Cognitive Diagnostic Models (CDMs): A Gentle Introduction and Exploration of Their Potential Use for Formative Assessment

Harold Doran, Steve Ferrara, Hye-Jeong Choi, and Nnamdi Ezike of HumRRO

"Nira had a scaled score of 627 on the fall formative assessment and her scaled score on the spring assessment is now 642. Nira grew by 15 scaled score points!"

Statements like this are common on score reports. In many respects, they have little instructional utility. Drawing meaningful inferences from these numbers regarding an examinee in terms of what they can and cannot do is virtually impossible. Often, score reports use a cut score and classify examinees into performance levels using language like "Nearing Proficiency" or "Proficient." These classifications are indeed a bit more helpful, but they are often very broad statements about what students know and can do and do not provide specific, targeted statements about examinee abilities.

Enter Cognitive Diagnostic Models (CDMs), a class of psychometric methods that provide a different way to describe how an examinee performed on a test.[1] Rather than reporting a scaled score, CDMs provide a score profile that describes what specific skills and concepts on the test an examinee appears to have mastered or not mastered. The entire idea is to provide diagnostic information to educators and parents that can guide instruction instead of providing a broad statement about performance like scaled scores do. CDMs can be particularly useful for formative assessments (e.g., interim and through-year assessments). Interest in CDMs in school districts and the measurement community is growing, as evidenced by recent blogs and a webinar sponsored by the Diagnostic Measurement special interest group of the National Council on Measurement in Education (see https://www.ncme.org/community/ncme-sigimie/diagnostic-measurement).

In this blog, we describe CDMs and offer considerations relevant to practical, operational use of CDMs in state and district assessment programs. We start from the proposition that the term "diagnostic" in the title, Cognitive Diagnostic Models, refers to providing **actionable feedback** on skills and concepts that students likely have mastered, and their specific learning needs. To us, this is about formative assessment, in that teachers can use the feedback from CDMs to design instruction around learners' needs.

We introduce CDMs and describe (a) how they differ from Item Response Theory (IRT) models; (b) how they differ from what is most familiar to all of us, state accountability and district formative assessment; (c) common CDM terminology; (d) some potential applications; and (e) some general considerations for their use. We also describe CDMs in the context of educational assessments; however, their applications in certification and selection testing are natural extensions.

[1] Sometimes called Diagnostic Classification Models (DCMs)

Overview

Cognitive Diagnostic Models (CDMs) are a class of psychometric methods that support instructionally useful assessments in ways that state accountability in their most common form do not. CDMs are used in a wide array of contexts, such as K-12 formative assessment and professional certification testing.

While CDMs are not a new development in psychometrics (e.g., Junker & Sijtsma, 2000; Tatsuoka, 2009; Leighton & Gierl, 2007), they are now emerging as a popular supplement, or alternative, to traditional IRT ways of assessing. Why? Probably because they provide actionable feedback on student learning needs and because of their potential to support through-year assessment.

Perhaps the most direct application of CDMs is in of Through Year Assessments (TYAs) and other formative testing models now becoming more dominant in the US. Because CDMs report scores in very different ways from traditional IRT-based test score reporting, they can be used by educators in very different ways as well.

What Are Cognitive Diagnostic Models? How Do They Differ from IRT Models and Current State and District Testing Practice?

The main result of CDMs is an examinee *skill profile* which indicates whether an examinee appears to have mastered or not mastered specified content measured by the test. CDMs do not produce a total test score, as in IRT applications. We provide an example of a skill profile in the next section. The main point is that the skill profile is intended to be diagnostic and very specific whereas a total test scaled score is intended to be general and summative.

It may be helpful to relate CDMs to the more widely familiar IRT as a starting point and then compare the two methods. First, both IRT and CDMs estimate item parameters. IRT models estimate item difficulty parameters and item discrimination and guessing parameters (in two-parameter and three-parameter IRT models). In contrast, CDMs estimate item *slipping* and *guessing* parameters. We define these parameters below.

Table 1

Comparison of IRT Analysis and CDM Analysis

	IRT	CDM
Requires response data	Yes	Yes
Estimates item parameters	Yes	Yes
Generates total scores	Yes	No
Generates skill profiles	No	Yes

Both IRT and CDMs use students' correct and incorrect responses to test items to estimate item parameters and produce total test scores (in IRT) or skills and concepts profiles (in CDM). Finally, test design for IRT and CDM applications differ in important ways. Both should be guided by principled assessment design (PAD), of course. The starting place in PAD is what you want to be able to say about examinees, based on their test performances, and what you will assess. In state and district assessments where IRT is applied, test designs are described in test blueprints. Test blueprints indicate the numbers of items that are allocated to each content standard that is included in a test. In CDM, a Q-Matrix defines the skills and concepts that will be assessed, and that make up the skills and concepts profiles that are reported. A Q-Matrix may contain a small collection (e.g., 3-8) of state content standards (e.g., adding and subtracting fractions) or more granular skills and concepts like adding and subtracting fractions with like and unlike denominators.

Examinee Skills and Concepts Profiles

Let's first examine the main result of a CDM—a student *skill profile*. Table 2 shows a simple skills and concepts profile for two examinees and an assessment that measures four skills. We use the generic labels of "Skill A," "Skill B," and so on for convenience. In a real-world context, these skills (and concepts) would be meaningful labels (e.g., subtracting two-digit numbers, decoding multi-syllabic words), as in Table 2.

Notice the "1s" and "0s" in the profile in Table 2, rather than the total test score usually found on test score reports. In this context, a "1" indicates that there is evidence that the examinee has demonstrated mastery of the skill or concept and a "0" indicates that there is no evidence of content mastery. In this case, Examinee 1 appears to have mastered Skills A, B, and D, but has not yet mastered Skill C. Examinee 2 appears to have mastered only Skill D. Other profiles might replace the 1s and 0s with the probabilities of mastery estimated by the CDM.

Table 2

Simple Skills and Concepts Profile

Examinee	Skill A	Skill B	Skill C	Skill D
Examinee 1	1	1	0	1
Examinee 2	0	0	0	1

This method of reporting examinee test performance is different from what state and district assessments report. When traditional IRT psychometric methods are used, a total score is generated—usually a scaled score. This scaled score is then mapped to a performance level which has been defined in a standard setting process that indicates performance that is Proficient or Nearing Proficiency (for example). In addition, IRT sub-domain scores often are developed to describe an examinee's performance on smaller parts of a test, such as Number Sense and Geometry.

In contrast, CDMs do not produce overall scaled scores in this way. The skills and concepts profile for each examinee is the final output. Profiles indicate the presence or absence of specified skills and concepts. They are intended to guide instructional planning. Using the profiles in Table 2, Examinee 1 appears to need additional instruction only on Skill C while Examinee 2 needs assistance with Skills A, B, and C.

What is a Q-Matrix?

Before a test can be administered and before CDMs can generate skills and concepts profiles, a few prerequisites must be addressed. First, of course, a pool of test items themselves that will appear on a test form must exist or must be developed for the CDM analysis. For this discussion, we assume that content experts have developed these items specifically to assess Skills A, B, C, and D. Further, we assume that the content experts have worked together to label the skills and concepts that are required for examinees to answer each item correctly. Some items are linked with only one skill or concept required; other items may be linked with two, three or even four. The content experts may even have consulted learning sciences research related to how students learn the skills and concepts to be assessed, their progression from novice to mastery in each area, and the misconceptions they develop along the way.[2] Their goal is to identify a set of skills and concepts and items that assess that knowledge at levels of complexity consistent with the skills and concepts.

The content experts will have followed a PAD approach to designing the Q-Matrix and developing items. They will decide what they want to know and be able to say about students' skills and concepts in a specific area of the curriculum. In this made-up example, they have selected four skills and concepts that define that area of the curriculum and developed items that are closely aligned with the four skills.

Simply to illustrate, suppose we now have a test form consisting of 10 total test items. This implies two important concepts. First, content experts must define the set of skills that the test is designed to assess. In this made-up example, experts have determined that this formative assessment intends to assess four skills labeled Skills A, Skill B, Skill C, and Skill D. Then, content experts work together to determine which skill and concept is required by each test item. In this example, the experts have determined that Item 1 requires Skills A and B but not C or D. Item 2 assesses Skills A and C but not B or D, and Item 3 assesses only Skill D.

The end result of this process is a complete Q-Matrix that has a set of 1s and 0s for all items that were assigned by experts. Table 3 provides on sample of a Q-Matrix. This Q-Matrix is then passed to a psychometric team that uses it to complete two tasks: (a) Calibrate the slipping and guessing parameters for all items, and (b) It will be used when generating the skills and concepts profiles.

[2] In addition, they could conduct cognitive labs, in which students think out loud as they process and respond to items. And the psychometricians can help: CDM calibration may indicate that some items don't fit the Q-Matrix as intended. Content experts can use the psychometric feedback to revise or replace misfitting items to maximize item alignment to the Q-Matrix.

A Comment on Q-Matrices

People who talk about CDMs almost always refer to assessing student mastery of skills and competencies—thus the term *skill profile*. This term appears to overlook the other significant part of school teaching and learning, concepts. For example, a **concept** in elementary school mathematics is the necessity of finding a common denominator in order to add or subtract fractions. The **skill** that is required to find a common denominator requires applying the *least common multiple* or *cross multiplication* approaches. It is easy enough to think about other examples in mathematics, English language arts, science, and social studies. The Common Core Standards in ELA and mathematics and Next Generation Science Standards make the skills and concepts distinction clear—and emphasize that learning and applying skills and concepts go hand in hand.

CDM Item Parameters

As in IRT, psychometricians estimate CDM item parameters. In IRT, we typically estimate difficulty, discrimination, and guessing parameters (for unidimensional models, such as the 3 PL model). For CDMs, we use examinee responses provided to the test items to estimate *guessing* and *slipping* parameters.

The guessing parameter enables a straightforward interpretation. It represents the probability that an examinee will answer an item correctly when they actually **do not** possess the skills and concepts required to do so. The slipping parameter is the opposite—it represents the probability that an examinee will not answer an item correctly even though they **do** in fact possess the skills and concepts needed to answer the item correctly.

CDMs: Compensatory and Non-Compensatory Models

Let's assume we now have the following things in hand:

1) Examinee responses to items on a test

2) A Q-Matrix that assigns those items to a specified set of skills and concepts

We almost have what we need to finish the task and score the examinees. What we need now is a decision on which type of model to use. There are many CDMs. Here, we consider just two common versions known as the deterministic inputs, noisy "and" gate (DINA) and the deterministic inputs, noisy "or" gate (DINO) model.

The DINA model is considered a *non-compensatory* CDM because the examinee is expected to possess all required skills and concepts in order to respond correctly to an item. For example, in the DINA CDM model, an examinee is expected to possess both Skills A and B in order to correctly respond to item 1 in our Q-Matrix above (see Table 3). In contrast, in the DINO model, an examinee is expected to have at least one of the skills needed to correctly answer the item, but not all are required to determine mastery of a skill or concept in a Q-Matrix. Again, referring to Item 1, DINO is *compensatory* in the sense that an examinee must have either Skill A or Skill B (or both) in order to correctly respond to item 1.

Table 3

A Q-Matrix with 10 Test Items, Each Labeled as a Requirement (1=yes) for Four Skills/Concepts

Test Item	Skill A	Skill B	Skill C	Skill D
Item 1	1	1	0	0
Item 2	1	0	1	0
Item 3	0	0	0	1
Item 4	0	1	0	0
Item 5	…	…	…	…
Item 6	…	…	…	…
Item 7	…	…	…	…
Item 8	…	…	…	…
Item 9	…	…	…	…
Item 10	…	…	…	…

Note. In a real-world example, the content experts would complete the entire Q-Matrix for all 10 items.

Once the model is selected, the psychometric team has all they need to (a) estimate the CDM guessing and slipping parameters, and then (b) generate the individual student skills and concepts profiles.

CDMs for Formative Assessment: Interim and Through-Year Assessments

You have probably noticed by now that we have not discussed applying CDM analysis to statewide accountability tests. That is because we think the most fruitful applications are for formative assessments that are tailored to instruction, that can provide actionable feedback to teachers about which of their students have probably mastered which content standards—that is, learning outcomes—and their current learning needs. The most prevalent commercial offerings can be referred to as interim assessments. The hottest topic in discussions these days is through-year assessments.

Periodic assessments offered by vendors as *interim assessments* are intended for administration in the fall, winter, and spring. At each time of administration, the test blueprint covers the full range of content standards that is covered in the spring statewide accountability test. Scores from interim assessments typically are intended to inform teachers and school and district leaders about areas where students are performing well and areas where additional learning is needed. And they typically are intended to predict student performance on the spring accountability test.

Feedback on student performance and learning needs might be useful at the school and district levels but may be less useful for classroom instructional decisions.

Through-year assessment (TYA) designs are receiving widespread attention these days. The prevailing design concept here is that a fall version of this formative assessment, for example, would assess only those content standards that are covered during the fall instructional period. The TYA administered in the winter might assess only those content standards that were covered during that instructional period and so forth. TYAs are often referred to as "curriculum-aligned" or "instructionally-aligned" and follow school district curriculum pacing guides. CDM analysis and reporting may be most appropriate for TYAs, where the Q matrix can focus on skill and concept profiles that are consistent with the content standards covered during a specific period of instruction. Vendors must offer considerable flexibility in creating Q-Matrices that align with differences among school district pacing guides.

Final Notes on Operational Uses of CDMs

We have proposed that CDMs have great potential as part of formative assessment practices in schools and districts. These include interim and TYAs. CDMs also have been used effectively in professional credentialing and selection testing.

There are challenges to using CDMs for formative assessments, as there were in the early 1980s, when assessment programs began implementing IRT models. In the early 1980's, IRT was a novel idea and it was rarely used in operational testing programs. Now, IRT is the *de facto* standard and used in virtually every educational testing program. Today, CDMs are emerging as a novel application and, like IRT in the 80s, it is rare to find them in operational use.

One particular challenge is the reliability and validity of the skill profiles. If the total test length is small, as we might find in a formative test, then the skill profiles might be based on a very small number of items per dimension. This is a test design issue to consider and a topic that psychometricians continue to explore. Templin and Bradshaw (2013) compared IRT and CDM test score reliabilities. They argued that CDM requires fewer items to achieve score reliability that is similar to a test with more items and calibrated using the 2PL IRT model.

A final challenge is that of software. IRT software is widely available, well-studied and documented, and all psychometricians are trained to use a wide array of IRT estimation programs. CDM software is an emerging area of development and new software applications are growing and becoming more widely available.

References

Junker, B., & Sijtsma, K. (2000). Cognitive assessment models with few assumptions, and connections with nonparametric IRT. *National Research Council*.

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

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*(2), 251-275.