

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/332073622>

Diagnostic Classification Models: Recent Developments, Practical Issues, and Prospects

Article in *International Journal of Testing* · March 2019

DOI: 10.1080/15305058.2019.1588278

CITATIONS

38

READS

2,810

2 authors:



Hamdollah Ravand

Vali-e-Asr University Of Rafsanjan

27 PUBLICATIONS 691 CITATIONS

[SEE PROFILE](#)




Purya Baghaei

Islamic Azad University Mashhad Branch

97 PUBLICATIONS 1,582 CITATIONS

[SEE PROFILE](#)

Diagnostic Classification Models: Recent Developments, Practical Issues, and Prospects

Hamdollah Ravand 

*English Department, Faculty of Literature & Humanities, Vali-e-Asr
University of Rafsanjan, Rafsanjan, Iran*

Purya Baghaei 

*Department of English, Mashhad Branch, Islamic Azad University,
Mashhad, Iran*

More than three decades after their introduction, diagnostic classification models (DCM) do not seem to have been implemented in educational systems for the purposes they were devised. Most DCM research is either methodological for model development and refinement or retrofitting to existing nondiagnostic tests and, in the latter case, basically for model demonstration or constructs identification. DCMs have rarely been used to develop diagnostic assessment right from the start with the purpose of identifying individuals' strengths and weaknesses (referred to as *true* applications in this study). In this article, we give an introduction to DCMs and their latest developments along with guidelines on how to proceed to employ DCMs to develop a diagnostic test or retrofit to a nondiagnostic assessment. Finally, we enumerate the reasons why we believe DCMs have not become fully operational in educational systems and suggest some advice to make their advent smooth and quick.

Keywords: attribute, DCM application, diagnostic classification models, model fit, Q-matrix

Diagnostic classification models (DCMs) provide multiple discrete proficiency scores which make them apt for situations where fine-grained feedback is required. In contrast to the more traditional item response theory (IRT) models, which usually scale test takers according to a continuous unidimensional attribute,

TABLE 1
A Sample Q-Matrix

Item	Lexical Meaning	Cohesive Meaning	Paragraph Meaning	Summarizing	Inferencing
1	0	1	1	1	0
2	1	1	1	0	0
3	1	0	0	0	1
4	1	0	0	1	0
5	1	0	0	0	0

Note: Reproduced from Ravand (2019).

DCMs classify test takers according to multiple categorical attributes with mastery/nonmastery statuses. Thus, the classification objective makes DCMs more amenable to the requirements of criterion-referenced assessment. DCMs predict probability of an observable categorical response from unobservable (i.e., latent) categorical variables. These discrete latent variables have been variously termed as *skill*, *subskill*, *attribute*, *knowledge*, and *ability*. In the present article, the term “attribute” is used to refer to the discrete latent predictor variables.

DCMs have been defined by Rupp and Templin (2008) as “... probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure” (p. 226). They are probabilistic models in that each DCM expresses a given respondent’s performance level in terms of the probability of mastery of each attribute, separately, or the probability of belonging to each latent class (Lee & Sawaki, 2009). Cognitive diagnostic models are also confirmatory in nature, like confirmatory factor analysis models, in the sense that latent variables in DCMs are defined a priori through a Q-matrix. A Q-matrix (Tatsuoka, 1985) is the loading structure of a DCM. It is a hypothesis about the required skills for getting each item right. Q-matrices are formulated in tables with as many rows as there are items and as many columns as there are attributes (see Table 1). At the intersection of each item and attribute is a 1 if the item measures the attribute, otherwise a 0. The Q-matrix presented in Table 1, reproduced from Ravand (2019), shows that Item 1, for example, measures cohesive meaning, paragraph meaning, and summarizing.

DCMs are notably different from multidimensional IRT models in that the latent variables in DCMs are discrete or categorical (e.g., masters/nonmasters), whereas ability estimates (θ) in multidimensional IRT models are continuous. For the purpose of the DCMs, each item typically requires more than one attribute. This leads to a complex loading structure where each item is specified in relation to multiple attributes. This complex loading structure, in terms of multidimensional IRT, is known as *within-item multidimensionality* (e.g., Baghaei, 2012).

Although DCMs have been around for more than a decade, they have rarely been applied to provide feedback to tailor instruction to the needs of learners. The comparative dearth of DCM applications can be blamed on the following factors (Ravand & Robitzsch, 2015): (1) their lack of accessibility to a broad audience interested in their application, (2) fast growth of the models which makes it hard for practitioners to keep up with the latest developments, and (3) unresolved issues such as sample size in DCMs, which hinder their applications.

There are articles that have attempted to present a user-friendly portrait of the DCMs (e.g., Ravand, 2016, 2019; Ravand & Robitzsch, 2015). Nevertheless, to keep up with the latest developments in DCMs, interested readers must review many articles in diverse sets of journals. To the best knowledge of the authors, there have been only two review articles on the important issues related to the theory and practice of DCMs (e.g., DiBello, Roussos, & Stout, 2007; Rupp & Templin, 2008). However, since then, there have been a lot of extensions to the models that have addressed some of the concerns raised by these researchers. Further, there is no study which addresses practical concerns in DCM applications. To remedy the void in the literature, the present article aims to encapsulate the most important attempts at extending and methodologically advancing the DCMs, which readers should otherwise search through different literature to review.

Furthermore, the present article attempts to address some of the major challenges which may have discouraged applications of DCMs. To name a few, the challenges are what DCM model to use, how many items are enough, how many attributes should be extracted/included, what are the steps involved in DCM applications, etc. By reviewing the existing literature, the present article provides some practical suggestions on how DCMs can be implemented. In so doing, to meet Rupp and Templin's (2009) concern, the authors present a set of best practices and codes of conduct, which will hopefully help expedite the slow pace of DCM application to keep up with its methodological advancement. It should be noted that in the review of the methodological advances, the year 2008 was set as the point of departure, since although the history of DCMs dates to many years before 2008, Rupp and Templin's (2008) paper and also the special issue of *Language Assessment Quarterly* journal in 2009 allocated to DCM, set benchmarks in the DCM studies.

APPLICATIONS OF DCMS

DCMs have been popular mainly with psychometricians and researchers with strong statistical backgrounds. However, the proponents of DCMs have hoped that the models become available to educational practitioners so that they can be used to tailor instruction to students' needs. Ironically, they have rarely

found their ways into low-stakes situations to provide feedback to promote teaching and learning. From the “proof of the pudding is in its eating” perspective, until DCMs have not benefited their true beneficiaries (i.e., teachers and learners), hardly can it be claimed that they have delivered the good that their proponents have long raved about.

DCMs can be used to serve any or a combination of the following three main purposes: (1) to develop tests for diagnostic purposes and consequently use the tests thus developed to glean fine-grained information as to the strengths and weaknesses of the test takers. This is the prototypical aim set forward by early DCM promoters. As a result, studies carried out to meet this purpose are referred to as *true DCM studies* in this article, (2) to extract diagnostic information from the existing nondiagnostic high-stakes tests, a practice referred to as *retrofitting*. This is the-measure-of-last-resort purpose of DCMs. Here, studies of this type are referred to as *retrofitting studies* and the first two purposes are collectively referred to as *DCM applications*, (3) to build the methodological infrastructure for DCM applications. This purpose has involved attempts to address technical issues such as model fit, growth DCMs, model selection issues, etc., which have already been addressed by rival models such as continuous IRT models. Studies intended to extend the methodological foundation of DCMs are referred to as *methodological studies* here. Besides, there are two other purposes DCMs might be used for, which are corollaries to the above three purposes: (4) to demonstrate that the methodological issues, addressed through the simulation studies mentioned previously, also work with real data. These applications have been add-ons to simulation studies. We refer to these types of studies as *example-of-methodology studies* and (5) to investigate the attributes underlying educational constructs. This type of use is a by-product of the studies mentioned in Purposes 1 and 2.

Googling the two most popular labels of cognitive diagnostic models (CDMs) and DCMs (the label preferred in the current study) in early 2018 returned over 240 hits, over 95% of which were methodological, about 4% were retrofitting, and less than 1% were true DCM studies. The true and retrofitting DCM studies embody the features which DCM promoters have been bragging about. The conspicuously lopsided makeup of the studies in favor of methodological studies raises serious concerns about the original promises of the DCMs.

CATEGORIZATION OF DCMs

In DCMs, item responses are predicted by a set of discrete latent variables called attributes, subskills, or processes. DCMs make varying assumptions as to how the predictor latent attributes combine to generate a response to the item. DCMs have been categorized into conjunctive/disjunctive or

compensatory/noncompensatory dichotomies. According to the compensatory/disjunctive models, mastery of one of the required attributes can compensate for nonmastery of the other attributes. In these models, mastery of more attributes does not increase the probability of providing the item with a right answer; that is, under these models, mastery of any subset of attributes is the same as the mastery of all the required attributes. In noncompensatory DCMs, on the other hand, the presence of all the required attributes results in a high probability of a correct answer. More recently, additive DCMs have been presented as another category of DCMs. Unlike compensatory DCMs, which do not credit test takers for the number of attributes mastered, in additive DCMs presence of any one of the attributes affects the probability of a correct response independent of the presence or absence of other attributes.

Lately, a new categorization of DCMs has been proposed: specific vs. general. Specific DCMs are models in which only one type of relationship is possible within any assessment: disjunctive, conjunctive, or additive. In contrast, in general DCMs (G-DCMs) such as the generalized deterministic noisy “and” gate (G-DINA) model (de la Torre, 2011), multiple DCMs are possible within the same assessment. The G-DCMs do not assume any prespecified relationships among the attributes underlying any assessment. Rather, each item can select its own model *a posteriori*. de la Torre (2011) showed that many of the specific DCMs, regardless of whether they are conjunctive, disjunctive, or additive, can be derived from the G-DINA by introducing constraints in the parameterization of the G-DINA.

A more recent extension of the DCMs such as the hierarchical log-linear CDM (HLC-DCM; Templin & Bradshaw, 2013) has led to a new category of DCMs: hierarchical vs. nonhierarchical. In the hierarchical DCMs (HDCMs), structural relationships among the required attributes are modeled. In instructional syllabi, some teaching materials are presented prior to others. The sequential presentation of skills might be reflected in test takers’ responses to items that require those skills. HDCMs are able to capture the effect of sequential teaching of materials where learning new skills builds upon prerequisite skills.

With the preceding discussion in mind, the categorization in Table 2 is suggested. At a global level, DCMs are divided into general and specific and, at a local level, specific DCMs are divided into disjunctive, conjunctive, and additive. Furthermore, HDCMs form a new category in both the general and specific DCMs. A further point that needs to be made before wrapping up the discussion of how DCMs can be categorized is that many DCMs are reparameterizations of each other. Changing the *link function* in a DCM would result in the parameterization of another one. For instance, the additive cognitive diagnostic model (ACDM, de la Torre, 2011) with the *identity* link function,

TABLE 2
DCM Categorization

DCM Type		Examples	Author(s)
Specific	Disjunctive	1. Deterministic-input, noisy-or-gate model (DINO)	Templin and Henson (2006)
		2. Noisy input, deterministic-or-gate (NIDO) model	
	Conjunctive	1. Deterministic-input, noisy-and-gate model (DINA)	Junker and Sijtsma (2001)
		2. Noisy inputs, deterministic “and-gate (NIDA)	DiBello, Stout, and Roussos (1995); Hartz (2002)
	Additive	1. Additive CDM (ACDM)	de la Torre (2011)
		2. Compensatory reparameterized unified model (C-RUM)	DiBello et al. (1995); Hartz (2002)
		3. Noncompensatory reparameterized unified model (NC-RUM) ^a	Hartz (2002)
		4. Linear logistic model (LLM)	Maris (1999)
	Hierarchical	1. Hierarchical DINA (HO-DINA) model	de la Torre (2008)
	General	Disjunctive, Conjunctive, and Additive	1. General diagnostic model (GDM)
2. Log-linear CDM (LCDM)			Henson, Templin, and Willse (2009)
3. Generalized DINA (G-DINA)			de la Torre (2011)
Hierarchical		Hierarchical diagnostic classification model (HDCM)	Templin and Bradshaw (2013)

^aOriginally, the NC-RUM has been parameterized as a non-compensatory model. However, Ma, Iaconangelo, and de la Torre (2016) showed that it is an additive DCM with log link function.

turns into the linear logistic model (LLM; Maris, 1999) and into the noncompensatory reparameterized unified model (NC-RUM; Hartz, 2002) when the link function is changed into *logit* and *log*, respectively. As another example, the G-DINA turns into LCDM by the change of the identity link function to *logit*. Therefore, it seems that the traditional distinctions between the DCMs are getting blurred, however, for the ease of reference and continuity with the DCM literature, the categorization in Table 2 is suggested.

METHODOLOGICAL ADVANCES IN DCMS

Most DCM studies have tried to address the technical questions already addressed by classical test theory (CTT) or continuous IRT models (Henson, 2009). Back in 2008, Rupp and Templin made a list of to-do tasks for the researchers in the field of DCM such as the question of growth modeling

within DCM framework, the issue of parameter bias and classification accuracy when local person dependence (LPD) and local item dependence (LID) is present and how missing data and differential item functioning (DIF) would affect DCM results. In the recent years, researchers have tried to address some of these tasks and other technical issues already established in the competing psychometric models such as IRT and CTT. In what follows, some of the major methodological advances in DCMs are discussed.

Extensions to the Models

A very recent attempt in strengthening the methodological infrastructure of DCMs concerns the development of models which could handle longitudinal trends (e.g., Kaya & Leite, 2017; Li, Cohen, Bottge, & Templin, 2016; Madison & Bradshaw, 2018; and Wang, Yang, Culpepper, & Douglas, 2018). These growth DCMs have combined latent transition analysis (LTA; Collins & Wugalter, 1992) and the deterministic input noisy and gate (DINA) model (Junker & Sijtsma, 2001) into a LTA-DINA to analyze changes in the mastery status of the attributes over time. LTA can be used to identify probability that subjects belonging to a given latent class will remain in that group or move into other latent groups. LTA is suitable for studying developmental changes as stipulated in theories such as that of Piaget. The combination of the LTA with the DINA can account for transition statuses on several latent discrete variables.

Another new development is the possibility of doing differential item functioning (DIF) within the context of DCMs. DIF in the context of DCMs (DCM DIF) occurs when probabilities of correct responses to any given item are different for test takers with the same attribute profiles but from different observed groups (Hou, de la Torre, & Nandakumar, 2014). In other words in DCM DIF, the matching criterion is the attribute profile of the test takers. DCM DIF is different from traditional DIF in two ways: (1) Unlike IRT and CTT DIF where the ability estimate and observed total score are the conditioning/matching variables, respectively, in DCM DIF, the attribute profiles of the test takers are the matching/conditioning variables. (2) In IRT and CTT DIF, item difficulties and discriminations are compared for the matched groups whereas in DCM DIF guessing and slipping parameters are compared across the groups matched on the attribute profiles. Hou, de la Torre, and Nandakumar (2014) proposed a DIF detection procedure based on the DINA model in which they adapted the Wald test (Morrison, 1967) to explore both uniform and nonuniform DIF. Currently, the CDM package (Robitzsch, Kiefer, George, & Uenlue, 2017) and GDINA package (Ma & de la Torre, 2018) in R conduct DCM DIF.

The birth of the general DCMs (G-DCMs) is a highly influential methodological advancement which has led to the unification of specific DCMs within the framework of general models. With the G-DCM such as the generalized deterministic-input noisy-and-gate (G-DINA; de la Torre, 2011), log-linear CDM (LCDM; Henson, Templin, & Willse, 2009), and general diagnostic model (GDM; von Davier, 2005), the once sharp boundaries between the specific DCMs have become fuzzier. Most of the popular specific DCMs such as deterministic-input noisy-and-gate (DINA) model (Junker & Sijtsma, 2001), additive CDM (ACDM; de la Torre, 2011), noncompensatory reparameterized unified model (NC-RUM; Hartz, 2002), etc. can be derived from the G-DCMs by introducing some constraints to the G-DCMs in their saturated form. Even at times, as pointed out before, the same constraints but different link functions would lead to a different specific DCM.

Another recent development that is closely related to the advent of G-DCMs and has been touted for sparing the researchers the toil of a priori model selection is the possibility of investigating model fit at item level proposed by de la Torre and Lee (2013). According to this procedure, the LCDM or the G-DINA is first applied to the data and in a next step fit of the specific models is compared against that of the general model. If the specific model does not worsen the fit it is retained as the best model for the respective item. Through these procedures, researchers are no longer forced to impose a single model on all items of a given assessment, a practice that has been argued not to be viable. In other words, this new development makes item selection at item level possible.

Another important development is the introduction of DCMs that can take into account possible multiple strategies that test takers might employ to reach the correct answer. One of the common assumptions of the conventional DCMs is that all the test takers follow uniform strategies/attributes to solve the items on any given assessment. However, this assumption might not hold in some educational contexts where alternative plausible strategies can be used to solve any given item. Students might be taught different ways of solving the same problem or they may devise their own ways of doing the problems. Conventional DCMs are not able to distinguish these alternative strategies from lucky guessing captured by the guessing parameter in these models. de la Torre and Douglas (2008) extended the DINA model to capture use of alternative strategies to solve test items. Later, Huo and de la Torre (2014) extended and improved upon the multistrategy DINA (MS-DINA). They suggested the use of MS-DINA on utility considerations because it can better capture strategies the students (especially the advanced ones) use to solve complex problem.

All the extensions discussed so far assume there is no hierarchical structure among the attributes, an assumption that might not be plausible in educational arenas. Instructional syllabi design teaching materials sequentially where some

prerequisite skills are presented first and then new skills build upon the previous ones. This hierarchical presentation of teaching materials may be reflected in test takers' responses to those items measuring the skills taught. If so, appropriate DCMs are required to capture the underlying structure of students' knowledge reflected in the item responses. Templin and Bradshaw (2013) presented an adaptation of the LCDM (called hierarchical diagnostic classification model [HDCM]) which bridges DCMs with the Attribute Hierarchy Method [AHM; Gierl, Leighton, & Hunka, 2007] and the Rule Space Method [RSM; Tatsuoka, 1983]). Unlike AHM and RSM, which do not have a statistical test to check the presence of attribute hierarchies, HDCM provides statistical tests to explore such hierarchies and does so within the flexible framework of the LCDM. For an application of HDCM to language assessment data see Ravand (2019).

Q-Matrix Validation Extensions

A pivotal element of any DCM is the Q-matrix (Tatsuoka, 1983). In most DCM studies (e.g., Jang, 2009; Lee & Sawaki, 2009; Li, 2011; Ravand, 2016) Q-matrices have been specified through qualitative analysis using expert judgement. Gorin (2009) called into question the wholly subjective process of Q-matrix development in DCMs. More recently, some empirical methods of Q-matrix validation have been proposed (e.g., Barnes, 2010; Chen, Liu, Xu, & Ying, 2015; Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016; De Carlo, 2012; Desmarais & Naceur, 2013; Liu, Xu, & Ying, 2012; Templin & Henson, 2006). Some of these methods are completely data-driven (e.g., Barnes, 2010; Chen, Li, Liu, & Ying, 2017; Chen, Liu, Xu, & Ying, 2015; Liu et al., 2012). In these methods the underlying attributes are derived from test takers' responses. On the other hand, some others have been developed to detect misspecifications in the expert-defined provisional Q-matrices.

The methods proposed by De Carlo (2012) and Templin and Henson (2006) are suitable for situations where the potential misspecifications in the Q-matrix can be identified. Desmarais and Naceur (2013) proposed a factorization method which could assist to derive a Q-matrix from test takers' responses using an expert-defined Q-matrix as the initial Q-matrix. de la Torre (2008) proposed a validation method for identifying misspecifications in Q-matrices. de la Torre's method is compatible with data that conform with the DINA. de la Torre and Chiu (2016) developed a generalization of the discrimination index in de la Torre (2008), which is compatible with the G-DINA and all the specific DCMs subsumed under it. Chen et al. (2017) proposed the regularized latent class analysis (RLCA), a method which does not need a provisional Q-matrix and assumes that both the true model and Q-matrix are unknown; the only thing that should be known is the number of the

TABLE 3
DCM Software Programs/Codes

Software/Code	Models Estimated	Access and Cost
Arpeggio (Bolt et al., 2008)	RUM and RRUM	A former commercial software but is free of charge now, available by contacting Lou DiBello at ldibello@uic.edu
Mplus code (Muthen & Muthen, 2013)	LCDM and constrained model	Commercial software, code downloadable from http://jonathantemplin.com
MDLTM (von Davier, 2006)	GDM and constrained models, latent class models, IRT models	Free research license, available by contacting Matthias von Davier at mvondavier@yahoo.com
R-package CDM (Robitzsch, et al., 2016)	G-DINA, LCDM, GDM, and constrained models, GDM, latent class models, IRT models	Freely downloadable from the R website
R-package GDINA (Ma & de la Torre, 2016)	G-DINA and constrained Models	Freely downloadable from the R website

Note. Adapted from Li, Hunter, and Lei (2015).

latent classes, which in turn is dependent on the number of attributes underlying performance on any given test.

Software Developments

A major breakthrough in DCM applications is the development of software programs that can estimate a wide variety of DCMs. One of the proposed reasons for under-application of the DCMs to real problems of education has been the issue of inaccessibility of DCM software (Ravand & Robitzsch, 2015). Formerly, DCM software programs have been mostly proprietary and tied to just one DCM (e.g., Arpeggio). As Table 3 shows, R packages are both free and can handle multiple DCMs.

In their review of the GDINA and CDM packages in R (R Core Team, 2018), Rupp and van Rijn (2018), concluded that despite the striking similarities between the routines in both packages, the CDM package is preferable in terms of user-friendliness, the coverage of DCM extensions, and model fit indices, etc. which have been implemented into the package. They also found that the CDM package is more time-efficient.

The CDM package is the most comprehensive software program (Ravand & Robitzsch, 2015), which is capable of estimating all the general models such as the G-DINA, LCDM, and GDM, and many of the specific models including but not limited to DINA, deterministic-input, noisy-or-gate model (DINO;

Templin & Henson, 2006), ACDM, reparameterized unified model (C-RUM; Hartz, 2002), NC-RUM (Hartz, 2002), LLM, higher-order LCDM (HDCM; Templin & Bradshaw, 2013), and higher-order DINA (de la Torre, 2004). The package also generates a lot of fit indices both relative and absolute, which can be used to evaluate model-data fit or compare models with each other. The possibility to estimate a wide array of DCMs by a single software program facilitates comparison of different DCMs within the same framework using the same estimation methods.

Despite all the methodological advances made, there are few studies on issues such as the effect of sample size (e.g., Kunina-Habenicht, Rupp, & Wilhelm, 2012), grain size (e.g., Skaggs, Hein, & Wilkins, 2016), and performance of fit indices (e.g., Chen, de la Torre, & Zhang, 2013; Hu, Miller, Huggins-Manley, & Chen, 2016), among others. Also, to keep up with the competing alternative models such as continuous IRT models, DCMs still need to address some more methodological concerns such as the question of parameter bias and classification accuracy when local person dependence and item dependencies are present and how missing data would affect DCM results. Another equally important concern that needs to be addressed is the issue of linking in DCMs. Methodological infrastructure should be devised through which one can make sure that attribute mastery takes the same level of knowledge and expertise across different administrations of a test.

SUGGESTED STEPS FOR THE APPLICATION OF DCMs

Here, steps for applying DCMs to develop diagnostic tests and to retrofit existing nondiagnostic tests for diagnostic purposes are suggested, separately. It should be noted that the procedures can be transferred across tests of different constructs including, but not limited to, math and language skills. However, the procedures of Q-matrix development for constructs such as writing, as practiced so far (e.g., Kim, 2011) are somehow different from the common drill for constructs such as math and reading comprehension.

Common to applications of DCMs, be they *true DCM studies* or *retrofitting studies*, is a cognitive theory underlying test performance. Because such cognitive theories in educational assessments are few and far between, researchers should make do with an implicit theory of test performance.

To develop a diagnostic test from the beginning using DCMs, the following steps are suggested:

1. Consult the relevant theories, expert judgment, and the previous DCM studies on the construct under study to draw a list of attributes (i.e., to construct the implicit theory).

2. Because in educational assessment there is no consensus as to exactly what attributes underlie the constructs, ask at least five expert judges to decide on the importance of the attributes according to your definition of the construct. For example, you could ask them to rate the importance of the attributes on a scale of 1 to 5 and include those which were rated at least 4 by at least two thirds of the judges. See below for considerations concerning the number of attributes in the test and the number of attributes per item.
3. Construct the Q-matrix (decide on the number of items, attributes, how many attributes each item should measure, how many times an attribute should be measured).
4. Construct the test.
5. At this stage, to make sure whether the items measure the subskills you have intended, you can use expert judges and think-aloud technique.
6. Revise items according to the results obtained.
7. Administer the test.
8. So far you have only used qualitative analysis in item development. After the test data have been collected, you can use the empirical procedures suggested by de la Torre and Chiu (2016) or Chen et al. (2017) to ensure that the item-by-attribute relationships as specified in the Q-matrix have been reflected in the items of the test.

As to model selection, there are two possible lines of actions: In the first line of action, which is the popular one, the researcher applies a single specific model to all items of any given test whereas in the second line of action, he applies different specific models to different items of any given test. It should be noted that the second line of action entails applying a general DCM to the whole test first and then applying different specific DCMs to different items of the test. Model selection can proceed in two ways: in a confirmatory manner (when there is good theoretical evidence) and in an exploratory manner (when the researcher does not have any sound reason why a given DCM should fit). Regardless of what line of action is taken, model application can be either confirmatory or exploratory. If the first line of action (i.e., application of a single model to all items) is taken and the researcher is dealing with math items, for example, he might have substantive reasons for a wholesale application of a single noncompensatory model.

However, if he is dealing with reading comprehension items, based on the available substantive evidence in the literature, he might decide to apply a single compensatory DCM to all the items. Otherwise, if there is no substantive reason for wholesale application of either a compensatory or noncompensatory model, the researcher may decide to rely on statistics and apply both compensatory and noncompensatory models and let the fit indices decide on the model to be selected. In the same vein, if the researcher is interested in model

selection at item level rather than test level (for the discussion of model selection at item level see Ravand, 2016 and Ravand & Robitzsch, 2018), depending on whether there are substantive reasons for how the attributes measured by each single item interact (i.e., compensatory, noncompensatory, or additive), he might intend to apply different DCMs to different items of the test in a confirmatory manner or let the DCM software program select the best-fitting model for each items in an exploratory manner.

In a nutshell, if the first line of action is taken and the approach is confirmatory:

9. Choose the specific DCM based on the extant theoretical evidence.
10. Check model fit.

If the first line of action is taken and the approach is exploratory:

9. Apply DCMs of different types, i.e., general, compensatory, noncompensatory, and additive.
10. Compare fit of the models.

A point worth mentioning is that because there are no cutoffs or significance tests for most of the fit indices in DCMs, many DCM studies (e.g., Lei & Li, 2016; Ravand, 2016; Ravand & Robitzsch, 2018; Yi, 2012) have checked fit of the specific models against that of the G-DINA model. As noted by Chen, de la Torre, and Zhang (2013), “any saturated DCM will always fit the data better than any reduced DCM because of their more complex parameterization.” This practice might be justified on the grounds that when working with real data, the true model is ordinarily unknown.

However, if the second line of action is taken and the approach is confirmatory:

9. Select different DCMs for different items.
10. Run the multi-DCM model and check model fit.

If the second line of action is taken and the approach is exploratory:

9. Run the G-DINA and let each single item select its own model a posteriori.
10. Check the fit of the model.

Both lines of action are possible with the CDM package (Robitzsch, Kiefer, George, & Uenlue, 2017). For an example of how the two lines of action can be applied see Ravand and Robitzsch (2018). Because currently explicit

cognitive theories that underlie design and development of educational assessments are few and far between, it may be a long time before truly diagnostic assessments can be developed (Liu, Huggins-Manley, & Bulut, 2018). Therefore, in coming years, most of the applications of the DCMs might involve retrofitting.

For retrofitting practices the following steps are suggested:

1. Identify the attributes underlying the test. The process of developing a DCM starts with a cognitive processing model. However, in retrofitting DCMs “some type of implicit substantive model is generated *post hoc* by reviewing the existing item” (Gierl & Cui, 2008). To come up with the “implicit substantive model” in retrofitting contexts, first, a list of attributes is drawn from any or the combination of the following sources (1) the existing literature, the theories of the construct under study, or construct models, (2) verbal reports or protocol studies, (3) eye tracking research, (4) expert panels, and (5) test specifications. Expert judgment is the most frequently employed source of attribute identification in the literature (e.g., Kim, 2015; Lee & Sawaki, 2009; Ravand, 2016).
2. Specify attribute-by-item relationships in a Q-matrix. After the list of attributes being measured by the assessment under study has been drawn, a group of experts or alternatively students similar in characteristics to whom the test was intended for (or a combination of both), may be asked to read the items on the test carefully and identify the attributes they use to answer items of the test. DCM studies have mostly relied on expert judgment to identify item-by-attribute relationship in a Q-matrix (Tatsuoka, 1985). There is one caveat about using expert judgment in identifying attributes measured by each item. Expert judges’ abilities are usually well above those of the students, and the students do not necessarily follow the same processes as specified by expert judges.
3. Empirically validate the Q-matrix (optional).

In the vein of true DCM studies, Steps 9–10 can be followed to apply DCMs in a retrofitting mode.

KEY ISSUES IN Q-MATRIX DEVELOPMENT

The importance of the Q-matrix in DCMs cannot be overstated since the validity of the inferences made about test takers’ performance hinges upon the accuracy of the Q-matrix. The most critical and challenging step in DCMs is Q-matrix development (Gorin, 2009). A Q-matrix contains information as to what attributes are measured by each item. In designing a Q-matrix the following considerations

need to be taken into account. (1) Correct specification of the Q-matrix: What attributes each item measures should be accurately specified, (2) design of the Q-matrix: what is the configuration of the attributes in the Q-matrix, and (3) the grain size of the attributes: how finely the attributes should be specified.

Misspecifications of the Q-matrix decrease classification accuracy (Kunina-Habenicht et al., 2012; Rupp & Templin, 2008). Rupp and Templin (2008) found that incorrect deletion of attributes from the Q-matrix resulted in high slipping¹ parameters and addition of attributes to the Q-matrix led to underestimation of guessing² parameter. They also found that deletion of certain combinations of attributes resulted in misclassification. Since the main objective of all the DCMs is classification of test takers, all the variables that might impact accuracy of classifications should be carefully taken into account.

Besides the accuracy of the Q-matrix, the design of the Q-matrix would also affect classification accuracy. De Carlo (2012) and Chiu, Douglas, and Li (2009) showed that the DINA model requires a Q-matrix where each attribute is measured in isolation at least once. Attributes that are not measured in isolation and those which are always measured in conjunction with other attributes can be causes for concern. Madison and Bradshaw (2015) found that keeping the number of items constant, the more an attribute is measured in isolation, the higher the accuracy of classifications in the LCDM will be. On the contrary, they found classification accuracy degenerated when two attributes were always measured together. Consequently, when two attributes are always measured together, they recommended combining them into a composite attribute.

Another key consideration in Q-matrix development is the level of specificity of the attributes or their grain size. The more specifically the attributes are defined, the better they can inform instruction, the more computationally intensive they get, and the more difficult to interpret they will be (Embretson & Yang, 2013; Xu & Zhang, 2016). From a diagnostic perspective, the attributes should be as specific as possible. However, as the number of attributes increases there will be an exponential increase in the number of latent classes (for k attributes there are 2^k latent classes) and in turn larger number of items and large sample sizes are required. As a rule of thumb, de la Torre and Minchen (2014) recommended 10 attributes at the most. However, with 10 attributes there will be $2^{10} = 1024$ latent classes. If we have a sample as large

¹Slips are aberrant responses. Simply put, slips are careless errors. Slipping occurs if a respondent who has mastered all the attributes required by a given item, slips and answers the item incorrectly.

²Guesses are another type of aberrant responses. They are lucky guesses. Guessing occurs if a person provides a correct answer to the item although he has not acquired all the attributes measured by the item.

as 1000, on average, there will be $1000/1024 = 0.98$ test taker in each class. As previous research has shown (e.g., Lee & Sawaki, 2009; Li, 2011; Ravand, 2016) the majority of the test takers are usually assigned to one of the two flat profiles, i.e., latent classes where either all the attributes or none of them have been mastered. Thus with 1024, most of the latent classes will be either empty or assigned to very few test takers. In a scenario such as this, classification consistency and accuracy might also be compromised. Most DCM studies have specified up to five attributes (e.g., Lee & Sawaki, 2009; Li, Hunter, & Lei, 2015; Ravand, 2016; Ravand, 2019; von Davier, 2005).

As to the specificity of the attributes, it should be born in mind that when coding items for attributes, one needs to take the level of the ability of the test takers into account. It may be argued that lower level attributes of vocabulary and syntax are required for every item of reading comprehension, for example. However, if language proficiency of the test takers is high and basic English grammar and vocabulary knowledge are required to deal with the items of the test, coding the items for vocabulary and syntax might lead to low discriminations for the respective attributes, hence low-quality items. In other words, the items might not be able to discriminate between masters and nonmasters of the respective skills. High levels of item discrimination, which is the index of item quality, have been shown to play an important role in diagnostic assessment. Madison and Bradshaw (2015) found that high item discrimination can mitigate potential problems that might arise out of Q-matrix design.

One final note as to model and Q-matrix selection should be made. In model or Q-matrix comparison situations, fit indices help select the model or the Q-matrix, which is the most appropriate among the competing ones, even though the true Q-matrix and model may not be among models and matrices studied (Lei & Li, 2016). Therefore, Lei and Li (2016) suggested that model and Q-matrix selection be informed by interpretability of the attributes as well as the fit indices. To ensure that the selected model and Q-matrix are as close to the true ones as possible, one should replicate the selected model and Q-matrix with other samples.

MODEL SELECTION

With the wide array of the available DCMs, selecting the most appropriate model for any given assessment situation has become a challenge (Ravand, 2019). Choice of the most suitable model has been taken for granted in most applications of DCMs. Model selection has mostly been driven by software availability and familiarity rather than the degree of match between the assumptions of the models and how the attributes underlying the test are assumed to interact (Ravand & Robitzsch, 2018). Relationships among the

attributes required by any given item can be either compensatory, or noncompensatory. Choice of the right model is of critical importance because model selection affects classification of test takers (Lee & Sawaki, 2009), which is the primary purpose of the DCMs.

G-DCMs, resting at the pinnacle of the evolutionary lineage of DCMs (Templin, 2009), may provide a solution to the challenge of model selection. G-DCMs such as the G-DINA and LCDM have been hailed for their flexibility to allow each item on an assessment to pick its own model depending on how the attributes required by the item combine to generate an observed response. In other words, with G-DCMs researchers need not to apply a single DCM across the board, rather several DCMs are possible within the same assessment. According to Ravand (2016) it sounds more viable to hypothesize that due to the complexity of the cognitive processes underlying successful performance on items and the variety of factors affecting performance, the difficulty of the attributes, the domain of the construct tapped by the items, the cognitive load of the attributes (e.g., whether they tap higher or lower order thinking), etc., the relationships among the attributes might change across items. Therefore, one cannot assume the same relationship across all items of a test.

However, flexibility of the G-DCMs comes at a price. G-DCMs in their saturated form estimate more item and person parameters that may result in overly complex models with the following ramifications: (1) interpretation of the model output may be difficult, (2) more estimation time may be needed, (3) there might be convergence issues, (4) large sample sizes are required, and finally (5) as Yi (2012, p. 49) noted: "... if the model is too complex for the given data, it may result in overfitting ..." which is inconsistent with the parsimony principle. In contrast, specific DCMs have been lauded for their more straightforward interpretation, the smaller sample size they require, their consistency with the parsimony principle, and provision of more accurate classifications when sample size is small (Ma, Iaconangelo, & de la Torre, 2016). It should be noted that the possibility of model selection at item level as a corollary of G-DCMs application renders the interpretability concern unwarranted.

MODEL EVALUATION

DCMs can be evaluated from different perspectives: fit, classification consistency and accuracy, item discrimination, and congruence of attribute difficulty with substantive expectations.

Fit in DCMs can be studied from three aspects: model fit, person fit, and item fit. Model-data fit can be evaluated at the level of test or item (i.e., item fit). Most of the fit indices reported for DCM are test-level model fit. Person fit refers to the degree to which test takers' observed responses deviate from

what is expected based on their attribute profiles (Liu, Douglas, & Henson, 2009). Nonmasters of any one of the required attributes who correctly answer a lot of items and masters of all the attributes who get many items wrong are examples of aberrant-behaving persons. Item fit is judged based on a discrepancy measure obtained from the difference between the actual responses to a given item and predictions made by a DCM. To this end, test takers are classified into different proficiency groups and the mean discrepancy between the observed and predicted responses for each group is calculated. From among the three aspects of fit, most fit studies in DCMs have addressed model fit. In what follows considerations regarding model fit at test and item levels are discussed.

Test-Level Model Fit

As with other statistical models, before interpreting parameter estimates of DCMs, fit of the models should be explored. Misfit in DCMs could be due to any or a combination of the following reasons: (1) assumptions of the selected DCM (i.e., compensatory/noncompensatory) do not match those assumed by the researcher and (2) the Q-matrix is misspecified. Q-matrices may be under-specified (a 1 has been erroneously specified as 0, indicating an attribute is not measured by a given item when it really is) or over-specified (a 0 has been erroneously specified as 1). *Relative fit* indices compare fit of different DCMs to a given data set and are appropriate for model selection purposes. *Absolute fit* indices are used to evaluate fit of any given model to the data. Relative fit indices that are usually used to compare different DCMs are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Both AIC and BIC impose penalty for the number of parameters in the model. BIC introduces larger penalty for overparameterization and also for larger sample sizes. The smaller values for both indices indicate a better model-data fit. Studies have shown that AIC has a better accuracy rate than BIC in selecting the best model when the Q-matrix was specified correctly and the generating model was a complicated model such as general models whereas performance of BIC is better when the generating model is simpler and sample size is larger (Kunina-Habenicht et al., 2012; Lei & Li, 2016).

Absolute fit indices are based on the residuals obtained from the difference between the observed and model predicted values. However, some researchers have resorted to absolute fit indices in model comparison (e.g., Li, Hunter, & Lei, 2015; Ravand, 2016; Ravand & Robitzsch, 2018; Yi, 2017). Kunina-Habenicht et al. (2012) suggested that the use of absolute indices for model comparison purposes could yield useful information especially when cutoffs are not available for these indices. On the other hand, some other researchers

have found that absolute fit indices are not suitable for model or Q-matrix selection for model comparison purposes (e.g., Chen et al., 2013; Hu et al., 2016; Lei & Li, 2016).

The following fit indices have been used in the literature:

1. X^2 (Chen & Thissen, 1997) is a measure of local dependence (LD) used in IRT. It is an index of independence of pairwise item response frequencies. The test-level MX2 is an adaptation of the item-level X^2 , which is averaged over all the item pairs. It is the mean difference between the model-predicted and observed response frequencies. Large differences are taken as evidence that there are dependencies between the items. Because respondents draw upon the same cognitive processes to respond to the items, dependencies are expected. However, if a given DCM fits the data well, “the X^2 -test statistic is expected to be 0 within each latent class as the attribute profile of the respondents would perfectly predict the observed response patterns” (Rupp, Templin, & Henson, 2010, p. 269).
2. The mean absolute difference for the item-pair correlations (MADcor) statistic (DiBello et al., 2007) is also a measure of LD, which is averaged across all the item correlation residuals: the difference between the observed and the model-predicted item correlations.
3. The mean residual covariance (McDonald & Mok, 1995; MADRES) is another LD index that is averaged over all the item covariances residuals: the mean difference between matrices of observed and reproduced item covariances.
4. The Q3 statistic (Yen, 1984) is another LD index used in IRT contexts. It is calculated by subtracting the model-predicted responses from the observed responses of the respondents and computing the pairwise correlation of these residuals. The average of Q3s over all the items’ residuals (MADQ3) is used as another test-level absolute fit index.
5. The root mean square error (RMSEA) is the mean difference between response proportions predicted by the model and those observed for each response category within each latent class weighted by the proportion of the test takers within the respective latent class.
6. The standardized root mean squared residual (SRMSR) is a fit index borrowed from factor analysis. For any item pair, SRMSR is the observed correlation between the items minus the expected correlation. Maydeu-Olivares (2013, p. 84) considered SRMSR values of below 0.05 as indicating “a substantively negligible amount of misfit.”

It should be noted that Lei and Li (2016) found that the above fit indices except for the MX2 were very much sensitive to sample size. MX2 was the

least impacted by sample size under the condition that the selected model and Q-matrix were the true ones.

Classification Consistency and Accuracy

As the Standard 5.12 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) requires, reliability, and validity of scores should be established before they are reported. Thus, the burden of establishing reliability and validity of attribute classifications before reporting the DCM profile scores lies on the researcher. To this end, classification consistency and accuracy have also been used in the literature to evaluate models (Lee & Sawaki, 2009; Ravand, 2016; Ravand & Robitzsch, 2018).

Cui, Gierl, and Chang (2012) presented two indices namely classification consistency (P_c) and accuracy (P_a) to refer to the reliability and validity of the examinees' classification into the latent classes or master/nonmaster of each separate skill. P_c is an indicator of the degree to which an examinee is consistently classified into the same latent class or will be indicated as master/nonmaster of the same attribute on readministration of the same or a parallel form of the test, and P_a refers to the degree to which an examinee's classification matches his true latent class or he is truly identified as master/nonmaster of any given attribute.

Discrimination Indices

According to Rupp et al. (2010), assessments with higher discrimination indices are expected to lead to more reliable classification of test takers as master or nonmasters of any given attribute. Thus, a DCM with higher discrimination indices should be preferred because it results in more accurate classifications. Discrimination definitions in educational assessment in general are either grounded in CTT or IRT. In both approaches, discrimination can refer to global item discrimination and attribute discrimination. In the CTT-based definition, it is a matter of the degree to which an item can discriminate between the test takers who have mastered all the attributes necessary to get the item right and those who have mastered none of the measured attributes. In other words, in the CTT-based definition, item discrimination compares the probability of getting a given item right for those who have mastered all the attributes required by the item to the probability of getting the same item right for those who have not mastered any of the required attributes.

Attribute discrimination is a more specific measure of discrimination which measures the degree to which probability of getting an item right is different between those who have mastered any given attribute and those who have not mastered that attribute. The IRT approach zeros in on the notion of statistical information. In this approach an index of discrimination is derived through Kullback-Leibler information (KLI), which gauges how informative a diagnostic assessment is for the classification of test takers. Nontechnically speaking, in the KLI approach, the predicted response patterns for any given item are compared for test takers from two different latent classes. An assessment that provides more pronounced discriminations between test takers from different attribute profiles is considered a better assessment. In the KLI approach, for any given item, the discrimination indices are calculated for any pairs of possible attribute profiles for that item. For a complete discussion of KLI see Rupp et al. (2010).

Congruence of Attribute Difficulties with Substantive Expectation

DCMs provide the prevalence of attributes among the subjects studied which could indicate how difficult the attributes are for the given subjects. To check the degree of match between the order of attribute difficulty generated by the DCMs and what is substantively expected, one can ask expert judges to rate the involved attributes in terms of difficulty and then compare the order with the DCM output.

Model Selection at Item Level

As it was alluded to before, a very recent development in DCM is the possibility of fitting specific models to individual items, because test-level model fit takes into account all the items on the test. In cases of misfit, it is not clear which subset of items should shoulder the blame. Henson et al. (2009) introduced the first method of item-level model fit by which the best specific model could be selected through visual inspection. Later on, de la Torre and Lee (2013) successfully used the Wald test to compare fit of the DINA, DINO, and ACDM against the G-DINA. In two other very recent studies, Ma, Laconangelo, and de la Torre (2016) and Sorrel, Abad, Olea, de la Torre, and Barrada (2017) also investigated model selection at item level. There are two R-packages, GDINA (Ma & de la Torre, 2018) and CDM (Robitzsch, Kiefer, George, & Uenlue, 2017) that implement the Wald test to compare the fit of DCMs at the item level.

To find the best DCM for each item first, the G-DINA is fitted to the data in its saturated form. As it is a general model, the G-DINA is expected to

show better fit than the constrained models. Therefore, it can be used as a metric against which fit of the constrained models could be compared. In the second step, reduced models are fitted to each item. If fit of the reduced model to each item does not result in a worse fit as compared to that of the G-DINA, the reduced model is preferred, otherwise the G-DINA is the best model. According to Ma et al. (2016), the process of model selection at item level can be described as follows: First, the Wald statistic for the reduced models for individual items is calculated. The null hypotheses are the fit of the reduced model equals the fit of the general model. If the null hypothesis is rejected ($p < .05$) the reduced model is rejected. If more than one reduced models are retained and the DINA or DINO are among the retained models, the DINA or DINO with the largest p value are selected, but if the DINA and DINO are not among the retained models, the reduced model with the largest p value is selected. It should be noted that when several reduced DCMs have p values larger than .05, DINA or DINO are preferred over the other specific DCMs as they are statistically the least complex DCMs (Rupp & Templin, 2008).

OPTIMAL NUMBER OF ITEMS, GRAIN SIZE, AND SAMPLE SIZE

As the number of attributes required per item increases, issues of identifiability of the DCMs might occur (DiBello et al., 2007). DCMs may be made complex to make them represent the cognitive theory of test performance, which might lead to estimation error and identifiability issues. A review of the rather scant literature on the optimal number of attributes (i.e., granularity) measured by an assessment shows that as the number of attributes increases, the number of items and the sample size should increase as well. Also, the selected model can impact considerations regarding the number of attributes and sample size. When the selected DCM is, for example, a general model in its saturated form, more attributes for each item mean more n -way interactions among the attributes, which in turn requires larger sample sizes and more items. Model identification, computational time and resource considerations may limit the number of attributes an item can measure (Templin & Bradshaw, 2013). Skaggs, Hein, and Wilkins (2016) found that as the number of attributes measured by the whole test and the number of attributes measured by each item increased, the standard errors of item parameters increased as well. Skaggs et al. also found that item parameter bias tended to be larger as the number of attributes measured by the whole test and those measured by each single item increased.

In a simulation study, Kunina-Habenicht et al. (2012) investigated the effect of sample size on parameter recovery. They applied the LCDM and found that with a sample size of 1,000 and one or two attributes per item, the main effects

were measured accurately whereas estimation of two-way interactions were unreliable. They also found with sample size of 1,000 and three attributes measured by each item, neither the main nor the interaction effects were accurately estimated. Furthermore, they found that the three-way interactions could not be reliably estimated even with sample sizes of 10,000. Especially, they found when the sample size is small and the number of items equals or is less than the number of latent classes, recovery of parameter estimates for the interaction terms is seriously impaired. Accordingly, they suggested the use of simple constrained DCMs when sample size is small.

In another study, Galeshi and Skaggs (2016) examined parameter recovery and classification accuracy within the framework of the compensatory reparameterized unified model (Hartz, 2002). The study included sample sizes of 50, 100, 500, 1,000, 5,000, and 10,000 and different combinations of attribute-item. They found that accuracy of both parameter recovery and classification was a function of the number of items and attributes as well as the sample size. The results of the study showed that as the number of attributes required by an item increases and the number of items decreases, larger sample sizes are required to obtain accurate estimates. They argued that with the same number of items, the smaller the number of attributes measured by an item the smaller the required sample size is.

Lei and Li (2016) studied performance of fit indices in choosing correct DCMs and Q-matrices and found sample size was the most influential factor. They found that sample size had a negligible effect on classification accuracy but a substantial effect on performance of fit indices. Increase in sample size resulted in increase in the relative fit indices but decrease in the absolute fit indices.

As to the grain size of the attributes, Jang (2009) suggested that the issue be considered from at least three aspects: first, theoretical (construct representativeness). Measurement of any attribute with as few as three to five attributes, for example, is likely to lead to content coverage/construct representation issues. According to Jurich and Bradshaw (2014), when the breadth of an attribute is underrepresented test takers may be misclassified. Second aspect is technical (number of items per attribute). From a statistical point of view, attributes measured by more items are expected to have higher attribute reliabilities. Besides, if few items measure an attribute and just one of the items, for example, discriminates highly between masters and nonmasters of the attribute (i.e., the attribute has a large main effect), yet the other items poorly discriminate, performance on the highly discriminating item determines classification of the test takers: those who have responded correctly to the item with the large main effect for that attribute are classified as the master of the attribute, yet those who have missed the item may be classified as nonmasters and, third, practical (usefulness for diagnostic feedback).

As a safe rule of thumb, it can be suggested that more fine-grained attributes should only be used when we are applying simpler DCM models and when there are larger number of items and test takers. As to the number of items, Hartz (2002) recommended at least three items for any given attribute. However, Jang (2009) suggested at least five items per attribute. It should be noted that high-quality items, that is, items which discriminate highly between masters and nonmasters of the skills are required for successful application of DCMs.

OUTLOOK

As mentioned earlier, DCMs have not yet penetrated into the mainstream classroom assessment for the purposes of providing diagnostic feedback to improve learning and teaching. The main reason perhaps is that diagnostic testing in general does not have a well-defined place in education. Diagnostic assessment has a separate identity from the psychometric models referred to as DCMs. Long before DCMs were introduced, diagnostic assessment existed. However, its existence was limited to a definition in test development textbooks. In the classification of different types of tests, diagnostic testing is usually defined as a test designed to identify learners' strengths and weaknesses (Hughes, 2003). Nevertheless, few, if any, attempts have been made to design diagnostic tests in education. Most educational tests are achievement or competency tests developed to locate individuals on a continuum of ability for the purpose of comparison and pass or fail. DCMs can complement and inform the practice of diagnostic assessment. As long as diagnostic testing has not been fully implemented in educational systems, the merits DCMs add to this enterprise will not be known. We believe the reason why DCMs have not been applied in true diagnostic situations is that educators do not need them because little diagnostic testing takes place in schools. If diagnostic testing existed, we believe, educators would have embraced DCMs as extremely useful developments to help them in their endeavor.

Other reasons include the complexity of the models and their inaccessibility for classroom teachers, lack of user-friendly programs to estimate the models, and, most importantly, the sample size requirements of the models that make their applications in small-scale classroom contexts almost impossible.

To integrate DCMs into the classroom activities so that they become the stock-in-trade of the teachers, more orchestrated efforts should be made on at least two fronts: (1) methodological and (2) practical. Methodologically, more studies on the effect of sample size on the performance of DCMs such as classification accuracy, model fit, etc. should be conducted. Practically, the problems that would get in the way of DCM applications should be eased. We

believe the complexity issue is not very serious. Although the mathematical bases of DCMs are extremely sophisticated, practitioners do not need to get involved with them. On the contrary, we think that understanding what DCMs do and their benefits are easy to explain for teachers. All school teachers with minimum teaching certification requirements are familiar with diagnostic testing, providing feedback for improved learning, and the topic of subskills underlying basic skills in math, languages, sciences, etc. Therefore, it should be a lot easier for teachers to grasp and appreciate the applications of DCMs in their career than the application of IRT models or structural equation models.

Therefore, the major phase of DCM analysis, namely, Q-matrix development, can efficiently be conducted by teachers. Even in theoretical applications teachers are always invited as experts to code the items to construct Q-matrices. DCM outputs do not seem to pose any difficulty for teachers either. Understanding which subskills a student has mastered and which s/he has not or which attributes are the hardest and which are the easiest, the main useful messages of DCMs are not difficult for teachers, students, their parents and other stakeholders to follow.

If the application of the models for classroom diagnostic purposes is truly desired, theoretical DCM researchers and model developers should give top priority to two other major issues. The first one is the advancement of estimation and model evaluation methods which do not require large sample sizes for stable and reliable estimates. A common feature of classroom settings is small sample sizes. As long as DCMs need sample sizes in the magnitude of 1,000–2,000, they never leave the psychometric laboratories, and school teachers and other stakeholders never take advantage of them. The second important issue that should be addressed is the development of more user-friendly software programs that can easily be used by school teachers. The current programs, whether R-packages or other commercial programs are difficult to work with and the outputs are rather hard to interpret. These future programs should be “click and point” programs with no code writing and with simple outputs that clearly depict the most important things a school teacher needs to know about their students and the syllabi they have taught.

Although the second issue is easier to deal with and can easily be addressed by some interested and creative program writers the first one may take a couple of decades to materialize. Meanwhile, for practical purposes, practitioners can use the available less-complex models that require smaller sample sizes. Depending on the assumptions regarding how the attributes underlying any given test interact to generate responses, one can select models with fewer parameters, hence smaller sample size requirements. If the attributes interact in a noncompensatory way, the DINA model, which is one of the simplest

available DCMs, may be appropriate. The DINA estimates two parameters per item, regardless of the number of attributes required. If the attributes interact in a compensatory way, the additive cognitive diagnostic model (ACDM; de la Torre, 2011) is recommended. The ACDM estimates only main effects of the attributes. Research (e.g., Kunina-Habenicht et al., 2012) has shown that estimation of main effects requires smaller sample sizes than estimation of the interaction effects.

ARE CONCERNS RAISED 10 YEARS BACK STILL IN PLACE?

Our article attempts to portray the current status of the DCMs in terms of both their method and practice. Methodologically, we reviewed some of the most important extensions to DCMs developed post-2008. We tried to suggest some future directions in light of the current concerns and also consider the concerns raised in Rupp and Templin (2008) and the commentaries it received.

Method-wise, it seems that DCMs have firmed their foundation so that they can compete with their rival predecessors such as continuous IRT models. However, despite all the methodological advancements, there are still other issues that need to be addressed in the future studies: the effect of sample size, missing data, and item and person local dependence on DCM results. Furthermore, algorithms that could yield stable estimates with sample sizes typically found in normal classroom is what is most wanted. On the contrary, application-wise, DCMs have moved at a pace much slower than their methodology. It seems that after 10 years still most of the commentators' concerns as to the practice of DCMs linger on.

There have been relatively few applications of DCM (in the sense we used the term application in the present study). Most of these applications have concerned retrofitting to high-stakes tests. Most of the commentaries (e.g., Gierl & Cui, 2008; Gorin, 2009; Jiao, 2009; Sinharay & Haberman, 2009; Tatsuoka, 2009) on Rupp and Templin's article advised against retrofitting DCMs to assessments supposed to be unidimensional. To deliver their original mission, they have to be used in true DCM studies to develop tests from scratch. Therefore, it can be concluded that, methodologically, DCMs have been fruitful whereas, practically, they seem to be still as lagging behind as they were about 10 years ago. To fill in the void in the practice of DCMs, a set of measures, some which also echoed in Rupp and Templin (2009) and Henson (2009), are suggested: accessible software programs, didactic materials in order to establish a set of best practices and codes of conduct, redefining test development procedures to align them with the requirements of DCMs, among others. Studies such as the present serve to suggest a set of hands-on procedures for the implementation and reporting

the results of DCMs. As to the availability of DCM software program, compared to 10 years back, there are more software programs available. Nonetheless, the available programs are not of the point-and-click nature that appeals to most researchers.

As to the alignment of test development procedures, many of the commentators (e.g., Gierl & Cui, 2008; Henson, 2009) believe that full potentials of the DCMs will be realized only when test development procedures are redefined so that they are in line with requirements of the DCMs. To this end, Gierl and Cui suggest some requirements for test design and analysis besides the standard practices of test development. Their *principled test design and analysis* starts with a cognitive model which describes the processes, strategies and attributes needed to be applied to solve tasks or problems. Items are developed to measure the knowledge and skills specified in the cognitive model and finally the DCMs are applied to analyze the data. Thus, the cognitive theory plays a key role in test construction and score interpretation. Without a detailed cognitive theory of response processes, “that can support and add practical benefit of DCMs over alternative measurement models, we do have somewhat of an identity crisis indeed” (Rupp & Templin, 2008).

Making do without a cognitive theory, a common practice in retrofitting, would compromise the quality of DCM results. A corollary of the above concern associated with the example-of-methodology studies is that they have invariably been retrofitting cases. It is not clear whether new extensions work with true DCM studies in the same way. Moreover, most of these retrofitting studies have used the popular fraction-subtraction data. It should be noted that educational testing and assessment is not confined to just math. DCMs hold promise for other areas such as language testing. Future application studies should be of true DCM studies and to demonstrate that the results obtained in simulation studies can also be replicated with real data, data from true DCM studies should be used. That being said, it seems that a pressing need is development of cognitive theories of response processes, which is beyond the realm of DCM research.

As to model selection, we share the concern of Wilhelm and Robitzsch (2009) that choice of the DCMs is commonly an arbitrary process rather than being informed by substantive considerations. As to model selection, there are at least three possibilities: (1) blanket imposition of a single specific DCM on all the items of a test, a practice we advise against, (2) running several specific models and selecting the one fitting the data the best, and (3) running a G-DCM and letting each item pick its own model, hence several specific models within the same DCM. When sample size is large enough ($> 5,000$) we recommend the third possibility. However, if the sample size is not large enough to run a G-DCM, the second possibility is what we recommend.

Finally, it may be too soon to expect DCMs become the stock-in-trade of classroom teachers. The evolution of DCM is reminiscent of how another latent trait model namely structural equation modeling (SEM) has evolved. In its early stages of development, SEM was so arcane that only scientists well-versed in matrix language and computer programming, had the privilege of benefiting from its robust modeling capabilities. However, along their evolutionary path, the matrix language was translated into a language shared by a wider group of researchers which could be implemented in several user-friendly software programs. Above all, more recently, with the development of the partial least squares SEM (PLS-SEM, see Ravand & Baghaei, 2016), working with small sample sizes, which was once a wild dream, has turned into a kitchen appliance for the SEM community. In the same vein, as Henson (2009) aptly reminded us, more pedagogical things including but not limited to easy-to-use software programs, articles that set out best practices and a set of good conduct should be produced to make the DCMs accessible to a wider community of researchers. More importantly, new algorithms need to be programmed so that DCMs can be reliably estimated in low-stakes situations with sample sizes as small as those normally found in a classroom.

FUNDING

The second author has been supported by Alexander von Humboldt Foundation via Group Linkage Program.

ORCID

Hamdollah Ravand  <http://orcid.org/0000-0002-8757-3850>

Purya Baghaei  <http://orcid.org/0000-0002-5765-0413>

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology*, 10, 233–252.
- Barnes, T. (2010). Novel derivation and application of skill matrices: The Q-matrix method. In C. Ramero, S. Vemtoro, M. Pechemizkiy, & R. S. J. de Baker (Eds.), *Handbook of educational data mining* (pp. 159–172). Boca Raton, FL: Chapman & Hall.
- Bolt, D., Chen, H., DiBello, L., Hartz, S., Henson, R. A., & Templin, J. L. (2008). *The Arpeggio Suite: Software for cognitive skills diagnostic assessment {computer software and manual}*. St. Paul, MN: Assessment Systems.

- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50(2), 123–140. doi:10.1111/j.1745-3984.2012.00185.x
- Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika*, 82(3), 660–692. doi:10.1007/s11336-016-9545-6
- Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, 110(510), 850–866. doi:10.1080/01621459.2014.934827
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289. doi:10.3102/10769986022003265
- Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37(8), 598–618. doi:10.1177/0146621613488436
- Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74(4), 633–665. doi:10.1007/s11336-009-9125-0
- Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, 27(1), 131–157. doi:10.1207/s15327906mbr2701_8
- Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, 49(1), 19–38. doi:10.1111/j.1745-3984.2011.00158.x
- De Carlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, 36(6), 447–468.
- 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(4), 343–362. doi:10.1111/j.1745-3984.2008.00069.x
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179–199.
- de la Torre, J., & Chiu, C.-Y. (2016). General method of empirical Q-matrix validation. *Psychometrika*, 81(2), 253–273. doi:10.1007/s11336-015-9467-8
- de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, 73(4), 595. doi:10.1007/s11336-008-9063-2
- de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, 50(4), 355–373. doi:10.1111/jedm.12022
- de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, 20(2), 89–97. doi:10.1016/j.pse.2014.11.001
- Desmarais, M., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In Lane, H. C., Yacef, K., Mostow, J., & Pavlik, P. (Eds.), *Proceedings of the 6th International Conference on Artificial Intelligence in Education* (pp. 441–450). Heidelberg: Springer.
- DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In Rao, C. R., Sinharay, S. (Eds.), *Handbook of statistics*, vol. 26: Psychometrics (pp. 979–1030). Amsterdam, Netherlands: Elsevier.
- DiBello, L. V., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Hillsdale, NJ: Lawrence Erlbaum.
- Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika*, 78(1), 14–36. doi:10.1007/s11336-012-9296-y

- Galeshi, R., & Skaggs, G. (2016). Parameter recovery of a cognitive diagnostic model: Evidence from a simulation study. *International Journal of Quantitative Research in Education*, 3(4), 223–241. doi:[10.1504/IJQRE.2016.082386](https://doi.org/10.1504/IJQRE.2016.082386)
- Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives*, 6(4), 263–268. doi:[10.1080/15366360802497762](https://doi.org/10.1080/15366360802497762)
- Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242–274). New York, NY: Cambridge University Press
- Gorin, J. S. (2009). Diagnostic classification models: Are they necessary? Commentary on Rupp and Templin. *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 30–33. doi:[10.1080/15366360802715387](https://doi.org/10.1080/15366360802715387)
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Doctoral dissertation. University of Illinois at Urbana-Champaign.
- Henson, R. A. (2009). Diagnostic classification models: Thoughts and future directions. *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 34–36. doi:[10.1080/15366360802715395](https://doi.org/10.1080/15366360802715395)
- 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(2), 191–210. doi:[10.1007/s11336-008-9089-5](https://doi.org/10.1007/s11336-008-9089-5)
- Hou, L., la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, 51(1), 98–125. doi:[10.1111/jedm.12036](https://doi.org/10.1111/jedm.12036)
- Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. -H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, 16(2), 119–141. doi:[10.1080/15305058.2015.1133627](https://doi.org/10.1080/15305058.2015.1133627)
- Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, 38(6), 464–485. doi:[10.1177/0146621614533986](https://doi.org/10.1177/0146621614533986)
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing*, 26(1), 031–073. doi:[10.1177/0265532208097336](https://doi.org/10.1177/0265532208097336)
- Jiao, H. (2009). Diagnostic classification models: Which one should I use? *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 65–67. doi:[10.1080/15366360902799869](https://doi.org/10.1080/15366360902799869)
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272. doi:[10.1177/01466210122032064](https://doi.org/10.1177/01466210122032064)
- Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing*, 14(1), 49–72. doi:[10.1080/15305058.2013.835728](https://doi.org/10.1080/15305058.2013.835728)
- Kaya, Y., & Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and Psychological Measurement*, 77(3), 369–388. doi:[10.1177/0013164416659314](https://doi.org/10.1177/0013164416659314)
- Kim, A.-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, 32(2), 227–258. doi:[10.1177/0265532214558457](https://doi.org/10.1177/0265532214558457)

- Kim, Y. H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing*, 28(4), 509–541. doi:10.1177/0265532211400860.
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, 49(1), 59–81. doi:10.1111/j.1745-3984.2011.00160.x
- Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, 6(3), 172–189. doi:10.1080/15434300902985108
- Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement*, 40(6), 405–417. doi:10.1177/0146621616647954
- Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spain Fellow*, 9, 17–46.
- Li, F., Cohen, A., Bottge, B., & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, 76(2), 181–204. doi:10.1177/0013164415588946
- Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33, 391–409. doi:10.1177/0265532215590848
- Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, 33(8), 579–598. doi:10.1177/0146621609331960
- Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, 78(3), 357–383. doi:10.1177/0013164416685599
- Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, 36(7), 548–564. doi:10.1177/0146621612456591
- Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, 75(3), 491–511. doi:10.1177/0013164414539162
- Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, 83(4), 963–990. doi:10.1007/s11336-018-9638-5
- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, 40(3), 200–217. doi:10.1177/0146621615621717
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2), 187–212. doi:10.1007/BF02294535
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, 11(3), 71–101. doi:10.1080/15366367.2013.831680
- Ma, W. & de la Torre, J. (2018). GDINA: The generalized DINA model framework. R package version 2.1. Retrieved from <https://CRAN.R-project.org/package=GDINA>
- McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30(1), 23–40. doi:10.1207/s15327906mbr3001_2
- Morrison, D. F. (1967). *Multivariate statistical methods*. New York, NY: McGraw-Hill.
- Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, 34, 782–799. doi:10.1177/0734282915623053
- Ravand, H. (2019). Hierarchical diagnostic classification models in assessing reading comprehension. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment II: Advanced methods* (pp. xxx). New York, NY: Routledge.

- Ravand, H., & Baghaei, P. (2016). Partial least squares structural equation modeling with R. *Practical Assessment, Research & Evaluation, 21*(11). Retrieved from <http://pareonline.net/getvn.asp?v=21&n=11>.
- Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation, 20*, 112.
- Ravand, H. & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology, 38*(10), 1255–1277. doi:10.1080/01443410.2018.1489524
- Robitzsch, A., Kiefer, T., George, A., & Uenlue, A. (2017). CDM: Cognitive diagnosis modeling. R package version 3.1-14: Retrieved from the Comprehensive R Archive Network [CRAN] website <http://CRAN.R-project.org/package=CDM>.
- Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*(4), 219–262. doi: 10.1080/15366360802490866
- Rupp, A., & Templin, J. (2009). The (un)usual suspects? A community in search of its identity. *Measurement, 7*, 115–121. doi:10.1080/15366360903187700
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Rupp, A. A., & van Rijn, P. W. (2018). GDINA and CDM packages in R. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 71–77. doi:10.1080/15366367.2018.1437243
- Sinharay, S., & Haberman, S. J. (2009). How much can we reliably know about what examinees know? *Measurement: Interdisciplinary Research and Perspectives, 7*, 46–49. doi:10.1080/15366360802715486
- Skaggs, G., Hein, S. F., & Wilkins, J. L. (2016). Diagnostic profiles: A standard setting method for use with a cognitive diagnostic model. *Journal of Educational Measurement, 53*(4), 448–458. doi:10.1111/jedm.12125
- Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 41*(8), 614–631. doi:10.1177/0146621617707510
- Tatsuoka, C. (2009). Diagnostic models as partially ordered sets. *Measurement: Interdisciplinary Research and Perspectives, 7*(1), 49–53. doi:10.1080/15366360802715510
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354. doi:10.1111/j.1745-3984.1983.tb00212.x
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*(1), 55–73. doi:10.3102/10769986010001055
- Templin, J. (2009). On the origin of species: The evolution of diagnostic modeling within the psychometric taxonomy (Powerpoint). International meeting of Psychometric Society pre-conference workshop, Cambridge University, England.
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251–275. doi:10.1007/s00357-013-9129-4
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3):287–305. doi:10.1037/1082-989X.11.3.287
- von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series, 2005*(2). doi:10.1002/j.2333-8504.2005.tb01993.x
- Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics, 43*(1), 57–87. doi:10.3102/1076998617719727

- Wilhelm, O., & Robitzsch, A. (2009). Have cognitive diagnostic models delivered their goods? Some substantial and methodological concerns. *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 46–49.
- Xu, G. and Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, 81(3), 625–649 doi:[10.1007/s11336-015-9471-z](https://doi.org/10.1007/s11336-015-9471-z)
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145. doi: [10.1177/014662168400800201](https://doi.org/10.1177/014662168400800201)
- Yi, Y.-S. (2017). In search of optimal cognitive diagnostic model(s) for ESL grammar test data. *Applied Measurement in Education*, 30(2), 82–101. doi:[10.1080/08957347.2017.1283314](https://doi.org/10.1080/08957347.2017.1283314)
- Yi, Y. (2012). *Implementing a cognitive diagnostic assessment in an institutional test: a new networking model in language testing and experiment with a new psychometric model and task type*. Unpublished doctoral dissertation. University of Illinois at Urbana Champaign, Urbana-Champaign, IL.