaning out of context Determine word meaning out of context with recourse to background knowledge	9, 27, 10, 29, 19, 21, 7
⚑	**Skill 3: Comprehend text through syntactic and semantic links** Comprehend relations betweens parts of text through lexical and grammatical cohesion devices within and across successive sentences without logical problems	3, 26, 12, 36, 4, 2, 22, 33, 24
⚑	**Skill 4: Comprehension of text-explicit information** Read quickly across sentences within a paragraph and comprehend literal meaning of explicitly stated information.	22, 18, 30, 17, 8, 24, 36, 20, 12, 25, 14
⚑	**Skill 5: Comprehend text-implicit information at global level** Read selectively a paragraph or across paragraphs to recognize salient ideas paraphrased based on implicit information in text.	6, 34, 26, 4, 5, 35
?	**Skill 6: Infer major arguments or a writer's purpose** Skim through paragraphs and make propositional inferences about arguments or a writer's purpose with recourse to implicitly stated information or prior knowledge	31, 16, 23, 15, 28, 2, 11, 7, 32
⚑	**Skill 7: Comprehend negatively stated information** Read carefully or expeditiously to locate relevant information in text and to determine which information is true or not true.	22, 7, 28, 5
⚑	**Skill 8: Summarize major ideas from minor details** Analyze and evaluate relative importance of information in the text by distinguishing major ideas from supporting details.	13, 5, 17, 25, 20
⚑	**Skill 9: Determine contrasting ideas through diagrammatic display** Recognize major contrasts and arguments in the text whose rhetorical structure contains the relationships such as compare/contrast, cause/effect or alternative arguments and map them into mental framework	37, 23, 35

- **Not all example questions are equally informative in assessing related skills. Questions are listed in the order from most informative to least informative.**
- ⚑ **indicates that these skills are weak areas you need to improve. '?' indicates that your mastery is not determined.**

FIGURE 1 (*Continued*)

While most foundational work in cognitive psychology targeted at understanding human information-processing is dominated by completely randomized experiments in laboratory settings, cognitively diagnostic educational assessment targeted at explaining performance via information processing is dominated by observational or quasi-experimental studies in real-life settings. Even though both fields utilize standardized data-collection instruments, the claims sought to be validated by each discipline are different in terms of their rhetorical organization, complexity, and level of detail. Moreover, while research in applied cognitive psychology informs research and practice in educational assessment, the opposite flow of information is effectively nonexistent.

Despite their ubiquity in the educational assessment literature, it is important to realize, however, that DCM can be applied to contexts outside of educational assessment, because their estimation does not require that the existence of their constituent latent variables be justified by theories from applied cognitive psychology. For example, DCM can provide multidimensional classifications of respondents based on behavioural dispositions in clinical psychology (e.g., Templin & Henson, 2006). At a very general level, DCM are suitable whenever statistically-driven classifications of respondents according to multiple latent traits are sought.

SECTION 2: DEFINITIONAL BOUNDARIES

Rationale for Approach

There are two essential paths along which a determination of the more exact definitional boundaries of DCM for a review paper can proceed. The first path focuses on the maximum number of connections that a broad class of statistical models share with one another, as is done, for example, in the comprehensive review of generalized latent variable models by Skrondal and Rabe-Hesketh (2004) and the comprehensive review of diagnostic latent variable models by Fu and Li (2007). This path is maximally inclusive and has the advantage that the resulting definition encourages readers to explicitly envision new models that fit this general framework. However, it has the disadvantage that the resulting definition of a family of models is often too broad to be of much use for practitioners and applied measurement specialists. These readers typically want to understand how particular models differ from one another in more detail rather than what the largest statistically family is that they can be subsumed under.

Therefore, an alternative path focuses on determining definitional boundaries based on an exhaustive set of defining characteristics or building blocks of existing models (Rupp, 2002) and restricting oneself to a particular subset of core models that share common characteristics that are meaningful to practitioners. This has the decided advantage that a resulting definition allows for a more detailed discussion of these models and that it provides a clear and manageable frame of reference for learning about these models. It has the disadvantage that the definitional boundary is set in a somewhat arbitrary manner, which some researchers may disagree with. Since this paper is a review paper aimed at a broad audience of practitioners and measurement specialists, and since a variety of excellent review papers on latent variable models exist, the second path was chosen.

Review of Existing Labels

There exist a variety of alternative labels for DCM that have been used in the literature including *cognitive diagnosis models or cognitively diagnostic models* (Henson & Douglas, 2005), *cognitive psychometric models* (Rupp, 2007), *multiple classification (latent class) models* (Macready & Dayton, 1977; Maris, 1999), *latent response models* (Maris, 1995), *restricted latent class models* (Haertel, 1989), *structured located latent class models* (Xu & von Davier, 2006, 2008), and *structured IRT models* (Rupp & Mislevy, 2007). Each of these labels carries with it a specific connotation that highlights a particular aspect of these models, which will be discussed in the following.

The labels cognitive diagnosis models, cognitively diagnostic models, and cognitive psychometric models refer to the *theoretical grounding* of the application of these models, particularly in educational assessment. They underscore the belief that any application of these models demands an elaboration of a theory of response processes that is grounded in cognitive psychology—specifically, applied cognitive psychology (for a comprehensive review of assessment design principles in educational assessment grounded in applied cognitive psychology see Mislevy, 2007). The word *psychometric* in the latter label further emphasizes that these models contain latent variables, rather than observed variables, as predictors. The word *diagnostic* in the former two labels highlights that the models are applied to a particular problem that requires a *diagnosis* of respondents (i.e., a fine-grained analysis of their strengths and weaknesses in some domain).

In contrast, the label multiple classification (latent class) models denotes the *statistical purpose* of these models, which is to develop a *multivariate profile* of respondents' traits that is based on *classifying* them according to their degree of mastery or disposition on each of the traits. The remaining four labels similarly focus on the *statistical properties* of these models. The term latent response models denotes the fact that response processes, when modelled with DCM, are typically decomposed into their constituent elements and that a *latent response* for each of the components is explicitly included in the model. In some models, an individual latent response is deterministic, whereas the overall response across all components is probabilistic. In other models the individual latent response is probabilistic, whereas the overall response is deterministic (i.e., the errors in responding are modelled at different levels). However, all models combine or *condense* these latent responses to predict the probability of an observable response.

The label restricted latent class models underscores that these models are used to group respondents into *unobserved (i.e., latent) classes*. Moreover, it reflects the fact that there are *restrictions* on the number of latent classes that are estimated, which in turn lead to restrictions of values for model parameters across different latent classes. The label structured located latent class models underscores that each latent class is represented in a multidimensional latent space via values on the individual scales that create these classes. As with any latent variable model, the exact location of the scale values is generally arbitrary (e.g., for a binary latent variable it could be set at –1 and +1, at –2 and +2, or at other values).

Finally, the term structured IRT models most broadly relates these models to the family of latent variable models, in particular IRT models. Unstructured IRT models are designed for homogeneous sets of items and respondents and contain one or more respondent and item parameters, depending on the complexity of the model and whether it is unidimensional or multidimensional. In contrast, structured IRT models have additional elements that allow for a representation of heterogeneity within these models, just like structured located latent class models. These elements include, but are not limited to, parameters reflecting observed group membership in a multigroup model, parameters reflecting unobserved class membership in a mixture model, parameters reflecting different response strategies, and parameters reflecting testlet dependencies.

Definition

Based on a review of the above labels of DCM, as well as a review of additional key features of these models in the literature, which are not reflected in their labels per se, the following definition is put forth for this paper.

> *Diagnostic classification models (DCM) are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCM are further able to handle complex sampling designs for items and respondents, as well as heterogeneity due to strategy use.*

Note that this definition specifically *excludes* any multidimensional latent variable model with continuous latent variables as a DCM, which is a deliberate choice. Many practitioners who read about DCM in the educational assessment literature in particular are exposed primarily to latent variable models with discrete latent variables that produce multidimensional classifications. Based on numerous discussions that the authors have had with colleagues, graduate students, and specialists from other disciplines at conferences, in technical advisory groups, or during joint projects, it seemed most critical to provide a detailed, accessible review of DCM as defined above.

Several useful reviews that include some or all of the DCM that are discussed in this review paper and the latent variable frameworks from which they emanate have appeared in the literature over the years (e.g., DiBello, Roussos, & Stout, 2007; NRC, 2001; Tatsuoka, 2002; Junker, 1999; Fu & Li, 2007; see also Mislevy, 2007; Nichols, Chipman, & Brennan, 1995; Rupp, 2007; Rupp & Mislevy, 2007). Hence, readers interested in an overview of a larger class of multidimensional latent variable models are encouraged to go directly to these sources or the additional references therein. Readers interested in a more extensive, didactically oriented introduction and discussion of DCM are similarly encouraged to read the book by Rupp, Templin, and Henson (2010).

SECTION 3: DCM AND OTHER LATENT-VARIABLE MODELS

The definition offered at the end of the previous section can be used as a basis for comparing DCM with other latent variable models that are probably more familiar to readers. Specifically, the following nine defining characteristics of DCM will be discussed in turn:

1. their multidimensional nature
2. their confirmatory nature
3. the complexity of their loading structure
4. the types of observed response variables for which they are suitable
5. the types of latent predictor variables they contain
6. the nature of the interactions of the latent predictor variables
7. the criterion-referenced interpretations they allow
8. the diagnostic nature of the interpretations
9. the types of heterogeneity they can model

As far as the word *probabilistic* in the definition is concerned, suffice it to say that the DCM in this review paper are all probabilistic models. Hence, they can be contrasted with models representing *deterministic Guttman response patterns* (Mokken, 1997) or *deterministic knowledge space formulations* that do not rely on latent variables at all (Doignon & Falmagne, 1999; Ünlü, 2006; Schrepp, 2005).

In terms of terminology, the word *skill* will be used in this review paper to generically denote the meaning of the discrete latent variables in DCM, even though other labels such as *component, characteristic, trait, knowledge, ability, disposition*, or *attribute* are used in the literature also. In addition, the term *diagnostic assessment* will be used to generically refer to the actual data-collection instrument whose data are being analyzed with a DCM.

Criterion 1: Their Multidimensional Nature

Just like multidimensional FA models and multidimensional IRT models, DCM contain *multiple* latent predictor variables, each indexing one of the postulated skills for the diagnostic assessment. The number of latent variables depends on the number of skills that researchers hope to numerically separate in a reliable manner with the assessment. Just like with other multidimensional models, larger numbers of skills are more challenging to separate empirically than smaller number of skills (Haberman, 2008; Sinhary, Haberman, & Puhan, 2007; Yao & Boughton, 2007; Rupp, 2008a). In addition to simply including multiple latent variables at the same levels, DCM for hierarchies of latent variables have also been proposed (de la Torre & Douglas, 2004; von Davier, 2007) similar to hierarchical FA models (McDonald, 1999) and hierarchical IRT models (Sheng & Wikle, 2008).

When DCM are used to model component skills in a response process whose definition is grounded in applied cognitive psychology, one can argue that they seek to map a real-life fluid, cognitive response process onto a mathematically-constructed static multidimensional model. Even though this is, to some degree, always the case when latent variable models are applied to response data, the requirements for a highly accurate mapping of the process onto multiple component latent variables are perhaps more stringent for those DCM applications. Furthermore, one can argue that typical multidimensional analyses in FA or IRT include latent variables that operationalize different constructs or different aspects of the same construct, but not elementary mental components and their interaction.

Criterion 2: Their Confirmatory Nature

DCM are confirmatory in nature. This notion, as simple as it may seem, has two important distinct shades to it. These shades arise from the difference between a *hypothesis-testing perspective* in research design, which is confirmatory in nature by definition, and a *statistical modelling perspective*, which can involve an exploratory or a confirmatory statistical model. Specifically, a confirmatory statistical model is one that contains parameter restrictions compared to its exploratory counterpart. For example, a multidimensional confirmatory measurement model is one in which certain latent variable loadings are set to "0," or one in which certain loadings are (additionally) constrained to be equal, such as in the true-score equivalence model (McDonald, 1999).

Statistically speaking, the loading structure for a DCM is referred to as its *Q-matrix* (Tatsuoka, 1983). It reflects a particular substantive hypothesis about the response process through the pattern of 0s and 1s that indicates which latent variables are associated with which items (i.e., which skills are required by which items). Of course, Q-matrices exist for any confirmatory model, and even for any exploratory model where they are filled with 1s because all observed variables are allowed to load on all latent variables. For FA models they are often simply labelled *factor loading matrices* and for traditional IRT models they are often not given a specific name at all. A very informative discussion of how various latent variable models, including certain DCM, can be compared via their Q-matrices is given in Hartz (2002).

Another way in which the confirmatory nature of a DCM analysis is visible is rarely discussed, however, perhaps because it is rather subtle. It refers to the fact that in applications that seek a close link between the cognitive response process and the modelling process of the data, the structure of the DCM could, ideally, be chosen so that it reflects the manner in which the skills interact in the response process (see criterion 5 below). For example, in order to test a hypothesis that all postulated skills for an item are actually necessary to respond correctly to an item, a *conjunctive or noncompensatory DCM* would be the natural choice, which would exclude *compensatory DCM*, as well as traditional multidimensional FA and IRT models as potential scaling options. While this does not make the DCM automatically confirmatory in a statistical sense, it makes the choice for the model reflect a confirmatory approach to modelling.

As another example, consider the *multicomponent latent trait model* (Embretson, 1980; Embretson, Schneider, & Roth, 1986) and its extension to multiple strategies (Embretson, 1985), both essentially noncompensatory, multidimensional Rasch models. In contrast to the generalized version of the first model (Embretson, 1984, 1997), these models can be used to confirm the hypothesis that the response process for each item involves all postulated skills, but they are exploratory statistical models because they do not place restrictions on model parameters compared to an unrestricted model. Nevertheless, the choice of these models reflects a belief about the structure of the response process, which gives the investigation a confirmatory flavour. In contrast, *componential IRT models*, as formulated by Hoskens and de Boeck (1995, 2001) and the model family of the *multidimensional random coefficient multinomial logit model* (Adams, Wilson, & Wang, 1997), can be used as confirmatory statistical models. Arguably, confirmatory models are preferred in statistical theory, because the power to test for model fit and parameter differences across subgroups is larger than in their exploratory counterparts.

Similarly, confirmatory FA models are quite commonly used to test substantive hypothesis about how constructs are divided into subconstructs and how constructs relate to one another. Confirmatory multidimensional IRT analyses are frequently used for the scaling of large-scale achievement tests such as PISA (e.g., OECD, 2003) or TIMSS (Mullis, Martin, & Foy, 2005). In practice, multidimensional IRT analyses are much rarer than multidimensional FAs across disciplines, however, and it is also not uncommon for them to be exploratory in nature outside of large-scale testing applications. Consequently, multidimensional analyses with DCM, which are derivatives of multidimensional IRT models, are currently even rarer in practice.

Criterion 3: Their Complex Loading Structure

Confirmatory FA and IRT models, in practice, typically possess a *simple* loading structure in the specific sense that each item only loads on one dimension (for a more detailed discussion of the technical definition of simple structure in the Thurstonian sense, see McDonald, 1999). This reduces the statistical complexity of the model and is often an artefact of the structure of the assessments to which these models are applied. For models that have simple structure, one typically finds that since each item indexes only one dimension, this dimension typically represents a rather coarsely defined construct such as “proficiency in working with numbers, equations, and functions” or “ability to verbalize negative emotions.” Therefore, it is practically sensible and feasible to instruct item developers to write items that tap only one of such broadly defined dimensions.

In contrast, DCM utilize latent variables that typically operationalize more narrowly defined constructs—for example, latent skills that are constitutive of a response process—so that each item typically requires multiple component skills. This leads to a more *complex* loading structure that is also known as *within-item multidimensionality* (e.g., McDonald, 1999; Wilson, Adams, & Wang, 1997); it is reflected in multiple “1s” in rows of the Q-matrix. This makes DCM similar in structure to componential IRT models and other *probability matrix decomposition models* (Maris, deBoeck, & van Mechelen, 1996; Meulders, deBoeck, & van Mechelen, 2003) such as the *linear logistic test model* for decomposing item difficulty parameters via multiple latent component variables (Fischer, 1997; see Adams & Wilson, 1996; Adams, Wilson, & Wang, 1997). Indeed, it is in the situation when complex loading structures exist and classifications of respondents are desired that DCM can function up to their theoretical potential best. Naturally, in applications where DCM have been *retrofitted* to assessments that were originally created to represent a rather coarsely defined single construct so that data could be scaled with a unidimensional measurement model, convergence problems during estimation, as well as poor item, respondent, or model fit, are common results.

Similarly, when certain DCM are applied to data with a simple loading structure for which analogous multidimensional FA or IRT models with continuous latent variables could be fit meaningfully also, the result is often viewed as an *informational loss*, because the multidimensional continua are merely discretized and split up into adjacent categories. This allows for classification, but reduces statistical precision. In those cases, traditional multidimensional FA or IRT models might be much more appealing, unless, of course, the classifications that result from a DCM analysis are the aspect of the analysis that is desired most.

Criterion 4: The Nature of the Observed Response Variables

There are DCM designed to handle dichotomous and polytomous response data and some that can handle dichotomous response data only. In this respect, DCM are similar to traditional IRT models or to FA models that utilize tetrachoric or polychoric correlation matrices. It should be noted that several alternative methods for treating polytomous data have been proposed for some DCM. For example, in the *reduced noncompensatory reparameterized unified model* (reduced NC-RUM), one approach utilizes a cumulative score probability formulation and another approach utilizes a binomial model (Bolt & Fu, 2004; Henson, Templin, & Porch, 2004; Templin, He, Roussos, & Stout, 2003).

Criterion 5: The Nature of the Latent Predictor Variables

Since DCM are restricted latent class models they contain categorical latent variables that allow for the creation of such classes. Of course, models with discrete latent variables can be used to approximate models with continuous latent variables and are, in some cases, statistically equivalent to them (Haertel, 1990). Most DCM and associated estimation routines allow only for dichotomous latent variables. A notable exception is the MDLTM software program for the *general diagnostic model* (GDM; von Davier, 2006) and the Arpeggio software program for the full NC-RUM, both of which also allow for polytomous latent variables (Templin, Roussos, & Stout, 2003). Hence, classifications of respondents into multiple performance categories such as “insufficient performance,” “sufficient performance,” and “outstanding performance” are possible. Thus, while DCM are conceptually similar to confirmatory multidimensional FA and IRT models, as well as componential IRT and conjunctive or disjunctive Rasch models, they differ from them in terms of the measurement scales of the latent variables they contain.

Multidimensional FA models, in particular, often further incorporate hierarchical factor structures where a higher-order factor (e.g., general intelligence) is postulated to account for the covariation of a set of lower-order factors (e.g., crystallized intelligence, fluid intelligence, spatial and verbal reasoning, working-memory capacity). This idea has also been adapted in applications of *Bayesian inference networks (*BIN; Levy & Mislevy, 2004) and has been directly incorporated into the *higher-order DINA* (HO-DINA) model (de la Torre & Douglas, 2004). While the higher-order latent variable in the latter model is continuous—just as in hierarchical FA or IRT models—the lower-order latent variables are categorical, unlike hierarchical FA or IRT models. In BIN, even the higher-order latent variables are typically categorical so that these models represent a completely categorical hierarchy of latent variables. At the same time, probabilistic relationships are not allowed to restrict the conditional probabilities in a traditional BIN. Consequently, the conditional independence of correctly applying all required skills for an item given latent class membership, for example, cannot be modelled (Hartz, 2002).

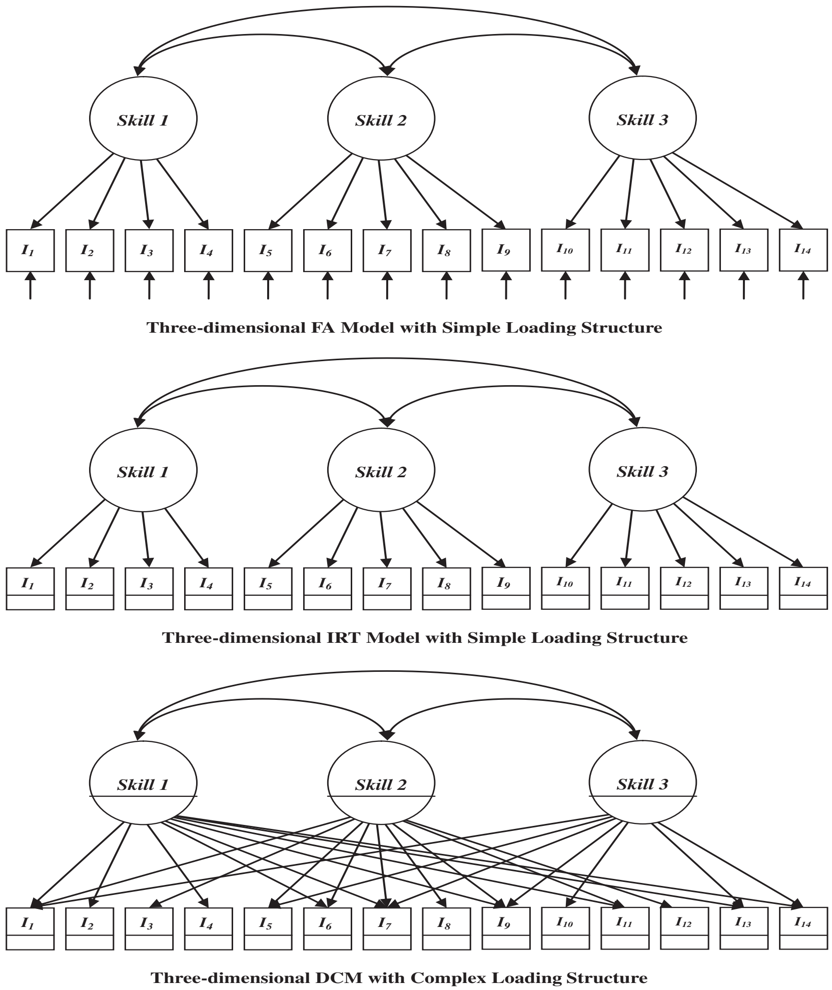
Criterion 6: The Interaction of the Latent Variables

To understand how the latent responses interact to produce an observed response, the concept of a *condensation rule* is critical. A condensation rule prescribes how the responses to the individual latent variables are combined (i.e., condensed) to produce an observed response. As stated above under criterion 2, traditional FA models and IRT models are compensatory latent variable models. This means substantively that respondents are able to compensate for a deficit on one skill by a surplus on another skill. In contrast, DCM include both compensatory and noncompensatory models. The latter are models that reflect the assumption that all skills have to be mastered for a respondent to produce a correct response. Thus, the multicomponent trait model and its extension (Embretson, 1991) and noncompensatory multidimensional IRT models (Bolt & Lall, 2003), which contain continuous latent variables, are structurally similar to noncompensatory DCM, even though DCM contain discrete latent variables.

However, currently estimable DCM do not involve *interactions between components*, even though authors such as Maris (1999) discuss this possibility. In contrast, componential IRT models explicitly allow for such interactions (Hoskens & de Boeck, 1995), which links these models to locally dependent conjunctive measurement models (Jannarone, 1997). Maris (1999) further argues that the use of a condensation function makes DCM different from latent variable models in a “narrow” sense, because DCM map the latent responses via a *function* to the observed responses, so that a joint probability distribution for all variables does not exist. Yet, one can construct what this distribution would look like from the general definition of the latent class model, so this argument is not without contention. Finally, as discussed by Haertel (1989), DCM are furthermore formally equivalent to latent distance models (Lazarsfeld & Henry, 1968) if and only if the latent response patterns generated from the latent variables form a latent Guttman scale.

Figure 2 summarizes the key characteristics of DCM vis-à-vis multidimensional FA and multidimensional IRT models. It shows a three-dimensional compensatory DCM with a complex loading structure and contrasts it with three-dimensional FA and IRT models with simple structures.

Of course, other comparisons can be easily constructed, but are omitted here for space reasons. For example, one could compare a compensatory multidimensional



Three-dimensional FA Model with Simple Loading Structure

Three-dimensional IRT Model with Simple Loading Structure

Three-dimensional DCM with Complex Loading Structure

FIGURE 2 Comparison of a three-dimensional FA and IRT model with a three-dimensional DCM.

This figure shows three models, each with three latent skill variables indicated by circles and 14 observed item variables indicated by rectangles. The two-sided arrows between the latent skill variables indicate pair-wise correlations, the one-directional arrows from the latent skill variables to the item variables indicate loadings corresponding to *Q*-matrix entries of 1, the arrows pointing to the rectangles in the first model indicate measurement error, the horizontal lines in the rectangles in the second and third model indicate latent thresholds (i.e., binary observed variables), and the horizontal lines in the circles in the third model indicate latent thresholds also (i.e., binary latent skill variables). The position of the horizontal lines is identical for simplicity purposes, but could indicate different marginal percentages correct for the item variables and different marginal mastery proportions for the latent skill variables.

FA and IRT model with a DCM where all three models would have a complex loading structure. In that case, the main difference between the FA and IRT models and the DCM would be that the latter would contain categorical latent variables, while the former two would contain continuous latent variables. It should be noted that this juxtaposition is somewhat oversimplified, however, because models can also differ in more subtle ways (e.g., in terms of the parameter restrictions they include, in terms of their likelihoods, and in terms of the estimation approaches that are available for them), but it serves as a useful didactic device.

Criterion 7: The Criterion-Referenced Interpretations They Allow

Broadly speaking, DCM are statistical models that allow for multiple *criterion-referenced* interpretations. This broadly contrasts them with multidimensional FA or IRT models with continuous latent variables, most of which allow for multiple *norm-referenced* interpretations. The simplest case for criterion-referenced interpretations is the case of a unidimensional scale that contains one cut-point that divides the latent continuum into two adjacent categories (e.g., “nonmastery” vs. “mastery,” or “not clinically depressed” vs. “clinically depressed”). However, recent developments in large-scale, standards-based assessments have seen the emergence of multiple cut points (e.g., “below standard,” “meets standard,” “exceeds standard”) for scales established via traditional FA or IRT methods (Cizek, Bunch, & Coons, 2004; Zieky & Perie, 2006). These multiple classifications along individual dimension can be captured via categorically polytomous latent variables in DCM (Almond, Yan, & Mislevy, 2007; Templin, Poggio, Irwin, & Henson, 2007).

Thus, *multiple* criterion-referenced interpretations are specifically possible in DCM in two senses. On the one hand, multiple cut-points are numerically established for each dimension in DCM that contain polytomous latent variables. On the other hand, each DCM naturally contains multiple categorical latent variables representing multiple skills, such that a final classification of respondents leads to an interpretation of this classification with reference to multiple criteria. Apart from the fact that DCM may, in some cases, be better suited to handle a complex loading structure than traditional multidimensional FA or IRT models, it is the direct numerical derivation of the cut-points leading to the multidimensional classifications that makes these models attractive to users with diagnostic needs.

Importantly, since interpretations about skill profiles with DCM are typically made at the level of the individual rather than at an aggregate level, there are important implications for the data-collection design of studies that seek to apply DCM to data from diagnostic assessments. For example, in the context of a standardized large-scale assessment for accountability purposes such as NAEP, PISA, TIMMS, or PIRLS, it is often sufficient to have a few hundred students answer each subset of questions from an item pool in linked booklets to make reliable inferences about the entire student population at the school, state, or national level. Notably, this is also a result of the fact that the resulting data are typically calibrated with separate unidimensional Rasch models for each subscale or one multidimensional Rasch model (Sheng & Wikle, 2007), which are rather simple models compared to most DCM. In contrast, it may be necessary to have thousands of students respond to questions on a diagnostic assessment if it is desired to fit a sufficiently flexible DCM to the data so that the person-level inferences about the multiple skills are substantively meaningful and reliable.

Criterion 8: The Diagnostic Nature of the Interpretations

As stated above, the prevalent purpose for a DCM analysis—and the multiple, criterion-referenced interpretations that follow from it—is *diagnosis*. A diagnostic decision-making process typically consists of first applying a coarser screening instrument whose objective is the general identification of the particular problem area that is most crucial for a successful *treatment* of the individual. A diagnostic assessment that is fine-tuned to providing more detailed information in the identified problem area is subsequently administered as a more precise tool to further investigate the *nature of the problem*. As a consequence of the diagnostic outcome, suitable *remedial interventions* are then selected or designed that can provide the best treatment for an individual with a certain diagnosis. Of course, the data from the diagnostic assessment would not necessarily have to be analyzed with a DCM, but diagnostic benefits may arise. This was demonstrated by Templin and Henson (2006), who applied a DCM to a pathological gambling inventory, which helped to carve out more complex patient profiles. These provided additional information, which were hidden by the coarser two-group classification.

This brief description makes clear that the context for applying a DCM to data from a diagnostic assessment differs in important ways from that of applying a FA or IRT model to data from an assessment for placement, admission, or certification purposes. Consequently, an analysis of data from a diagnostic assessment with a DCM is meaningful only if their collection and interpretation is embedded within a *comprehensive diagnostic system* consisting of repeated cycles of diagnosis, treatment, and evaluation. This crucially requires that the resulting classifications can be summarized in such a way that treatments can be practically implemented and monitored afterwards, which is different from an accountability purpose where the predominant function is the rank-ordering of respondents without any direct instructional implications.

Criterion 9: The Types of Heterogeneity They Can Model

DCM are typically utilized for investigating hypotheses about a singular response process for *all* respondents. However, they can also be useful for learning about *different response strategies of different groups of respondents for the same set of tasks*, about *different response strategies of single individuals across different tasks*, and about *different response strategies of single individuals within the same task*, if data can be collected that provide sufficient information about these cases (NRC, 2001).

As Maris (1999) points out, even a disjunctive model without any additional parameters can be interpreted as a model capturing multiple strategies for responding. Typically, however, strategy selection will be explicitly modelled through separate parameters in a DCM. In the full NC-RUM, for example, an additional product term, which reflects an interaction between a continuous latent skill variable and the categorical latent skill variables, provides an indication about the *completeness of the Q-matrix* for each item. It also provides an indication of the degree to which each respondent seems to make use of a strategy for that item that differs from that proposed via the Q-matrix. A different path was chosen by de la Torre and Douglas (2005), who represent different solution strategies within a DINA model via different Q-matrices. This is conceptually similar to the approach taken by Embretson (1997, p. 307), where a weighted mixture of conjunctive Rasch models represents the different solution strategies.

In addition, in the *mixed Rasch model* (Rost, 1990) and its adaptation to cognitive modelling (Bolt, 1999; Mislevy, Wingersky, Irvine, & Dann, 1991), different strategies are reflected in the differing item parameters across the various latent groups. In the *Hybrid model* (Gitomer & Yamamoto, 1991) an additional unidimensional class is utilized for further alternative strategies akin to the residual term in the NC-RUM. Similarly, Mislevy and Verhelst (1990) utilize cognitive theories to differentiate between different response strategies of respondents and model the responses within each of the strategy classes using different basic IRT models. However, as pointed out by Hartz (2002), the reduction of the cognitive multidimensionality without an explicit modelling of the interaction between individual skills in such models does not lead to a multiple criterion-referenced respondent classification, because continuous latent variables are utilized in these approaches.

Heterogeneity can also arise from the fact that respondent populations are diverse, with the result that different item parameter values hold for different groups (i.e., item parameters are not invariant across subpopulations), which is a common path of inquiry in research on bias at the item, testlet, or assessment level. Other types of heterogeneity arise from a complex sampling and assessment design structure for items, which is common in assessments that use multiple forms. Similarly, a complex sampling structure for respondents (e.g., students nested in schools that are nested in districts) induces heterogeneity at the respondent level. Both types of heterogeneity are common in national, large-scale accountability studies such as NAEP and international large-scale accountability studies such as PISA, TIMMS, or PIRLS. Consequently, some DCM have been proposed to handle these types of heterogeneity (von Davier, 2006, 2007), even though the research frontiers leave room for optimization at this point (Gonzales, 2008; von Davier, 2008).

Finally, Sijtsma and Verweij (1999), as well as researchers such as Leighton (2004), place a strong emphasis on the actual development of the (cognitive) response theory—and therefore the strategy choices made by individuals—based on a qualitative investigation via think-aloud protocols to capture different types of heterogeneity. Rich information from these studies can then help triangulate inferences from traditional FA or IRT analyses, as well as inferences from DCM analyses of diagnostic assessment data (Böhme & Rupp, 2008).

SECTION 4: A TAXONOMY OF CORE DCM

Notation

In this section, an organization of core DCM is presented, both nominally and analytically, which requires a consistent notation. Respondents (e.g., learners, patients) are indexed by $i = 1, \ldots, I$, stimuli (e.g., assessment items, objects for judgment) are indexed by $j = 1, \ldots, J$, and component skills (e.g., borrowing numbers, generating synonyms) are indexed by $k = 1, \ldots, K$. Observed responses of respondent i to item j are denoted $X_{ij,}$ while the latent variable / skill profile vector of a respondent is denoted $\boldsymbol{\alpha}_{i,}$ such that α_{ik} indexes whether respondent i has mastered skill k ($\alpha_{ik} = 1$) or not ($\alpha_{ik} = 0$). The pattern of 0s and 1s that indicates which skills are required for which items is captured in a Q-matrix that typically consists of dichotomous entries, such that entry q_{jk} indicates whether item j requires skill k ($q_{jk} = 1$) or not ($q_{jk} = 0$). In addition, ξ_{ij} is a latent response variable defined at the item level that denotes whether respondent i has mastered all necessary skills for item j while ζ_{ijk} is a latent response variable defined at the item × skill level that denotes whether respondent i has mastered skill k for item j. Since respondents are grouped by DCM into latent classes, which are indexed by $c = 1, \ldots, C$, all DCM treat respondents in the same latent class as indistinguishable. Thus, at the latent class level, all subscripts i for individual respondents get replaced by subscripts c for the latent classes.

Overview

The DCM that can be subsumed under the definition in section 2 include *probabilistic knowledge-space approaches* (Ünlü, 2006), the *rule-space methodology* (RSM; Tatsuoka, 1983, 1995), the *attribute hierarchy method* (AHM; Leighton, Gierl, & Hunka, 2004), the *DINA* and *NIDA* models (Junker & Sijtsma, 2001), the *higher-order DINA* (HO-DINA) model (de la Torre & Douglas, 2004), the *multi-strategy DINA* (MS-DINA) model (de la Torre & Douglas, 2005), the *DINO* and *NIDO* models (Templin, 2006; Templin & Henson, 2006), the *full noncompensatory reparametrized unified model* (full NC-RUM)/*fusion* model (Roussos et al., in press; Hartz, 2002), the *reduced NC-RUM* (Templin, 2006), the *compensatory RUM* (Templin, 2006), the *random effects reparameterized unified model* (RE-RUM; Templin & Henson, 2005), the *multiple classification latent class model* (MCLCM; Maris, 1999), the *general diagnostic model* (GDM; von Davier, 2005; Xu & von Davier, 2006), the *loglinear cognitive diagnosis model* (LCDM); (Henson, Templin, and Willse, 2007), and *Bayesian inference networks* (BIN; Yan, Mislevy, & Almond, 1993).

Importantly, the RSM as well as the AHM—which was developed as an extension of the RSM—are essentially classification algorithms and not unified statistical models that are completely embedded within a fully probabilistic framework. For these methods, an estimation of the multidimensional skill profiles typically proceeds in several distinct steps rather than in one joint estimation process. Specifically, the RSM uses a unidimensional IRT model as a starting point to obtain respondent parameters that are used for later classification. That classification is done via a Bayes classification rule that utilizes a multidimensional residual function and the Mahalanobis distance measure (Tatsuoka, 1983, 1995). Similarly, the AHM originally used a likelihood-based pattern matching approach (Leighton, Gierl, & Hunka, 2004) and has recently been extended to include a neural network approach for classification (Gierl, Cui, & Hunka, in press). Yet, neither method involves a direct statistical link between the individual latent responses and the probability of an observed response as represented by a likelihood function for the complete data that depends on all item parameters and discrete skill vectors.

Taxonomy

In order to broadly differentiate these core DCM, it is useful to jointly consider three of their defining characteristics: (1) the measurement scales of the observed response variables they can model (dichotomous vs. polytomous), (2) the measurement scales of the latent predictor variables they contain (dichotomous vs. polytomous), and (3) the manner in which the latent predictor variables are combined (compensatory vs. non-compensatory manner), which leads to the classification in Table 1.

TABLE 1
A Taxonomy of DCM

Observed Response Variables	*Latent Predictor Variables*		*Model Type*
	Dichotomous	*Polytomous*	
Dichotomous	RSM		Noncompensatory
	AHM		
	DINA		
	HO-DINA		
	MS-DINA		
	NIDA		
	BIN	BIN	
	MCLCM	MCLCM	
	NC-RUM	NC-RUM	
	RERUM		
	DINO		Compensatory
	NIDO		
	BIN	BIN	
	MCLCM	MCLCM	
	C-RUM	C-RUM	
	GDM	GDM	
	LCDM	LCDM	
Polytomous	RSM		Noncompensatory
	AHM		
	BIN	BIN	
	MCLCM	MCLCM	
	NC-RUM	NC-RUM	
	BIN	BIN	Compensatory
	MCLCM	MCLCM	
	C-RUM	C-RUM	
	GDM	GDM	
	LCDM	LCDM	

Notes. RSM = Rule-space method. AHM = Skill hierarchy method. BIN = Bayesian inference network. DINA = Deterministic inputs, noisy ‘and’ gate. HO-DINA = Higher-order DINA. MS-DINA = Multi-strategy DINA. LCDM = Loglinear Cognitive Diagnosis Model. DINO = Deterministic inputs, noisy ‘or’ gate. NIDA=Noisy inputs, deterministic ‘and’ gate. NIDO = Noisy inputs, deterministic ‘or’ gate. RUM = Reparametrized unified model / Fusion model. C-RUM = Compensatory RUM. NC-RUM = Non-compensatory RUM. GDM = General diagnostic model. LCDM = Loglinear cognitive diagnosis model. MCLCM = Multiple classification latent class model.

A few things are noteworthy about Table 1. First, the largest variety of models is available for dichotomously scored observed response variables; most of these models utilize dichotomous latent predictor variables also. Second, the C-RUM and NC-RUM, as well as the LCDM and GDM, appear in several cells of Table 1 showing that they are more flexible DCM because they are parameterized for dichotomous and polytomous data, as well as dichotomous and polytomous latent variables. The LCDM and GDM can be viewed more generally as representing *model families* consisting of a variety of compensatory DCM that arise out of restrictions placed on the parameters in the model. Third, the MCLCM and BIN appear in every cell of the table, because they represent modelling families that are even more flexible, because they can accommodate compensatory as well as noncompensatory interactions between latent variables of different types.

As stated earlier in the paper, the difference between *compensatory* and *noncompensatory* models reflects how the latent predictor variables are combined across the different skills to produce the observed responses. Recall that compensatory models allow that a deficit in one skill can be compensated for by a surplus in another skill, while noncompensatory models require that each skill is present in order to produce a correct response or the highest graded response. Recall also that the combination of a set of latent variables is formally known as a *condensation* (Maris, 1992, 1995, 1999; Maris, de Boeck, & van Mechelen, 1996), because *multiple* latent response variables are combined (i.e., condensed) to generate a *single* observed response. Condensation rules can be viewed as elementary building blocks of DCM so that other models could be theoretically relatively easily constructed. The two most commonly utilized *condensation rules* involving products of latent variables are those of *conjunction* and *disjunction*, and they relate to the concept of compensation. Note, however, that other condensations functions such as a *drop-off rule* can be flexibly defined to suit the particularities of the postulated cognitive response process (Maris, 1995); all three condensation rules are shown in Table 2.

As Table 2 shows, in a *conjunctive condensation rule*, all required skills need to be present to produce an observed response, while in a *disjunctive condensation rule*, any one required skill needs to be present to produce a maximally correct observed response. Consequently, a disjunctive condensation rule can be considered as an extreme case of a compensatory rule—even though it contains a product and not a sum—because the presence of *one* skill is able to compensate for the lack of *all other* skills. Examples of models with a conjunctive condensation rule are the DINA and NIDA models, the original MCLCM estimated in Maris (1999), the full and reduced NC-RUM, as well as the RE-RUM; in contrast, the DINO model contains a disjunctive condensation rule. Finally, note that all DCM technically assume, in the language of Maris, de Boeck, and van Mechelen (1996), that the items *dominate* the respondents, which means

TABLE 2
Three Common Condensation Rules

Label	*Rule*	*Types*
Conjunctive	$P(X_{ij}=1)=\prod_{k=1}^{K}P(\zeta_{ijk}=1)$	Noncompensatory
Disjunctive	$P(X_{ij}=1)=1-\prod_{k=1}^{K}\left(1-P(\zeta_{ijk}=1)\right)$	Compensatory
Drop-off	$P(X_{ij}=0)=1-P(\zeta_{ij1}=1)$ $P(X_{ij}=1)=P(\zeta_{ij1}=1)\cdot\left(1-P(\zeta_{ij2}=1)\right)$ $P(X_{ij}=2)=P(\zeta_{ij1}=1)\cdot P(\zeta_{ij2}=1)\cdot\left(1-P(\zeta_{ij3}=1)\right)$ $\vdots$ $P(X_{ij}=K)=\prod_{k=1}^{K}P(\zeta_{ijk}=1)$	Noncompensatory

Note. The parameter ζ_{ijk} denotes a latent response.

that the condensation rules are formulated with the required skills of the items as the reference point.

Mathematical Formulas of Core DCM

The most commonly presented DCM in current research and practice that are arranged nominally in Table 2 are presented analytically in Table A1; the only exceptions are the HO-DINA and MS-DINA models, as well as BIN. The reason for this is that a BIN is a general modelling framework for representing different kinds of latent variable models, rather than a single model, and the HO-DINA and MS-DINA models were derived from the DINA model, which is included in the table.

Note specifically that atypical responses for DCM are denoted by parameters representing *slipping processes* (i.e., unexpected incorrect responses) and *guessing processes* (i.e., unexpected correct responses). These parameters are defined either as s_j and g_j at the item level—with identity restrictions across skills—as s_k and g_k at the skill level—with identity restrictions across items—or as s_{jk} and g_{jk} at the item × skill level—with separate values for each item and skill. In addition, new parameters that are related to the above parameters are defined in the GDM and RUM.

A closer inspection of the formulas for the DCM in Table A1 shows that the DINA and DINO models both model slipping and guessing processes at the item level *with equality restrictions across skills*. Consequently, the latent response variable is defined at the item level and only one slipping and guessing parameter is estimated for each *item*. In contrast, the NIDA and NIDO models both model slipping and guessing processes at the skill level *with equality restrictions across the items*. Consequently, the latent response variable is defined at the latent skill level and only one slipping and guessing parameter is estimated for each skill. The MCLCM, C-RUM, full and reduced NC-RUM, RE-RUM, and GDM can model slipping and guessing processes at the skill level *without equality restrictions across items*. Consequently, the latent response variable is defined separately for each skill and each item so that one latent response parameter can be estimated *for each entry of 1 in the Q-matrix*. Except for the MCLCM, these parameters are technically not identical to slipping and guessing parameters, but can be related to them.

Sometimes, formulas for different DCM may appear to be identical at first sight, but are slightly different upon closer inspection. For example, Table 3 shows the kernels for the NIDO, C-RUM, and GDM models, which are all

TABLE 3
Comparison of Formulas for Three Compensatory DCM

Model	*Formula*
NIDO	$P(X_{ij}=x_{ij})=\dfrac{\exp\left(x_{ij}\sum_{k=1}^{K}(\beta_k+\gamma_k\alpha_{ik})q_{jk}\right)}{\left[1+\exp\left(\sum_{k=1}^{K}(\beta_k+\gamma_k\alpha_{ik})q_{jk}\right)\right]}$
C-RUM	$P(X_{ij}=x_{ij})=\dfrac{\exp\left(x_{ij}\left\{\beta_j+\sum_{k=1}^{K}\gamma_{jk}\alpha_{ik}q_{jk}\right\}\right)}{\left[1+\exp\left(\beta_j+\sum_{k=1}^{K}\gamma_{jk}\alpha_{ik}q_{jk}\right)\right]}$
GDM	$P(X_{ij}=x_{ij})=\dfrac{\exp\left(\beta_{xj}+\sum_{k=1}^{K}x_{ij}\gamma_{jk}\alpha_{ik}q_{jk}\right)}{1+\sum_{m=1}^{M_j}\exp\left(\beta_{mj}+\sum_{k=1}^{K}m\gamma_{jk}\alpha_{ik}q_{jk}\right)}$

compensatory models that are constructed by summing across individual component skills.

A closer look at the kernels reveals that they combine the constituent parameters in different ways despite structural similarities. For example, the NIDO model weighs the entire expression in the parenthesis by the score assigned to a category, whereas the GDM only weights a part of the expression by the score. Similarly, both the C-RUM and the GDM provide a baseline probability reflected by the intercept before the sum. Yet, in the C-RUM, this value is independent of the score for a response category, whereas it depends on the score category in the GDM. In fact, what this shows is that there are relationships between some DCM that can be exploited to define families of DCM, so that some models can actually be expressed by placing restrictions on parameters in more general models. Moreover, through the choice of DCM and the skill specifications in the Q-matrix, different equality constraints are imposed on model parameters across latent classes for different DCM.

DCM as Constrained Latent Class Models

An alternative approach to investigate the relationship between different DCM is to explicitly focus on their resemblance within the family of *latent class models*. Letting η_c denote the proportion of respondents within each latent class—also called the *mixing proportion/mixing parameter* of the model—a general latent class model for dichotomous responses can be represented as follows:

$$P(\mathbf{X}_i = \mathbf{x}_i) = \sum_{c=1}^{C} \eta_c \prod_{j=1}^{J} \pi_{jc}^{x_{ij}} (1 - \pi_{jc})^{1-x_{ij}} \tag{1}$$

In this representation, which is a product over *J* Bernoulli random variables for the *J* items that is summed over all *C* latent classes, the product portion can be regarded as the *measurement component* of the model, whereas the summation portion with the mixing proportions can be regarded as the *structural component* of the model. Utilizing this representation, one can view the different DCM coarsely as providing different parameterizations of the response probabilities π_{jc} in the measurement component of the model.

For example, in the DINA model

$$\pi_{jc} = P(X_{ij} = 1 \mid \xi_{jc}, s_j, g_j) = (1 - s_j)^{\xi_{jc}} g_j^{(1-\xi_{jc})} \tag{2}$$

while in the reduced NC-RUM

$$\pi_{jc} = P(X_{ij} = 1 \mid \pi_j^*, r_{jk}^*) = \pi_j^* \prod_{k=1}^{K} r_j^{*(1-\alpha_{ck})q_{jk}} \tag{3}$$

Comparing these formulas to the ones in Table A1, one notices that the subscript *i* for an individual respondent is replaced here with the subscript *c* for a latent class, because all DCM effectively group a large number of *I* respondents indexed by $i = 1, \ldots I$ into a smaller set of *C* latent classes indexed by $c = 1, \ldots, C$. In each latent class, respondents are then statistically exchangeable, because they possess the same skill mastery pattern and, thus, they all have the same probability of correct response to an individual item. This reduction is formally represented through the mixing proportion parameter via $\eta_c = P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)$ where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$ and $\boldsymbol{\alpha}_c = (\alpha_{c1}, \alpha_{c2}, \ldots, \alpha_{cK})$. Thus, the mixing proportion for an individual latent class *c* is the proportion of respondents in the population of interest with latent skill vector $\boldsymbol{\alpha}_c$.

Families of DCM

As Henson, Templin, and Willse (2007) and von Davier (2005) have shown, many of the core DCM reviewed in this paper, as well as novel models that lie in-between these models, can be represented via more general model families. Specific DCM are obtained as special cases within the more general model families by placing restrictions on model parameters. Despite their generality, these model families do not automatically subsume every last DCM that is currently available, but they generally subsume a large number of them. Thus, they hold great potential for unifying the estimation of several DCM, thereby allowing different DCM to be used for different items on the same diagnostic assessment.

SECTION 5: MODEL ESTIMATION AND FIT FOR DCM

Even though the mathematical formulas for core DCM that have recently appeared in the literature were presented in Table A1, such a presentation is incomplete without discussing their estimability with software programs or specially written code. Crucial issues in model estimation for DCM concern

1. their identifiability
2. the parametrization of the latent skill space
3. the availability of estimation software
4. challenges in obtaining convergence with the software
5. the ability of the software to handle complex data structures
6. the ability of the software to provide indices of global and local model misfit

Each of these issues will now be discussed in turn.

Model Identifiability

Model estimation first and foremost requires that DCM be identified, which means that every parameter can be estimated by a unique value. For example, the extensive research conducted by Hartz (2002) on the full NC-RUM was necessary, because the original unified model (DiBello, Stout, & Roussos, 1995) on which it was based was not identified, even though it possessed immense theoretical appeal. Therefore, Hartz (2002) reparameterized the model by combining and redefining parameters and developed extensive Bayesian estimation routines to estimate its reparameterized version.

An additional problem may arise whenever identifiability can be ensured through parameter restrictions, but they are not commensurate with the cognitive theory about the response processes. For example, Maris (1999) showed how, depending on the condensation function used in the DCM, it was necessary to either restrict all slipping or guessing parameters to 0 or 1 for all items in models involving more than two skills. Such restrictions imply substantively that slipping or guessing processes may never or always occur for all items, which may be difficult to reconcile with a theory of response processing.

Therefore, identifiability of a model should be carefully considered when novel models are proposed. For example, Fu and Li (2007) developed a very general measurement model that arises from combining numerous defining features of existing models in various combinations. The authors carefully acknowledge that not all models that can be derived from this general model will be identified, estimable with real data sets, or even theoretically desirable (for a similar argument in traditional latent variable modeling see Rupp 2002).

Similarly, the generality of the MCLCM family has been extensively discussed by Maris (1999), who formally introduced it, but provided only one numerical example for dichotomous latent and observed variables. Some extensions were subsequently programmed in Pascal, as well as a specialized programming language (Maris, 2005, personal communication), but they cover far fewer models than the theoretical realm of possibilities allows for. Consequently, as measured by the number of publications that reference the MCLCM framework, there has been little interest in applying MCLCMs, even in psychologically-oriented disciplines.

Parameterization of the Latent Skill Space

Due to the multidimensional nature of DCM, the number of parameters that are needed to estimate the skill profiles for all respondents can explode very quickly, making alternative parameterizations of the latent skills space vis-à-vis the saturated parameterization necessary.

Saturated parameterization

Under a *saturated parameterization* the mixing proportions for all possible 2^k latent classes are estimated for each respondent, except for one proportion due to the summation constraint $\sum_{c=1}^{C} \eta_c = 1$; this amounts to estimating a very large number of parameters for most models.

The primary advantage of the saturated parameterization is that it allows one to detect logical skill hierarchies, which postulate that certain skills must be mastered before others (Leighton, Gierl, & Hunka, 2004; Tatsuoka, 1995). Under the saturated parameterization, such hierarchies are detectable by examining whether the postulated skill hierarchies are reflected in the skill pattern distributions, because patterns with nonadmissable skill combinations should have mixture proportions of zero in the population.

Even though such hierarchies may not be detectable when alternative parameterizations of the skills space are used, alternative parameterizations may be necessary to reduce the number of model parameters that need to be estimated. These parameterizations include: (a) a general loglinear model (Henson & Templin, 2005; Xu & von Davier, 2007), (b) a general tetrachoric model (Hartz, 2002), and a structured tetrachoric model (Templin et al., 2007; Templin & Henson, in press).

Loglinear parameterization

Under a *loglinear parameterization* of the skills space, the skill associations are modelled using a loglinear model that contains main effects and interactions up to a degree that remains estimable with current estimation routines. Furthermore, the model needs to be chosen to yield a sufficiently parsimonious representation for the unbiased and precise estimation of model parameters. For example, Henson and Templin (2005) proposed a general loglinear model that contains main effects associated with each latent skill variable, as well as all possible interactions between the latent skill variables. Formally,

$$\log(I \bullet \boldsymbol{\eta}_c) = \lambda_{(0)} + \sum_{k=1}^{K} \lambda_{(1)k}\alpha_{ck} + \sum_{k=1}^{K}\sum_{l=2}^{K} \lambda_{(l)}\alpha_{ck}^{l} + \sum_{k=1}^{K-1}\sum_{l=k+1}^{K} \lambda_{(2)kl}\alpha_{ck}\alpha_{cl} + \ldots + \lambda_{(K)}\prod_{k=1}^{K}\alpha_{ck} \tag{4}$$

where the subscripts of λ refer to the level of the interaction term, and sum-to-zero constraints are placed on the λ parameters for model identifiably. This representation shows that the model contains one intercept, one set of main effect terms, one set of product terms reflecting the interactions of the latent skill variables with themselves, and multiple sets of product terms reflecting the higher-order interactions of the latent skill variables with each other. With all main and interaction effect terms included, this general loglinear model is identical to the saturated model described in the preceding subsection, but by leaving terms out of the model the complexity of the parameterization can be significantly reduced. Deciding on which interaction effect terms to leave out depends on how much information about the latent skill space is desired.

For example, Xu and von Davier (2007) present a reduced version of this model that contains all main effects and all two-way interaction terms, but only one three-way interaction term representing the cube of each latent skill variable, because they want to capture only up to the first three moments of each latent skill variable:

$$\log(I \bullet \boldsymbol{\eta}_c) = \lambda_{(0)} + \sum_{k=1}^{K} \lambda_{(1)k}\alpha_{ck} + \sum_{k=1}^{K}\sum_{l=2}^{3} \lambda_{(l)}\alpha_{ck}^{l} + \sum_{k=1}^{K-1}\sum_{l=k+1}^{K} \lambda_{(2)kl}\alpha_{ck}\alpha_{cl} \qquad (5)$$

In a simulation study and real-data application to language-test data with the GDM, the authors show that their reduced loglinear parameterization leads to almost identical parameters as the more complex saturated specification, but results in a reduction of the number of parameters to be estimated. Specifically, only $1 + 3K + \binom{K}{2}$ parameters need to be estimated when all three moments are identified, which is the case when polytomous latent skill variables are used, and only $1 + K + \binom{K}{2}$ parameters need to be estimated when the latent skill variables are dichotomous, because neither quadratic nor cubic terms (i.e., neither $\{\lambda_{(2)k}\alpha_{ck}^{2}\}_{k=1}^{K}$ nor $\{\lambda_{(3)k}\alpha_{ck}^{3}\}_{k=1}^{K}$) are identified.

Tetrachoric parameterization

Under the *tetrachoric parameterization* of the structural part of the latent class model, the discrete latent skill variables α_{ck} for a given latent class *c* are related to underlying continuous skill variables $\tilde{\alpha}_{ck}$ via latent threshold parameters τ_k such that

$$\alpha_{ck} = \left\{ \begin{matrix} 0 & \text{if} & \tilde{\alpha}_{ck} < \tau_{ck} \\ 1 & \text{if} & \tilde{\alpha}_{ck} \geq \tau_{ck} \end{matrix} \right\} \qquad (6)$$

In other words, the discrete latent skill variable is 1 if the value on the underlying continuous latent skill variable exceeds its threshold; otherwise, the value is 0. Consequently, a multivariate normal distribution with a zero mean vector and a *tetrachoric correlation matrix* Ξ that contains the correlations between the underlying latent continuous skill variables can be used to estimate the latent threshold parameters (Hartz, 2002; Templin and Henson, 2006, Henson & Templin, 2007).

Consequently, the mixture proportion of the model can be represented as

$$\eta_c = \int_{a_1}^{b_1}\int_{a_2}^{b_2}\ldots\int_{a_k}^{b_k} \frac{1}{(2\pi)^{K/2}\,|\,\Xi\,|^{1/2}} \exp\left(-\frac{1}{2}\tilde{\boldsymbol{\alpha}}^T\,\Xi^{-1}\tilde{\alpha}\right) d\tilde{\alpha}_K \ldots d\tilde{\alpha}_2 d\tilde{\alpha}_1 \tag{7}$$

where

$$a_k = \begin{Bmatrix} \tau_k & \text{if} & \alpha_{ck} = 1 \\ -\infty & \text{if} & \alpha_{ck} = 0 \end{Bmatrix}, \quad b_k = \begin{Bmatrix} \infty & \text{if} & \alpha_{ck} = 1 \\ \tau_k & \text{if} & \alpha_{ck} = 0 \end{Bmatrix} \tag{8}$$

In other words, the mixture proportion is a multidimensional integral with dimensional lower and upper bounds of $-\infty$ and τ_k, respectively, for skills that have not been mastered and with lower and upper bounds of τ_k and ∞, respectively, for skills that have been mastered. In this model, only $K + \binom{K}{2}$ parameters need to be estimated, which are K threshold parameters and $\binom{K}{2}$ tetrachoric correlations.

Constrained tetrachoric parameterization

The fourth parameterization for the structural part of the latent class model, the *constrained tetrachoric parameterization*, builds on the previous unconstrained tetrachoric parameterization and imposes a hierarchical factor structure upon the tetrachoric correlation matrix Ξ. Formally, $\Xi = \Lambda\Phi\Lambda^{\mathbf{T}} + \Psi$ where $\Lambda_{K \times F}$ is the patterned loading matrix containing the loadings of the F hierarchical latent factors onto the K latent skill variables, which are treated as indicators, $\boldsymbol{\Phi}_{F \times F}$ is the matrix of the higher-order correlations of the F latent factors, and $\boldsymbol{\Psi}_{K \times K}$ is the diagonal matrix of the of the F latent factors.

That is, even though the idea is the same as a linear decomposition of a tetrachoric correlation matrix for discrete indicator variables in traditional FA (McDonald, 1999), the indicator variables in this context are not observed but latent. As a result of this parameterization, the number of parameters that needs to be estimated is further reduced from the unstructured tetrachoric parameterization. For example, for a higher-order model with one factor, the number of parameters to be estimated is reduced from $K + \binom{K}{2}$ to $2K$, which are one loading λ_k and one threshold parameter τ_k for each latent skill variable (Templin, Henson, Templin, & Roussos, 2008).

Using real data on pathological gambling, Templin and Henson (2006) used the DINO model for the measurement component of the latent class model and estimated a constrained log-linear model with three higher-order factors for the structural component of the latent class model. The number of estimated parameters for the 10 latent skill variables included a total of 33 parameters (i.e., 15 factor loadings, 15 factor thresholds, and 3 higher-order inter-factor correlations), which was much smaller than either 55 parameters (i.e., 10 thresholds and 45 tetrachoric correlations) under the unstructured tetrachoric parameterization or $2^{10} - 1 = 1023$ parameters under a saturated parameterization.

In sum, the general loglinear, general tetrachoric, and structured tetrachoric parameterizations of the structural part of the latent class models are used to reduce the number of the parameters for the skills space vis-à-vis the saturated model. The primary motivation for this is to make the estimation computationally feasible, especially for data structures that contain a large number of skills. To illustrate this reduction, Table 4 shows the number of parameters that need to be estimated under the different parameterizations for dichotomous items generally and for assessments with 4, 8, and 12 skills specifically.

Table 4 shows specifically that a reduced loglinear and unconstrained tetrachoric parameterization lead to very similar reductions of the number of parameters to be estimated, while a structured tetrachoric parameterization reduces this number significantly.

Availability of Estimation Routines or Software Programs

As already alluded to, model estimation also requires that appropriate software programs or estimation routines be available. In comparison with many excellent

TABLE 4
Number of Parameters under Different Skill Space Parameterizations

Model description		*# Parameters for sample tests*		
Parametrization	*# of Parameters*	*4 skills*	*8 skills*	*12 skills*
Saturated model (general or loglinear)	$2^K - 1$	15	255	4095
Loglinear model (reduced form)	$1 + K + \frac{K(K-1)}{2}$	11	37	145
Tetrachoric model (unconstrained)	$K + \frac{K(K-1)}{2}$	10	36	144
Tetrachoric model (constrained one-factor model)	$2K$	8	16	24

software programs that are at an analyst's disposal for traditional latent variable models, only a few programs are easily accessible and well documented for DCM. Table 5 lists the most important currently available options.

Notably, not included in Table 5 are programs for estimating BIN, because there are numerous programs available on the market. These include freeware programs such as *Winbugs* (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml), *MSBNx* (http://research.microsoft.com/adapt/MSBNx), and *Genie* (http://genie.sis.pitt.edu); commercial programs such as *Hugin* (http://www.hugin.com), *Netica* (http://www.norsys.com), and *ERGO* (http://www.noeticsystems.com), as well as programs with a research license such as *StatSho*p (ralmond@ets.org).

Applications for BIN are becoming more popular (Almond, Williamson, Mislevy, & Yan, in press; Pourret, Naïm, & Marcot, 2008), but they may be accompanied by substantial estimation challenges that may require estimation approaches that are specifically tailored to a particular application (Levy & Mislevy, 2004). Moreover, they may also require prohibitively large sample sizes for reliable parameter estimation if prior distributions, which are normally provided by expert committees, need to be empirically estimated and if statistical tests for the omission of paths and effects need to be reliably conducted.

While the existence of routines in programming languages such as *R* (www.r-project.org) and *Ox* (www.doornik.com/products.html), and the availability of the freeware interface for *Mplus* (Muthén & Muthén, 2006) as listed in Table 5 are laudable, application problems may equally remain. For example, the routines in *R* and *Ox* are not yet accompanied by user-friendly manuals and graphical user interfaces that make them attractive for people with less experience in estimation and programming. In contrast, the DCM interface for Mplus

TABLE 5
Software for Estimating DCM

Software	*Type of software(Contact)*	*Estimated models*
Arpeggio	Commercial(www.assess.com)	Full NC-RUM, reduced NC-RUM
DCM	Freeware (requires the commercial version of Mplus)(jtemplin@uga.edu)	DINA, NIDA, DINO, NIDO, reduced NC-RUM, C-RUM
DCM in R	Freeware (requires freeware R) (alexander.robitzsch@iqb.hu-berlin.de)	DINA, DINO
DINA in Ox	Freeware (requires freeware Ox) (j.delatorre@rutgers.edu)	DINA, HO-DINA, MS-DINA, G-DINA
LCDM	Freeware (requires the commercial version of Mplus)(jtemplin@uga.edu)	LCDM family
MDLTM	Research license(mvondavier@ets.org)	GDM family
BUGLIB	Research license(tatsuoka@prodigy.net)	RSM
AHM	Research license(mark.gierl@ualberta.edu)	AHM

has a short but concise user manual. Other programs such as BUGLIB or MDLTM are currently only available as research licenses to researchers upon request; BUGLIB, furthermore, requires a licensing fee to be paid. The Arpeggio program for estimating the full NC-RUM has recently been made available commercially, and other research-only programs will hopefully make this step also, so that publication restrictions of research licenses will become obsolete in the future.

Challenges in Achieving Convergence

Even if estimation routines or software packages are available, DCM are identified, and suitable data have been collected, the estimation of DCM can be rather involved, especially as the complexity of the DCM and the test structure increases. On the one hand, DCM such as the DINA and DINO models or the Rasch-type GDM are relatively simple to estimate due to their equality restrictions of item parameters, such that convergence can often be obtained for sample sizes consisting of a few hundred respondents, especially when tests are of moderate length (e.g., 20 or 40 items) and the number of involved skills is moderate also (e.g., 4 or 6).

On the other hand, test data that researchers want to analyze with more complex models, such as the full NC-RUM or a complex BIN, require much larger sample sizes and number of items, because higher-dimensional latent skill spaces that are defined by items that have within-item multidimensionality require large a number of items for each skill. Of course, this curse of dimensionality is familiar to researchers working with traditional multidimensional FA and IRT models that contain simple structures, where it is equally unreasonable to expect reliable score estimation with only few pieces of information for each dimension (Rupp, 2008a). In contrast, a purely computational method such as the RSM or the AHM can handle a larger number of skills than a fully parameterized statistical model such as the full NC-RUM, but the drawbacks are that it provides less statistical power to assess model fit and provides few inferential statistics.

The speed of convergence of an estimation routine is also related to the estimation algorithm or approach that is used. For example, since the DINA routines in Ox and R (de la Torre, in press), as well as the GDM code in MDLTM (von Davier, 2006) use an EM-algorithm, convergence can often be achieved within seconds or minutes. In contrast, the full NC-RUM and BIN are estimated within a fully Bayesian framework (Gelman, Carlin, Stern, & Rubin, 1995; Kim & Bolt, 2007; Lynch, 2007; Rupp, Dey, & Zumbo, 2004), and analyses can take several hours or even days until convergence is achieved. Moreover, interpreting convergence of the estimation routines requires advanced knowledge of Bayesian estimation theory, which involves concepts such as autocorrelation, burn-in periods, as well as proposal and stationary distributions (Sinharay, 2003, 2004). Similarly, the routines in Mplus that are activated by the free DCM interface (Templin, 2006) utilize a combination of EM and Quasi-Newton optimization approaches that are currently much too slow to be of practical use for analyses with large data sets. Recently genetic algorithms have been proposed as flexible alternatives (van der Maas, Raijmakers, & Visser, 2005), which appear to be a promising alternative to these traditional estimation frameworks and approaches.

Ability of Software to Handle Complex Data Structures

In many testing situations, diagnostic assessments contain *testlets*. These testlets induce method effects leading to correlated errors that can be viewed as undesirable multidimensionality of the instrument (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Wang, 2007). To our knowledge, the impact of the magnitude of the error correlations within a testlet on the bias and precision of parameter estimates, as well as on the classification accuracy in DCM, has not been studied, even though it is known that it can lead to an underestimation of the standard error in unidimensional IRT models (Wainer et al., 2006).

Additional complexities may arise if the data have a complex sampling structure such that, for example, students are nested within schools that are, perhaps, nested within different provinces or states. It is always important to check whether the algorithm can handle data that are missing either by design, or data that are missing at random. While data from large-scale accountability studies such as NAEP have been analyzed with the GDM (von Davier, 2005; Xu & von Davier, 2008), these estimations have thus far not proceeded within a hierarchical framework. However, recently, data from the TOEFL assessment have been analyzed successfully with the GDM (von Davier, 2007), but little information about the impact of nonzero, intraclass correlations on the bias and precision of item or skill parameter estimates, as well as the classification accuracy for respondents, is currently publicly available.

From an assessment design-perspective, it is important to realize that not all latent classes that are generated by the latent predictor variables in a DCM may be statistically distinguishable if the assessment design does not contain items tapping all potential skill combinations. This is a problem that arises more frequently in DCM than in simpler latent class models (Haertel, 1989; Rupp & Templin, 2008). Specifically, if there are 2^K possible skill patterns, and all of these latent classes are truly occupied with respondents in the population (i.e., if there are no skill hierarchies in the population), then a test needs to consist of items that tap all $2^K - 1$ skill requirements to accurately classify respondents into these classes.

Model Fit

Just as with any other latent variable or statistical model, an important aspect of data analysis with a DCM, which is a prerequisite for the interpretation of model parameters, is the assessment of the fit between the postulated model and the data structure. Crucial considerations in model misfit concern the types of misfit that are possible, as well as the development of different fit statistics that can detect each of these.

Types of misfit

Misfit could be due to the selected model structure (e.g., a noncompensatory combination of latent variables was fit, whereas a compensatory combination led to the generation of the data), the postulated loading structure within the selected model structure (e.g., certain skills are missing from the Q-matrix or certain loading specifications are not supported by the data), misspecified constraints of model parameters (e.g., some slope parameters were restricted to be equal across items with similar stimulus material, but this equality is not empirically supported), or the fact that the population of respondents is not homogeneous (e.g., a mixture of two subpopulations exists in the data that is reflected in different item parameters across these subpopulations). Of course, the fit of a DCM to data will never be perfect, but hopefully it supports the postulated hypothesis inherent in the model choice and Q-matrix for a given population to a sufficient degree.

Assessment of misfit via fit statistics

The assessment of the degree of model misfit can take the different foci of *model fit*, *item fit*, and *respondent fit*. A variety of global model-fit statistics exist for factor-analytic models (Hu & Bentler, 1999), which can be classified coarsely according to whether they allow for the assessment of absolute model fit, or relative model fit for nested models. While these indices are not directly applicable to DCM, information indices for relative model fit such as Akaike's information criterion or a corrected Bayesian information criterion can be used to compare models that are not nested, both within factor-analytic frameworks and for DCM (von Davier, 2005).

Moreover, a slew of indices emanating from nonparametric and parametric frameworks within IRT has been proposed for item fit (Orlando & Thissen, 2000) and person fit (Meijer & Sijtsma, 2001) in the second half of the last century and has recently been adapted to the realm of DCM (Sinharay & Almond, 2007). Many of the popular indices that are available for these models are based on normalized squared residuals that follow either χ^2 or standard normal distributions within frequentist or Bayesian estimation frameworks.

Rupp and Templin (2008) and Rupp (2008b), following other research on DCM model robustness, investigated the sensitivity of item parameters and classification accuracy in the DINA model to misspecification in the Q-matrix, which was further formalized into an item-fit index by de la Torre (2007). Similarly, Liu and Douglas (2007) proposed a likelihood-ratio for the DINA model that compares the baseline model with a model that includes separate parameters that model the propensity of a respondent to respond aberrantly.

Hartz (2002), in collaboration with Louis Roussos, proposed item-fit and respondent-fit statistics for the full NC-RUM based on extensive simulation work. Put simply, the item mastery statistics compare the observed item scores for respondents that have either mastered many skills relevant for an item with those who have mastered very few; the differences are then plotted graphically. The respondent mastery statistics compare the observed total scores for groups of items for which the respondents have either mastered many skills or very few; a statistical significance test is then conducted on the differences between the groups.

The most influential state-of-the-art approach that has been developed and refined in recent years, however, is *posterior predictive model checking (*PPMC; Levy, Mislevy, & Sinharay, 2006, 2007; Sinharay, 2005; Sinharay, Johnson, & Stern, 2006). In PPMC, the posterior predictive distribution based on the data (i.e., the distribution of new data predicted from the model under a Bayesian framework) is used to simulate a large number of data sets and a test statistic of interest is computed for each data set. The observed value of the test statistic from the sample data is then compared to the empirical sampling distribution, so that critical values and credible intervals can be computed. Based on these values it is decided whether the observed value of the statistic is unlikely or not and, thus, whether there is evidence for item or respondent misfit.

The first advantage of PPMC is that the uncertainty in the parameter estimates from the model calibration is taken into account in the computation of the empirical sampling distribution through the posterior predictive distribution. The second advantage is that it is a general approach and can be applied to almost any statistical model, including DCM, such as the DINA and HO-DINA (de la Torre & Douglas, 2004), the NC-RUM (Henson, Templin, & Willse 2007), and BIN (Sinharay, 2004). The third advantage is that it forces analysts to be judicious about the selection of existing fit statistics or the definition of new fit statistics to match a fit statistic to the specific source of misfit, because research studies with PPMC have effectively demonstrated how a comprehensive assessment of fit requires a careful synthesis of information from multiple statistics.

FINAL OUTLOOK

Before closing this paper, a few words about research areas where relatively little empirical research has so far been conducted for DCM should be said. It was already addressed earlier how relatively little is known about parameter bias and precision of item parameters as well as classification accuracy for respondents under complex sampling designs, different types of missing data, and testlet effects. Specifically, little research has investigated item bias (e.g., differential item or bundle functioning) within the family of DCM, even though Xu and von Davier (2007) estimate various models for observed groupings of sex and ethnicity.

An important extension of DCM concerns the explicit modelling of longitudinal trends. As Fu and Li (2007) discuss, several latent variable models have been proposed that can be used for this purpose; not all models that are reviewed in this paper are formally suited for the modelling of longitudinal data. Related to this is the issue of cross-validation, which has received little attention by researchers so far. One notable exception is the paper by Anozie and Junker (2007), who compute the test-retest reliability of the skill classifications based on a repeated calibration of student response data with the DINA model within the context of an online tutoring system. Similarly, Kunina, Rupp, and Wilhelm (2008) have empirically compared the skill profiles from a multidimensional Rasch model with those from a discrete DCM approximation to it and cross-validated them with school grades. Given that cross-validation is so frequently addressed in multivariate statistical analysis, that the complexity of some DCM is so often high, and that many practitioners doubt their practical benefit in comparison to traditional multivariate latent variable models or multivariate clustering algorithms, it seems indispensable that more empirical research is published that focuses on the temporal stability of classification results developed from DCM.

CONCLUSION

This review paper has presented a comprehensive review of the current state-of-the-art of DCM from the perspective of defining, estimating, and applying them to data from diagnostic assessments. DCM are multidimensional confirmatory models which contain categorical latent variables that create multiple latent classes. They are developed to classify respondents based on discrete response data from diagnostic assessments. It was shown how DCM share various characteristics with latent class models, multidimensional FA models, multidimensional IRT models, and loglinear models. Despite their theoretical potential, it was noted how applications of DCM across disciplines have been sparse, which is partly a result of the stringent demands on the response theories from applied cognitive psychology and related disciplines. It is also a result of the lack of data from appropriate research designs matched to these hypotheses, as well as a lack of attractive software programs needed to calibrate such data.

The field of measurement specialists interested in DCM is thus called upon to communicate the complexity and resource demands of these models in a clear and nuanced manner to anyone interested in working with these models. It is key that practitioners develop a realistic picture of the balance between the theoretical possibilities of DCM and the practical limitations regarding their implementation, especially considering that alternative multidimensional scaling alternatives with larger research bases are available to them.

There are numerous understandable reservations that specialists have toward these models. Therefore, it is advisable to take these reservations seriously and to consider each of their aspects carefully. As a consequence, the enthusiasm that measurement experts have for DCM might be somewhat dampened by the real-life constraints of their application contexts. At the same time, it will, most likely, also generate a healthy dose of optimism that can help practitioners realize the additional benefits that a formative diagnostic assessment, which is carefully designed and subsequently calibrated with a DCM, can have for their particular assessment context. It is hoped that such a balanced perspective will lead to more insightful illustrations of the potential and realized power of DCM vis-à-vis different scaling alternatives.

REFERENCES

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory in practice*. (pp. 143–166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Almond, R. G., Yan, D., & Mislevy, R. J. (2007). *Using anchor sets to identify scale and location of multiple latent variables*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL, April.

Almond, R. G., Williamson, D. M., Mislevy, R. J., & Yan, D. (in press). *Bayes nets in educational assessment*. New York: Springer.

Anozie, N., & Junker, B. (2007). Investigating the utility of a conjunctive model in Q-matrix assessment using monthly student records in an online tutoring system. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL, April.

Böhme, K, & Rupp, A. A. (2008, March). *Diagnostic assessment of orthographic competence in a large-scale assessment*. Presented at the annual meeting of the American Educational Research Association (AERA), New York, NY, March 24–28.

Bolt, D. M. (1999). *Applications of an IRT mixture model for cognitive diagnosis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, QC, Canada, April.

Bolt, D., & Fu, J. (2004). *A polytomous extension of the fusion model and its Bayesian parameter estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA (April).

Bolt, D. M., & Lall, F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*, 395–414.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–50.

de la Torre, J. (in press). Parameter estimation of the DINA model via an EM algorithm: A didactic. *Journal of Educational and Behavioral Statistics.*

de la Torre, J. (2007). *A large-scale assessment application of an empirically-based method of Q-matrix validation*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL, April.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

de la Torre, J., & Douglas, J. A. (2005). *Modeling multiple strategies in cognitive diagnosis*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Montréal, QC, Canada (April).

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Hillsdale, NJ: Erlbaum.

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of Statistics (*Vol. 26, *Psychometrics)* (pp. 979–1027). Amsterdam: Elsevier.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer.

Embretson, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479–494.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics*. New York: Academic Press.

Embretson, S. E. (1991). Multidimensional latent trait models in measuring fundamental aspects of intelligence. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 117–132). Hillsdale, NJ: Erlbaum.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3,* 380–396.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Embretson, S. E., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23,* 13–32.

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 221–224). New York: Springer.

Fu, J., & Li, Y. (2007). *An integrated review of cognitively diagnostic psychometric models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Gierl, M. J., Cui, Y., & Hunka, S. (in press). Using connectionist models to evaluate examinees' response patterns on tests. *Journal of Modern Applied Statistical Methods*.

Gitomer, D. H., & Yamamoto, K. (1991). Performance modelling that integrates latent trait and class theory. *Journal of Educational Measurement, 28,* 173–189.

Gonzalez, E. J. (2008, June). *Current operational analyses of large scale survey data and publicly available data bases*. Presented at the annual meeting of the Psychometric Society (IMPS), June 29-July 2.

Haagenars, J., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.

Haberman, S. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–323.

Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika, 55,* 477–494.

Hartz, S. M. (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, Urbana-Champaign, IL.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262–277.

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

Henson, R. A., Templin, J., & Willse, J. (in press). Defining a family of cognitive diagnosis models, *Psychometrika*.

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

Henson, R. A., Templin, J., & Porch, F. (2004). *Description of the underlying algorithm of the improved Arpeggio*. Unpublished manuscript.

Hoskens, M., & de Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement, 32*, 364–384.

Hoskens, M., & de Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement*, *25*, 19–37.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG-TOEFL*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign.

Jannarone, R. J. (1997). Models for locally dependent responses: Conjunctive item response theory. W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 465–481). New York: Springer.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Unpublished manuscript. Accessed November 28, 2006, from http://www.stat.cmu.edu/~brian/nrc/cfa

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.

Kim, J.-S., & Bolt, M. D. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38–51.

Kunina, O., Rupp, A. A., & Wilhelm, O. (2008). Convergence of skill profiles for cognitive diagnosis models and other multidimensional scaling approaches: An empirical illustration with a diagnostic mathematics assessment. Presented at the annual meeting of the Psychometric Society (IMPS), Durham, NH, June 29–July 2.

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

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15.

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

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

Liu, Y., & Douglas, J. A. (2007). *Testing person fit in cognitive diagnosis*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL, April.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4,* 333–369.

Levy, R., Mislevy, R. J., & Sinharay, S. (2006). *Posterior predictive model checking for multidimensionality in item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA, April.

Levy, R., Mislevy, R. J., & Sinharay, S. (2007). *Posterior predictive model checking for conjunctive multidimensionality in item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL, April.

Lynch, S. (2007). Introduction to applied Bayesian statistics and estimation for social scientists. New York: Springer.

Macready, G. B, & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*, 99–120.

Maris, E. (1992). *Psychometric models for psychological processes and structure*. Unpublished doctoral dissertation, University of Leuven, Belgium.

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*, 523–547.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Maris, E., de Boeck, P., & van Mechelen, I. (1996). Probability matrix decomposition models. *Psychometrika*, *61*, 7–29.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25,* 107–135.

Meulders, M., de Boeck, P., & van Mechelen, I. (2003). A taxonomy of latent structure assumptions for probability matrix decomposition models. *Psychometrika, 68,* 61–77.

Mislevy, R. J. (2007). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th edition) (pp. 257–305). Portsmouth, NH: Greenwood Publishing Group.

Mislevy, R. J. (2008, September). *Some implications of expertise research for educational assessment*. Keynote address at the 34th International Association for Educational Assessment (IAEA) Conference, Cambridge University, Cambridge, September 8.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different respondents employ different solution strategies. *Psychometrika*, *55*, 195–212.

Mislevy, R. J., Wingersky, M. S., Irvine, S. H., & Dann, P. L. (1991). Resolving mixtures of strategies in spatial visualization tasks. *British Journal of Mathematical and Statistical Psychology*, *44*, 265–288.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–368). New York: Springer.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2005). *IEA's TIMSS 2003 international report on achievement in the mathematics domain: Findings from a developmental project*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthén, L. K., & Muthén, B. O. (2006). *Mplus Version 4.1* [Computer software]. Los Angeles, CA: Muthén & Muthén.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

OECD. (2003). *PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*. Retrieved November 28, 2006, from www.pisa.oecd.org/dataoecd/46/14/33694881.pdf

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.

Pourret, O., Naïm, P., & Marcot, B. (2008). *Bayesian networks: A practical guide to applications*. New York: Wiley.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Roussos, L., DiBello, L. V., Stout, W., Hartz, S., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnosis system. In J. P. Leighton, & Gierl, M. J. (Ed.), *Cognitively diagnostic assessment for education: Theory and practice*. (pp. 275–318). Thousand Oaks, CA: SAGE.

Rupp, A. A. (2002). Feature selection for choosing and assembling measurement models: A building-block-based organization. *International Journal of Testing*, *2*, 311–360.

Rupp, A. A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing*, *7*, 95–125.

Rupp, A. A. (2008a, April). *Psychological vs. psychometric dimensionality in diagnostic reading assessment: Challenges for creating integrated assessment narratives based on multidimensional profiles*. Presented at the ETS / IEA conference entitled “Assessing Reading in the 21st Century: Aligning and Applying Advances in the Reading and Measurement Sciences”, Philadelphia, PA, April 16–19.

Rupp, A. A. (2008b). *Why simulate? An analytical solution for predicting item parameter values under misspecifications of the Q-matrix in the DINA model*. Manuscript submitted for publication.

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and practice* (pp. 205–241). Cambridge: Cambridge University Press.

Rupp, A. A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, *68,* 78–98.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications*. New York: Guilford Press.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modelling. *Structural Equation Modeling*, *11*, 424–521.

Schrepp, M. (2005). About the connection between knowledge structures and latent class models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 93–103.

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *67*, 899–919.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68,* 413–430.

Sijtsma, K., & Verweij, A. C. (1999). Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement*, *23*, 55–68.

Sinharay, S. (2003). *Assessing convergence of the Markov Chain Monte Carlo algorithms: A review* (Research report RR-03–07). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2004). *Model diagnostics for Bayesian networks* (Research report RR-04–17). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375–394.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*, 239–257.

Sinharay, S., Haberman, S., & Puhan, G. (2007) Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice 26* (*4*), 21–28.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298–321.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. New York: Chapman & Hall / CRC.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society (Series C, Applied Statistics)*, *51*, 337–350.

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. (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. 327–360). Hillsdale, NJ: Erlbaum.

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

Templin, J., & Henson, R. A. (2005). *The random effects reparametrized unified model: A model for joint estimation of discrete skills and continuous ability*. Unpublished manuscript.

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

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. L., & Ivie, J. L. (2006). *Analysis of the Raven's Progressive Matrices (RPM) scale using skills assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA, April.

Templin, J., Roussos, L., & Stout, W. (2003). *An extension of the current fusion model to treat polytomous skills* (Technical Report Draft Copy). Urbana-Champaign: University of Illinois at Urbana-Champaign.

Templin, J., He, X., Roussos, L., & Stout, W. (2003). *The pseudo-item method: A simple technique for analysis of polytomous data with the fusion model*. Unpublished manuscript.

Templin, J., Poggio, A., Irwin, P., & Henson, R. (2007). *Latent class model-based approaches to standard setting*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL, April.

Templin, J. L., Henson, R. A., Poggio, A., Irwin, P., Poggio, J., & Yang, X. (2007). *Searching for cognitive structure in Kansas*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL, April.

Ünlü, A. (2006). Estimation of careless error and lucky guess probabilities for dichotomous test items: A psyhometric application of a biometric latent class model with random effects. *Journal of Mathematical Psychology*, *50*, 309–328.

van der Maas, H. L. J., Raijmaker, M. E. J., & Visser, I. (2005). Inferring the structure of latent class models using a genetic algorithm. *Behavior Research Methods*, *37*, 340–352.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05–16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2006). *Multidimensional latent trait modelling (MDLTM)* [Software program]. Princeton, NJ: Educational Testing Service.

von Davier, M. (2007). *Hierarchical general diagnostic models* (Research Report No. 07–19). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008; July). *Psychometric modeling of educational survey data*. Presented at the annual meeting of the Psychometric Society (IMPS), Durham, NH, June 29–July 2.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its application*. Cambridge: Cambridge University Press.

Wainer, H., Brown, L. M., Bradlow, E. T., Wang, X., Skorupski, W. P., Boulet, J., & Mislevy, R. J. (2006). An application of testlet response theory in the scoring of a complex certification exam. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 169–200). Mahwah, NJ: Erlbaum.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (Research Report No. RR-06–08). Princeton, NJ: Educational Testing Service.

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

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment* (Research Report No. RR-03–32). Princeton, NJ: Educational Testing Service.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83–105.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service.