CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Cognitive Diagnostic Models and how they can be useful

Research Report

Joanna Williamson

07 December 2023

Author contact details:

Joanna Williamson
Assessment Research and Development,
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

joanna.williamson@cambridge.org
https://www.cambridge.org/

How to cite this publication:

Williamson, J. 2023. *Cognitive Diagnostic Models and how they can be useful*. Cambridge University Press & Assessment.

Contents

Introduction

There is a lot of interest in providing detailed reports to schools indicating which skills pupils have mastered and which still need development – and, more broadly, the knowledge, skills and understanding that pupils have acquired and not yet acquired. Cognitive diagnostic assessment is an approach designed to provide this kind of insight. The core assumption is that correctly answering a question depends on having mastery of a specific set of latent skills and/or knowledge. By asking test takers to attempt carefully constructed sets of well-designed items, and then analysing their responses, we can make detailed inferences about the attributes they have and have not mastered (Leighton & Gierl, 2007b). Rather than producing a score or grade (or even multiple scores), the aim of a cognitive diagnostic assessment is to assign test takers to discrete classes based on their mastery of the different attributes (e.g., skills) being measured in the test. In this way, cognitive diagnostic assessment provides both fine-grained information about the skill profiles of individual test-takers, and the identification of latent subgroups in the test-taking population (Ma et al., 2023).

Deriving such fine-grained inferences from test-takers' responses is not a trivial task (Leighton & Gierl, 2007b, pp. 11-14). Cognitive diagnostic assessment requires a detailed theory or model of the knowledge and skills to be measured, and how they relate to items. This theory then guides the assessment design, which may follow a formalised approach such as Evidence Centred Design (Mislevy et al., 2003) to ensure sufficiently robust validity. Cognitive diagnostic assessment then requires specific technical approaches to modelling and interpreting test-takers' responses. Cognitive Diagnostic Models (CDMs) — also known as Diagnostic Classification Models (DCMs) — are an important mainstream formal approach to this modelling. CDMs are a subset of multidimensional latent variable models, designed to "diagnose" students according to their mastery of the various skills or attributes being measured in a particular domain.

Leighton and Gierl's comprehensive *Cognitive Diagnostic Assessment for Education* (2007a) summarised the state of cognitive diagnostic assessment and raised its profile. At this time, CDMs presented attractive possibilities, but there existed many competing approaches and some obstacles to practical implementation. In recent years, there have been major developments in cognitive diagnostic assessment and the field of CDM research specifically. Meanwhile, the demand for detailed reporting – for the purposes of formative assessment but also to investigate and evidence assessment validity – has continued to grow. There is also increasing interest in estimating learners' mastery of specific skills in order to facilitate adaptivity and personalisation within digital learning and assessment products – for example, to drive recommendation systems that suggest pages of interactive learning material based on the skills shown to have been mastered (or not mastered) so far. Cognitive diagnostic models are a significant area of overlap between psychometrics and the techniques employed within advanced digital learning systems, such as intelligent tutoring systems.

The overarching aim of this report is to give an up-to-date view on cognitive diagnostic models and what they might offer. In particular, it considers how assessment organisations could use CDMs to provide something of meaningful value to schools, and what would be required to make this work.

Research questions

The core questions to address are:

1. In a "best-case" scenario (sufficient resources to meet sample size, item construction and test design requirements), how would reporting to schools based on CDMs be an improvement upon (i) raw sub-score reporting by topic, (ii) topic or domain scaled scores (e.g., Cambridge English scale scores, Cambridge Checkpoint scores), or (iii) raw item level results?
2. What would be the minimum requirements in terms of sample sizes, number of items and test design considerations to allow valid reporting based on CDMs?
3. Are CDMs a technology that assessment organisations could use to report outcomes from existing assessments?
4. What are the considerations around score interpretation and transparency?
5. For which purposes, subjects and customers might assessment organisations want to use CDMs?

Background: what counts as a CDM?

Surveying the field of cognitive diagnostic modelling is not an entirely straightforward exercise. In their introduction to the *Handbook of Diagnostic Classification Models*, von Davier and Lee (2019b, p. 1) recognised a need for clarification, but did not sound overly optimistic: "We are attempting to organize the growing field somewhat systematically to help clarify the development and relationships between models. However, given the fact that DCMs have been developed based on at least two, if not three traditions, not all readers may necessarily agree with the order in which we put the early developments." Complicating factors include that different authors use the term "CDM" with different scope[1], and that research has regularly shown equivalencies between previously separate lines of model development. That is, different models "may be later understood as variants of one common more general approach" (von Davier & Lee, 2019b, p. 2). Besides these specific factors, the field of CDMs may also be described differently depending on whether authors take an historical view; organise modelling approaches by technical features (e.g., statistical assumptions); offer an account framed by one unifying model; or concentrate on the practicalities of applications.[2]

Zhang et al. (2023) offer a useful account of statistical approaches to cognitive diagnostic assessment, and the relation of mainstream CDM methods to predecessor approaches and other psychometric models. They propose that the two main precursors to cognitive diagnostic testing were "mastery testing based on the IRT framework" (pp. 653-654), and the mastery model based on latent class models (e.g., Macready & Dayton, 1977). This mastery

[1] This also applies to the alternative term "DCM".

[2] The challenge of defining CDMs is far from new: Rupp and Templin (2008b) dedicate a substantial section of their review paper to "the more exact definitional boundaries of DCM" (p. 224). They argue that applied assessment practitioners and specialists will "typically want to understand how particular models differ from one another in more detail rather than what the largest statistically [sic] family is that they can be subsumed under". Their discussion also highlights what different choices of labels prioritise (Rupp & Templin, 2008b, pp. 225-227). For reference, Appendix A lists different definitions of "CDM" from major researchers in the field, as well as definitions for other key terms and abbreviations used in this report.

model assumes that test-takers belong to one of a set of pre-specified latent classes with a hierarchical structure, and models the relationship between class membership and observed test responses. The rule space method (RSM - Tatsuoka, 1983) and knowledge space theory (Doignon & Falmagne, 1985, 2012) are credited by Zhang et al. as independently developed approaches that between them established the framework for cognitive diagnostic assessment as we now know it. Zhang et al. (2023, p. 655) formally define this framework as follows:

1. "Performance on a test is assumed to depend on the mastery status, or knowledge state (Tatsuoka, 2009), of a collection of K attributes, denoted $\boldsymbol{\alpha} = (\alpha_1, \dots \alpha_K)'$"
    a. "For simplicity, we assume binary latent attributes where $\alpha_k \in \{0,1\}$ for each k, indicating mastery/nonmastery of the kth attribute, but this can be extended to ordinal attributes where different levels of mastery are assumed (e.g., Chen & de la Torre, 2013)."
2. "The attributes can be skills, procedures, and knowledge that are required for solving an item, and the set of attributes assessed by a test is typically identified by domain experts (Tatsuoka, 1990)."
3. "The attributes may be either parallel or hierarchical, where one is the prerequisite to mastering another … In the latter case, the number of permissible attribute patterns can be less than 2^K."
4. "Consider a J-item test. The relationship between items and attributes is defined via a Q-matrix, a J × K incidence matrix indicating the presence (q_{jk} = 1) or absence (q_{jk} = 0) of a connection between each item and each attribute. We say an attribute k is a requisite skill of item j if q_{jk} = 1."

Arithmetic test example

Throughout this report, I will return to the same example to illustrate the concepts introduced wherever possible. The example is a mini "test" consisting of four arithmetic items, adapted from the example presented by Rupp et al. (2010, Chap. 10), which will be analysed within the cognitive diagnostic framework defined above. Success on this test is assumed to depend on mastery of two attributes, addition (α_1) and subtraction (α_2), which are coded on dichotomous scales and not considered to have any hierarchical relationship or dependency. The number of possible attribute profiles (and hence latent classes) is therefore 2^2, and these possible profiles are listed in **Table 1**. The relationship between the four items and two attributes is given by the Q-matrix in **Table 2**; this shows that attribute 1 (addition) is considered a requisite skill for items 1, 3 and 4, and attribute 2 (subtraction) is a requisite skill for items 2, 3 and 4.

Table 1: Possible attribute profiles for example arithmetic test.

Attribute profile	α_1: addition	α_2: subtraction
$\boldsymbol{\alpha}_1$ = [0,0]	0	0
$\boldsymbol{\alpha}_2$ = [0,1]	0	1
$\boldsymbol{\alpha}_3$ = [1,0]	1	0
$\boldsymbol{\alpha}_4$ = [1,1]	1	1

Table 2: Q-matrix for example arithmetic test.

Item	α_1: addition	α_2: subtraction
Item 1: 3 + 5 = ?	1	0
Item 2: 6 – 2 = ?	0	1
Item 3: 2 + 3 – 1 = ?	1	1
Item 4: 9 – 5 + 2 = ?	1	1

CDMs are an example of the probabilistic models built upon the cognitive diagnostic assessment framework. They are examples of restricted latent class models in which both the number of latent classes involved and their interpretations are known in advance (de la Torre & Minchen, 2019). What motivated their development was the need "to account for the stochastic relationship between the theorised attribute profile and the observed responses" (Zhang et al., 2023, p. 656).

Many different CDMs have been developed, and they share the core assumption that "the solution of an item depends on the availability of a set of latent attributes, and for different items different albeit partly overlapping subsets of latent attributes may be required" (von Davier & Lee, 2019b, p. 29). They differ from each other in their complexity, in how or where measurement error is accounted for in the model, and the assumptions they make about how skills combine – that is, "how different skills are combined to affect the correct response probability for an item" (Zhang et al., 2023, p. 656). In models assuming a compensatory relationship between attributes, a test taker's probability of success on an item is generally higher when they possess a higher number of the required attributes. By contrast, when the assumed relationship is conjunctive, the test taker is assumed to need mastery of all required attributes to successfully answer an item. In other words, "it does not matter whether a person does not master any attributes, or if a person has all but one required attribute, the probability of a success is the same (and typically low)" (von Davier & Haberman, 2014, p. 340).

Definitional complications

Different authors have used the term "CDM" to include and exclude slightly different subsets of models. In this report, the default meaning of "Cognitive Diagnostic Models" is that given by George and Robitzsch (2021, p. 107): "a class of multidimensional categorical latent variable models that integrate theoretical assumptions about skills and then estimate the students' possession of these skills." This definition is consistent with the statistical account by Zhang et al. (2023), and usage by major researchers in the field, for example de la Torre and Minchen (2019, p. 155), Sen and Cohen (2021, p. 1), Deonovic et al. (2019, p. 444) and Ravand and Baghaei (2020, p. 25). It also aligns with distinctions that are consistently drawn by researchers writing about cognitive diagnostic assessment, even where definitions are absent. For example, to explain the motivations for research into higher-order CDMs and multidimensional IRT (MIRT), Min et al. (2021) contrast these approaches with (implicitly "ordinary") CDMs.

Unlike the other authors cited in this report, Bradshaw and Levy (2019) draw a distinction between "CDMs" and "DCMs". They describe CDMs informally as models that "categorize examinees according to mastery levels for a set of hypothesized latent skills", before stating that (p. 79) "These classification-based models, collectively termed cognitive diagnosis models (CDMs), can be organized into four major frameworks: rule space methodology (RSM; Tatsuoka, 1983), the attribute hierarchy method (AHM; Leighton et al., 2004), diagnostic classification models (DCMs; e.g., Bradshaw, 2016; Rupp et al., 2010), and Bayesian networks (BNs; e.g., Almond et al., 2015).” Roughly speaking, Bradshaw and Levy (2019) are using the term “CDM” to refer to a subset of what other authors call “cognitive diagnostic assessment” methods, and using the term “DCM” where other authors may use “DCM” or “CDM” (or both). **Figure 1** and **Figure 2** show where the terms “DCM” and “CDM” sit within the overall domain of cognitive diagnostic assessment, for clarity. Bradshaw and Levy’s usage of these terms does not appear to have been adopted more widely.

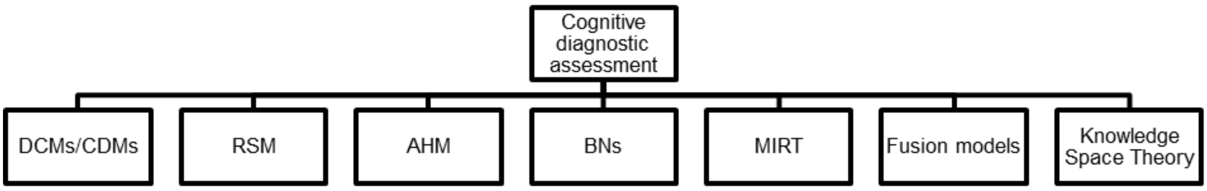


Figure 1: Mainstream terminology usage in cognitive diagnostic assessment.

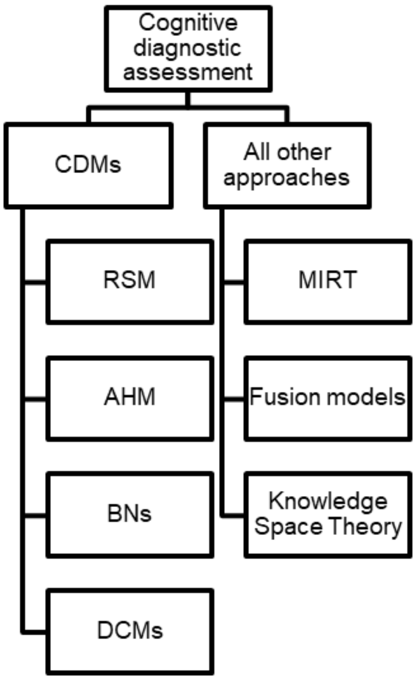


Figure 2: Use of terms "CDM" and "DCM" by Bradshaw and Levy (2019).

Attribute scales

Theoretically, the attributes in a CDM can be coded on any discrete scale. However, presentations and explanations of CDMs often appear to assume a dichotomous scale (i.e., mastered/not mastered), and in practice, attributes are indeed usually coded as dichotomous (Rupp, 2023, pp. 4-5; von Davier & Haberman, 2014). Bradshaw and Levy (2019) note this is true for all four of the modelling approaches they consider under the label “CDM” (DCMs, Bayesian Networks, RSM, and AHM). An obvious practical consideration is that allowing an attribute to have a finer-grain scale rapidly increases the total number of latent classes. For example, assuming two attributes represented by latent variables with 4-point polytomous scales (e.g., 0, 1, 2, 3) instead of dichotomous scales, the initial[3] number of latent classes is 16 rather than 4.

Discrete vs continuous latent variables

Most of the authors referenced so far draw a clear distinction between CDMs (multivariate discrete latent variable models) and other models that can be used for cognitive diagnostic assessment on the basis that CDMs use discrete and not continuous latent variables. For example, George and Robitzsch (2021, p. 108) describe multidimensional IRT (MIRT) as an "Instead of CDMs…" option, while Ravand and Baghaei (2020, p. 25) spell out that "DCMs are notably different from multidimensional IRT models in that the latent variables in DCMs are discrete or categorical". Most emphatically of all, Rupp and Templin (2008b, pp. 226-227) note that their definition of DCM "specifically excludes any multidimensional latent variable model with continuous latent variables as a DCM, which is a deliberate choice."

Writing more recently, however, Rupp (2023, p. 2), seems a little more vague about where the boundaries lie, and states that "It is best to think of DCMs as a family of models that allow analysts to represent different aspects of how item responses can be modeled as a function of the characteristics of learners/respondents and items/response patterns." This more inclusive definition differs from recent common usage in the field (e.g., de la Torre & Minchen, 2019; Deonovic et al., 2019; George & Robitzsch, 2021; Sen & Cohen, 2021) and there is potential for ambiguity. As a result, the broader definition necessitates terms like "core models" or "traditional CDMs" (or simply lists of models), in order to draw the distinctions that authors in the field regularly refer to between CDMs and other cognitive diagnostic assessment approaches.

A major upside of the more inclusive definition is neatly including "models that combine different ideas" (Rupp, 2023, p. 2), in particular, models that include a continuous/scale-score dimension in addition to the discrete latent variables. Examples of these include the Reparameterised Unified Model (RUM) for binary data, which includes a continuous latent variable η "encapsulating the combined influence of any influential skills not built into the specified latent class space" (Stout et al., 2019, p. 50), previously discussed as the "fusion model" (Roussos et al., 2007); and the higher-order DINA model (HO-DINA) in which the (discrete) latent attributes are themselves modelled as "arising from a broadly defined latent trait resembling the θ of item response models" (de la Torre & Douglas, 2004).

Literature review: major developments since 2007

There has been a large amount of relevant research and development activity in cognitive diagnostic modelling since the era of *Cognitive Diagnostic Assessment for Education* (Leighton & Gierl, 2007a). The purpose of this review is to draw attention to major results and trends, rather than exhaustively listing developments.

The topics addressed have not been given equal attention in the research literature. In particular, researchers have written a large amount about the statistical relationship of

[3] As alluded to by Zhang et al. (2023, p. 655), the total number of latent classes may be reduced if structural relationships between attributes (e.g., mastery of attribute A is a prerequisite for mastery of attribute B) mean that the classes defined by certain attribute combinations are not considered possible.

attributes to item responses[4], and comparatively little about the other aspects involved in actually using CDMs to classify students (e.g., attribute pattern identifiability, methods of classification, test length requirement, and Q-matrix specification).

Models and their categorisation

A major area of CDM development in recent years has been categorisations and unifying frameworks that “subsume many (albeit technically still not all) models” (Rupp, 2023). Relatedly, research has also shown that many CDMs are reparameterisations of other models, all of which means that “the traditional distinctions between the DCMs are getting blurred” (Ravand & Baghaei, 2020, p. 29). Since many CDMs are still referred to by name, however (whether or not subsequently subsumed), **Table 7** shows a summary of well-known CDMs and their classifications, for reference.

The examples that Rupp (2023) lists as common CDM “frameworks” (e.g., G-DINA) are described by other authors as belonging to the category of “general” CDMs. Models in the general CDM category allow attributes to combine differently across different items. In particular, they allow attributes to combine in both compensatory and non-compensatory ways for different items within the same test. By contrast, “specific” (also known as “reduced” or “constrained”) CDMs are those models where only one type of relationship between attributes is possible within the assessment (whether that relationship between attributes is disjunctive, conjunctive, or additive). For example, the deterministic-input noisy-“and”-gate (DINA) model specifies a non-compensatory relationship between attributes (indicated by the “and”-gate in the name) for all items. That is, *all* attributes assessed by an item must be mastered in order to expect a correct response. By contrast, the deterministic-input noisy-“or”-gate (DINO) model specifies a compensatory relationship between attributes.

For both DINA and DINO, the “deterministic input” part of the name describes the fact that a test-taker’s attribute profile is deterministic of the latent response vector $\boldsymbol{\eta}_i$ (for DINA, η_{ij} = 1 wherever test-taker i has mastered all the attributes assessed by item j and η_{ij} = 0 otherwise; for DINO, the latent response vector takes the value 1 wherever the test-taker has at least one of the attributes assessed by the item). This vector $\boldsymbol{\eta}_i$ represents an ideal response pattern, and noise is only introduced for DINA and DINO by modelling item-level slip and guess parameters. For DINA, the final probability that test-taker i with the attribute profile $\boldsymbol{\alpha}_c$ answers item j correctly is given by:

$$P\left(X_{ij} = 1 \middle| \boldsymbol{\alpha}_c\right) = g_j^{1-\eta_{ij}}\left(1 - s_j\right)^{\eta_{ij}} \tag{1}$$

Where s_j is the slip parameter giving the probability of an incorrect response despite mastery of the necessary attributes, and g_j is the guess parameter giving the probability of a correct response despite not having mastered the sufficient attributes ($s_j = P\left(X_{ij} = 0 \middle| \eta_{ij} = 1\right)$, and $g_j = P\left(X_{ij} = 1 \middle| \eta_{ij} = 0\right)$).

[4] This is a useful observation from the perspective of practical assessment developers (Dynamic Learning Maps Consortium, 2016, p. 165).

Arithmetic test example

Table 3 shows the ideal response patterns for each attribute profile $\boldsymbol{\alpha}_c$ in the arithmetic test example, using the DINA model, and **Table 4** shows the ideal response patterns under DINO, for comparison.

Table 3: Ideal response patterns for example arithmetic test, DINA model.

Attribute profile	Item 1: 3 + 5 = ?	Item 2: 6 – 2 = ?	Item 3: 2 + 3 – 1 = ?	Item 4: 9 – 5 + 2 = ?
$\boldsymbol{\alpha}_1$ = [0,0]	0	0	0	0
$\boldsymbol{\alpha}_2$ = [0,1]	0	1	0	0
$\boldsymbol{\alpha}_3$ = [1,0]	1	0	0	0
$\boldsymbol{\alpha}_4$ = [1,1]	1	1	1	1

Table 4: Ideal response patterns for example arithmetic test, DINO model.

Attribute profile	Item 1: 3 + 5 = ?	Item 2: 6 – 2 = ?	Item 3: 2 + 3 – 1 = ?	Item 4: 9 – 5 + 2 = ?
$\boldsymbol{\alpha}_1$ = [0,0]	0	0	0	0
$\boldsymbol{\alpha}_2$ = [0,1]	0	1	1	1
$\boldsymbol{\alpha}_3$ = [1,0]	1	0	1	1
$\boldsymbol{\alpha}_4$ = [1,1]	1	1	1	1

After calibrating the mini arithmetic test using a suitable dataset of candidate responses, we would obtain slip and guess parameters for each item. **Table 5** shows a set (based on dummy data) for the DINA model.

Table 5: Slip and guess parameters for example arithmetic test, DINA model.

Item	s_j: slip	g_j: guess
Item 1: 3 + 5 = ?	0.10	0.15
Item 2: 6 – 2 = ?	0.10	0.18
Item 3: 2 + 3 – 1 = ?	0.12	0.13
Item 4: 9 – 5 + 2 = ?	0.15	0.09

The slip and guess parameters are used in equation (1) to calculate the probabilities of correct responses for each item j, for respondents in each attribute class (**Table 6**). For respondents in attribute class c, these probabilities are typically notated π_{jc}.

Table 6: Probabilities of correct responses for example arithmetic test, DINA model.

Attribute profile	Item 1: 3 + 5 = ?	Item 2: 6 – 2 = ?	Item 3: 2 + 3 – 1 = ?	Item 4: 9 – 5 + 2 = ?
$\boldsymbol{\alpha}_1$ = [0,0]	0.15	0.18	0.13	0.09
$\boldsymbol{\alpha}_2$ = [0,1]	0.15	0.90	0.13	0.09
$\boldsymbol{\alpha}_3$ = [1,0]	0.90	0.18	0.13	0.09
$\boldsymbol{\alpha}_4$ = [1,1]	0.90	0.90	0.88	0.85

Many of the specific CDMs shown in **Table 7** can be obtained by specifying constraints to one of the general CDMs. Ravand (2016) argues that if we consider the complexities (particularly, the actual cognitive processes underlying success on varied items) then allowing attributes to combine in different ways seems a far more plausible assumption than insisting on the same relationship throughout an assessment.

Three of the general models shown in **Table 7** have themselves been shown to be reparameterisations of each other (de la Torre, 2011; von Davier, 2014). These are the log-linear CDM (Henson et al., 2009), the general diagnostic model (von Davier, 2005) and the generalised DINA model (de la Torre, 2011). The General Diagnostic Model (GDM) contains both the log-linear diagnostic model (LCDM) and G-DINA with logistic link function (von Davier, 2018, pp. 62-63).

A further useful categorisation of CDMs is into hierarchical and non-hierarchical categories. In hierarchical CDMs, structural relationships among the attributes are modelled: for instance, mastery of attribute A being a pre-requisite for mastery of attribute B (meaning that latent classes in which B is mastered but not A can be ruled out a priori). von Davier and Haberman (2014) strongly criticised this use of the term “hierarchical” as it results in potential confusion between on the one hand CDMs that model structural relationships between attributes (as just defined), such as the “hierarchical DCM” (Templin & Bradshaw, 2014) and on the other hand CDMs such as the “hierarchical GDM” (von Davier, 2007, 2010) that are hierarchical in the sense of multilevel models (i.e., the data is clustered, and the modelling takes account of this hierarchical structure). The usage that von Davier and Haberman (2014) objected to, however, remains standard in CDM contexts.

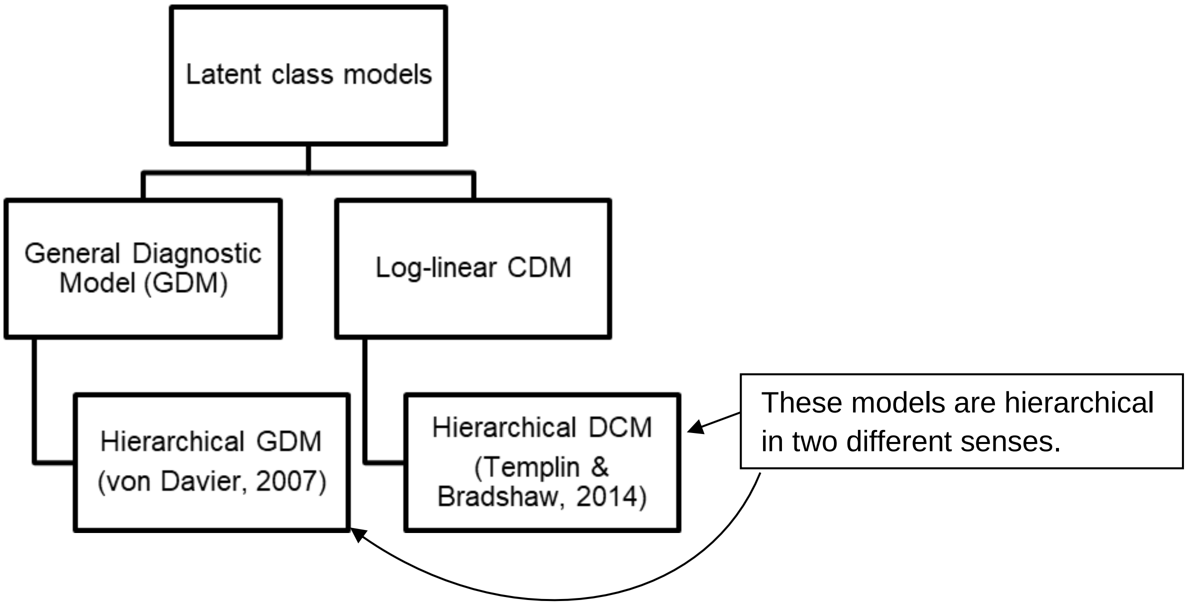

Latent class models
General Diagnostic Model (GDM)
Log-linear CDM
Hierarchical GDM (von Davier, 2007)
Hierarchical DCM (Templin & Bradshaw, 2014)
These models are hierarchical in two different senses.

Figure 3: Relationship of four important CDMs. Each model or model class is a restricted subset of the class above.

Table 7: Categorisation of well-known CDMs, adapted from Ravand and Baghaei (2020, p. 29).

CDM category	Attribute relationship(s)	Abbreviation	Name	Authors	
Specific or constrained	Disjunctive	DINO	Deterministic-input noisy-“or”-gate model	Templin and Henson (2006)	
		NIDO	Noisy input, deterministic-“or”-gate model		
	Conjunctive	DINA	Deterministic-input noisy-“and”-gate model	Junker and Sijtsma (2001)	
		NIDA	Noisy input, deterministic-“and”-gate model	DiBello et al. (1995); Hartz (2002)	
	Additive	ACDM	Additive CDM	de la Torre (2011)	
		LLM	Linear logistic model	Maris (1999)	equivalent to C-RUM
		C-RUM	Compensatory reparameterised unified model	DiBello et al. (1995); Hartz (2002)	equivalent to LLM
		NC-RUM	Noncompensatory reparameterised unified model	Hartz (2002)	
	Hierarchical	HO-DINA	Higher-order DINA	de la Torre and Douglas (2004, 2008)	
General or saturated	Disjunctive, conjunctive and additive	GDM	General diagnostic model	von Davier (2005)	de la Torre (2011) and von Davier (2014) showed these three models are equivalent
		LCDM	Log-linear CDM	Henson et al. (2009)	
		G-DINA	Generalized DINA	de la Torre (2011)	
	Hierarchical	HDCM	Hierarchical diagnostic classification model	Templin and Bradshaw (2013)	Is a constrained version of the LCDM

G-DINA

The G-DINA model (or framework, in Rupp's terms) has been particularly influential in recent years. As the name suggests, the G-DINA model is a generalised version of the deterministic-input noisy-"and"-gate (DINA) model (de la Torre, 2011), for dichotomous latent attributes, in which all interactions between attributes are considered. The G-DINA model is therefore a "saturated" CDM: if we assume that item k requires the first D_k^* attributes (out of a possible set of D attributes measured by the test), then there are $2^{D_k^*}$ latent classes for item k, and G-DINA estimates all $2^{D_k^*}$ parameters, as shown below.

For each latent class ($l = 1, \ldots, 2^{D_k^*}$) we can write the reduced attribute vector $\boldsymbol{a}_{lk}^*$: this is a vector of ones and zeroes with length D_k^* recording the attributes mastered by test-takers in that class. The item response function for the G-DINA model is then the following:

$$g[P(X_k = 1|\boldsymbol{a}_{lk}^*)] = \varphi_{k0} + \sum_{d=1}^{D_k^*} \varphi_{kd} a_{ld} + \sum_{d'=d+1}^{D_k^*} \sum_{d=1}^{D_k^*-1} \varphi_{kdd'} a_{ld} a_{ld'} + \cdots + \varphi_{12\ldots D_k^*} \prod_{d=1}^{D_k^*} a_{ld} \qquad (2)$$

Where g is a link function (identity, log or logit); φ_{k0} is the intercept; φ_{kd} is the main effect from mastering attribute φ_{kd}; and the remaining $\varphi_{k\cdot}$ parameters represent "all possible higher-order interaction effects, ranging from two-way to D_k^*-way" (de la Torre & Minchen, 2019, p. 157)[5].

When g is the identity function, then the probability that test-takers with the attribute profile $\boldsymbol{a}_{lk}^*$ answer item k correctly is just the sum of the effects due to attributes and their interactions. In this scenario, φ_{k0} represents the baseline probability of answering correctly when none of the attributes have been mastered; φ_{kd} is the change in probability of correctly answering item k from having mastered attribute d; and $\varphi_{12\ldots D_k^*}$ is the change in the probability of answering correctly (over and above the main and lower-order interaction effects) due to mastery of all the required attributes (de la Torre, 2011, p. 181).

When the link function g is the logit function, the G-DINA item response function is equivalent to the LCDM (de la Torre & Minchen, 2019, p. 157).

When all parameters except the intercept and highest-order interaction term are set to zero, the result is equivalent to the item response function for DINA:

$$g[P(X_k = 1|\boldsymbol{a}_{lk}^*)] = \varphi_{k0} + \varphi_{12\ldots D_k^*} \prod_{d=1}^{D_k^*} a_{ld} \qquad (3)$$

Similarly, other well-known CDMs can be obtained through constraining the G-DINA function.

[5] The total number of φ parameters in equation (2) is $2^{D_k^*}$, equal to the number of latent classes, because each φ parameter corresponds to a possible combination of the D_k^* dichotomous attributes.

There are also variants or extensions of G-DINA that have been designed to address specific contexts:

- The polytomous G-DINA (pG-DINA) is a version that can handle polytomous attributes. Importantly, pG-DINA reduces the number of latent classes to $2^{D_k^*}$ for item k (i.e., to the same number as in the usual G-DINA for dichotomous variables) by assuming that for each item and attribute, test-takers can be classified as "at or above" or "below" the level of mastery actually required for that item (de la Torre & Minchen, 2019, p. 158).
- The sequential G-DINA (sG-DINA) for responses scored using ordered polytomous categories.
- The continuous G-DINA (cG-DINA) for continuous responses.

Model selection

The consensus from more recent CDM work is that rather than selecting a CDM for an entire assessment, different CDMs should be allowed for different items within the same test (de la Torre & Minchen, 2019; Deonovic et al., 2019; Ravand & Robitzsch, 2018; Shafipoor et al., 2021).

Since it is difficult to determine which model should be chosen for each item in advance, the recommended strategy is to make it an empirical choice. de la Torre (2011) showed that many specific CDMs can be obtained by adding constraints to G-DINA, making the models nested. For this reason, a common route is to fit G-DINA, and then use the Wald test to compare the fit (at item level) with the fit of a number of constrained models, to determine the best CDM structure for each item. This approach is clearly explained by de la Torre and Minchen (2019, pp. 165-166), and an example of its use in an actual assessment development context is given by Deonovic et al. (2019, p. 445), in their presentation of the "Education Companion App" development by ACTNext, the innovation arm of education and assessment organisation ACT.

More practical discussion of model selection, and how this interacts with sample size and context, can be found under research question 2 (requirements for valid reporting).

Methods of classification

As alluded to earlier, the details of classifying test-takers based on a CDM have received comparatively little attention in the literature, relative to parameterising the relationship between attributes and items.

Rupp et al. (2010, p. 233) write that "the full process of estimating a DCM consists of estimating the item parameters, which statistically characterize the measurement properties of the diagnostic assessment, along with estimating the respondent parameters". Rupp et al. clarify that the term "respondent parameters" is used to mean test-takers' "statistically driven classification into one of the distinct latent classes of the DCM" or more simply "attribute profile" – that is, a vector of ones and zeroes indicating which attributes have been mastered (p. 232). At the same time, Rupp et al. write that "the respondent parameter estimates that DCMs provide are the probabilities that a respondent belongs to any of the *C* latent classes in the model" (p. 233). To be completely explicit, if there are 16 different latent classes then "there will be 16 different probabilities of latent class membership for each respondent with the probabilities across all latent classes summing to 1” (p. 233). For the simple arithmetic test example, there are four different latent classes (attribute profiles), and so there will be four different probabilities of class membership for each test-taker (e.g., **Table 8**).

Table 8: Probabilities of latent class membership for test-takers of example arithmetic test.

Attribute profile	Ali	Benni	Cami
$\boldsymbol{\alpha}_1$ = [0,0]	0.03	0.26	0.91
$\boldsymbol{\alpha}_2$ = [0,1]	0.07	0.73	0.06
$\boldsymbol{\alpha}_3$ = [1,0]	0.04	0.01	0.03
$\boldsymbol{\alpha}_4$ = [1,1]	0.86	0.00	0.00
	1.00	1.00	1.00

In practice, authors at times present both final latent class classifications and the probabilities underlying the classification as “the output” of a CDM. Reporting may consist of a single classification (into a latent class with a particular attribute profile, e.g., “Ali is classified as having attribute profile $\boldsymbol{\alpha}_4$, mastery of both addition and subtraction”), or the probabilities associated with multiple latent classes, or the probabilities of mastery of individual attributes.

Huebner and Wang (2011) give a particularly clear account of classification in CDMs. They write that test-takers are “often” classified either via maximum likelihood estimation (MLE), maximum a posteriori (MAP), or expected a posteriori (EAP) estimates, and the examples reviewed for this report consistently used one of these three methods (or a close variant). The MLE and MAP methods are parallel to those in IRT, but it is worth spelling out exactly what is meant in CDM contexts.

The MLE classification is the latent class that maximises the likelihood of the test-taker’s observed responses (Huebner & Wang, 2011, p. 410). This involves:

- calculating the likelihood of the observed responses $\boldsymbol{X}_i$ given each possible latent class $\boldsymbol{\alpha}_c$ (i.e., each possible vector of ones and zeroes indicating mastery or non-mastery of the K attributes),
- assigning the test-taker i the class $\widehat{\boldsymbol{\alpha}}_{MLE}$ that maximises the likelihood,
- more formally, we write: $\widehat{\boldsymbol{\alpha}}_{MLE} = \underset{c}{\operatorname{argmax}}\{L(\boldsymbol{X}_i|\boldsymbol{\alpha}_c)\}$.

The MAP classification is the latent class that maximises the posterior probability $P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i)$ (p. 410). This Bayesian approach requires:

- prior probabilities (e.g., proportions of test-takers with mastery of each attribute estimated from an earlier test administration),
- using Bayes’s theorem, calculating the posterior probability $P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i)$ for each class $\boldsymbol{\alpha}_c$,
- assigning test-taker i the latent class that maximises this posterior probability,
- more formally, $\widehat{\boldsymbol{\alpha}}_{MAP} = \underset{c}{\operatorname{argmax}}\{P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i)\}$.

Since the MLE estimates are equivalent to MAP estimates with flat priors (as usual), Huebner and Wang (2011) reserve the term MAP for situations where the prior probabilities $P(\boldsymbol{\alpha}_c)$ are non-equal for at least two values of c.

The EAP classification is generated slightly differently, and described as "in a sense" averaging over the posterior probabilities (Huebner & Wang, 2011, p. 411; Rupp et al., 2010, p. 239). The steps taken are:

- calculating the posterior probabilities $P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i)$ as in the MAP approach,
- aggregating the posterior probabilities to find the marginal probability $\tilde{\alpha}_k$ for each attribute $k = 1, \dots, K$ – this is the sum of the posterior probabilities corresponding to mastery of attribute k,
- more formally, $\tilde{\alpha}_k = \sum_{c=1}^{C} P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i)\, I\big(\alpha_{c,k} = 1\big)$ where $I\big(\alpha_{c,k} = 1\big)$ is an indicator function with value 1 if element k of the c^{th} pattern is equal to 1, and zero otherwise,
- from each probability $\tilde{\alpha}_k$, a binary classification of mastery or non-mastery for attribute k is obtained, most commonly by rounding at 0.5.

Arithmetic test example

The (hypothetical) student Ali answered the first three arithmetic items correctly, but not the fourth, so we have response data $\boldsymbol{x}_A = (1,1,1,0)$. For all classification methods, it is necessary to first calculate the likelihood of the observed responses $\boldsymbol{x}_A$ for each of the four possible attribute profiles, that is, $P(\boldsymbol{X}_A = \boldsymbol{x}_A|\boldsymbol{\alpha}_c)$ for all classes c. These probabilities are calculated using the following[6]:

$$P(\boldsymbol{X}_A = \boldsymbol{x}_A|\boldsymbol{\alpha}_c) = \prod_{j=1}^{4} \pi_{jc}^{x_{jA}} \big(1 - \pi_{jc}\big)^{1-x_{jA}} \tag{4}$$

Where π_{jc} is the probability of a correct response to item j for a respondent with attribute profile $\boldsymbol{\alpha}_c$ (these probabilities were calculated and shown in **Table 6**), and x_{jA} is the score Ali achieved on item j. The results are shown in the final column of **Table 9**, and the MLE estimate of Ali's classification can be taken directly from this column; it is latent class or attribute profile $\boldsymbol{\alpha}_4$, since this has the highest likelihood value.

Table 9: Likelihood of Ali's responses, arithmetic test example.

	Ideal response patterns				Observed responses from Ali				$P(\boldsymbol{x}_A|\boldsymbol{\alpha}_c)$
					1	1	1	0	
					Likelihood of observed responses				
Attribute profile	Item 1	Item 2	Item 3	Item 4	Item 1	Item 2	Item 3	Item 4	Product
$\boldsymbol{\alpha}_1$ = [0,0]	0	0	0	0	0.15	0.18	0.13	0.91	0.003
$\boldsymbol{\alpha}_2$ = [0,1]	0	1	0	0	0.15	0.90	0.13	0.91	0.016
$\boldsymbol{\alpha}_3$ = [1,0]	1	0	0	0	0.90	0.18	0.13	0.91	0.019
$\boldsymbol{\alpha}_4$ = [1,1]	1	1	1	1	0.90	0.90	0.88	0.15	0.107

[6] See Rupp et al. (2010, pp. 234-235). For a parallel extended example demonstrating respondent classification using the LCDM (instead of DINA), see Rupp et al. (2010, pp. 235-240).

Ali's MAP classification is the latent class that maximises the posterior probability $P(\boldsymbol{\alpha}_c|\boldsymbol{x}_A)$. The posterior probability for each class is calculated using Bayes' theorem:

$$P(\boldsymbol{\alpha}_c|\boldsymbol{x}_A) = \frac{P(\boldsymbol{x}_A|\boldsymbol{\alpha}_c)P(\boldsymbol{\alpha}_c)}{P(\boldsymbol{x}_A)} \tag{5}$$

This requires prior probabilities of latent class membership ($P(\boldsymbol{\alpha}_c)$). For this example, the prior probabilities are assumed to be those in **Table 10**:

Table 10: Prior probabilities, arithmetic test example.

Attribute profile	Prior probabilities of class membership, $P(\boldsymbol{\alpha}_c)$
$\boldsymbol{\alpha}_1$ = [0,0]	0.35
$\boldsymbol{\alpha}_2$ = [0,1]	0.20
$\boldsymbol{\alpha}_3$ = [1,0]	0.10
$\boldsymbol{\alpha}_4$ = [1,1]	0.35

The denominator of equation 5 is the total probability of observing the responses $\boldsymbol{x}_A$ (considering all four possible attribute classes), which is calculated using equation 6:

$$P(\boldsymbol{X}_A = \boldsymbol{x}_A) = \sum_{l=1}^{4} P(\boldsymbol{\alpha}_l) \prod_{j=1}^{4} \pi_{jl}^{x_{jA}} \left(1 - \pi_{jl}\right)^{1-x_{jA}} \tag{6}$$

Combining equations 4, 5 and 6 gives the following final equation for the posterior probabilities:

$$P(\boldsymbol{\alpha}_c|\boldsymbol{x}_A) = \frac{P(\boldsymbol{x}_A|\boldsymbol{\alpha}_c)P(\boldsymbol{\alpha}_c)}{P(\boldsymbol{x}_A)} = \frac{P(\boldsymbol{\alpha}_c)\prod_{j=1}^{4} \pi_{jc}^{x_{jA}} \left(1 - \pi_{jc}\right)^{1-x_{jA}}}{\sum_{l=1}^{4} P(\boldsymbol{\alpha}_l)\prod_{j=1}^{4} \pi_{jl}^{x_{jA}} \left(1 - \pi_{jl}\right)^{1-x_{jA}}} \tag{7}$$

The calculations for Ali are shown in **Table 11**: the numerator of the fraction is in the penultimate column, and the denominator of the fraction is the sum (0.044). The MAP classification is the latent class that maximises the posterior probability $P(\boldsymbol{\alpha}_c|\boldsymbol{x}_A)$, so in this case it is the class corresponding to attribute profile $\boldsymbol{\alpha}_4$, with posterior probability 0.86.

Table 11: Calculation of posterior probabilities, arithmetic test example.

	Observed responses				Product of response likelihoods	Prior probabilities	Product of likelihoods with $P(\boldsymbol{\alpha}_c)$	Posterior probabilities
	1	1	1	0				
	Likelihood of observed responses							
Attribute profile	Item 1	Item 2	Item 3	Item 4	$P(\boldsymbol{x}_A\|\boldsymbol{\alpha}_c)$	$P(\boldsymbol{\alpha}_c)$	$P(\boldsymbol{x}_A\|\boldsymbol{\alpha}_c)P(\boldsymbol{\alpha}_c)$	$P(\boldsymbol{\alpha}_c\|\boldsymbol{x}_A)$
$\boldsymbol{\alpha}_1$ = [0,0]	0.15	0.18	0.13	0.91	0.003	0.35	0.001	0.03
$\boldsymbol{\alpha}_2$ = [0,1]	0.15	0.90	0.13	0.91	0.016	0.20	0.003	0.07
$\boldsymbol{\alpha}_3$ = [1,0]	0.90	0.18	0.13	0.91	0.019	0.10	0.002	0.04
$\boldsymbol{\alpha}_4$ = [1,1]	0.90	0.90	0.88	0.15	0.107	0.35	0.037	0.86
					Total	1.00	0.044	1.00

Ali's EAP classification is found by aggregating the posterior probabilities corresponding to mastery of each individual attribute. In this example, $\tilde{\alpha}_1 = P(\boldsymbol{\alpha}_3|\boldsymbol{x}_A) + P(\boldsymbol{\alpha}_4|\boldsymbol{x}_A)$ since $\boldsymbol{\alpha}_3$ and $\boldsymbol{\alpha}_4$ are the two attribute profiles where addition is mastered, and $\tilde{\alpha}_2 = P(\boldsymbol{\alpha}_2|\boldsymbol{x}_A) + P(\boldsymbol{\alpha}_4|\boldsymbol{x}_A)$ since $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_4$ are the two attribute profiles corresponding to mastery of subtraction. Hence $\tilde{\alpha}_1$ = 0.04 + 0.86 = 0.90 and $\tilde{\alpha}_2$ = 0.07 + 0.86 = 0.93. These values are rounded at 0.5 and the estimated binary classification is 1 for both attributes. That is, we estimate that Ali has mastered both addition and subtraction.

Comparing classification methods

Rupp et al. (2010, p. 239) point out that "MAP estimates can sometimes be hard to interpret, because they do not provide direct probability estimates for each attribute separately." Perhaps obviously, they also note that "if probabilities for individual attributes for individual respondents are desired" then EAP estimates will be more useful. Generating EAP estimates does however require an additional decision, specifically "which cutoff value should be used for deciding on the mastery status of an attribute to arrive at an overall classification of a respondent into a particular latent class" (p. 241). The cut-off generally chosen is 0.5 (i.e., classification as non-master for marginal probability values below 0.5 and master for values of at least 0.5), as reported by Huebner and Wang (2011), which is the statistically optimal[7] cut-off (Bradshaw & Levy, 2019, p. 86). EAP classifications can however be generated based on other values, which may be more appropriate when the costs of over- and under- misclassification are not equal. In particular, "in a formative assessment context, it may be more costly to misclassify a non master as a master" (p. 86) since a student lacking mastery of a skill may miss out on necessary support. Following this logic, EAP classifications of mastery in Dynamic Learning Maps assessments are based on a marginal probability of at least 0.8 (Clark et al., 2017, p. 6; Dynamic Learning Maps Consortium, 2016, p. 170)[8].

Huebner and Wang (2011) carried out a simulation study to compare the classification accuracy of MLE/MAP and EAP when modelling using DINA. Besides the classification method, they varied the number of attributes measured, discrimination of individual items, and attribute mastery distribution in the test-taker population. Their results showed consistent results across all conditions (p. 417):

- MLE/MAP classification resulted in the highest number of test-takers classified completely correctly (i.e., correct mastery/non-mastery recorded for all K attributes)
- EAP classification resulted in the highest total number of attributes classified correctly and fewer severe misclassifications.

Huebner and Wang (2011) conclude that the choice of method should be based on the purpose of the assessment (and classification), and that EAP seems "very suitable for proposed uses of CDMs in a practical setting" (p. 417). Their argument, which seems reasonable, is that "it may be desirable to classify ''most'' students ''mostly'' correctly rather

[7] In the sense that when we use 0.5 as the cut-off, this "assigns each examinee to their most likely class and will yield the fewest total errors in classifying examinees" (Bradshaw & Levy, 2019, p. 86).
[8] The DLM classifications in fact allocate mastery for marginal probability of at least 0.8, or, if the student has answered 80% of all items assessing the relevant attribute correctly (Dynamic Learning Maps Consortium, 2016, p. 170).

than to classify a maximum number of students exactly correctly while having a higher number of severe misclassifications" (p. 418).

This review did not find further empirical comparisons of classification methods. More recent studies have commented that Bayesian methods "offer a natural framework" (Culpepper & Hudson, 2018, p. 100) and that classification is "generally done within a Bayesian framework" (Maas et al., 2022, p. 5). In terms of deciding between MAP and EAP, Maas et al. appeal directly to Huebner and Wang's (2011) simulation results and arguments.

Metrics for evaluation

CDM implementations can be can be evaluated from different perspectives, depending on the weight given to model fit, classification consistency and accuracy, item discrimination, and the extent to which attribute difficulty aligns with substantive expectations (Ravand & Baghaei, 2020, p. 40). Rupp (2023, p. 10) lists the different categories of CDM statistics that can usually be obtained in fitting CDMs, and that can help evaluation (from various perspectives):

1. Estimates of item difficulty and discrimination parameters.
2. Estimates of differential item functioning statistics for particular subgroups.
3. Estimates of attribute mastery probabilities for individual learners.
4. Estimates of latent class membership probabilities for individual learners.
5. Estimates of attribute mastery probabilities for a sample or subgroup.
6. A distribution of learners in a sample across latent classes.
7. Estimates of parameters for the effects of learner or item characteristics.
8. Estimates of classification consistency for each latent class and attribute.
9. Estimates of item-level, absolute, and relative fit of models.

Measures of classification accuracy and classification consistency are the most commonly found examples of reliability-type measures. Sinharay and Johnson (2019) discuss these in some detail, and note that CDMs lacked these measures for a fairly long time. Sinharay and Johnson (2019) focus particularly on the following:

- Measures of accuracy (P_a) and consistency (P_c) for an entire vector of attributes, as proposed by Cui et al. (2012). P_a is defined very simply as "the probability of accurately classifying a randomly selected student based on his or her responses to test items" (p. 24), and P_c as "the probability of classifying a randomly selected student consistently on two administrations or forms of the test" (p. 24). Cui et al. stress that it is vital to establish whether a CDM adequately fits observed student data before attempting to calculate P_a or P_c, because their computations "rely heavily" on however student responses are modelled by the CDM (p. 24).
- Measures for individual attributes, as proposed by Templin and Bradshaw (2013), Wang et al. (2015), Johnson and Sinharay (2018).

Methods based on classification accuracy can have unsatisfactory results. Sinharay and Johnson (2019, p. 369) give as an example the case where 90% of the population is known to possess an attribute, and the CDM fitted "provides absolutely no information about the attribute" – but will still be accurate for 90% of the population. Alternative methods were proposed by Johnson and Sinharay (2018) include:

- Youden's statistic

- Goodman & Kruskal's lambda
- Cohen's kappa
- Tetrachoric correlation
- Sensitivity (true positive rate) and specificity (true negative rate).

Sinharay and Johnson (2019, p. 371) also mention the cdm.est.class.accuracy function in the R package CDM (George et al., 2016; Robitzsch et al., 2014). However, they note that at the time of writing there was no peer-reviewed publication supporting the methods that this package implements and that the function gave "substantially" different results from the indices proposed by Wang et al. (2015) and Johnson and Sinharay (2018).

In simulation studies, the key metrics for CDM evaluation are item parameter recovery (i.e., successful retrieval of parameters used in the simulation of the data), and correct classification rate (i.e., test-takers assigned to the same class assumed in the simulation of their responses).

Q-matrix development and refinement

The Q-matrix is a two-dimensional matrix that records which combination of attributes is assessed by each item. In general terms, it specifies "the dependence structure between the observed variables and the latent attributes" (Ma et al., 2023, p. 175), and in educational measurement contexts, can be thought of as "a hypothesis about the required skills for getting each item right" (Ravand & Baghaei, 2020, p. 25).

Traditionally, the Q-matrix was developed by domain experts, but the process is "quite challenging" (Wang et al., 2023, p. 11). Wang and colleagues point out that the attributes required in the fractions subtraction dataset introduced by Tatsuoka (2002) have "been debated for two decades, and no definitive conclusion has been agreed upon so far" (Wang et al., 2023, p. 11).

Q-matrix mis-specification has long been understood to be a point of vulnerability for CDMs, with the potential for not just inefficient modelling but actively misleading results (de la Torre, 2008; Rupp & Templin, 2008a). An important development in recent years has been the use of data-driven methods to validate or even construct the Q-matrix. de la Torre and Chiu (2016) presented a general method suitable for any specific model subsumed within the G-DINA family. The method uses a measure called the G-DINA model discrimination index (GDI), which can be calculated for any q-vector proposed for item k. The GDI is equal to the variance (across attribute patterns) of the success probabilities for item k, given that q-vector. A given q-vector is considered "appropriate" if it maximises the GDI, and among the appropriate q-vectors[9], the one considered "correct" is the q-vector that specifies the lowest number of attributes for item k (de la Torre & Minchen, 2019, pp. 163-164). de la Torre and Chiu (2016) also promoted the use of mesa plots to visualise the GDIs of multiple Q-matrix structures, to avoid reliance on a single GDI cut-off. **Figure 4** shows an example of a mesa plot for item 1 (3 + 5 = ?) from the simple arithmetic test example. The y-axis plots the GDI

[9] Multiple q-vectors may result in the same maximal GDI – there is no expectation that the maximum level of variance of success probabilities (across attribute patterns) should be achieved uniquely by one q-vector.

or PVAF[10] values for multiple q-vectors being considered for the item, in ascending order of GDI. The recommendation is to choose a q-vector from the region where the plot plateaus, that is, at the edge of the "mesa". In the case of **Figure 4**, there is only one such q-vector, so the recommended choice would be [1,0], corresponding to mastery of single-digit arithmetic but non-mastery of single-digit subtraction. The mesa plot shows that the addition attribute contributed most of the variance of item success probabilities, and that mastery of subtraction did not contribute much. The mesa plot strongly suggests the choice of the q-vector [1,0] for item 1, whereas a single cut-off value (e.g., 0.95, as shown by the dotted line) may have led to the choice of q-vector [1,1]. Deonovic et al. (2019) demonstrate how empirical Q-matrix validation was used in development of ACTNext's Education Companion App, using a method based on de la Torre and Chiu (2016).

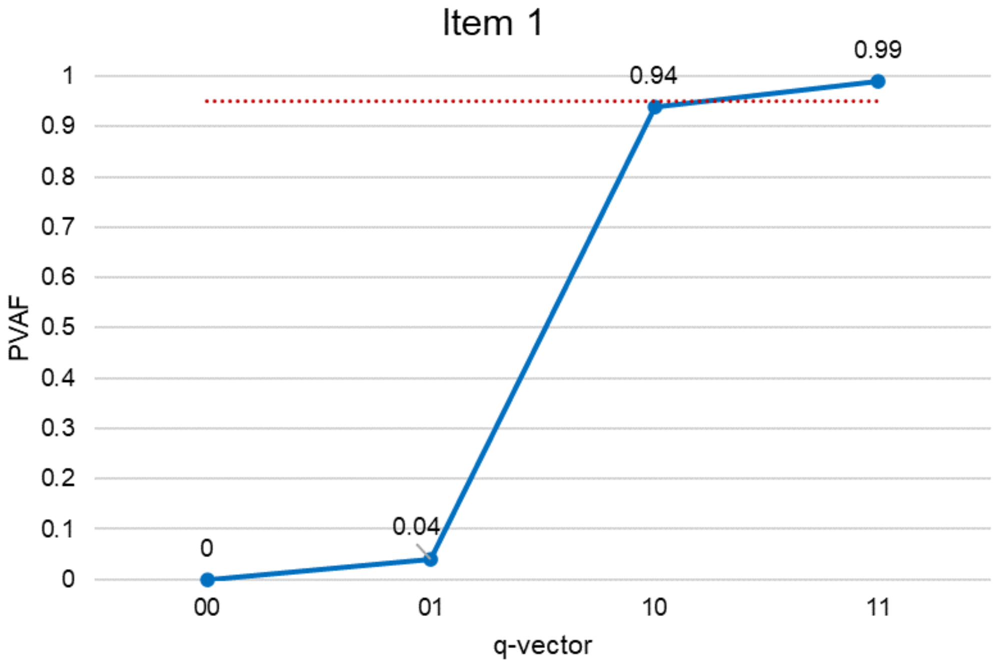


Figure 4: Mesa plot for item 1, arithmetic test example.

Using data-driven methods to actually learn or construct (not just validate or refine) the Q-matrix, has been demonstrated by multiple authors, including Liu and Kang (2019).

Wang et al. (2023) demonstrated a data-driven method for constructing a Q matrix for cognitive diagnostic assessment using a Bayesian network (BN – see p.29 onwards). This method depends on statistical independence testing of the BN structure, and for this reason, Wang and colleagues caution against relying "solely" on their data driven approach due to dependence on p-values as evidence. Their recommendation is that "the opinions of the domain experts are needed" (Wang et al., 2023, p. 11).

[10] Proportion of variance accounted for (PVAF) is defined as GDI as a proportion of the item's maximum GDI; expressing the GDI in this way can be more useful when comparing q-vectors.

CDM applications

CDMs can be applied for different purposes. The application generally considered the “true” purpose of CDMs is the development of new tests for diagnostic purposes, where the aim is to gather fine-grained information about the strengths and weaknesses of the test takers, in order to support their further development – that is, with the goal of informing action (Ravand & Baghaei, 2020, p. 27). CDMs may also be used to extract diagnostic information from existing assessments (not necessarily designed to be cognitive diagnostic assessments at all), a process called “retrofitting”. Investigation of the structure of educational constructs can take place within both “true” CDM studies and retrofitting studies. CDMs are generally regarded as confirmatory, but Mislevy and Bolsinova (2021, p. 103) note that CDMs can be used in both confirmatory and exploratory psychometric modelling when the Q-matrix is estimated from data.

Something that does not appear to have developed over recent years is widespread use of CDMs in assessment practice. Bradshaw and Levy stated in 2019 that CDMs were “only recently being used in operational settings to provide results to examinees and other stakeholders” (Bradshaw & Levy, 2019, p. 79). Ravand and Baghaei, similarly, noted around the same time that “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” (2020, p. 26). In Ravand and Baghaei’s view, the stark imbalance in the academic literature on CDMs[11] reflects the trends in their usage. That is, CDMs are far more likely to be used in methodological research than for actual diagnostic assessment. Further, despite the serious objections to retrofitting (described as “the-measure-of-last-resort” (p. 27); see also Bradshaw et al. (2014); Rupp and Templin (2009)), examples of retrofitting outnumbered “true” CDM applications four to one. More recently, in a 2023 NCME webinar on CDMs, Matthew Madison appeared to agree that examples were still sparse: “DCMs are still new, so we don't have many [examples]” (Madison, 2023).

Zhang et al. (2023) note that besides cognitive diagnostic assessments and retrofitting studies, cognitive diagnostic modelling can be integrated into a number of products or technologies in digital learning settings:

1. Cognitive diagnostic computerized adaptive testing.
2. Longitudinal models for learning, e.g., to track learning over a course.
3. Recommendation systems for adaptive learning, see for example Chen et al. (2018).

Commercial assessment products using CDMs at scale

1. Navvy assessments – created by Dr Laine Bradshaw, then acquired by Pearson in 2022. The Navvy assessments are pre-set formative assessments, designed to provide reliable diagnostic information on mastery of individual learning objectives (“standards”). The assessments are short (6-8 items per standard) and designed to

[11] “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.” (Ravand & Baghaei, 2020, p. 27)

be convenient for regular classroom use, so that a teacher can regularly check and diagnose understanding. The key facts are outlined in a white paper by Bradshaw (2022). The publicly available documentation on Navvy assessments does not specify what form of cognitive diagnostic modelling they use, but in earlier projects Bradshaw has applied the LCDM (Bradshaw et al., 2014; Madison & Bradshaw, 2018).

2. Dynamic Learning Maps (DLM) assessments – developed by Dr Neal Kingston and colleagues at the ATLAS centre, University of Kansas. These assessments are again "true" CDM applications: the DLM assessments measure what students know and can do in terms of specific elements of knowledge, skills and understanding, which are linked to college and career readiness standards. The system is designed for students with serious cognitive difficulties, and the assessments are delivered in short "testlets" consisting of "an unscored engagement activity and three to nine items" (Karvonen et al., 2020). Extensive details are available in Clark et al. (2017) and the DLM technical manuals (e.g., Dynamic Learning Maps Consortium, 2016).

Other assessment development examples

1. The Education Companion App (ECA) developed by ACTNext, as reported by Deonovic et al. (2019). The goal of the ECA app development was to make use of the vast amount of learner data held by ACT across tests and platforms, for example to recommend learning resources for a student to practice a skill they had not yet mastered. Student proficiency was modelled (from existing ACT test data) using the Linear Logistic Test Model (LLTM), an extension of the Rasch model incorporating a Q-matrix and skill easiness parameters (in addition to unidimensional student ability θ and item difficulties). The Q matrices and LLTM approach was then validated using an "intensive analysis of the data using the standard CDM approach" (p. 449). Further details can be found in Appendix B.
2. Wang et al. (2023) developed a cognitive diagnostic assessment for the concept of buoyancy in physics. As noted earlier, they demonstrated a data-driven approach for developing and improving the Q-matrix via statistical independence testing of Bayesian networks (BNs). They then compared classification using the original BN, a 3-level hierarchical BN, G-DINA and the HDCM.

Using CDMs to investigate the structure of constructs

1. Comparing the fit of G-DINA with specific CDMs on a specially designed cognitive assessment in English as a second language (Shafipoor et al., 2021). The study concluded that "relationships among the attributes of grammar and vocabulary are not 'either-or' compensatory or noncompensatory but a combination of both" (p. 1), and that model choice at item-level rather than test-level is therefore preferable.
2. Retrofitting different implementations of the DINA model to PIRLS data to empirically validate assumptions about reading (George & Robitzsch, 2021).
3. Retrofitting G-DINA to TIMSS data in order to provide more informative feedback (Delafontaine et al., 2022). The results showed that using country-specific Q-matrices led to better fit compared to a single (expert-designed) universal Q matrix.
4. Retrofitting CDMs to PISA data in order to investigate trajectories of learning in statistics, across 14 countries (Jia et al., 2021). Six models were compared: DINO, DINA, GDINA, ACDM, LLM, RRUM and LCDM.

5. A similar retrofitting study investigating mathematics learning across 10 countries, again using PISA data (Wu et al., 2020).
6. Retrofitting CDMs to investigate the structure of construct measured by the Singapore/Cambridge O Level English listening test (Aryadoust, 2018). The study fitted five different CDMs: DINA, DINO, HO-DINA, G-DINA and RRUM, with a notably small sample size (n=205). Q-matrix development was informed by the theoretical framework of the study (i.e., expert domain knowledge), a think-aloud study with participating students, and an eye-tracking study.
7. Retrofitting G-DINA to IELTS data to investigate the reading construct (Mirzaei et al., 2020).

Relevant non-CDM topics

Besides research and development in CDMs, there have been significant developments in recent years in other areas with high relevance to CDMs. In particular, it is worth outlining some non-parametric approaches to diagnostic classification, and some important connections between CDMs and network models.

Nonparametric approaches

CDMs – the mainstream approach to diagnostic classification – require estimation of model parameters. Nonparametric methods for cognitive diagnosis were developed originally by researchers seeking to circumvent technical difficulties in parameter estimation: Chiu and Köhn (2019) note that in the early days of CDMs, publicly available implementations of estimation algorithms were “scarce” and computationally costly (and sometimes not actually feasible). The main relevance of nonparametric approaches today, however, lies in their suitability for small sample sizes.

Nonparametric approaches can broadly be grouped into clustering and classification methods. For the context of cognitive diagnosis, clustering approaches aggregate each test-taker’s responses into sum scores for each attribute (using the Q-matrix), and use these to identify clusters via a clustering algorithm such as K-means. The obvious drawback of clustering for cognitive diagnosis is that clustering is not a method of classification; it “identifies groups that do not come with inherent interpretations” (Zhang et al., 2023, p. 661). The clusters need to be labelled before they can be of use; specifically, “their underlying attribute vectors must be reconstructed from the chosen input data” (Chiu & Köhn, 2019, p. 117). While methods to do this have been developed, they are not without risks or complications (pp. 117-118).

True nonparametric classification methods, unlike clustering approaches, aim (like CDMs) to classify test-takers into predefined classes based on their mastery of attributes. The nonparametric classification (NPC) method developed by Chiu and Douglas (2013) is an important example, which can be used with samples of any size whatsoever. The core idea of the NPC is simple: to classify a test-taker to whatever class minimises the distance between the test-taker’s observed response vector $\boldsymbol{x}_i$ and the ideal response vector $\boldsymbol{\eta}_c$ for that class. More formally,

$$\hat{\boldsymbol{\alpha}}_i = \arg \min_{c \in \{1, \dots C\}} d(\boldsymbol{x}_i, \boldsymbol{\eta}_c) \tag{8}$$

The usual measure of distance d is the Hamming distance, which counts the number of disagreements between the observed and ideal response vectors. Another option is to weight the Hamming distance according to the inverse of the item's response variability (so that a disagreement 'counts' more where the item variance was low). An advantage of the weighted approach is that it reduces the number of ties.

Table 12 shows how the NPC approach would be applied to Ali's responses from the simple arithmetic test example. In this case, there are no ties, and the NPC approach would classify Ali as having attribute profile $\boldsymbol{\alpha}_4$, that is, mastery of both addition and subtraction.

Table 12: Observed and ideal response patterns, arithmetic test example.

	Observed response pattern from Ali				Number of disagreements between observed and ideal
	1	1	1	0	
	Ideal response patterns (DINA)				
Attribute profile	Item 1	Item 2	Item 3	Item 4	
$\boldsymbol{\alpha}_1$ = [0,0]	0	0	0	0	3
$\boldsymbol{\alpha}_2$ = [0,1]	0	1	0	0	2
$\boldsymbol{\alpha}_3$ = [1,0]	1	0	0	0	2
$\boldsymbol{\alpha}_4$ = [1,1]	1	1	1	1	1

The ideal response vector $\boldsymbol{\eta}_c$ is straightforward to define for the entirely conjunctive DINA model, and the entirely disjunctive DINO. Wang and Douglas (2015) showed that the NPC approach resulted in consistent estimation for several other conjunctive models, while Chiu et al. (2018) introduced a generalized NPC (GNPC) approach suitable for any data-generating model. The GNPC achieves this by creating an ideal response vector that is a weighted combination of the purely conjunctive and disjunctive ideal response vectors. The weights are estimated from the data after first using some initial estimate of test-taker classification; Zhang et al. (2023, p. 662) suggest classification based on using the entirely conjunctive ideal response vector for this purpose.

Chiu and Douglas (2013) compared classification via parametric and nonparametric approaches, for a variety of conditions, using both simulated and actual test-taker data (the frequently analysed fractions dataset from Tatsuoka (2002)). Overall, their results showed that in conditions where it was possible to estimate the parametric model and the data-generating model was known, the model-based classification was more efficient and more reliable. However, the nonparametric approach was and is viable for a much wider range of conditions. For the fractions dataset specifically (N=536, with 20 items and 8 parameters), Chiu and Douglas (2013) compared maximum-likelihood estimation using parameters obtained from HO-DINA, with nonparametric classification using the weighted Hamming distance. The results showed that around 45% of test-takers were classified into the same class under both methods, but 87% of individual attributes of test-takers were classified the same way (p. 247).

A separate nonparametric approach to mention is the use of artificial neural networks (ANNs), which can be applied for the purposes of cognitive diagnosis in both supervised and unsupervised forms (Cui et al., 2016; Paulsen & Valdivia, 2021). Cui et al. explain that the most commonly used supervised neural network model, the multilayer perceptron (MLP), functions in statistical terms like a non-linear multivariate regression model (Cui et al., 2016, pp. 1066-1067). The MLP requires an observed input layer (akin to independent variables), and an observed output layer (akin to dependent variables). In between is a hidden layer of network nodes, and the ANN is “trained” by repeatedly connecting the input and output layers until it can predict output layer values with the desired level of precision[12]. For the purposes of nonparametric cognitive diagnosis, the different possible ideal response vectors are set as the input layer, and the output layer consists of attribute profiles, meaning that an ANN can be estimated without any observed data at all (Paulsen & Valdivia, 2021, p. 921). This is a useful feature when working with small samples, but a downside is that the ANN does not attempt to account for any deviations from ideal response vectors. Consequently, this approach does not tend to perform well when actual response data deviates frequently from ideal response behaviour (Cui et al., 2016, p. 1080).

Network psychometrics

The field of network psychometrics takes a fundamentally different approach to studying and understanding observed phenomena in psychology, in comparison with a traditional latent variable or common-cause framework (Marsman et al., 2018). Specifically, “Instead of attributing the relationships between observed variables as arising from the underlying latent variable related to all the observed variables, the set of observed variables is seen as a network with causal interactions among these variables” (Mislevy & Bolsinova, 2021, p. 103). In practice, this is achieved by using network models that make use of (mathematical) graph structures[13]: these network models “view observed variables as nodes and the strength of conditional association between two variables after controlling for all other variables as edges” (p. 103). In network psychometrics, the motivation behind such modelling is “an attempt to map out the complex interplay between psychological, biological, sociological, and other components” (p. 103). Marsman et al. (2018) illustrate this with an example from psychopathology, showing the modelling of a set of symptoms (e.g., irritation, sleep problems, depressed mood) associated with major depression (MD) and generalized anxiety disorder (GAD). Modelling these symptoms under a traditional common cause model posits that these symptoms develop as a consequence of underlying illness represented by the latent variables MD and GAD). By contrast, “the direct interaction model suggests that a symptom develops under the influence of other symptoms or (observable) external factors” (p. 18).

One important category of network models is pairwise Markov random field (MRF) models. These include Gaussian graphical models (GGMs), also known as partial correlation networks, which are the most commonly used network models in network psychometrics, and Ising models, which are used to estimate pairwise interactions between binary variables (Briganti et al., 2022; Marsman et al., 2018). The graph structures of pairwise MRF models

[12] Cui et al. provide much more detail on ANNs and the iterative process of training. To be more precise about the goal of training, “To predict the values of the output nodes for the MLP, the unknown parameters that must be estimated are two sets of connections weights, one linking input nodes to hidden nodes and the other linking hidden nodes to output nodes.” (Cui et al., 2016, p. 1068)

[13] A mathematical graph is a finite set of elements (often called nodes) accompanied by a set of ordered pairs of nodes (usually called edges).

are undirected[14], and may contain cyclic paths. Undirected graph models such as GGMs are widely applied in part because of the relatively few assumptions required. However, because the edges linking nodes are undirected, the scope for causal interpretations is limited (Briganti et al., 2022).

CDMs and network models

von Davier (2018, pp. 66-67) links CDMs and network models via the log-linear skills model class presented by von Davier and Yamamoto (2004). This log-linear skills model can be used to model observed or latent variables, and when used to model binary or ordinal attributes, it can be re-written as a log-linear model with main effects and first-order interactions. When applied to binary response variables, it is equivalent to the Ising model. This is the network psychometrics connection alluded to by von Davier and Lee (2019a, p. 3) in their introduction to the *Handbook of Diagnostic Classification Models*. **Figure 5** shows a schematic map of how the Ising model forms a point of connection between network psychometrics, CDMs and IRT.

The LCDM is an extension of the log-linear model, and further details on this relationship can be found in Henson et al. (2009, pp. 196-198). The details of the relationship of the Ising model to IRT are set out by Marsman et al. (2018). Marsman and colleagues in fact present far more than this, in a systematic account relating network models to IRT: “even though the conceptual framework that motivates the statistical representation in a psychometric model may be strikingly different for network models and latent variable models, the network models and latent variable models turn out to be strongly related; so strongly, in fact, that we are able to establish a general correspondence between the model representations and, in certain cases, full statistical equivalence.” (Marsman et al., 2018, p. 16)

[14] An undirected graph is one in which the edges linking pairs of nodes have no order: the two nodes connected have equal ‘status’ (in particular, there is no start/end or parent/child hierarchy in the pair).

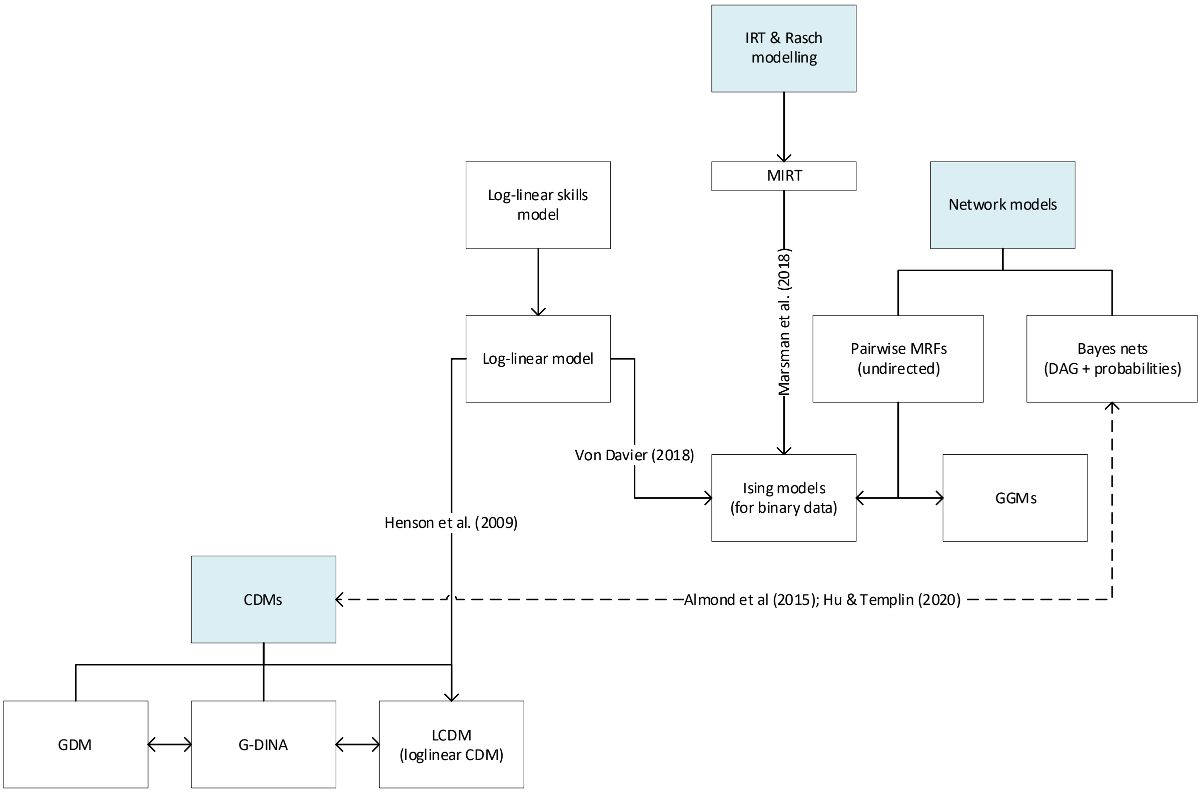


Figure 5: Linking CDMs, IRT and network models – see also Appendix C, **Figure 7**.

Bayesian networks

Bayesian inference networks, also known as Bayesian networks or Bayes nets (BN) for short, are a different important category of network model[15]. A Bayes net consists of “a joint probability distribution over a collection of discrete[16] variables” that can be represented as an acyclic directed graph[17] with a corresponding set of conditional probability distributions (Mislevy & Bolsinova, 2021, p. 94). Bayesian networks can be used in contexts like psychopathology where network psychometrics developed, but are less commonly found there than pairwise MRF models. Since all BN edges are directed, BNs are better placed for modelling causal relationships, but require numerous strong assumptions (see Briganti et al., 2022, pp. 1-2; McNally, 2023, pp. 2-3).

In the educational measurement context, BNs are a useful and flexible alternative approach to assessing sub-domains, to be considered instead of multidimensional IRT or CDMs

[15] Briganti et al. (2022) highlight an important possible confusion to avoid: “BNs [Bayesian Networks] should not be confused with pairwise Markov random fields estimated with Bayesian methods” (p. 1). By definition, the graph structures of BNs are different to the graph structures of pairwise MRFs.

[16] Culbertson (2016, p. 6) notes that: “…most applications of BN have used discrete latent variables. However, there is no theoretical restriction to discrete variables; and advances in computing power and new computational algorithms … render BN with continuous latent variables more feasible.”

[17] A directed acyclic graph (DAG) is a finite set of nodes accompanied by a set of directed edges (i.e., a set of ordered pairs of nodes), in which there exist no cyclic paths – that is, it is not possible to find a closed loop of directed edges.

(Culbertson, 2016; Wang et al., 2023). Culbertson notes that although psychometricians in educational assessment have paid relatively little attention to BNs, in the artificial intelligence field BNs have been used "extensively" in intelligent tutoring systems (ITS) to model student knowledge[18] (2016, p. 4).

Culbertson (2016) emphasises that Bayesian networks "are not a type of model, per se" but rather, "represent a framework or approach to model building". Capitalising on the relationship between a graph representation and the complex joint probability distribution makes BNs "a convenient and intuitive" way to model such distributions (p. 4). The "convenience" aspect stems from the fact that nodes are conditionally independent given the separating set[19]. The conditional independencies implied by the BN graph structure then simplify considerably (via the rules of probability), with the result that "complex relationships between many variables can be described through conditional relationships between much smaller subsets of variables, which may be read directly from the graph" (see explanation by Culbertson, 2016, pp. 4-5).

Bayes nets are highly flexible: they can model both latent and observed variables, they are well-suited to modular model development, and a node's conditional distribution can be specified using "any" probability distribution (Culbertson, 2016, p. 6). Within an ITS, it is usual for both latent attributes and items to be represented as nodes, with the latent attributes "conventionally represented as hidden nodes, and test items as observed nodes" (Hu & Templin, 2020, pp. 303-304). In educational measurement contexts generally, including ITS, it is usual for the directed graph edges to flow from the latent attribute nodes to the test item nodes, that is, for the conditional probability distributions of observed variables to be specified in terms of latent attribute "parent" nodes. It is also permissible, however, to have edges between latent attribute nodes and between test item nodes (Almond & Zapata-Rivera, 2019, p. 83; Culbertson, 2016, p. 5). The probabilities associated with each possible value of an observed variable node in educational measurement BNs are commonly generated by IRT models or CDMs – on which basis, Culbertson points out that "traditional IRT and DCM can be viewed as special cases of BN with a single (potentially multi-dimensional) hidden node" (Culbertson, 2016, p. 6).

A sense in which BNs are *less* flexible than other approaches to cognitive assessment is that the between-attribute structure is pre-specified: "unlike in DCMs and MIRT models, which often use an unstructured model for the structural model (i.e., the relationship among latent variables), BayesNets are predominantly (but not exclusively) built upon prespecified attribute hierarchies expressed as a series of marginal and conditional distributions" (Hu & Templin, 2020, p. 304). This point is also made by Culbertson (2016). Almond and Zapata-

[18] In ITS contexts, this is referred to as the "student model", as in the original Evidence Centred Design formulation (more recent ECD iterations refer to the "proficiency model"). The term "student model" refers to "the detailed tracking of the student on their knowledge on a topic, their skills, and various psychological attributes, including personality, motivation, and emotions" (Graesser et al., 2018, p. 249). The "primary goal" of a student model is "to take the data on the previous performance of a student and use it to provide an estimate of knowledge and predictions of future performance. Specifically, for a student s and an item i, a model predicts the probability the student s will answer the item i correctly." (Pelánek, 2018, p. 210)

[19] See Almond p. 84 for details of when nodes in a directed graph are separated.

Rivera (2019, p. 88) are unequivocal: “the Bayes net requires the relationship among the proficiency variables (attributes) to be explicitly stated and modeled”.

CDMs and Bayesian networks

Bayesian network models are closely connected to cognitive diagnostic models. Almond et al. (2015) show in detail how model structures known by CDM names (e.g., DINA, DINO – pp. 250-254) can be expressed as BNs; Mislevy and Bolsinova (2021) note that this is true for many CDMs (and latent class models in general).

Hu and Templin (2020) showed that Bayesian networks “can be parameterized to provide proficiency models that are equivalent to the saturated LCDM and to the HDCM” (p. 304) – in other words, that it is possible to formulate a “saturated proficiency model under [a] BayesNets framework”, and that this is equivalent to a saturated model under LCDM (p. 305). Hu and Templin’s motivation for looking at the saturated BN model was both to compare it to saturated DCMs, and to look at comparisons with the other possible (not saturated) BN proficiency models nested within the saturated BN model (pp. 304-305).

The Dynamic Learning Maps (DLM) technical documentation gives an insight into what the relationship between CDMs and BNs can offer assessment developers. Firstly, the technical documentation states explicitly that DLM assessments were informed by research and development from both CDM and BN traditions. In fact, not only the ideas but the representations and language used are intermixed by the DLM developers: “Since the latent variables (called nodes) from Bayesian inference networks and the latent variables (called attributes) from diagnostic classification models are **mathematically equivalent**, this document blends research and terminology from the two measurement paradigms from which such methods have evolved” (Dynamic Learning Maps Consortium, 2016, pp. 159-160, emphasis in original).

The cognitive diagnostic assessment developed by Wang et al. (2023), assessing buoyancy, demonstrated ways in which the flexibility of the BN structure allowed the researchers to combine aspects from multiple CDMs. Specifically, the final 3-level hierarchical BN that the researchers developed (which outperformed the original BN, G-DINA and HDCM models in classification) incorporated a high-level proficiency variable such as can be modelled in the HO-DINA model (which cannot model hierarchical attribute relationships) and the hierarchical relationships that can be modelled by the HDCM (which cannot incorporate a high-level proficiency variable that’s not present in the Q-matrix).

In assessment contexts that require real-time scoring - such as an ITS or adaptive testing system – Almond and Zapata-Rivera (2019, p. 89) make the case for implementing CDMs via BNs more generally, since it makes “scoring” individual students (and hence classifying them, or reporting probabilities of mastery) straightforward[20]. Specifically, they argue that "first translating the estimated [CDM] models to a Bayes net and then using the Bayes net for scoring is an attractive method for using those models in embedded applications" (p. 89).

Why are CDMs not used more frequently?

Ravand and Baghaei (2020) attribute the paucity of operational CDM applications to three essentially practical factors:

1. The fact that CDMs have a "lack of accessibility to a broad audience interested in their application" (p. 26).
2. The fact that the field is rapidly developing, which makes it difficult and time-consuming for practitioners to keep abreast of changes and new developments.
3. The remaining presence (despite extensive research and development activity) of practical barriers, most notably sample size requirements.

Supporting the final point, Zhang et al. (2023, p. 670) emphasise that "Real-world implementations of cognitive diagnostic testing often face a difficult trade-off between statistical (or psychometric) soundness and practical feasibility", in particular, because reliable measurement often requires repeated measures, and parameter identifiability requires that attributes are assessed (at least once) in isolation. The practical requirements for using CDMs are described in research question 2 (page 36).

Principled objections

In contrast to these practical concerns, von Davier (2018) develops an account ("Diagnosing Diagnostic Models") that questions the value of CDMs at a more fundamental level, and, in particular, questions whether they have any value over and above IRT models and network psychometrics. The strongest point admitted in favour of CDMs is that they are conceptually attractive: "discrete skills representing nuggets of knowledge, success in learning specific content or particular fine-grained proficiencies are a natural representation of hypothesized probabilistic causes … of observed behavior" (von Davier, 2018, p. 68).

von Davier's core argument against CDMs is that "Most diagnostic models assume that observed behaviors can be explained by a multidimensional skill attribute vector that is binary … this assumption cannot be made without being challenged" (p. 68). The article is structured around making this challenge in some detail, specifically, addressing four "implicit claims or assumptions" many diagnostic models have made (p. 59):

1. Skills can be represented as latent structures with two possible outcomes.
2. Multiple skill attributes are needed to explain the observed response data.

[20] "A completely specified Bayesian network (one whose conditional probability tables are all known) is a description of the joint probability of the latent and observable variables for the population of interest. Scoring a single student is straightforward and the required operations are supported by almost all Bayesian network software. First, a student specific copy of the network is made. Next, all of the observed variables are instantiated (set) to their observed values. Then a simple message passing algorithm is used to update the marginal probabilities for all nodes in the network. Statistics of these marginal probability distributions can be reported as scores." (Almond & Zapata-Rivera, 2019, p. 89)

3. The exact set of required skill attributes can be determined for each item.
4. The rate-response function of how skill attributes add up, or interact, or both, can be determined.

The general CDMs – which potentially include all possible attribute interactions – are highly parameterised models. There is an obvious risk of overfitting (e.g., an LCDM with 5 skills may require up to 32 parameters per item), and also that parameters will lack identifiability[21] (p. 61). With regard to CDMs involving attribute hierarchies, von Davier and Haberman (2014) proved that hierarchically related binary attributes can be replaced by a polytomous latent variable (because the allowable attribute patterns correspond to a perfect Guttman pattern, meaning "which" attributes is not informative, only "how many"). For these scenarios, von Davier argues that "using multiple attributes only obscures the validity of a much simpler model" (2018, p. 65).

It is essential to at least consider ordinal as well as binary variables, or compare binary-variable to continuous-variable models, to test whether there is actually support for the binary variable assumption. Studies that have carried out this kind of comparison show in many cases that alternative conceptions (e.g., a unidimensional IRT model) explain the observed responses equally well (or better). Hence, "instead of assuming many binary skills in a latent variable model, fewer ordinal or even one continuous or ordinal skill can potentially be assumed without loss of accuracy of model predictions" (von Davier, 2018, pp. 66-67).

von Davier (2018) notes that modelling associations between observed variables can be done without assuming latent variables at all (using a log-linear skills model or network psychometrics approach), but offers a caution about the kinds of conclusions this could lead to. Specifically, "assuming that there are no underlying causes (latent variables) and that the observed indicators are all that is needed for a comprehensive explanation could lead to simplified explanation patterns that only look at what we can easily observe, assuming that all we observed is all there is" (von Davier, 2018, p. 68).

In the introduction to the *Handbook of Diagnostic Classification Models* von Davier and Lee (2019b) offer a similar overall verdict with respect to CDMs: "In many cases, latent class analysis, customary IRT, and other latent variable models can directly be considered alternatives to diagnostic models, as these are often more parsimonious (in the case of IRT)

[21] Mathematically, we mean that it is not possible to satisfactorily estimate the model parameters, because the likelihood of the observed data under the model is the same for two or more different sets of parameters. von Davier gives an illustration (pp. 64-65) of this occurrence in hierarchical CDMs where attributes are related in a so-called "linear hierarchy". In this example "the interaction of these two skills cannot be distinguished from the additive effect of the two hierarchically ordered attributes, because the higher order attribute can never be mastered by a person not having mastered the lower order attribute" (von Davier, 2018, p. 65). Marsman and colleagues make some helpful comments, noting in particular that "mapping from statistical association structure to generating causal structure is typically one-to-many, which means that many different underlying causal models can generate the same set of statistical relations" (Marsman et al., 2018, p. 26). Their framing emphasises the un-remarkableness of the consequences: "Because each of these models implies the same probability distribution for the data, one cannot conclude from the fit of the statistical model that the conceptual model is accurate. Thus, causal interpretations do not follow from the statistical model alone, as indeed they never do."

or do not make as strong (parametric) assumptions about the latent structures and how these structures are related to the conditional response probabilities in the levels of the latent variables.” As a result, von Davier and Lee recommend that researchers always compare their results from more complex modelling approaches (CDMs) to a baseline of more standard approaches: “customary standard examples of latent variable models such as IRT or LCA”. This will enable researchers to determine whether there is in fact added value from the increased model complexity, either in terms of better model fit to the data, or in terms of “more useful derived quantities such as estimated mastery states” (von Davier & Lee, 2019b, p. 15).

Relevant example

Hong et al. (2015) developed two hybrid models which they trialled on the Tatsuoka (2002) fractions dataset. The first (“DINA-NIRT”) included both a noncompensatory IRT term and a DINA model. The second (the Continuous Conjunctive Model - CCM) was a conjunctive model inspired by NIDA, but for continuous latent variables. Its key innovation was minimising the number of model parameters as far as possible (notably, no item-level parameters) in order to make it practical for situations with a very high number of dimensions. By directly addressing the granularity of latent traits, Hong et al. (2015) address many of the objections made by von Davier (2018). For instance, Hong et al. (2015, p. 41) state that “Whether classification or scoring is desired, it is worth considering the interpretation of the latent variables involved and modeling them appropriately. In some instances, it may be reasonable to assume that the latent variables are a mixture of discrete and continuous variables, and may be modeled accordingly.” The two models developed offer a practical way forward: “Hybrid models may afford a chance to fit more realistic models, rather than misspecifying the types of the latent variables, merely for the purpose at hand”, and in “the difficult case where many continuous latent traits are required along with a noncompensatory assumption” the CCM may offer a way forward where other NIRT models cannot be fit (p. 42).

Research questions

1. Reporting to schools based on CDMs

RQ1: In a 'best-case' scenario (sufficient resources to meet sample size, item construction and test design requirements), how would reporting to schools based on CDMs be an improvement upon (i) raw sub-score reporting by topic, (ii) topic or domain scaled scores (e.g., Cambridge English scale scores, Cambridge Checkpoint scores), or (iii) raw item level results?

The claimed benefit of reporting to schools based on CDMs is that CDMs provide information that is more readily actionable than information from other assessments, and thus better placed to have positive effects on teaching and learning. This is because CDMs assign students to easily interpretable mastery states – preferably linked to intuitively discrete concepts, skills or learning outcomes. A teacher does not need to work out (either 'upward' from item level results, or 'downward' from topic-level results) which students have mastered which concepts, because this information is immediately provided, and can inform teaching immediately.

The level of granularity is core to this argument. If CDMs are applied to higher-level skills or groups of skills (e.g., moving away from the highly granular maths and science models demonstrated in the research literature, or the highly granular DLM and Navvy products available commercially), the claimed benefit of CDMs seems to diminish, as it is no longer plausible to claim a direct link between reporting and "action required", a point clearly argued by Bradshaw (2022). Bradshaw presents "standards-level information" – that is to say, information at the level of individual learning objectives such as "Divide multi-digit numbers" – as the useful and actionable level of granularity. The "Navvy classroom" products provide information at this level (Pearson Assessments US, 2023b).

The claim that reliable information at an actionable level is useful is sound. However, this claim does not directly address the question of how CDM results would improve upon raw score or scaled score reporting at the same level of granularity (e.g., the "standards" level used in Navvy). Instead, the studies that compare CDM outputs to other types of assessment make comparisons to random classification, other CDM outputs, MIRT, or Bayesian networks.

In theory, advantages offered by CDM in comparison with simply reporting raw or scaled scores at attribute level include:

- capitalising on complex loading of (potentially) multiple attributes on a single item (Rupp et al., 2010, pp. 83-84),
- making use of information about how other students performed on items assessing the same attribute, and information (if any) about the proportion of respondents in the population that have the same attribute profile (Rupp et al., 2010, p. 235),
- being able to reach an acceptable level of accuracy and reliability with a lower number of test items, due to the above factors in combination with the coarse (dichotomous) reporting scale at attribute level (Paulsen & Valdivia, 2021, p. 929; Rupp, 2023, p. 7).

Some authors in the research literature are cautious (von Davier, 2018; von Davier & Lee, 2019b), and argue that although the use of CDMs is conceptually attractive, the actual benefits over other approaches to reporting on sub-domains is still not clear.

2: Valid reporting from CDMs

RQ2: What would be the minimum requirements in terms of sample sizes, number of items and test design considerations to allow valid reporting based on CDMs?

In terms of the numbers of items and learners required for CDM use and reporting, there is of course “no single answer because, as always, it depends” (Rupp, 2023, p. 9). Factors affecting the requirements include:

- how many attributes each item measures
- the scale on which attributes are coded
- how many score points are available on each item (e.g., dichotomous or polytomous)
- the strength of discriminatory power of items in relation to the attributes they require
- the match between item difficulty and the test-takers
- the level of classification consistency desired
- the (true) distribution of test-takers across the possible latent classes.

Rupp (2023, p. 9) offers helpful reminders that “the number of high-quality items for each attribute/ dimension helps with the classification challenge for learners”, while accurately estimating parameters for the items is helped by “the number of learners with distinct, item-relevant profiles” (i.e., not just a higher total number of test-takers). In particular, “as designs get more complex (e.g., adaptive testing, matrix sampling with multiple forms) it is particularly important to keep in mind not the numbers overall but the interaction of these numbers”.

Sample size requirements

Sen and Cohen (2021) carried out a useful simulation study that systematically investigated the effect of varying sample size on classification accuracy and parameter recovery, for four CDMs: the reduced LCDM, the DINA and DINO models, and the C-RUM. The factors manipulated were:

1. Sample size: 50, 100, 200, 300, 400, 500, 1000 and 5000 simulated test-takers.
2. Test length: 12, 24 and 36 items.
3. Number of attributes: three and five; maximum of two attributes per item.
4. Base rate (proportion of test-takers who have mastered an attribute): 0.25 and 0.50.
5. Model used to generate the simulated data: reduced LCDM, DINA, DINO, and C-RUM.
    a. For the specific models (DINA and DINO models, and the C-RUM) the underlying CDM structure was the same for all items.
    b. For the reduced LCDM, different underlying DCM structures were used to generate the data for different items. For the 12 item tests, the structures used were 3 x DINO, 3 x DINA, and 3 x C-RUM.

The simulations were set up so that item quality, tetrachoric correlations between each pair of attributes, and Q-matrices were held constant. Specifically, item discrimination was set to be 0.60 and tetrachoric correlations between each pair of attributes set to be 0.70, based on plausible values found in the literature (Sen & Cohen, 2021, p. 3).

The results for each model are reported at length (Sen & Cohen, 2021, pp. 6-14). As a high-level summary, the main results were:

1. In common with previous studies, the results showed improved item parameter recovery for larger sample sizes. Sen and Cohen (2021, p. 14) state: “In general, it appears that sample sizes should be at least 500 for the four DCMs considered in this study in order to obtain precise estimates.” Results were particularly poor for sample sizes < 200. For DINA, DINO and C-RUM, sample sizes “as small as *N*=1000 would be sufficient to adequately recover all model parameters, under all the given conditions”. This was not true for LCDMREDUCED, which required larger sample sizes.
2. Again, in common with previous studies, the findings showed that item parameters were estimated more accurately as the test length increased from 12 to 36 items.
3. Item parameters were recovered *less* accurately when the number of attributes being measured increased from three to five. As Sen and Cohen (2021, p. 14) note, this is important to emphasise, because “most of the studies in DCM literature use more than three attributes” – and in fact, CDMs have been recommended over alternatives such as MIRT precisely for those contexts in which the number of latent attributes is high (Hong et al., 2015). Previous results on required sample sizes should be considered in light of this finding.
4. For a fixed number of items measuring an attribute, classification accuracy increased as the number of items measuring the attribute in isolation increased. Conversely, “classification accuracy suffered most when a pair of attributes was measured” (p. 14). The ranges of classification accuracy percentages reported (pp. 12-13) were:
    a. C-RUM: 21.6 to 70.3
    b. DINA: 47.5 to 80.2
    c. DINO: 51.0 to 88.3
    d. LCDMREDUCED: 32.8 to 77.5.

This study represents the most thorough contribution so far to the CDM sample size literature, but has two major limitations (both acknowledged by the authors). The first is that it did not investigate any general CDM, and the second is that the generalisability of the results is “necessarily limited to the conditions manipulated in this study” (p. 14). Hence, although Sen & Cohen manipulated a fairly extensive set of factors important to CDM estimation, results remain unknown for assessment designs that differ otherwise (even while sample size, number of attributes, number of items, and base rate remain in the range of the simulation). In particular, previous studies have shown that classification accuracy “varied markedly for different Q-matrix designs” (Madison & Bradshaw, 2015; Sen & Cohen, 2021, p. 14).

Sample size and model choice

The lower sample size requirements of specific CDMs (due to the fewer parameters estimated) is one of their advantages over general CDMs such as G-DINA (de la Torre & Minchen, 2019; Deonovic et al., 2019). However, this has to be weighed against the disadvantages of estimating specific rather than general models (**Table 13**).

Table 13: Pros and cons of general vs specific CDMs (Deonovic et al., 2019).

General CDMs	Specific CDMs
Require fewer assumptions	Require more assumptions
More likely to fit the data	Less likely to fit the data
More likely to have identification issues	**Less likely to have identification issues**
Requires larger sample sizes	**Can be used with smaller sample sizes**
Less straightforward interpretation	**More straightforward interpretation**

Ravand and Baghaei (2020, p. 48) argue that CDM research should develop estimation methods able to achieve stable and accurate results with smaller samples. They appear to recommend specific CDMs as a stop-gap approach: “Meanwhile, for practical purposes, practitioners can use the available less-complex models that require smaller sample sizes.” The tension between “practical purposes” and good practice is maintained in their conclusions. Ravand and Baghaei (2020, p. 50) state that they share concerns about choice of CDM being “commonly an arbitrary process rather than being informed by substantive considerations”, and list three possible routes. The first, “blanket imposition of a single specific DCM on all the items of a test” is “a practice we advise against”. The preferred route is to run a general CDM, then let each item select its own model based on fit. However, this recommendation is dependent on the availability of a “large enough (> 5,000)” sample size. If the sample size does not permit running a general CDM, then the recommendation is “running several specific models and selecting the one fitting the data the best” (p. 50).

Sample size and context

Ravand and Baghaei (2020) state that sample size requirements make applications of CDMs in classroom contexts “almost impossible” (p. 47). In particular, they argue that “As long as DCMs need sample sizes in the magnitude of 1,000–2,000, they never leave the psychometric laboratories” (p. 48). This perspective appears to ignore the role of CDM-based assessments that are designed, estimated, and calibrated (using large representative samples) by assessment organisations, and then delivered at classroom level.

Paulsen and Valdivia (2021) argue explicitly that there are important advantages to estimating CDMs at classroom level. They acknowledge that “vendor-designed assessments targeted to particular sets of state standards could be designed for classroom use and calibrated at a large sample level” (pp. 918-919). However, their view is that “such assessments are at a greater distance from the classroom dynamics and may not achieve all the benefits documented above” – that is, the high relevance to teaching and learning cited as the benefit of CDM-based assessment (p. 919). Paulsen and Valdivia’s position is clearly stated: “To achieve the benefits of CDMs at the classroom level, they need to be able to function at the sample size of a classroom.”

This position motivated a simulation study focusing specifically on diagnostic classification with small sample sizes. The study compared the use of DINA and two nonparametric approaches: nonparametric diagnostic classification (NPC - Chiu & Douglas, 2013) and supervised artificial neural networks (SANN - Cui et al., 2016). The factors investigated by Paulsen and Valdivia (2021) in the simulation study were:

- **sample sizes:** 25, 1000

- **number of attributes:** 2, 4, and 6
- **number of items**: 10, 20 and 50
    - Q-matrix complexity held constant at 1.6 attributes per item
    - no item measured more than 2 attributes
    - each Q-matrix contained at least one item measuring each attribute in isolation
    - each Q-matrix contained at least 3 items assessing each attribute
- **Q-matrix misspecification:** 0%, 10% and 20% misspecification
- **item discrimination:** variable. The researchers randomly generated more and less effective discrimination parameters for the simulation.

Data were generated using the RRUM model, and the results for the different models and conditions were compared using attribute and profile classification accuracy rates (Paulsen & Valdivia, 2021, p. 924). The main findings were:

- The authors concluded that use of DINA at classroom levels is fundamentally a feasible approach. In contrast to previous studies indicating much larger sample sizes were necessary, “the DINA model converged at N=25 across 8,399 out of 8,400 replications.” (p. 929) The authors credit this finding to improved computing capacity, and specifically, “use of marginal maximum likelihood (MML) estimation with expectation-maximization (EM) algorithm”.
- The DINA model and NPCD consistently outperformed SANN.
- Classification accuracy was not improved by increasing sample size. Item discrimination, however, “impacted classification accuracy substantively across models, more so than any condition.” (p. 930) The number of attributes and number of items also had statistically significant impact on classification accuracy.
- Attribute classification rates (ACA) were better than profile classification rates (PCA), and the authors recommend “great care” making inferences about latent classes. In particular, “Under high item discrimination, ACA rates across all conditions for DINA and NPCD remained above 0.80” (p. 930). By contrast, “Assuming that a .80 classification accuracy rate was acceptable for low stakes classroom inferences, PCA rates only reached or exceeded this threshold for the DINA and NPCD model in high item discrimination, two-attribute or 50-item conditions with 0% or 10% Q-matrix misspecification. The SANN model never reached .80 PCA.”

Items per attribute

No hard-and-fast rule can guarantee the number of items required per attribute in a CDM analysis, because this figure will also depend on item quality, the distribution of attributes across items (including how many times each is measured in isolation), and the level of classification accuracy and consistency that is considered acceptable.

In the Pearson Navvy products, the diagnostic measurement model reports mastery at the level of individual “standards”. Bradshaw, the original creator of the Navvy assessments, states that working at “standards” level allows measurement “that is valid and reliable, but also only takes 6-8 items to get a diagnosis of either you’ve learned the standard or still need help on the standard” (Pearson Assessments US, 2023b). The level of reliability considered acceptable is not reported. In DLM assessments, items are arranged into testlets of 3-8 items assessing one “linkage level” – the unit for reporting[22].

Besides these two commercial assessment products, the item numbers and Q-matrix designs investigated by simulation studies are the most obvious source of information on necessary requirements. These support the rule of thumb stated by Matthew Madison in his NCME workshop, that “each attribute should be isolated at least once, preferably more than once, on an item for good model performance” (Madison, 2023).

An observation worth noting is that the requirements for CDM estimation and usage may be regarded differently depending on the frame of reference. In comparison with the requirements for achieving precise, valid and reliable but unidimensional measurement (reporting a sum score, single grade or scale score), CDM requirements can seem demanding. If reporting on multiple dimensions is taken as a given, however, and CDMs are viewed in comparison with confirmatory factor analysis or MIRT (i.e., reporting fine distinctions on multiple dimensions or attributes), then CDMs appear as the more achievable or realistic modelling choice, and the requirements of CDMs seem relatively low. This is the perspective voiced by Rupp (2023, p. 7):

“Essentially, they are multidimensional models that use distinctions amongst learners that are coarser for each of the dimensions (e.g., mastery/non-mastery) than those used in confirmatory factor analysis or multidimensional item response theory models (i.e., scale scores with fine distinctions). As a result, one requires fewer items to make coarser distinctions on each dimension although one still wants items that have strong discrimination power for each dimension.”

3: CDM outcomes from existing assessments

RQ3: Are CDMs a technology that assessment organisations could use to report outcomes from existing assessments?

As noted in earlier sections, there are serious objections in the literature to “retrofitting” CDMs to existing, non-diagnostic assessments (Bradshaw et al., 2014; Rupp & Templin, 2009). The practice has been described as a “the-measure-of-last-resort” (Ravand & Baghaei, 2020, p. 27), but despite this criticism, published examples of retrofitting outnumber “true” CDM applications by a large margin.

Retrofitting CDMs to existing assessments is considered particularly challenging “when the breadth of the domain is relatively wide” (Deonovic et al., 2019, p. 446), which may be a particularly relevant concern for high-stakes curriculum-based assessments. The immediate reason why breadth of domain causes problems for CDMs is that it forces either coarse granularity or a large number of attributes. Deonovic et al. also list poor item quality (i.e., poor discrimination) and Q-vectors lacking in variability as problems for retrofitting to relatively wide-domain assessments. A practical solution is to break down larger domains

[22] In the DLM system, “Each linkage level represents one or more nodes, or skills, in the learning map model that underlies the assessment system” and reporting contains a classification of mastery or non-mastery of each linkage level (Clark et al., 2017, p. 6). The DLM approach (and its terminology) are not the most straightforward to understand but are explained fully in the technical documentation (Dynamic Learning Maps Consortium, 2016).

where possible, as demonstrated by Deonovic et al. (2019) in the development of the Education Companion App (modelling “geometry”, “algebra”, and “number” separately, for instance, rather than “mathematics”).

Another heightened risk for retrofitting studies is that model parameters are not identifiable, due to the combinations of attributes targeted (and not targeted) by test items. Deonovic et al. (2019) state that “additional information (e.g., extra test items, ancillary variables) than can supplement test data, at least for some examinees, may be needed to ensure that every examinee is reliably classified” (p. 446). Within the modelling for the Education Companion App, Deonovic and colleagues created “nuisance” attributes to account for skills not included in the Q-matrices of the domain-focused CDM analyses[23].

4: Reporting considerations

RQ4: What are the considerations around score interpretation and transparency?

Bradshaw and Levy (2019) emphasise the need to first *interpret* (and agree on) the outcomes of probabilistic classification (such as carried out using CDMs) in a measurement sense, before deciding on the implications for reporting. They appear to assume that the probabilities of mastery will definitely be reported (in some form) to stakeholders. Their discussion then focuses on how to make clear what these probabilities are and what they mean. In general, classification of test-takers using CDMs is “based upon the probability of membership in each group, where the sum of the probabilities across all groups equals 1” (p. 81). As discussed earlier, this can involve classification based on probabilities of latent class membership (i.e., membership of a latent class which is defined by mastery and non-mastery across a set of multiple attributes). Alternatively (the EAP estimate approach), we may be interested in the probabilities of classifying the test-taker as “master” or “non-master” of a single attribute. In both cases, “probability of membership indicates the certainty— and really, the uncertainty—of the classification. The (un)certainty is predominantly driven by how consistently an examinee exhibited the attribute on the items across the assessment.”[24]

Bradshaw and Levy particularly emphasise the need to avoid any of four common misinterpretations (pp. 82-83) of CDM classification probabilities:

1. Probabilities as classifications by a different name

[23] Specifically, in analysing each domain (“geometry”, “algebra”, and “number”), the domain-specific attributes were the focus of analysis and the rest (e.g., number skills when looking at the algebra test) were collapsed into coarser nuisance attributes. So, the total attributes analysed in each domain were the domain-specific ones plus several nuisance attributes. For further details, see Appendix B.

[24] Bradshaw and Levy acknowledge immediately that this raises the questions “Whose certainty?” and “What is certainty?”, which they explore in some depth. Briefly, Bradshaw and Levy argue that certainty “belongs” to “all of the assumptions, beliefs, and decisions that contributed to the assessment enterprise” – the collective consisting of everything from attribute operationalisation to sample collection to the physical testing environment (p. 81). The word “certainty” is intended to “communicate a degree of sureness that is aligned with the understanding of the likelihood of an event occurring” (p. 81). Bradshaw and Levy suggest that in the case of CDMs, there is a case for communicating that “the results are more *clear* when probabilities of mastery are closer to 0 or 1 and less clear when they are near .5” (p. 81). They also note that a relative frequency metaphor may be helpful (i.e., where the probability of mastery is 0.75, “out of 100 students that responded to the items in this way, we expect that 75 of them have mastered the concept and 25 of them have not” (p. 82)).

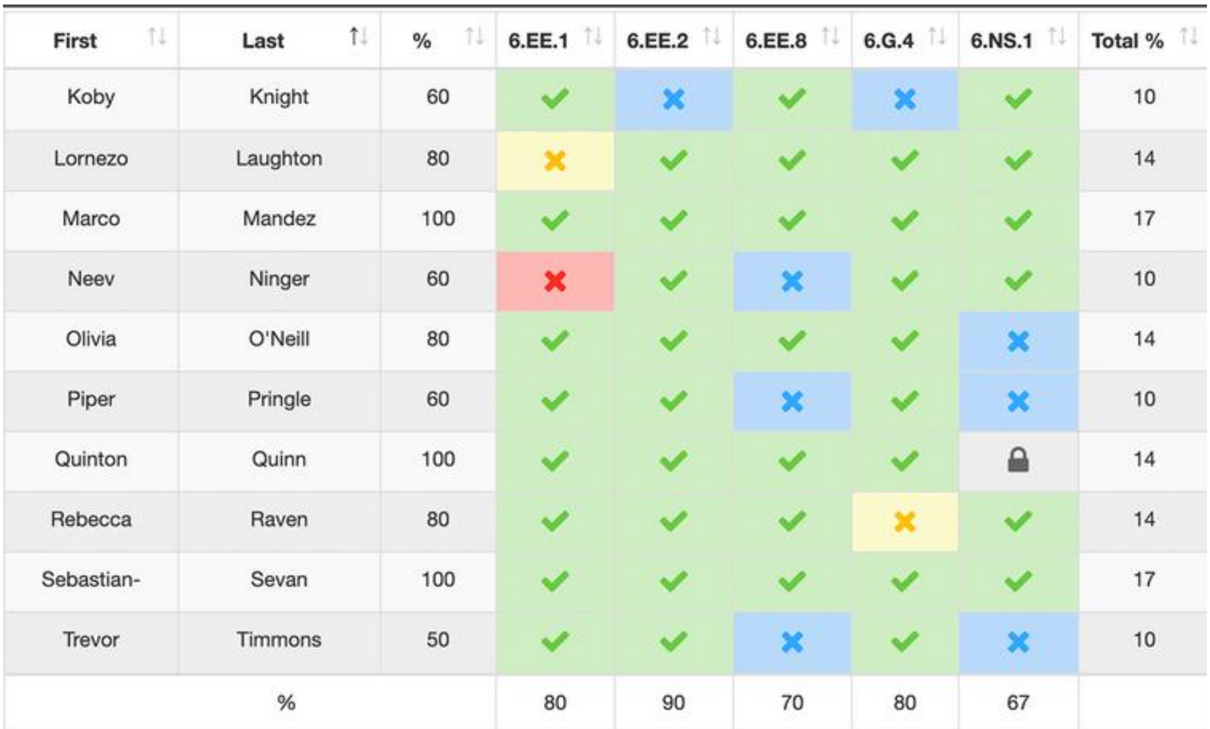
2. Probabilities as percent of items correct.
3. Probabilities as amount of mastery.
4. Probabilities as progress.

Bradshaw and Levy (2019) give two good arguments for taking special care when reporting from CDMs. First, that “the literature is filled with evidence that decision making based on probabilities is difficult for adults, including adults with high intelligence and successful careers” (p. 82). Second, that the results from CDMs – classifications with a measure of uncertainty – are “unfamiliar not only to end users, but also to much of the community of psychometric and assessment experts” (p. 83).

Clark et al. (2022) - colleagues from the University of Kansas ATLAS centre, which developed the Dynamic Learning Maps products - carried out a useful empirical investigation of teacher literacy in relation to diagnostic assessment systems. They collected data from a survey of teachers using DLM assessments, and focus groups with teachers who had been using DLM assessments for at least one year. The findings showed that teachers “were comfortable” using the term mastery to discuss assessment results, student knowledge, DLM score reports and their plans for action (p. 8). However, it was clear that teachers had gaps in their understanding and also some active misunderstandings, including those listed by Bradshaw and Levy. Clark et al. noted a lack of nuance: teachers “did not talk about mastery as reflecting skills that were likely mastered or as being based on probabilities of mastery”, but rather, as absolute facts. In addition, it was “evident from their comments that teachers were unsure how mastery was defined or determined”, teachers referred to mastery score reports “as what students ‘got right’”, “some misinterpreted mastery as representing a percent correct rather than a probability value”, some “shared confusion about how results were calculated”, and some referred to a “black box” (p. 8). Clark et al. emphasised that our focus should remain on how this affects assessment validity. The question is not whether mastery classifications were in some sense misunderstood (they were), but rather, whether “these more nuanced understandings … constitute fundamental understandings that are critical for appropriate use of results” (p. 8).

In contrast to the authors cited above, Ravand and Baghaei (2020, p. 48) assert that “DCM outputs do not seem to pose any difficulty for teachers” and that “the complexity issue is not very serious”. Their logic is that “Although the mathematical bases of DCMs are extremely sophisticated, practitioners do not need to get involved with them”. Further: “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.” (Ravand & Baghaei, 2020, p. 48). The comparative argument seems plausible, as does the assertion that teachers need not fully engage with the mathematical details of CDMs. However, Ravand and Baghaei do not appear to consider that CDM outputs (i.e., classifications, and probabilities) could in fact be misunderstood in ways important to assessment validity – precisely as illustrated by Clark et al. (2022) – whether or not they “seem” to pose difficulty for teachers.

In some implementations of CDMs, for instance in some of the Navvy product dashboards presented to teachers, probabilities of mastery are not in fact immediately reported. It may be that teachers can obtain the probabilities of mastery in reporting if they request them, but the probabilities are not presented in the dashboard outputs advertised as the “usable” assessment output in marketing and public-facing materials. Instead, the dashboards simply convey the classification decisions per student and attribute, in tables (or similar) with green and red shading. In the example dashboard shown in **Figure 6**, green indicates mastery; yellow, blue and red indicate non-mastery after varying numbers of attempts; and the padlock symbol indicates a not-yet-assessed attribute.



First	Last	%	6.EE.1	6.EE.2	6.EE.8	6.G.4	6.NS.1	Total %
Koby	Knight	60	✓	🟦 ✗	✓	🟦 ✗	✓	10
Lornezo	Laughton	80	🟨 ✗	✓	✓	✓	✓	14
Marco	Mandez	100	✓	✓	✓	✓	✓	17
Neev	Ninger	60	🟥 ✗	✓	🟦 ✗	✓	✓	10
Olivia	O'Neill	80	✓	✓	✓	✓	🟦 ✗	14
Piper	Pringle	60	✓	✓	🟦 ✗	✓	🟦 ✗	10
Quinton	Quinn	100	✓	✓	✓	✓	🔒	14
Rebecca	Raven	80	✓	✓	✓	🟨 ✗	✓	14
Sebastian-	Sevan	100	✓	✓	✓	✓	✓	17
Trevor	Timmons	50	✓	✓	🟦 ✗	✓	🟦 ✗	10
	%		80	90	70	80	67	

Figure 6: Example dashboard from Navvy products (Pearson Assessments US, 2023a).

5: Possible CDM purposes

RQ5: For which purposes might assessment organisations want to use CDMs?

Taking into consideration the aspects reviewed so far, the evidence suggests that (i) assessment organisations would not find it easy to use CDMs, but that (ii) there is potential to use CDMs for four different purposes:

1. To structure the design, scoring and reporting of diagnostic tests.
2. To drive adaptivity and/or personalisation within advanced digital learning and assessment products[25].

[25] Note, this refers to using the results of CDM analyses to drive adaptation at a higher level in the product, rather than introducing adaptivity within assessments themselves (which would add further complexity). For instance, the recommended pathway through pages of digital learning content can adapt based on the estimation of mastery (or non-mastery) of a core concept. This kind of approach is seen in Dynamic Learning Maps, where the system is “adaptive between testlets” – that is, testlets assessing concepts at particular levels are fixed (small) units, and the system adapts and personalises based on their results (Dynamic Learning Maps Consortium, 2016).

3. For validation work, especially to verify specific claims about the cognitive behaviours elicited by test items.
4. To research the nature of constructs we wish to assess in a more exploratory way, and to research their role in successfully answering items that we expect to measure these constructs.

Table 14 lists different kinds of assessment and the ways in which each could potentially make use of CDMs. Since the third and fourth purposes are closely related, they are combined in what follows.

For validation and construct research

The use of CDMs for validation and construct research is in some ways the most straightforward purpose to consider. There are substantial hurdles to overcome, as documented throughout this report, but there is a potential role for CDMs to assist validation and construct research to support any of the assessment types in **Table 14**.

High-stakes assessments have already been investigated using CDM methods in several retrofitting studies, for example on IELTS (Mirzaei et al., 2020) and the Cambridge/Singapore O Level English listening test (Aryadoust, 2018). Furthermore, the numerous examples of retrofitting studies in mathematics and science domains demonstrate how relatively broad domain assessments can be explored using CDM methods (e.g., Delafontaine et al., 2022; Deonovic et al., 2019; Jia et al., 2021). In comparison with multidimensional IRT, both CDMs and the AHM make it much more straightforward to analyse the hierarchical structures between skills (George & Robitzsch, 2021, p. 108). CDMs have the additional advantage, over the AHM, of more easily allowing comparisons between groups (as the AHM is not a likelihood-based approach).

Potential for "true" CDM applications

Where the objective is to classify test-takers according to mastery of discrete skills or attributes, it is possible to imagine CDMs being applied in many of the assessment types in **Table 14**. This is not (of course) to say that CDMs would be appropriate or helpful in all baseline assessments (for example), but that baseline assessment in general is not incompatible with CDMs. At least in some contexts, a CDM-based baseline assessment could be developed.

There are two assessment types in **Table 14** for which it is very difficult to imagine CDMs being useful in live assessment and scoring. Most fundamentally, this is because both "summative" and "selection" assessments are very likely to be used for ranking students, at least as one of their purposes. These assessment types are therefore not very compatible with CDMs, because "rank ordering examinees along a continuum is not what CDMs are designed to do" (Bradshaw & Levy, 2019, p. 83). Attempting to make a CDM generate a rank order would be "at odds with the fundamental assumptions of the model upon which the validity of the model interpretations relies": CDMs assume two underlying groups [per attribute], and do not attempt to make further distinctions within the "masters" and "non-masters" groups (p. 83). An additional factor is that since summative and selection assessments are often high-stakes for candidates (and for schools), the transparency and explainability of decisions are particularly important. While not all authors perceive a problem for CDMs here (see RQ4), it is certainly easier to imagine widespread CDM use in non-summative and non-selection assessments first.

Besides the *type* of assessment, the domain being assessed is important for whether CDMs are an appropriate choice. Multiple factors are involved: the breadth of the domain as defined by the assessment; whether that domain has been mapped in granular detail; and whether it is *possible* for the domain to be mapped in granular detail in the ways demonstrated in mathematics and science topics, where much CDM research and development has focused. Here it is again worth noting that the attributes required by items in the Tatsuoka (2002) dataset – a tightly focused dataset on fractions subtraction – have been researched and debated for two decades without consensus being reached. From the context of learning modelling (in ITS), Pelánek (2017) explicitly argues that results about model choice "probably do not generalize" and that both the domain and the purpose of modelling are important for ensuring "an appropriate choice of a modelling approach" (p. 328). For example, Pelánek (2017) states that logistic models are the better choice for modelling "memory and fluency building processes" where a student's knowledge state changes gradually. By contrast, Bayesian Knowledge Tracing (BKT) is "more appropriate for modeling understanding and sense making processes", as its assumptions are that learning involves a discrete transition from a state of not-knowing to knowing - but this is only appropriate "for fine grained knowledge components" (p. 328).

Potential to drive adaptivity and/or personalisation

The assessment types for which CDMs are most obviously relevant here are learning oriented assessment, and integrated learning and assessment, as these are assessment types with high likelihoods of being incorporated into a digital learning and assessment system where adaptivity and personalisation are desired. Diagnostic assessments and formative assessments, similarly, are assessment types more likely than others to be incorporated into such a system, and hence benefit from CDMs applied in this way.

Table 14: Kinds of assessment and relevance of CDMs.

Assessment type	Purpose 1: potential for "true" CDM application	Purpose 2: driving digital product functionality	Purpose 3 & 4: validation and construct research
Baseline assessment	Yes	Possibly?	Yes
Diagnostic assessment	Yes	Yes	Yes
Formative assessment	Yes	Yes	Yes
Learning Oriented assessment	Yes	Yes	Yes
Integrated learning and assessment	Yes*	Yes*	Yes
Summative assessment	Unlikely/no	No	Yes
Achievement test	Yes	Possibly?	Yes
Proficiency test	Yes	Possibly?	Yes
Aptitude/ability assessment	Possibly	Possibly?	Yes
Selection assessment	Unlikely/no	No	Yes
Progress / progression test	Yes (relative to baseline)	Possibly?	Yes
Placement assessments	Yes, potentially	Possibly?	Yes

References

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer.

Almond, R. G., & Zapata-Rivera, J.-D. (2019). Bayesian networks. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 81-106). Springer. https://doi.org/10.1007/978-3-030-05584-4_4

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge General Certificate of Education O-Level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening, 35*(1), 29-52. https://doi.org/10.1080/10904018.2018.1500915

Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Eds.), *Handbook of Cognition and Assessment* (pp. 297–327). Wiley-Blackwell.

Bradshaw, L. (2022). *Empowering personalized instruction with a three-tiered approach to learning evidence*. [White paper]. Pearson Education. https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/district-assessment/navvy-white-paper.pdf

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2-14. https://doi.org/10.1111/emip.12020

Bradshaw, L., & Levy, R. (2019). Interpreting probabilistic classifications from diagnostic psychometric models. *Educational Measurement: Issues and Practice, 38*(2), 79-88. https://doi.org/10.1111/emip.12247

Briganti, G., Scutari, M., & McNally, R. J. (2022). A tutorial on Bayesian networks for psychopathology researchers. *Psychological Methods*. https://doi.org/10.1037/met0000479

Camara, W., O'Connor, R., Mattern, K., & Hanson, M. A. (Eds.). (2015). *Beyond academics: A holistic framework for enhancing education and workplace success*. ACT Research Report Series 2015 (4). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2015-4.pdf.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-eefined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419-437. https://doi.org/10.1177/0146621613479818

Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Recommendation system for adaptive learning. *Applied Psychological Measurement, 42*(1), 24-41. https://doi.org/10.1177/0146621617697959

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*(2), 225-250. https://doi.org/10.1007/s00357-013-9132-9

Chiu, C. Y., & Köhn, H. F. (2019). Nonparametric methods in cognitively diagnostic assessment. In M. Von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification Models* (pp. 107-132). Springer.

Chiu, C. Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika, 83*(2), 355-375. https://doi.org/10.1007/s11336-017-9595-4

Clark, A. K., Nash, B., & Karvonen, M. (2022). Teacher assessment literacy: Implications for diagnostic assessment systems. *Applied Measurement in Education*, 1-16. https://doi.org/10.1080/08957347.2022.2034823

Clark, A. K., Nash, B., Karvonen, M., & Kingston, N. (2017). Condensed mastery profile method for setting standards for diagnostic assessment systems. *Educational Measurement: Issues and Practice, 36*(4), 5-15. https://doi.org/10.1111/emip.12162

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.

Cui, Y., Gierl, M., & Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology, 36*(6), 1065-1082. https://doi.org/10.1080/01443410.2015.1062078

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*, 19-38.

Culbertson, M. J. (2016). Bayesian networks in educational assessment: The state of the field. *Applied Psychological Measurement, 40*(1), 3-21. https://doi.org/10.1177/0146621615590401

Culpepper, S. A., & Hudson, A. (2018). An improved strategy for Bayesian estimation of the reduced reparameterized unified model. *Applied Psychological Measurement, 42*(2), 99-115. https://doi.org/10.1177/0146621617707511

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.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353. https://doi.org/10.1007/BF02295640

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), 595-624. https://doi.org/10.1007/s11336-008-9063-2

de la Torre, J., & Minchen, N. D. (2019). The G-DINA model framework. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 155-169). Springer. https://doi.org/10.1007/978-3-030-05584-4_7 de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199. https://doi.org/10.1007/s11336-011-9207-7

Delafontaine, J., Chen, C., Park, J. Y., & Van den Noortgate, W. (2022). Using country-specific Q-matrices for cognitive diagnostic assessments with international large-scale data. *Large-scale Assessments in Education, 10*(1). https://doi.org/10.1186/s40536-022-00138-4

Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., & von Davier, A. A. (2019). Application of cognitive diagnostic models to learning and assessment systems. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 437-460). Springer.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Erlbaum.

Doignon, J.-P., & Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies, 23*(2), 175-196.

Doignon, J.-P., & Falmagne, J.-C. (2012). *Knowledge spaces*. Springer Science & Business Media.

Dynamic Learning Maps Consortium. (2016). *2014-2015 Technical manual – Integrated model*. University of Kansas, Center for Educational Testing and Evaluation. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Technical_Manual_IM_2014-15.pdf

George, A. C., & Robitzsch, A. (2021). Validating theoretical assumptions about reading with cognitive diagnosis models. *International Journal of Testing, 21*(2), 105-129. https://doi.org/10.1080/15305058.2021.1931238

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2). https://doi.org/10.18637/jss.v074.i02

Graesser, A. C., Hu, X., & Sottilare, R. (2018). Intelligent tutoring systems. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences*. Taylor & Francis.

Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. [PhD thesis]. University of Illinois at Urbana-Champaign.

Heller, J., Stefanutti, L., Anselmi, P., & Robusto, E. (2015). On the link between cognitive diagnostic models and knowledge space theory. *Psychometrika, 80*(4), 995-1019. https://doi.org/10.1007/s11336-015-9457-x

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

Hong, H., Wang, C., Lim, Y. S., & Douglas, J. (2015). Efficient models for cognitive diagnosis With continuous and mixed-type latent variables. *Applied Psychological Measurement, 39*(1), 31-43. https://doi.org/10.1177/0146621614524981

Hu, B., & Templin, J. (2020). Using diagnostic classification models to validate attribute hierarchies and evaluate model fit in Bayesian networks. *Multivariate Behavioural Research, 55*(2), 300-311. https://doi.org/10.1080/00273171.2019.1632165

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407-419. https://doi.org/10.1177/0013164410388832

Jia, B., Zhu, Z., & Gao, H. (2021). International comparative study of statistics learning trajectories based on PISA data on cognitive diagnostic models. *Frontiers in Psychology, 12*, 1-9. https://doi.org/10.3389/fpsyg.2021.657858

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 55*(4), 635-664.

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. https://doi.org/10.1177/01466210122032064

Karvonen, M., Burnes, J. J., Clark, A. K., & Kavitsky, L. (2020). *Aligned academic achievement standards to support pursuit of postsecondary opportunities: Instructionally embedded model*. Technical Report No. 20-02. University of Kansas ATLAS (Accessible Teaching, Learning, and Assessment Systems). https://dynamiclearningmaps.org/sites/default/files/documents/publication/Aligned_Academic_Achievement_Standards_to_Support_Pursuit_of_Postsecondary_Opportunities_Instructionally_Embedded_Model.pdf

Leighton, J. P., & Gierl, M. J. (Eds.). (2007a). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.

Leighton, J. P., & Gierl, M. J. (2007b). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: theory and applications* (pp. 3-18). Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule‐space approach. *Journal of Educational Measurement, 41*(3), 205-237.

Liu, J., & Kang, H.-A. (2019). Q-Matrix learning via latent variable selection and identifiability. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 247-263). Springer. https://doi.org/10.1007/978-3-030-05584-4_12

Ma, C., Ouyang, J., & Xu, G. (2023). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika, 88*(1), 175-207. https://doi.org/10.1007/s11336-022-09867-5

Maas, L., Brinkhuis, M. J. S., Kester, L., & Wijngaards-de Meij, L. (2022). Diagnostic classification models for actionable feedback in education: Effects of sample size and assessment length. *Frontiers in Education, 7*, 1-17. https://doi.org/10.3389/feduc.2022.802828

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

Madison, M. (2023). *Introduction to diagnostic measurement models*. Diagnostic measurement SIGMIE webinar, NCME, March 7, 2023.

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.

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika, 83*(4), 963-990. https://doi.org/10.1007/s11336-018-9638-5

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187-212. https://doi.org/10.1007/BF02294535

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., Maas, H., & Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioural Research, 53*(1), 15-35. https://doi.org/10.1080/00273171.2017.1379379

McNally, R. J. (2023). Points of contact between network psychometrics and experimental psychopathology. *Journal of Experimental Psychopathology, 14*(1), 1-7.

Min, S., Cai, H., & He, L. (2021). Application of bi-factor MIRT and higher-order CDM models to an in-house EFL listening test for diagnostic purposes. *Language Assessment Quarterly, 19*(2), 189-213. https://doi.org/10.1080/15434303.2021.1980571

Mirzaei, A., Heidari Vincheh, M., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation, 64*(Mar 2020), Article 100817. https://doi.org/10.1016/j.stueduc.2019.100817

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centred design*. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., & Bolsinova, M. (2021). Concepts and models from psychometrics. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python* (pp. 81-107). Springer International Publishing. https://doi.org/10.1007/978-3-030-74394-9_6

Paulsen, J., & Valdivia, D. S. (2021). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of Experimental Education, 90*(4), 916-933. https://doi.org/10.1080/00220973.2021.1891008

Pearson Assessments US. (2023a). *Navvy assessments*. Pearson. Retrieved 19 May 2023 from https://www.pearsonassessments.com/large-scale-assessments/district-assessment/navvy-assessment.html

Pearson Assessments US. (2023b). *Navvy: A fresh way to navigate student learning [Video]*. YouTube. https://www.youtube.com/watch?v=P80f7AkDi-k

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction, 27*(3-5), 313-350. https://doi.org/10.1007/s11257-017-9193-2

Pelánek, R. (2018). The details matter: methodological nuances in the evaluation of student models. *User Modeling and User-Adapted Interaction, 28*(3), 207-235. https://doi.org/10.1007/s11257-018-9204-y

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782-799. https://doi.org/10.1177/0734282915623053

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing, 20*(1), 24-56. https://doi.org/10.1080/15305058.2019.1588278

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. *Educational Psychology, 38*(10), 1255-1277. https://doi.org/10.1080/01443410.2018.1489524

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2014). *CDM: Cognitive diagnosis modeling [Software-Handbuch]*. R package version 4.1. http://CRAN.R-project.org/package=CDM

Roussos, L. A., Di Bello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications.* (pp. 275-318). Cambridge University Press.

Rupp, A. A. (2023). *Primer on diagnostic classification models* (2.0 ed.). Center for Assessment. https://www.nciea.org/library/primer-on-diagnostic-classification-models-dcms/

Rupp, A. A., & Templin, J. (2008a). The effects of Q-Matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.

Rupp, A. A., & Templin, J. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219-262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., & Templin, J. (2009). The (un)usual suspects? A measurement community in search of its identity. *Measurement: Interdisciplinary Research & Perspective, 7*(2), 115-121. https://doi.org/10.1080/15366360903187700

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. The Guilford Press.

Sen, S., & Cohen, A. S. (2021). Sample size requirements for applying diagnostic classification models. *Frontiers in Psychology, 11*, 1-16. https://doi.org/10.3389/fpsyg.2020.621251

Shafipoor, M., Ravand, H., & Maftoon, P. (2021). Test-level and item-level model fit comparison of general vs. specific diagnostic classification models: A case of true DCM. *Language Testing in Asia, 11*(1). https://doi.org/10.1186/s40468-021-00148-z

Sinharay, S., & Johnson, M. S. (2019). Measures of agreement: Reliability, classification accuracy, and classification consistency. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 359-377). Springer.

Stout, W., Henson, R. A., DiBello, L., & Shear, B. (2019). The reparameterized unified model system: A diagnostic assessment modeling approach. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 47-80). Springer.

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

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error analysis. In N. Frederiksen (Ed.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–388). Lawrence Erlbaum Associates.

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

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*, 251-275. https://doi.org/10.1007/s00357-013-9129-4

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317-339. https://doi.org/10.1007/S11336-013-9362-0

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305. https://doi.org/10.1037/1082-989X.11.3.287

von Davier, M. (2005). *A general diagnostic model applied to language testing data*. ETS Research Report Series RR-05-16. ETS. https://files.eric.ed.gov/fulltext/EJ1111422.pdf

von Davier, M. (2007). *Hierarchical general diagnostic models*. ETS Research Report Series RR-07-19. ETS. https://doi.org/10.1002/j.2333-8504.2007.tb02061.x

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*(1), 8-28.

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)*. ETS Research Report Series RR–14-40. ETS. https://doi.org/10.1002/ets2.12043

von Davier, M. (2018). Diagnosing diagnostic models: From von Neumann's elephant to model equivalencies and network psychometrics. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 59-70. https://doi.org/10.1080/15366367.2018.1436827

von Davier, M., & Haberman, S. (2014, Apr). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies - a commentary. *Psychometrika, 79*(2), 340–346. https://doi.org/10.1007/s11336-013-9362-0

von Davier, M., & Lee, Y.-S. (Eds.). (2019a). *Handbook of diagnostic classification models*. Springer.

von Davier, M., & Lee, Y.-S. (2019b). Introduction: From latent classes to cognitive diagnostic models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 1-17). Springer.

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

Wang, L. L., Jian, S. X., Liu, Y. L., & Xin, T. (2023). Using Bayesian networks for cognitive assessment of student understanding of buoyancy: A granular hierarchy model. *Applied Measurement in Education*, 1-15. https://doi.org/10.1080/08957347.2023.2172014

Wang, S., & Douglas, J. (2015, Mar). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika, 80*(1), 85-100. https://doi.org/10.1007/s11336-013-9372-y

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for diagnostic assessment. *Journal of Educational Measurement, 52*, 457-476.

Wu, X., Wu, R., Chang, H. H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in Psychology, 11*, 1-13. https://doi.org/10.3389/fpsyg.2020.02230

Zhang, S., Liu, J., & Ying, Z. (2023). Statistical applications to cognitive diagnostic testing. *Annual Review of Statistics and Its Application, 10*(1), 651-675. https://doi.org/10.1146/annurev-statistics-033021-111803

Appendix A: definitions

For reference, this appendix provides definitions of key terms and abbreviations.

Table 15: specific model names and other abbreviations.

Abbreviation	Definition
A-CDM	Additive CDM (de la Torre, 2011)
AHM	Attribute Hierarchy Method (Leighton et al., 2004)
BKT	Bayesian Knowledge Tracing (Corbett & Anderson, 1995; Pelánek, 2017)
BN	Bayesian Network or Bayes Net (Almond & Zapata-Rivera, 2019; Culbertson, 2016)
CDM	Cognitive Diagnostic Model (equivalent to DCM for most authors)
cG-DINA	continuous G-DINA, for continuous responses (de la Torre & Minchen, 2019)
C-RUM	Compensatory Reparameterized Unified Model (DiBello et al., 1995; Hartz, 2002)
DAG	Directed Acyclic Graph
DCM	Diagnostic Classification Model (equivalent to CDM for most authors)
DINA	Deterministic Input, Noisy “And” gate model (Junker & Sijtsma, 2001)
DINO	Deterministic Input, Noisy “Or” gate model (Templin & Henson, 2006)
DLM	Dynamic Learning Maps® (Clark et al., 2017)
EAR	Element-wise Agreement Rate
GDI	G-DINA Discrimination Index (de la Torre & Chiu, 2016; de la Torre & Minchen, 2019)
G-DINA	Generalized DINA (de la Torre & Minchen, 2019; de la Torre, 2011)
GDM	General Diagnostic Model (von Davier, 2005)
GGM	Gaussian Graphical Model
GNPC	Generalized Nonparametric Classification method (Chiu et al., 2018)
HDCM	Hierarchical DCM: based on full LCDM with constraints added and considering attribute hierarchical relationships (Templin & Bradshaw, 2014)
HGDM	Hierarchical General Diagnostic Model (von Davier, 2007, 2010)
HMM	Hidden Markov Model
HO-DINA	Higher Order DINA (de la Torre & Douglas, 2004, 2008)
IAV	Intraindividual variation
IEV	Interindividual variation
KST	Knowledge Space Theory (Doignon & Falmagne, 1985)
LCDM	Log-linear CDM
LLM	Log-linear model (Maris, 1999)
LLTM	Linear Logistic Test Model
MIRT	Multiple IRT
MRF	Markov Random Field
MS-DINA	Multiple Strategies DINA (de la Torre & Douglas, 2008)

NC-RUM	Noncompensatory reparameterized unified model (Hartz, 2002)
NIDA	Noisy Input, Deterministic "And" gate model
NPC	Nonparametric Classification method (Chiu & Douglas, 2013)
PFA	Performance Factor Analysis
pG-DINA	polytomous G-DINA, for polytomous attributes (de la Torre & Minchen, 2019)
PVAF	Proportion of Variance Accounted For
rRUM	Reduced Reparametrized Unified Model (DiBello et al., 1995; Hartz, 2002)
RSM	Rule Space Method (Tatsuoka, 1983)
SANN	Supervised Artificial Neural Network (Cui et al., 2016)
sG-DINA	sequential G-DINA, for polytomous responses (de la Torre & Minchen, 2019)
VAR	Vector-wise Agreement Rate

Definitions of cognitive diagnostic assessment and CDMs

“Diagnostic classification models (DCM) are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables.” (Rupp & Templin, 2008b, p. 226)

“DCMs are confirmatory multidimensional latent-variable models. Their loading structure/Q-matrix can be complex to reflect within-item multidimensionality or simple to reflect between-item multidimensionality. DCMs are suitable for modeling observable response variables with various scale types and distributions and contain discrete latent predictor variables. The latent predictor variables are combined by a series of linear-modeling effects that can result in compensatory and/ or noncompensatory ways for predicting observable item responses. DCMs thus provide multivariate attribute profiles for respondents based on statistically derived classifications.” (Rupp et al., 2010, p. 83)

“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.” (Ravand & Baghaei, 2020, p. 25)

“Among possible multidimensional models containing latent variables, cognitive diagnosis models (CDMs; DiBello et al., 2007; also labeled as diagnostic classification models, see von Davier & Lee, 2019) recently gained some attention. Roughly spoken, CDMs are a class of multidimensional categorical latent variable models that integrate theoretical assumptions about skills and then estimate the students’ possession of these skills.” (George & Robitzsch, 2021, p. 107)

“Cognitive diagnosis models (CDMs) can be viewed as restricted versions of the more general latent class models. In particular, the number of latent classes, as well as their interpretation, are known a priori when CDMs are involved.” (de la Torre & Minchen, 2019, p. 155)

“Diagnostic classification models (DCMs), also known as cognitive diagnostic models (CDMs), can be viewed as restricted versions of general latent class models … These models provide one way of classifying respondents into different diagnostic states.” (Sen & Cohen, 2021, p. 1)

“CDMs are multivariate, discrete latent variable models developed primarily to identify the mastery, or lack thereof, of skills (or more generically, *attributes*) measured in a particular domain. Two features distinguish CDMs when compared to traditional item response models, namely, the finer-grained nature of the inferences that can be derived from the models, and the interpretability and relevance of these inferences to the student learning process.” (Deonovic et al., 2019, p. 444)

“… statistical models that are well-suited to categorize examinees according to mastery levels for a set of hypothesized latent skills or abilities. These classification-based models, collectively termed cognitive diagnosis models (CDMs), can be organized into four major frameworks: rule space methodology (RSM; Tatsuoka, 1983), the attribute hierarchy method (AHM; Leighton, Gierl, & Hunka, 2004), diagnostic classification models (DCMs; e.g., Rupp, Templin, & Henson, 2010; Bradshaw, 2016), and Bayesian networks (BNs; e.g., Almond, Mislevy, Steinberg, Williamson, & Yan, 2015).” (Bradshaw & Levy, 2019, p. 79)

Cognitive Diagnostic Assessment (CDA) is “designed to measure specific knowledge structures and processing skills in students so as to provide information about their cognitive strengths and weaknesses” (Leighton & Gierl, 2007a, p. 3).

(Heller et al., 2015)

Appendix B: the ACTNext Education Companion App (ECA)

The goal of the app development project was to make use of the huge amount of learner data held by ACT across tests/platforms, to make useful inferences: e.g., recommend learning resources for a student to practise a skill they haven't yet mastered.

The ECA app is built from six modules:

1. The Learning Analytics Platform (LEAP), a data repository.
2. A student data matching module.
3. A CDM-based diagnostic model.
4. A feedback model using the information from the diagnostic model.
5. A feedback dashboard providing usable and interpretable output.
6. Linking of the feedback to ACT resources.

Deonovic et al. (2019, p. 449) state that the ECA development programme had to solve three major challenges:

1. The mapping challenge: "how to leverage a large bank of assessment data which has been tagged and associated with multiple sets of attributes" (i.e., not one consistent taxonomy). The solution involved mapping from old to new (more comprehensive) taxonomies so that **all** data had rich metadata using a consistent holistic framework (HF) classification scheme[26].
2. The modelling challenge: "designing a model capable of drawing inference from the data available to the ECA about users' skills and attributes". The solution chosen for the ECA was the LLTM, an extension of the Rasch model which takes into account concepts from CDMs in the form of the Q-matrix.
3. Validation challenge: "To validate the approach taken by the ECA we performed an intensive analysis of the data using the standard CDM approach."

In terms of the mathematics topics in the ECA and Q-matrix development (Deonovic et al., 2019, pp. 451-452):

- content experts developed the Q-matrices for the maths test
- under consideration: 4 test forms, each with 60 items (=240 items in total)
- 24 attributes in total, across three domains:
  - 10 for Operations, Algebra and Functions (OAF)
  - 5 for Geometry (G)
  - 9 for Number (N)
- each domain was analysed separately, because there were a large number of attributes

[26] The holistic framework referred to is now used by ACT across many assessments and is available here: https://www.act.org/content/act/en/college-and-career-readiness/holistic-framework.html. Camara et al. (2015) described the underpinning research, and presented the different sections of the framework. The attributes defined by Deonovic et al. (2019, p. 453, Table 21.1) are granular statements that fit within this framework, while the domain and sub-domain structure used by Deonovic et al. is taken directly from the framework.

- In analysing each domain, the domain-specific attributes were the focus of analysis and the rest (e.g., number skills when looking at the algebra test) were collapsed into coarser nuisance attributes. So, the total attributes in each domain were the domain-specific ones reported above plus several nuisance attributes each.
    - The number of times target attributes were measured was variable: some as few as 3 times, others as many as 24 times.
    - Each item generally measured 1-3 target attributes + 0-1 nuisance attributes; this varied a bit by domain (higher ratio of nuisance to targets in Domain 2 – Geometry).
    - Not all 60 items were relevant to each domain: some measured only nuisance attributes.
    - On average, 54 items per test form were relevant to a given domain.
    - Nuisance attributes were the most frequently measured, for all three domains.

Modelling

- The G-DINA model was fitted to a subset of the data (N=5000 examinees).
- Q-matrix validation was then carried out using the GDI and mesa plot (data driven) approach (Deonovic et al., 2019, p. 452).
    - "A mesa plot shows the PVAF [Proportion of variance accounted for] for some possible q-vectors for a given item. It always starts with all-zero q-vector. The cutoff for a q-vector to be considered appropriate was set at PVAF = 0.85."
    - "The validation results given in Table 21.5 show that the attribute-wise agreement between the provisional and suggested Q-matrices across all test forms and domains was very high: the minimum was 93% and the average was 95%."
- The Wald test was used to conduct item-level comparisons of G-DINA and a number of reduced (specific) CDMs: DINA, DINO, LLM, rRUM and A-CDM. The purpose was to find the optimal set of CDMs for a given test.
- Patterns were found by domain:
    - G-DINA was selected most frequently, especially for Algebra and Number
    - LLM and rRUM also selected frequently. All of these relax the "two latent classes" constraint, to different extents.
    - Few items were selected to be DINA, DINO or A-CDM
    - For Geometry, G-DINA, LLM and rRUM were selected most frequently; A-CDM and DINA infrequently.
- The authors concluded that the results supported "the construct validity of the Q-matrices developed by content experts" (p. 455).

Appendix C: organising the field of cognitive diagnostic assessment

Figure 7 gives a graphical representation of the relationships between different models and model families described in the current report. The left hand side of **Figure 7** shows model relations within cognitive diagnostic assessment: the basic structure is from the statistical account by Zhang et al. (2023), and additional models and relationships have been added to this structure. The right hand side of **Figure 7** shows a simplified map of learner modelling techniques, structured by the overview from Pelánek (2017). While learner modelling was not the focus of this report, CDMs have close relationships to models applied in learner modelling, and the context is of particular interest given the importance of learner modelling for advanced digital learning and assessment products.

Figure 7 summarises multiple kinds of relationship:

- connecting lines with no arrows indicate that the lower model (or model family) is an example or case of the model/category above,
- connections with a single arrow indicate where one model is a development of or from another,
- connections with arrows at each end indicate a statistical equivalence. A solid line indicates that the equivalence is true generally, while a dotted line indicates equivalence for a subset or under certain conditions only. Where possible, references have been added to show where to find precise details.

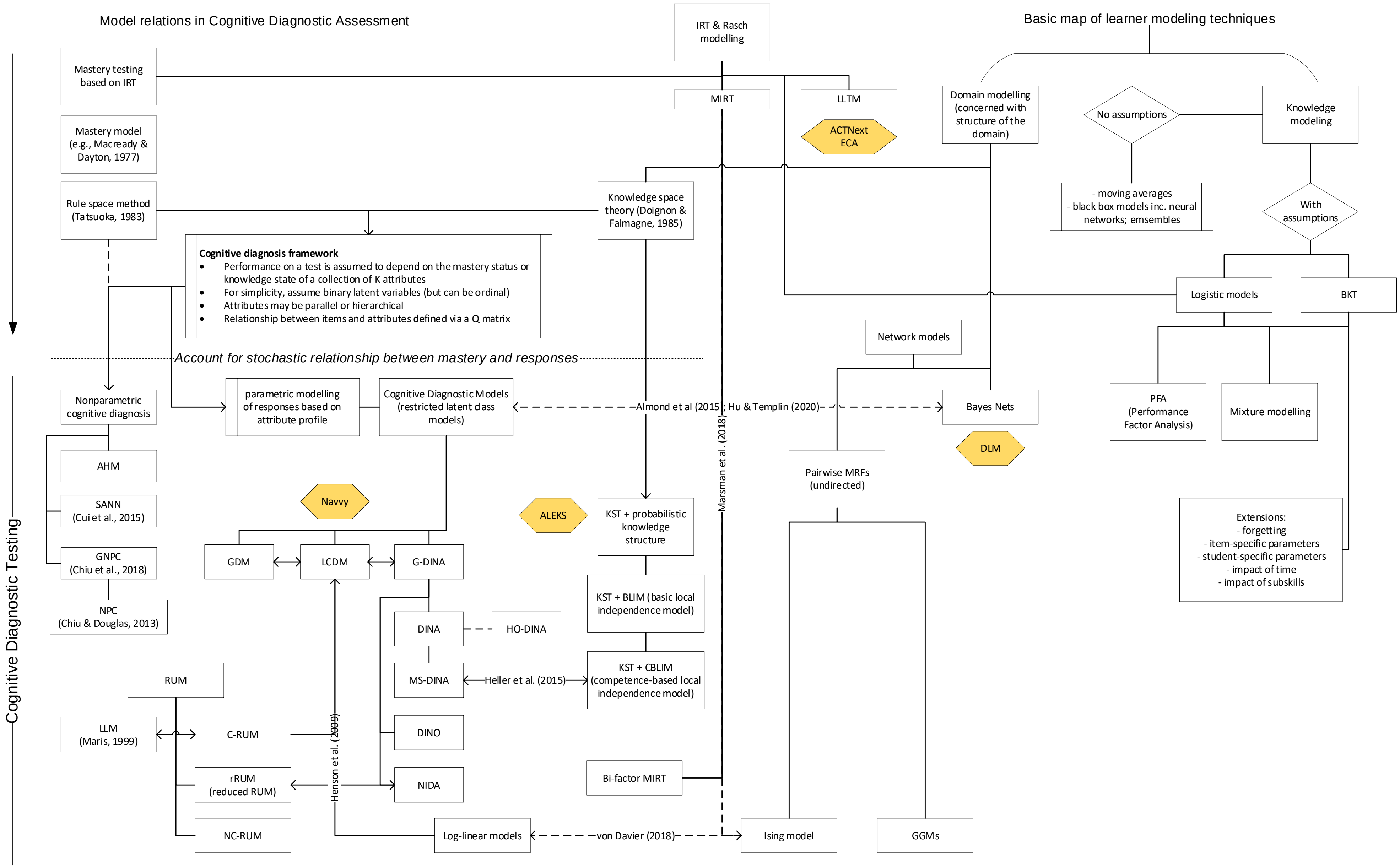

Model relations in Cognitive Diagnostic Assessment
IRT & Rasch modelling
Basic map of learner modeling techniques
Mastery testing based on IRT
MIRT
LLTM
Domain modelling (concerned with structure of the domain)
No assumptions
Knowledge modeling
ACTNext ECA
Mastery model (e.g., Macready & Dayton, 1977)
- moving averages
- black box models inc. neural networks; emsembles
With assumptions
Rule space method (Tatsuoka, 1983)
Knowledge space theory (Doignon & Falmagne, 1985)
Cognitive diagnosis framework
Performance on a test is assumed to depend on the mastery status or knowledge state of a collection of K attributes
For simplicity, assume binary latent variables (but can be ordinal)
Attributes may be parallel or hierarchical
Relationship between items and attributes defined via a Q matrix
Logistic models
BKT
Network models
Account for stochastic relationship between mastery and responses
Nonparametric cognitive diagnosis
parametric modelling of responses based on attribute profile
Cognitive Diagnostic Models (restricted latent class models)
Almond et al (2015); Hu & Templin (2020)
Bayes Nets
PFA (Performance Factor Analysis)
Mixture modelling
DLM
AHM
Marsman et al. (2018)
Pairwise MRFs (undirected)
SANN (Cui et al., 2015)
Navvy
ALEKS
KST + probabilistic knowledge structure
Extensions:
- forgetting
- item-specific parameters
- student-specific parameters
- impact of time
- impact of subskills
GNPC (Chiu et al., 2018)
GDM
LCDM
G-DINA
Cognitive Diagnostic Testing
KST + BLIM (basic local independence model)
NPC (Chiu & Douglas, 2013)
DINA
HO-DINA
RUM
MS-DINA
Heller et al. (2015)
KST + CBLIM (competence-based local independence model)
LLM (Maris, 1999)
C-RUM
Henson et al. (2009)
DINO
rRUM (reduced RUM)
NIDA
Bi-factor MIRT
NC-RUM
Log-linear models
von Davier (2018)
Ising model
GGMs

Figure 7: models for cognitive diagnostic testing and relevant nearby areas.