COGNITIVE DIAGNOSTIC MODELS

Executive Report for the Maryland State Department of Education

By the MARC Team

(Bob Lissitz, Hong Jiao, Ming Li, Daniel Yangsup Lee and Yoonjeong Kang)

Introduction

The Council of Chief State School Officers (CCSSO) announced in 2006, as part of the ongoing effort to improve educational achievement across states, to establish the Formative Assessments for Students (FAST) State Collaborative on Assessment and Students Standards (SCASS), to help push forward teaching and learning through innovative uses of formative assessments. The initiative has turned to experts together with Education Testing Service to come up with practical tools for diagnosis and improvement. In addition, FAST SCASS has turned to participating states such as Maryland as a member of the Partnership of Assessment Readiness for College and Careers (PARCC) to implement useful formative assessment tools in all classrooms.

Under the Race to the Top program, Maryland schools recently adopted the common-core state standards and will begin operational PARCC test based on these new standards starting in the 2014-15 school year. A new feature being implemented to go in tandem with the new standardized tests are computer adaptive diagnostic assessments called K-1 formative tools. For more information, the reader is referred to the PARCC website which provides additional detail on the upcoming development of this system (http://www.parcconline.org/non-summative-assessments). These assessments will help teachers target students, especially those at low ability levels, to focus on areas that may be hindering them from progressing. Such assessments would also help teachers spend less time planning for interventional activities.

In 2013, PARCC, on behalf of the state of Florida, issued a procurement of services for the establishment of diagnostic and formative assessment tools, to go with the new summative assessments to give additional attention to students who are behind relative to their peers. According to the proposal for a computerized diagnostic system, the mathematics portion of the diagnostic assessment design is required to diagnose students based on the skills within the mathematics domains as described by the Common-Core State Standards, to inform teachers of students that need additional work on specific sub-domains.

In response, PARCC has set out to implement a diagnostic system that will address these requirements. Their goal is to develop easy to use and understand online diagnostic assessments that will allow for immediate diagnostic results. For more detailed information on the goals of this new innovative diagnostic assessment, one may refer to the PARCC documentation on K-1 diagnostic assessments. These formative assessments are set to be field tested in early 2015 and available for use in the 2015-2016 school year.

Behind these ambitious goals to implement computerized diagnostic assessments are the theories that would allow for proper diagnosis of students' learning using Cognitive Diagnostic Models (CDM). CDMs are the building blocks to diagnostic assessments which can be used by teachers and students by giving them fine-grained information to help teachers differentiate instructional needs, dynamically group students according to mastery of skills, assign individualized assignments and develop individualized intervention plans based on these diagnostic profiles. For example, homework assignments for a mathematics class could be individualized by assigning problems that heavily focus on their weaknesses according to the skills that students do not master based on the diagnostic information provided by the formative assessment. Teachers could also group students according to the skills they lack and have sub-

lessons for these students as a way of differentiated instruction. Essentially, CDMs output information about students' strengths and weaknesses that could be used to help teachers and students focus on the skills that they really need help on.

The purpose of this document is to provide some theoretical background on these diagnostic models by introducing some technical terminology and then providing an overview of some models that could be used in practice. The goal is to prime the way for more analyses regarding these types of diagnostic assessments, in an overall effort to provide background information about some commonly studied diagnostic models that might be useful for the stakeholders of such tests in Maryland. Ultimately, it is hoped that this review will shed light on which models will serve best for giving students and teachers the more accurate and useful information that is in line with the definition of formative assessments as proposed by the CCSSO, PARCC, and Maryland public schools.

Basic Terminology of Cognitive Diagnostic Models (CDM)

The following is an overview of the basic terminology used in cognitive diagnostic models. This report is not an all-inclusive review of all CDMs. Some more theoretically complex models will be mentioned but not elaborated upon due to their impracticalities in practice. Before discussing the models in detail, a few common characteristics about the models will be described, including, compensatory vs. non-compensatory, latent trait vs. latent class models, their handling of slipping and guessing parameters, and finally, the specification of the skills or what is otherwise known as Q-matrix specification.

A skill is referred to as a single latent attribute that a person possesses. For example, within the context of mathematics, a skill would be being able to add two digit numbers. A skill

could be more complex than simple addition. For example, a skill in mathematics could also be "solving linear equations", which is a far more complex skill and would include the skill of adding two digit numbers. This level of complexity and relationship between skills needs to be defined when specifying what is called a Q-matrix before any actual analysis can even take place.

Compensatory and non-compensatory models. CDMs can be compensatory, non-compensatory, or both. These terms refer to how skills are related to modeling the probability of a correct response. In Item Response Theory, the probability of a correct response to a certain item by a certain student is modeled by using:

1) Information about the student based on their responses to the items in the overall test (referred to a person's ability), and

2) Information about the characteristics of the item (difficulty of the item based on how others perform on the item)

The probability of a correct response to an item is given by a student's ability and difficulty of the item. This relationship is modeled by:

$$P(X = 1|\theta) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)},$$

where P is the probability of a correct response, θ_i is student i's ability and b_j is the difficulty of the item.

CDMs specify which skills are necessary to have in order to increase the probability of a correct response on the item. In unidimensional IRT, the skill might be general mathematics ability. But CDMs allow for a more specific specification of skills that might allow one to assess the skills within mathematics ability that might lead to a correct response. Some items might need just one skill within mathematics ability. Others might require several skills. For example, a

mathematics question in the context of a word problem would require the student to know how to read, synthesize information, and have some basic mathematics skills to correctly answer the item. In this light non-compensatory models assume that the student must have mastered all the skills within the item in order to get the item correct.

Non-compensatory. The example that was just described is a type of non-compensatory model because one can imagine that if any one of the skills described were missing (say the student could not read), then the student will be less likely to get the item correct. In essence, a high competency in one skill cannot "compensate" for the lack or low competency of other skills. If the student lacks a single skill needed to answer the item correctly, the student is likely to get the item incorrect. Non-compensatory models are sometimes referred to conjunctive, meaning that all skills need to be mastered for a high probability of success on an item. In the example of a mathematics test that requires high level of language proficiency, having high reading ability will not compensate for the lack of basic math skills or vice versa, having high math skills will not compensate for low reading ability if the student cannot understand the question.

Compensatory. Sometimes, certain skills compensate for the lack of others. Models that allow for this type of relationship are called compensatory models. Compensatory models imply that the student's possession of one skill can compensate for the low proficiency or the lack of the other skills that are required for answering the item correctly. In the context of education, compensatory modeling is much more studied in the summative assessment setting for subdomain score reporting.. Such compensatory modeling is referred to as disjunctive.

Latent class and latent trait model based CDMs. CDMs have been developed along two different frameworks: latent trait models and latent class models. The main difference between the two frameworks is that whether the latent attributes are measured on a continuum or as discrete categories.

Latent class models. For latent class model based CDMs, the diagnostic information is specified at either dichotomous levels of mastery or non-mastery of skills or polytomous levels such as basic, proficient, and advanced. For the latent trait model based CDMs, the diagnostic information is provided along a continuous scale. Often a cut point or multiple cut values could be used to indicate different levels of proficiency. While the former provides grouping information but no information is available about how far a student is away from a target while latter provides more fine-grained information and quantifies the relative distance between a student's current stand and how far the student is from an achievement target.

Latent-class CDMs make explicit the probability of a correct response on an item based on an student's mastered attribute profile, which are known as the latent classes. The different combinations of attributes that have been mastered differentiate different latent classes. If for example, a test contains items that measure two different attributes, then there will be 4 different possible latent class memberships: students who have mastered neither attributes, students who have mastered all attributes, and two classes representing that only one attribute has been mastered. In essence, the number of classifications that a student can fall into is finite.
The first CDMs were developed from this framework. The outcomes from these models can inform whether a student has mastered a certain skill or not. In this regard, a disadvantage of this latent class approach is that students are only classified into a certain classification profile indicating mastery or non-mastery of each specific skill. Without detailed information about how

much of the skill a student possesses and whether a student is close to the master threshold or still far away from the threshold, it is still not informative enough to provide guidance for instructions to make up the deficiency.

Latent trait models. Latent trait models allow for estimation for each skill. This allows for more specific information, like how much of the skill a student possesses. In this regard, latent trait models are more useful in providing diagnostic feedback to teachers about planning a targeted remedial instruction. For example, a teacher could devise different remedial instructions to adapt to the needs of students who might have different levels of skill proficiency as we do expect that students who are about to master the skills need different amount of remedial instructions compared with those students who are still far away from the mastery of a latent skill.

Slipping and guessing. Different CDMs treat slipping and guessing parameters distinctively. Slipping refers to an instance where a student of supposedly high ability answers an easy item incorrectly. Guessing is the opposite: a student with low ability answers a difficult item correctly. Some CDMs model these parameters at the item level, others model them at the attribute level, and yet some CDMs take a combined perspective at both item and attribute level.

Q-matrix. CDMs function through what are called Q-matrices. These matrices contain rows for each item and columns for all the possible attributes found in a test. For each cell, a 1 indicates that the item measures the skill. The cell contains a zero if the item does not measure the skill. Cells can also contain polytomous values if one wishes to specify the degree of mastery of the attribute measured by the item. As a simple example, figure 1 shows some sample items.

item 1: 4+1

item 2: 4*3-2

item 3: (4/2)-1

Figure 1. Sample items for a mathematics test assessing basic algebra arithmetic skills.

These items require the following skills: addition, subtraction, multiplication and division. A Q-matrix for this simple example would look like the following:

Table 1. Sample Q-matrix corresponding to figure 1.

	Addition	**Subtraction**	**Multiplication**	**Division**
Item 1	1	0	0	0
Item 2	0	1	1	0
Item 3	0	1	0	1

Specifying the Q-matrix is analogous to specifying the loadings in a factor analysis framework. Models use Q-matrices together with response data to produce student profiles about which skills individual students may or may not possess. Therefore, specification of the Q-matrix is one of the most important parts of the CDM application, as misspecification can lead to inaccurate, incorrect, and/or meaningless diagnosis (Rupp & Templin, 2008; DeCarlo, 2011).

Constructing a Q-matrix is as complicated as the type of items that make up the exam. For the simple example above, a simple inspection of items would be sufficient. Other methods include multiple rater inspection or iterative procedures based on item parameters (Rupp, Templin, & Henson, 2010). However, one can imagine that the Q-matrix construction can become complicated and will bring in some subjectivity when items become complex. For

example, to get a sense of how complex it may become, sometimes the attributes may be hierarchically related, meaning that some skills will be pre-requisites of others. These Q-matrices are usually constructed by content experts based on their knowledge of the skills they believe are required to answer an item correctly. Because of the subjective nature of this task, several advanced methods have been developed to properly specify Q-matrices that are complex (DiBello, Stout, and Roussos, 1995; de la Torre, 2008; deCarlo, 2011). For example, some models that capture skills not specified by the Q-matrix have been developed (MLTM-D; Embretson, & Yang, 2012).

Incompleteness of a Q-matrix refers to the assumption that there exist additional strategies not represented by the Q-matrix, which allow for a correct response on an item. In some instances, a Q-matrix does not capture all the possible strategies that a student might take to correctly answer an item. Some models will assume that all the possible strategies are accounted for, while others allow for this assumption to be relaxed. Such models allow for more realistic situations, especially when items are complex and can be solved in many different ways.

Cognitive Diagnostic Models (CDM)

The models discussed in this section will be explained with minimal technical details. For mathematical formulas of these models, the reader is directed to the appendix B at the end of the report (Rule Space Model and Restricted Latent Class model not provided). All the models discussed share the common goal to diagnose students, taking on different approaches according to the type of information desired. Models developed earlier are specific to certain contexts while more recent models are generalized, flexible versions of these earlier models, which can be specified to become these earlier models. The advantage of these flexible models is that they fit

the realities of most testing situations that many of the simpler models cannot address. However, these models are oftentimes more difficult to estimate, they require very large sample sizes, and may sometimes be difficult to interpret. In reality, simpler models may suffice because they provide reasonably simpler output that allows one to draw meaningful interpretations of complex data structures if certain assumptions met. Many operational CDMs today still use simpler models despite the recent development of more sophisticated models.

Latent class models. The following models fall under the category of latent class models. For a more technical and comprehensive list of these models, reader can refer to a thorough review of these latent class models by Roussos, Templin and Henson (2010) and work by Rupp, Henson, and Templin (2010).

Rule space. One of the earliest classification models of non-compensatory cognitive diagnostic methods was introduced by Tatsuoka in 1984 called the Rule Space model (RSM; Tatsuoka, 1985, K. Tatsuoka & M. Tatsuoka, 1989). The model can be used with both dichotomous and polytomously scored items, but is limited to dichotomous classifications, making it a type of latent class CDM model. RSM first uses traditional IRT to estimate ability and item parameters. It then uses this information, along with a reduced Q-matrix (reduced in the sense that certain dependencies among skills allow for a smaller number of permissible attribute profiles) to compute an expected latent variable corresponding to the expected score pattern from the reduced Q-matrix.

Slipping and guessing issues are considered as aberrant response patterns that are addressed by an atypical/caution index for each person. This index is essentially a residual value that shows how far away from the expected set of responses a student response falls. Large

negative values of this index mean more incorrect responses than expected (slipping), whereas large positive values mean more correct responses than expected (guessing).

RSA can handle around 20 attributes and require sample sizes of approximately 1000 to 2000 student. A crucial disadvantage of the RSA is that because probabilities are only applied in separate steps, fits statistics that use common probability distributions to evaluate the entirety of the model do not exist. Also, use of the model is stymied due to the lack of easily accessible software. Many of the latent class models for CDM purposes will have similar properties to this model with some slight modifications. Software currently available for estimation using the rule space model can be done by BUGLIB.

RLCM. Latent class CDMs are considered restricted versions of the traditional latent class model (RLCM; Haertel, 1984, 1990). These models are restricted in the sense that the Q-matrix is limited to a certain number of response patterns. LCMs model the probability of a correct response on an item based on the response to the item, the student's response data across items, and the probability of a student being in a certain latent class. Levels of mastery are not continuous, but are instead modeled by vectors of 1's and 0's representing mastery and non-mastery of skills. The model says that if a student has mastered all the skills required by an item, then he/she will have a high probability of answering the item correctly, where as if the student has not mastered all the skills, this probability will be small. This probability stays the say no matter how many skills or which particular skills have not been mastered. This model has more recently referred to as the Deterministic input, noisy "and" gate mode (DINA; Junker and Sitjma, 2001), which will be subsequently discussed. Other latent class models have been derived from this framework.

DINA and DINO. The Deterministic Input; Noisy "And" Gate model (DINA) (Hartel, 1990; Junker & Sitjsma 2001) is a non-compensatory, conjunctive CDM. It allows for dichotomous and polytomous response variables but only dichotomous attribute variables. The model is used to deterministically determine the items a student will get correct and incorrect, giving a high probability of a correct response only if all the required skills have been mastered. In other words, the probability of success in an item depends on having all the required skills within the item. Those who lack at least one skill within the item are classified the same regardless of which combinations of skills they have and lack. The DINA model allows for slipping and guessing, and the model allows for only one possible strategy to obtain a correct response. The probability of a slip is assumed to be zero. The Deterministic Input; Noisy "Or" Gate Model (DINO) (Templin & Henson, 2006, Templin 2006) is the compensatory analog of the DINA model. Extensions of the DINA/DINO model include the higher-order DINA (HO-DINA; de la Torre & Douglas, 2004), which allows for a hierarchical structure for the items, and the multi-strategy DINA (MS-DINA; de la Torre & Douglas, 2005), which allow for multiple strategies to solve the items. Software currently available for estimation of the DINA and DINO model include: DCM and LCDM in Mplus, DCM in R, DINA in Ox, MDLTM, BUGLIB, and AHM.

***NIDA and NIDO*.** The Noisy Inputs, Deterministic "And" gate model (Maris, 1999; Junker & Sijtsma, 2001) is an extension of the DINA model. Like the DINA model, it is a non-compensatory conjunctive model, which allows for both polytomous and dichotomous responses, but only produces dichotomous attribute levels. Unlike the DINA model, the NIDA model differentiates between students who lack different combinations of skills required to answer an item correctly. For instance, one would assume that a student who possesses more skills within

the item would have a higher probability of answering the item than an student who possesses fewer skills. Also, NIDA models the slipping and guessing parameters at the attribute level, relaxing difficulty levels across attributes, allowing for more fine grained profiles among those who may not have all skills mastered but different combinations of them. The Noisy Input, deterministic-or-gate model (NIDO) (Templin, 2006) is the compensatory analog to the NIDA model. Estimation using NIDA/NIDO model can be done using the Mplus extension called DCM.

RUM. The Reperameterized Unified Model (RUM; Hartz, 2002), also known as the *fusion model,* further relaxes the restrictions made by the NIDA model. Specifically, it relaxes the constraint at the item level, allowing both different slipping and guessing parameters at the item and skill level. Essentially, the model captures the idea that the same skills within items can vary in difficulty. The RUM model is also explicitly acknowledges that the Q-matrix may not be complete, allowing for different strategies to solve the item correctly. Though these differences could certainly be specified in DINA/NIDA models, the RUM allows for a more parsimonious handling of such situations. RUM can be non-compensatory (Dibello et al., 1995; Hartz, 2002) as well as compensatory (Hartz, 2002; Templin, 2006).

Two types of RUMs exist. The full RUM and reduced RUM. Though both similar in regards to capturing differences at the item and skill level, the full RUM also addresses any model misfits that arise from an incomplete Q-matrix. This improves the chances of the data fitting a model, even with an incomplete matrix. This model, however, has been proven to be difficult to estimate with small sample sizes and number of items due to the extra parameters in the model.

General Latent Class DCMs. In addition to the models just described exist models that are more flexible in their parameterization, essentially allowing for the unification of the models that were just described. These models include the multiple classification latent class model (MCLCM**;** Maris, 1999), loglinear cognitice diagnostic model (LCDM; Henson, Templin, Willse, 2009), and general diagnostic model (GDM; von Davier, 2005; Xu & von Davier, 2006), which will be briefly discussed. These complex models are oftentimes difficult to estimate due to their complexity. They require large sample sizes and many items as the number of skills and complexities within the items increase. As a result, some models, like the MCLCM, have been rarely used in operational settings and sometimes even programs to handle such models have not been developed with success.

LCDM. The log-linear cognitive diagnostic model (Henson, Templin, Willse, 2009) is a flexible model that allows relationships between categorical variables to be modeled using a latent class model. Using binary item response data, it uses the log-odds of a correct response, although it is denoted as a log-linear function. The LCDM is a generalization of the RUM, reduced RUM, DINA and DINO and can take on compensatory or non-compensatory effects. The model also provides insight to a more appropriate model for some items. The LCDM provides a model that can describe the functionality between attribute mastery and the item response probability without having to restrict the model as being conjunctive or disjunctive. In addition, it has the benefit of estimating attribute mastery while providing empirical information regarding a more reasonable model. An extension called LCDM on Mplus can be used to run estimations using the LCMD model.

GDM. Another more generalizable model that can be modified into the simpler models is the General Diagnostic Model (von Davier & Yamamoto, 2007; von Davier, 2005, 2007). The

model allows for both compensatory and non-compensatory CDMs. GDM can be specified to include a variety of latent class models as well as models that include latent dimensions. It includes higher-level interactions, allowing for higher-order latent ability models. It can also be formulated to permit binary as well as ordered polytomous item responses. In this regard, the LCDM is a specific version of the GDM. Data from large-scale tests such as the NAEP and TOEFL have been analyzed under GDM, but have not succeeded within the hierarchical framework from which the data were derived (von Davier, 2005; Xu & von Davier, 2008). Little information about the bias and precision of item or skill parameter estimates and classification accuracy for respondents has been made publicly available.

Latent-trait models. Another alternative to CDMs are latent-trait models like multidimensional item response models (MIRT) that capture information about the multidimensional nature within items to make appropriate diagnostics. MIRT models can provide more information about the different skills by representing skills in a continuum rather than discrete levels of competence. Like latent class approaches, latent-trait models can also be specified to assume compensatory or non-compensatory relationships among skills within items, although most studies have focused on the compensatory models because they are easier to estimate (Stout, 2007). For a more comprehensive review of latent-trait models, the reader is referred to the article by Stout (2007).

LLTM. The linear logistic test model (LLTM) (Fischer, 1973) is a unidimensional IRT model where skills are mapped onto the predicted difficulty levels of items for diagnosis. Student latent traits can be used to obtain the probability of solving items that contain certain skills when organized on a single dimension. Item difficulty is modeled by the skill structure within the item as well as the relative difficulties of the skills by means of weights. The model does not provide

explicit diagnostic information, but rather, uses a Q-matrix to identify the skills that affect the difficulty of the item.

MIRT-C. The compensatory multidimensional item response theory model first introduced by Reckase and McKinley (1991), and then popularized by Adams, Wilson, and Wang (1996, 1997) using the Rasch version is a type of compensatory latent trait model. Low ability level on one trait can be compensated by high values in other traits. The MIRT-C assumes a complete Q-matrix, which allows for less parameters to be estimated at the expense of a less than realistic situation in which alternative problem solving strategies are not considered.

MIRT-NC. In some instances, skills required for a correct response to an item are undeniably non-compensatory. For this reason, the non-compensatory multidimensional item response theory model (MIRT-NC; Sympson, 1978) was introduced. The model shows a multiplicative relationship that accounts for an overall low probability of a correct response to an item when at least one skill is low. Like the compensatory model, the model assumes a complete Q-matrix. In certain situations, this assumption may not be desired, such as when there exist multiple alternative strategies to obtain a correct response on an item. Later models (MLTM; Embretson, 1985, 1997) allow for a relaxation of this assumption of a complete Q-matrix.

MLTM. The multicomponent latent trait model (MLTM; Embretson, 1985, 1997; Whitely, 1980) is an extension of the MIRT-NC. Unlike the MIRT-NC model, MLTM allows for alternative strategies that are part of the individual skill level, meaning that the Q-matrix is allowed to be incomplete. For example, the model might be specified to allow an examinee to employ one of two strategies to correctly answer an item, one where all the skills required for the item are successfully executed, and the other where at least one skill is not successfully executed and the examinee guesses the answer correctly. In general, MLTM allows for multiple cognitive

strategies that can be represented within the model. Examinees are assumed to go through the same order of strategy application. The order of each strategy application and the skills for each strategy are crucial to the model (DiBello, Roussos, & Stout 2007).

GLTM. The general component latent trait model (GLTM; Embretson, 1985, 1997) is a combination of the MLTM for a single strategy and the LLTM. The model estimates item and person parameters in the same was as MLTM, but adopts the complex attribute structure that influence the item difficulty of an item by defining the difficulty parameter similarly to the LLTM. The GLTM can be specified to model the LLTM and MLTM.

MLTM-D. The MLTM-D (Embretson and Yang, 2012) is a non-compensatory IRT model. It allows for a hierarchical relationship between components at one level and attributes nested within components at the second levels. It utilizes two matrices, a Q-matrix, used to represent components in one level, and a C-matrix, used to define the nested attributes within components. The model is a generalization of the MLTM and the GLTM model and can be applicable to measures of broad traits, such as achievement tests, in which the component structure varies between items. The number of parameters depends on how the exam structure is defined. The MLTM-D allows for a relatively small number of item parameters when attribute structures are defined within components because difficulty parameters can be represented by linear combinations of attributes and control variables that can be estimated using weights that estimate the impact of these attributes (Embretson & Yang, 2012). The MLTM-D is applicable for large-scale assessments with items that vary in the types of cognitive operations and skills needed to obtaining solutions, such as in end-of-year high stakes mathematics assessments that contain a broad range of concepts.

Applications of Diagnostic Models

MLTM-D with High-Stakes Mathematics Data (Embretson & Yang, 2012). The MLTM-D has been applied to a large-scale high-stakes test of 8th grade mathematics achievement. The test consists of 86 multiple choice items scaled with a 3-PL logistic IRT model. Competency levels were evaluated with respect to proficiency categories set by experts based on analysis of items. The test contains four standard areas: (1) Number/computation, (2) Algebra, (3) Geometry, and (4) data analysis. Each standard contains benchmarks, and benchmarks contain indicators (25 total indicators) that define the skills needed to master the benchmark. Fitting the MLTM-D resulted in better fit than the DINA and LCDM models. In addition, students with similar overall scores near the mathematics competency borderline were diagnosed to provide diagnosis at the standard level, showing which standards students need remedial instruction on. In addition, the analysis provided the level of mastery within each standard.

GDM in Language Testing Data (von Davier, 2005). The GDM was applied to an Internet-based TOEFL test containing partial credit data for two forms, A and B, and with two sections, reading and listening. Four skills were identified for reading and four for listening, which were used to construct four distinct Q-matrices (Form A and Form B reading and Form A and Form B writing). Another Q-matrix was constructed for a joint analysis of the reading the listening to see if the two sections, reading and listening, needed to be analyzed separately. This matrix contained eight skills. All Q-matrices were retrofitted to existing tests. The GDM model was compared to a 2-PL IRT and two-dimensional 2-PL IRT model (for the aggregate data) for fit. A 2-PL ability parameter calibration was used as a benchmark for comparison of classifications

from the diagnostic model. The analyses showed similarities between the skills across test forms. It also showed the need to clearly separate skills when making items used specifically for diagnostic test developments. Most specific information, such as examples of how students were classified was not provided.

Some Concluding Remarks

This report is by no means exhaustive of all the CDM literature that is currently available. The models that were presented here were given as examples as a basis to build on the more applicable and useful models that could be used for further research on diagnostic assessments with the new upcoming Maryland State Assessments under PARCC. We hope to build further evidence for using these models and find grounds for introducing them to educators.

Only a few of the many CDMs developed were mentioned in this report. For a more complete and comprehensive list of the models, the reader is directed to Roussos, Templin, and Henson (2007), Stout (2007), DiBello, Roussos, and Stout (2007), and Rupp, Henson, and Templin (2010).

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.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles Policy and Practice*, *5*, 7-73.

DeCarlo, L. T. (2011). The analysis of fraction subtraction data: the DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8–26.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343-362.

de la Torre, J., & 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. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood based classification techniques. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Erlbaum: Hillsdale.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. InC. R. Rao & S. Sinharay (Eds.),

Handbook of Statistics, Volume 26, Psychometrics (pp. 979–1030). Amsterdam, The Netherlands: Elsevier.

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

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometric s* (pp. 195-218). New York: Academic Press.

Embretson, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Erlbaum.

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

Embretson, S. E., Yang, X. (2012). A multicomponent latent trait model for diagnosis. *Psychometrika*, *78*, 14-36.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Formative Assessment For Students and Teachers (FAST) SCASS (2012). Distinguishing Formative Assessment from other Educational Assessment Labels. Published paper on CCSSO website. http://www.ccsso.org/Documents/FASTLabels.pdf

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement*, *8*, 333–346.

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

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

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

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

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.

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

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361–373.

Roussos, L., DiBello, L., Henson, R., Jang, E., & Templin, J. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S.E. Embretson (Ed.), *Measuring psychological constructs: advances in model-based approaches*. Washington: American Psychological Association.

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96.

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

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement*, *44*, 313–324.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 82 98). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55–73.

Tatsuoka, M. M., & Tatsuoka, K K (1989). Rule space. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (pp. 217- 220). New York: Wiley.

Templin, J. L. (2006). Generalized linear mixed proficiency models for cognitive diagnosis. Manuscript under review.

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

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

von Davier, M., & Yamamoto, K. (2007). Mixture distribution Rasch models and hybrid Rasch models. In M. Davier & C.H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models*. New York, NY: Springer.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, A19-A9A.

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.

Appendix I.

Table of Models Discussed in the Report.

	Compensatory	**Non-compensatory**
Latent Class models	DINA RSM NIDA LCDM GDM RUM	DINA NIDO LCDM GDM RUM RLCM
Latent Trait models	LLTM MIRT-C	MIRT-NC GLTM MLTM-D

Appendix B.

In the following formulas, the subscripts representing items are denoted by subscript *i*, persons by the subscript *j*, and attributes by the subscript *k*. In general, the following variables are defined as follows:

θ_j is the person parameter (unidimensional or multidimensional)
β_i is the difficulty of the item
η_k is the contribution of attribute *k* to the difficulty of the item
K is the number of dimensions of θ_j
q_{ij} is the score of item *i* on attribute *k* in the cognitive complexity of items

Latent class models

Deterministic input, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001)

$$P(X_{ij} = 1 \mid \xi_{ij}, s_i, g_i) = (1 - s_i)^{\xi_{ij}} g_i^{(1-\xi_{ij})}$$

where s_j is the probability slipping (possesses the skill, but gets item incorrect), g_i is the probability of guessing (does not possess the skill, but gets the item correct), and ξ_{ij} is the latent variable indicator, indicating whether examinee *j* has mastered all required attributes $(\xi_{ij} = 1)$ or has not $(\xi_{ij} = 0)$ for item *i*. Then,

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{jk}^{q_{ik}}$$

α_{jk} are latent vectors representing knowledge states. The probability of a response is only high when a student has mastered all required attributes.

Deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006)

$$P(X_{ij} = 1 \mid \omega_{ij}, s_i, g_i) = (1 - s_i)^{\omega_{ij}} g_i^{(1-\omega_{ij})}$$

Instead of using ξ_{ij} from the DINA model, the DINO model replaces this variable with ω_{ij} and is defined as follows:

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{jk})^{q_{ik}},$$

indicating whether examinee *j* has mastered at least one of the required attributes for item *i*. If examinee has mastered one or more of the attributes in item *i*, then $\omega_{ij} = 1$, otherwise $\omega_{ij} = 0$.

Noisy inputs, deterministic "and" gate model (NIDA; Maris 1999)

$$P(X_{tj}=1\,|\xi_{tj},s_t,g_t)=\prod_{k=1}^{K}\left[(1-s_k)^{\alpha_{jk}}g_k^{1-\alpha_{jk}}\right]^{q_{ik}}$$

where s_k and g_k are guessing and slipping probabilities defined as:

$s_k = P(\eta_{tjk}=0|\alpha_{jk}=1,q_{ik}=1)$,
$g_k = P((\eta_{tjk}=1|\alpha_{jk}=0,q_{ik}=1)$,
$P(\eta_{tjk}=1|\alpha_{jk}=a,q_{ik}=0)\equiv 1$, regardless of the value of a.

Noisy inputs, deterministic "or" gate model (NIDO; Templin & Henson, 2006)

$$P(X_{tj}=1|\,\alpha_{tj})=\frac{1}{[1+\exp\,(\sum_{k=1}^{K}(\tau_k+\beta_k\alpha_{tk})q_{jk})]}$$

where τ_k represents lack of mastery of attribute *k*, and β_k is mastery of attribute *k*.

Reparameterized Unified Model (RUM; DiBello et al., 1995; Hartz, 2002)

$$P(X_{tj}=1\,|\alpha_j,\theta_j)=\pi_i^*\prod_{k=1}^{K}r_{ik}^{*(1-\alpha_{jk})q_{ik}}\left(P_{\beta_i}(\theta_j)\right),$$

where $P_{\beta_i}(\theta_j)=\frac{1}{1+\exp[-1.7[\theta_j-(-\beta_i)]]}$

and $\pi_i^*=\prod_{k=1}^{K}\pi_{ik}^{q_{ik}}$ and $r_{ik}^*=\frac{r_{ik}}{\pi_{ik}}$,

with π_i^* being the probability that examinee will correctly execute all mastered required skills for item *i* and r_{ik}^* is a reduction for each non-mastered skill.

Log-linear cognitive diagnosis model (LCDM; Henson et al., 2009)
For binary item response data, the probability of a correct response is given by:

$$P(X_{tj}=1\,|\alpha_j,q_i)=\frac{1}{(1+\exp(-1(\lambda_i^T h(\alpha_j,q_i)-\pi_i)))}$$

λ_i^T is as vector of weights for item *i* and $h(\alpha_j,q_i)$ are linear combinations of attributes in the item and attribute possession, and π_i is the probability of a correct response for examinees who have not mastered any attributes. The model obtains generality from the $h(\alpha_j,q_i)$ term, which can be a single term, α_{jk}, and interactions of terms depending on q_i.

For compensatory model:

$$h(\alpha_j, q_i) = \lambda_1 \alpha_{j1} + \lambda_2 \alpha_{j2} + \cdots + \lambda_K \alpha_{jK}$$

and for the non-compensatory model:

$$h(\alpha_j, q_i) = \lambda_1 \alpha_{j1} \alpha_{j2} \alpha_{jK}$$

General Diagnostic Model (GDM; von Davier & Yamamoto, 2004, 2007; von Davier, 2005, 2008)

$$P(X_{ij} = x | \alpha_j, q_i) = \frac{1}{(1 + \exp(-1(\beta_{ix} + \lambda_{ix}^T h(\alpha_j, q_i))))}$$

π_{ix} is an intercept term, λ_{ix}^T is a vector of weights for item i where λ_{ix} specifies the impact of attribute *k* for category *x* in item *i*. Like LCDM, generality is obtained from the $h(\alpha_j, q_i)$ to be other models. LCDM can be specified using the GDM.

Latent trait models

Linear Logistic Test Model (LLTM; Fischer, 1973)

$$P(X_{ij} = 1 | \theta_j, q_i, \eta) = \frac{\exp(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k + \eta_0)}{1 + \exp(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k + \eta_0)}$$

where q_i is the score of item *i* on attribute *k*, and η_k is the weight of attribute *k* on the item difficulty.

Compensatory multidimensional item response theory (MIRT-C; Bock and Aitkin, 1981)
The logistic form of the MIRT-C model is given by:

$$P(X_{ij} = 1 | \theta_j, \lambda_i, \pi_i) = \frac{1}{(1 + \exp(-1.7(\beta_i + \sum_{m=1}^{M} \lambda_{im}\theta_{jm})))}$$

β_i is the difficulty, θ_{jm} is examinee *j*'s ability level on latent dimension *m*, and λ_{im} is the weight of dimension *m* on item *i*.

Non-compensatory multidimensional item response theory (MIRT-NC; Sympson, 1978)

$$P(X_{ij} = 1 | \theta_j, \beta_i) = \sum_{m=1}^{M} \frac{1}{(1 + \exp(-1.7(\theta_{jm} - \beta_{im})))}$$

β_{im} is the difficulty of item on component *m*, θ_{jm} is examinee j's ability level on dimension *m*.

General component latent trait model (GLTM; Embretson 1985, 1997)

Similar to the MIRT-NC, except β_{im} is replaced by:

$$\beta_{im} = \sum_{k} \eta_{mk} q_{ikm} + \eta_{m0}$$

so that,

$$P(X_{ij} = 1 | \theta_j, \beta_i) = \sum_{m=1}^{M} \frac{1}{(1+\exp(-1.7(\theta_{jm} - \sum_k \eta_{mk} q_{ikm} + \eta_{m0})))}$$

Multicomponent latent trait model for diagnosis (MLTM-D; Embretson & Yang 2012)

The probability that student *j* solves item *i* is given by:

$$P(X_{ij} = 1) = \prod_{m=1}^{M} P_{ijm}^{c_{im}}$$

and

$$P(X_{ij} = 1 | \theta_{jm}, q_{im}, \eta_m) = \frac{\exp(1.7(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}))}{1 + \exp(1.7(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}))},$$

where θ_{jm} is the ability for subject *j* on component m, q_{imk} is the score for attribute *k* in component *m* for item *i*, η_{mk} is the weight for attribute *k* on component *m*, η_{m0} is the intercept for component *m*, and c_{im} is a binary variable indicating involvement of component *m* in item *i*.