Primer on Diagnostic Classification Models (DCMs)

Andrė A. Rupp, Senior Associate

Version 2.0, February 6, 2023

This document lays out a few high-level principles and practices around *cognitive diagnosis models* (CDMs) / *diagnostic classification models* (DCMs) that are commonly of relevance for practical applications of these models. The document has a *frequently asked questions* (FAQ) format.

The information in this document is geared towards interdisciplinary teams composed of psychometricians/data scientists, content/curriculum specialists, and product developers/program managers. Consequently, it does not contain any formulas or overly technical details on nuances of model specification, estimation, or evaluation procedures even though some brief references to key ideas from these areas are made when appropriate.

It is treated as a living document to which updates can be made as new methodological questions arise, new scientific insights are gained in the field, or users want to see additional explanations, examples, or other representations included. If you have any suggestions for what questions to include or any comments on the provided explanations, please reach out to arupp@nciea.org!

General Features

1. What is the name for these models?

These models are predominantly known as *cognitive diagnosis models* (CDMs) or *diagnostic classification models* (DCMs) in the literature with both terms appearing to be equally popular; for simplicity, I will use the term DCM in this document since that is the one that we used in our book a few years back (Rupp, Templin, & Henson, 2010). Each term highlights a particular feature of the models more than the other. Specifically, the term "cognitive diagnosis models" highlights the idea that these models were originally designed to model finer-grained response processes grounded in applied cognitive psychology. The term "diagnostic classification models" highlights the fact that these models produce statistically-derived classifications of learners ("respondent" more generally if they are applied to data in other use contexts). In terms of the term "diagnosis" there is quite a bit of controversy around that as described next.

2. What is meant by the term "diagnosis"?

This term carries a variety of connotations with it that need to be carefully examined. Statistically speaking, the diagnosis that these models provide is simply a classification of learners into distinct groups ("latent classes") akin to what one might consider the idea behind classifications of the DSM-IV in clinical psychology or psychiatry for instance. The meaning of these groupings comes from the definition and operationalization of the attributes in these models, the items given to learners, the evidence extracted through the scoring process, and so on. Clearly, the classifications in education and their resulting interpretations are not akin to a thorough differential diagnosis in clinical psychology or psychiatry, which would take multiple sources of evidence, multiple instruments, parent-child-clinician conferences, and extended reports that may take 25+ hours to create. There are some common ideas and principles, however, that are shared in these practices. As a result, to avoid over-interpretations of classifications and other aspects of score meaning, so one has to be very clear on what one aspires to do with the information from the assessment through these models in a given context, especially considering the implementation supports for teachers, students, and parents around the models' output.

3. How many different DCMs are there?

In the early days of these models - in about the late 1990s to the early 2000s - there were quite a few models with different names out there that got distinct attention and that are often still referenced today. These included the DINA, DINO, NIDA, NIDO, C-RUM, NC-RUM, as core models as well as semiparametric approaches such as the rule-space methodology or the attribute hierarchy method, along with other techniques related to clustering. In later years - especially since the mid-2000s onwards - specialists recognized and underscored the connections between these models and other modeling approaches, which mirrored developments in other areas of statistics and measurement. This led to a few unified frameworks for model specification and estimation, which subsume many (albeit technically still not all) models. Common ones here are the LCDM, GDM, G-DINA, and GDCM frameworks. Then there are models that combine different ideas such as a model that has a higher-order variable that is like a unidimensional scale score and then an attribute structure underneath it so learners can be described in terms of both general ability/competency as well as, say, misconceptions (e.g., Kalkan et al., 2018). There are also models that include parameters for human raters to describe their relative severity and leniency for example (e.g., Li, Wang, & Xi, 2020). 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.

4. Are DCMs better than other measurement models?

That is a complex question because it depends on what one means by "better". If one desires statistically-driven classifications then these models are quite powerful and useful if one can estimate them well with the data that one has. Moreover, they do provide internally-anchored interpretations for their resulting classifications since each learner group (i.e., "latent class" in technical terms) is defined through the mastery states of the attributes (i.e., which attributes learners have mastered, which ones they still have to master, and which one there is insufficient information about). However, they are also just statistical tools and, as such, just part of the overall analytical toolbox. They still require strong item design, sufficiently large samples of learners and items for reliable classifications, and cannot do anything “magical” to compensate for weaknesses in item design or implementation models. For example, using DCMs does not create deeply insightful “diagnostic” insight out of nowhere. It is therefore helpful to think of them as one possibility for modeling and reporting but also to recognize their place within the overall assessment enterprise.

Test Design

5. What is an attribute?

This is a term for a skill, competency, or component of a cognitive process that is used in the literature around DCMs. Attributes are typically the elements that one wants to report on, so they may be related to a content strand or standard in a domain, specific skills or sub-skills in a technical area, or even broader competencies in educational surveys such as the reading ability related to informational texts or fictional text. Then there are applications in clinical psychology or behavioral science where these attributes represent dispositions (e.g., propensity to take certain kinds of drugs or display certain kinds of behaviors); in other words, they are defined through the use context in which DCMs are applied. Statistically speaking, an attribute becomes a dimension in these models as well as, by implication, a component of a latent class since these classes are defined through the attributes. For example, if one defines four attributes for reporting and item design, the resulting DCM will be a four-dimensional statistical model.

6. Is an attribute exactly the same as an ability/skill/competence then?

The simple answer to this question is “no”, just as with any other measurement model. All models are simply statistical tools that allow us to describe the variation in observed response patterns in systematic ways and then make predictions about item and person characteristics that have uncertainty associated with them. The variables that are in a model that correspond to the persons - learners typically in an educational setting - are mathematical representations. Essentially, the logic is that differences in the numerical values of these variables reflect differences in the underlying unobservable (i.e., “latent”) characteristics of the learners. This argument hinges on a variety of considerations that relate to how the items are designed, how tests are administered, how learners truly respond to the items, and so on, which is typically subsumed under the umbrella term of “validation argumentation”. In other words, the abilities that we use to describe learners are human-made constructions with some reasonable scientific basis - hence the term “construct” - and models help us to tell stories about our learners along these constructs. Another way of looking at this is to think through a prototypical cycle of assessment design, scoring, and inference:

- Constructs are identified: this happens through narrative descriptions, visualizations, tables and other representations, informed by learning science, experience, standards, and other resources [conceptual level]
- Tasks are created: this happens through systematic principled assessment design approaches where design information may be represented as design variables and the resulting tasks and underlying routines for their creation are physical or digital objects
- Tasks are administered to learners who respond to them: the processes that learners are engaging in are unobservable ("latent") although one can get "observed traces" of these through certain data-collection methods; the resulting responses are physical or digital artifacts
- Responses are scored: Relevant evidence about the utilization of the skills/abilities/competencies is identified in the observed work products and scores are assigned by human raters or automated routines; these scores become variables in a spreadsheet
- Responses are aggregated: Scores are combined using statistical models such as DCMs to produce aggregate representations of learners as well as empirical information about items; this information creates variables in spreadsheets
- Patterns of scores are analyzed: The resulting score patterns are analyzed and differences on individual variables, correlations between variables, and other systematic relationships are used empirically to underscore which sources of evidence converge and diverge
- Inferences about constructs are made: A process of reasoning backward is done, from the empirical patterns through the assessment design and implementation all the way back to the constructs and their hypothesized relationships in order to make claims about learners, instruments, and contextual conditions at different levels of aggregation

This reasoning process is quite nuanced if taken seriously and it is helpful generally to remember that constructs / skills / competencies are neither observable artifacts nor statistical variables. Moreover, latent variables in certain statistical models (e.g., latent attribute variables in DCMs) are not the same as latent characteristics of learners even though both are unobserved. One is a hypothesized aspect of people that we use to reason about them, the other is a component in a statistical model.

7. How can attributes be coded?

Technically, attributes can be coded on any scale; however, in most common applications of DMCs they are coded dichotomously (e.g., 0 / absent / not-yet-mastered vs. 1 / present / mastered) and sometimes polytomously (e.g., 0-1-2-3 for degrees of mastery). The finer the distinctions for each attribute, the more items are needed to classify people into these states on any given attribute, just like with any assessment. Moreover, the more attributes one has in the design, the more possible combinations of so-called "mastery states" there are. For example, if one had four attributes, each coded dichotomously (0-1), there would be a total of 2 x 2 x 2 x 2 = 2^4 = 16 possible mastery states (i.e., latent classes) ranging from [0,0,0,0] to [1,1,1,1]. Thus, if one wanted to provide distinct instructionally useful information in this context, one would have to be able to provide different information for each of the 16 latent classes. If these attributes were coded polytomously with three levels each one would have a total of 3 x 3 x 3 x 3 = 3^4 = 81 distinct mastery states (i.e., latent classes). Admittedly, it is unlikely – or even impossible - that one would/could observe all different latent classes in any given classroom, but this only underscores considerations about resources and pathways at different levels of a system. This is one of the reasons why large-scale applications rely on large item banks with differentiated resources and why these models are powerful for intelligent-tutoring-type systems that serve educational management functions.

8. What are attribute hierarchies?

Conceptually, attribute hierarchies are specifications of conditional relationships / dependencies / hierarchies amongst the set of attributes one wants to measure (e.g., “attribute 2 needs to be mastered generally in order for attribute 3 to be mastered”, “either attribute 2 or attribute 5 need to be mastered before attribute 6 can be mastered”.) Statistically, this reduces the number of potential latent classes (e.g., since one cannot have students who have mastered attribute 3 but not 2 in the first example above latent classes with this combination will not have any learners assigned to them).

9. What are different ways in which attributes can be combined for a given item?

The general idea is that an item can measure one or multiple attributes. Different DCMs technically allow for different mechanisms in which the attributes are combined for responding. Helpful examples are models that specify non-compensatory relationships (e.g., all attributes that an item is measuring are required to respond correctly to an item) or compensatory relationships (e.g., the lack of mastery on one attribute can be made up for by the mastery of a different item that the item is measuring; at least one attribute needs to be mastered for learners to provide a correct response). Statistically, this is all reflected in different ways in which the response probabilities are estimated by models.

10. What is a Q-matrix?

A Q-matrix is a technical term for the matrix (i.e., table) that specifies the relationship between items (typically listed in rows) and attributes (typically listed in columns). The most common way of indicating these relationships at this level is with a 0 and a 1 in the cells of that table where a 1 indicates that the item measures the attribute or, alternatively, that the attribute is required to respond correctly to the item. The Q-matrix is akin to the loading matrix in confirmatory factor analysis models where certain loadings are set to 0 (i.e., certain graphical paths between items and factors do not exist). Items that have multiple 1s in their rows are considered items with “within-item multidimensionality” (i.e., they measure multiple statistical dimensions) whereas those that have only one 1 in their row are items with “between-item multidimensionality” (i.e., they measure only one dimension). Q-matrices can technically contain other secondary information such as expected or observed item parameters such as difficulty or discrimination as well as design variables, which some call “augmented Q-matrices”. Note that the Q-matrix is separate from the item response matrix, which shows the relationship between learners and the items (i.e., the scores that learners have received on items). That matrix may also have only 0s and 1s in it (if all items are scored dichotomously/0-1) but it may also have other responses (e.g., polytomous scores for partial credit items, continuous scores for response time measurements). It may also include missing data in it due to different forms being administered to different people or adaptive testing approaches.

11. What can we do with item design information in augmented Q-matrices?

Generally speaking, capturing essential item design information is never a waste of time if the initial architecture for reporting and design is truly meaningful. The information can technically be used in a variety of ways that include using certain design variables:

- as formal reporting dimensions / attributes
- as secondary information that may be displayed in score reports
- as predictor variables within DCMs to explain variation in item difficulty or discrimination parameters to better understand how effective item design was (so-called “explanatory modeling” approaches)
- as predictor variables, outside of the DCM model proper, to explain variation item-level fit statistics to better understand whether there is a systematic relationship between design and fit
- to train future staff to write more targeted items or select items from existing banks in a more principled fashion
- to understand more deeply the relationship between human and automated scores that are used as input to DCMs for different item types

Importantly, not all design variables can serve all purposes simultaneously and there are a variety of secondary considerations that need to be attended to (e.g., using design variables to formally predict variation in item difficulty or discrimination-type parameters requires a relatively large number of items since the total number of items is essentially the “sample size” for these analyses.)

12. What should I do to make sure that my test design is suitable for DCMs?

The answer to this question depends on the assessment design architecture as well as the actual quality of the items and other information that is available. It is always helpful to see a representation of the architecture of the attributes (e.g., number, definition, relationships) as well as the rationale for how they were defined in the first place. With that in place, a common issue is to determine how test forms are created to make sure that the distribution of attributes within each form is such that sufficient information is available to classify each individual learner along all of the attributes. This is particularly the case in implementation models where fixed test forms are being administered at a given point in time, even with matrix-sampling setups, and long time windows exist between such administrations. If administration were more fluid then mastery states could be updated across each administration and current estimates could serve as prior information (i.e., ongoing collection of evidence). It is important to review the design from both cross-sectional perspectives and longitudinal perspectives to make sure the appropriate linkages are there for all intended use cases. Generally speaking, similar considerations as for other large-scale applications with multidimensional structures apply.

Methodological Connections

13. How do these models compare to other statistical models?

The answer to this question depends on how one looks at these models, which may include how they are structured, how they are estimated, how they are used in practice, and so on. Generally speaking, DCMs have quite a bit in common with factor analysis models, item response theory models, and unconstrained latent class models. 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. All of these models contain so-called latent variables to represent differences amongst learners whose values have to be estimated from the response patterns although there are nonparametric approaches that can use the information in a Q-matrix in other ways for clustering learners into groups.

14. Can we do longitudinal data analysis with DCMs?

There are now models available that provide promising starting points for doing this work (e.g., Pan, Qin, & Kingston, 2020). The particular model that is needed depends on the implementation design in that a model with regular data collection in, say, an online learning environment provides different kinds of opportunities for updating attribute mastery probabilities each time than an implementation model that consists of, say, three longer, fixed-form assessments that are temporally spaced out by several weeks or months. It is important to review these designs and the estimation demands of the DCMs (or related models such as Bayesian networks; see Choi & Mislevy, 2022; Mislevy, Almond, & Yan, 2019) to make principled choices about which models or modeling families might be most suitable for the targeted use contexts.

15. Can we do adaptive testing with DCMs?

Yes, that is possible and, in fact, with a large item bank and appropriate selection rules and delivery architecture, it is a valuable use of these models at a large scale. The ideas for adaptive testing in this space are similar to the ones in other spaces even though there are technical differences. For example, the statistical information functions that are used behind the scenes for DCMs have a different structure than those for item response theory models since items with appropriate levels of difficulty have to be selected that provide sufficient discriminatory power amongst possible latent classes for learners, which is a multidimensional problem with DCMs.

16. Can we handle missing data with DCMs?

Missing data can indeed cause issues for parameter estimation in DCMs if not properly treated. There are different ways of handling missing data, whose proper selection includes considerations about whether the data are missing by design (e.g., not all learners received all items due to different booklets) or through random or seemingly-random mechanisms (e.g., learners omit answers and/or do not reach certain items). The most common way to deal with missing data is to use estimation algorithms that leverage all available information once appropriate rescoring of certain responses has been done so that they can be included in the estimation (e.g., scoring omitted responses with a 0). Estimation can be done in traditional frequentist frameworks via variations of likelihood-based algorithms or within a Bayesian estimation framework using, commonly, Markov-chain Monte Carlo methods (e.g., Shan & Wang, 2020). Some studies also investigate different imputation methods for DCMs in which missing data is “filled in” before subsequent estimation (e.g., Dai & Valdivia, 2022).

17. Can we combine automated item generation and DCMs?

The area of automated item generation uses a variety of modern artificial intelligence approaches and is essentially independent from the choice of statistical model that is used. In the end, the items are presented to learners and the scores from these items are what is analyzed through DCMs. Thus, poor generation algorithms that produce poor-quality items will likely lead to very noisy or inappropriate score distributions that can cause estimation problems, but one would have to wrestle with these issues no matter what model one was using (and the model could do almost nothing about that problem in the end). That being said, in the totality of the ecosystem of the enterprise, automated item generation should be seamlessly integrated into the design-implementation-analysis lifecycle whenever it is used (e.g., Burstein et al., 2021); see also the next point.

18. Can we combine automated scoring and DCMs?

Similar to the previous question about automated item generation, considerations of automated scoring mostly pertain to the trustworthiness of the scores that are used to estimate any model, including DCMs. As such, this process can be thought of as an independent component of the modeling part although, in the totality of the ecosystem of the enterprise, it should be seamlessly integrated into the design-implementation-analysis lifecycle whenever it is used (e.g., Burstein et al., 2021); see also previous answer.

Technical Considerations

19. How many items and learners do I need for my test?

This question has no single answer because, as always, it depends on a few design factors. For example, it depends on how attributes are coded (e.g., mastery/non-mastery), how many attributes each item measures, how strong the discriminatory power of each item on each attribute / dimension is, how well the item difficulties are matched to the learner population, how many score points are available for each item (e.g., dichotomous/0-1 scores or polytomous/0-1-2-3), what kind of classification consistency is required for the targeted use(s) of the test, and so on. Intuitively, if there are three items that measure an attribute, each one is scored dichotomously, and the attribute is also coded dichotomously, one has a total of four score points from 0 (all items incorrect) to 3 (all items correct) available to make this simple (mastery/non-mastery) classification distinction. The discriminatory power of items acts effectively as weights for items. So with more high-quality items and/or more score points available for each item one would get more reliable/robust classification information but if one were to classify learners into more than two mastery states this would be traded off again.

The answer also depends on the true (and unknown) distribution of the learners across the latent classes, which includes knowing whether certain latent classes should be excluded a priori through attribute hierarchies. It is generally helpful to remember that the number of high-quality items for each attribute/ dimension helps with the classification challenge for learners while the number of learners with distinct, item-relevant profiles in turn helps with the estimation of the parameters for the items. Sufficient information is needed for both and, 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 for specific situations. For example, it does not matter per se whether a sample has 10,000 learners or not; what matters is how many of these learners have responded to each item. Similarly, it does not matter per se whether an item bank has 200 items or not; what matters is which items different sets of learners see on their form (in a matrix-sampling setting) and what the distribution of these forms across the attributes is (e.g., one would not want a notable imbalance of one attribute only being measured by one item while others are measured by five or six items). The good news is that there are quite a few simulation studies out nowadays (or that can be internally conducted as needed) that can provide more precise answers to these questions for particular scenarios of interest (see, e.g., Sen & Cohen, 2021). It is most important though to understand that the design of the study needs to match the targeted use context(s) of interest (as always).

20. What kinds of statistics can I get from DCMs?

Similar to other modeling approaches, there are variations on item statistics and learner statistics for these models as well as statistics that help one to evaluate the model-data fit. For example, one can get:

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

Depending on the model that one is estimating and the estimation approach that one is choosing one gets slightly different sets of parameter estimates. Helpful overviews can be found in different chapters within the most recent technical handbook (von Davier & Lee, 2019); for reliability-type indices see especially chapter 17 within the book.

21. How do we estimate DCMs?

As always, to create a sustainable data-collection, scoring, and reporting ecosystem with DCMs, the internal computational ecosystem of processing routines, databases, and other architectural components has to be designed / adapted from past practices carefully. These efforts are typically implemented by an interdisciplinary team composed of specialists in more traditional areas such as psychometrics, data science / engineering, and information technology as well as more recently matured areas such as artificial intelligence; however, it always depends on how an organization is structured. There are now several different packages available that can be used to estimate these models. Some of these approaches rely on independent / commercial software packages for which special code is created (e.g., Mplus, FlexMIRT, mdltm), others rely on freely available code packages in open-source environments (e.g., CDM, G-DINA, or mirt in *R*, specialized code in *OpenBUGS*). Reading descriptions of studies that estimate these models reveals a variety of cautions for estimation by authors, which should be carefully considered as stakes for score interpretations increase and more stringent audit / peer review requirements have to be met.

Further Learning

21. How can I learn more about DCMs?

There are two main volumes that speak to different audiences. The first book (Rupp, Templin, & Henson, 2010) is a more introductory book for applied practitioners (first half) and applied analysts (second half) that still contains valuable overall descriptions with a strong didactic flavor. However, it is also a bit dated in later chapters, especially in areas of model-data fit and modeling extensions where the toolbox has definitely expanded in the 12 years since its initial publication. The second book (von Davier & Lee, 2019) is an edited volume that represents an excellent snapshot of a more recent state of the art with quite a few concepts from the first book reappearing, but it is definitely geared at a more technical audience. It has multiple brief examples rather than the extended step-by-step walkthrough in *Mplus* of the first volume, which are more for illustrative purposes. The latter volume contains descriptions of how different programs can be used to estimate these models in a dedicated section, which include the independent / commercial programs *Mplus*, *FlexMIRT*, and mdltm as well as the *cdm* and *GDINA* packages in *R*.

Similarly, the literature is filled with a variety of simulation studies that get at different aspects of the estimation of these models. It is generally best to work internally across areas so that there is a minimum common understanding about the general properties of these models for content teams, product development teams, and psychometric teams and then let the psychometric teams educate the other teams about advances in the state-of-the-art that might affect what is possible. Identifying tension points between desired design, implementation, and reporting solutions and the data requirements that these models have early on can be valuable for foreseeing bottlenecks or roadblocks along the way.

Please cite this document as:

Rupp, A. A. (2023). *Primer on diagnostic classification models*. Dover, NH: Center for Assessment.

Select References

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

Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (2016). *Handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297-327). Malden, MA: Wiley-Blackwell.

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2021). *A theoretical assessment ecosystem for a digital-first assessment – the Duolingo English Test* [Duolingo Research Report DRR 21-04]. Retrieved Feb 6, 2023 from https://duolingo-papers.s3.amazonaws.com/other/det-assessment-ecosystem.pdf

Choi, Y., & Mislevy, R. J. (2022). *Evidence centered design framework and dynamic Bayesian network for modeling learning progression in online assessment system*. Frontiers in Psychology. Retrieved Jan 11, 2023 from https://www.frontiersin.org/articles/10.3389/fpsyg.2022.742956/full

Dai, S., & Validivia, D. S. (2022). Dealing with missing responses in cognitive diagnostic modeling. *Psych, 4*, 318-342. Retrieved Jan 31, 2023 from https://www.mdpi.com/2624-8611/4/2/28

Kalkan, Ö. K., Kelecioğlu, H., & Başokçu, T. O. (2018). Comparison of cognitive diagnosis models under changing conditions: DINA, RDINA, HODINA and HORDINA. *International Education Studies, 11*, 119-131. Retrieved Jan 31, 2023 from https://files.eric.ed.gov/fulltext/EJ1181058.pdf

Li, X., Wang, W.-C., & Xie, Q. (2020). Cognitive diagnostic models for rater effects. *Frontiers in Psychology*. Retrieved Jan 11, 2023 from https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00525/full

Ma, W., & de la Torre, J. (2019). Digital Module 05: Diagnostic Measurement – The G-DINA Framework. *Educational Measurement: Issues and Practice, 38*(2), 114-115.

Pan, Q., Qin, L., & Kingston, N. (2020). Growth modeling in a diagnostic classification model (DCM) framework: A multivariate longitudinal diagnostic classification model. *Frontiers in Psychology*. Retrieved Jan 11, 2023 from https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01714/full

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *19*, 1-33.

Rupp, A. A., & Templin, J. (2011). 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.

Rupp, A. A., Templin, J., & Henson, R. H. (2012). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford Press.

Sen, S., & Cohen, A. S. (2021). Sample size requirements for applying diagnostic classification models. Frontiers in Psychology. Retrieved Jan 1, 2023 from https://www.frontiersin.org/articles/10.3389/fpsyg.2020.621251/full

Shan, N., & Wang, X. (2020). Cognitive diagnosis modeling incorporating item-level missing data mechanism. Frontiers in Psychology, 11. Retrieved Jan 31, 2023 from https://www.frontiersin.org/articles/10.3389/fpsyg.2020.564707/full

Sessoms, J., & Henson, R. A. (2018) Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 1-17

von Davier, M., & Less, Y.-S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. New York, NY: Springer.