Diagnostic Test Design with Blueprint Specifications

Matthew J. Madison and Nancy Alila

University of Georgia

Paper presented at the annual meeting of the National Council on Measurement in Education in Denver, CO.

Diagnostic Test Design with Blueprint Specifications

Abstract

Diagnostic classification models (DCMs) are psychometric models designed to support diagnostic assessment score interpretations by providing proficiency classifications. To achieve their potential in providing valid and reliable classifications, a DCM should be paired with a suitable diagnostic assessment with predefined and operationalized attributes. Oftentimes, these attributes are defined within a diagnostic test blueprint, which can include specifications of content areas, levels of rigor or cognitive complexity, item type, and more, depending on the context. The purpose of this study is to systematically examine DCMs' ability to approximate prespecified test blueprints. Results from two simulated diagnostic test scenarios confirmed prior research suggesting that general DCMs struggle to approximate prespecified test blueprints. A recently developed constrained DCM, however, was successful in approximating the prespecified blueprints with high accuracy and precision. We discuss the implications of these results for diagnostic assessment operations.

Keywords: diagnostic classification model, cognitive diagnosis model, diagnostic assessment, test design, assessment blueprint, item influence, attribute information

Diagnostic Test Design with Blueprint Specifications

Introduction

Diagnostic classification models (DCMs; Rupp, et al., 2010), also known as cognitive diagnosis models (CDMs), are modern psychometric models that are designed to support diagnostic assessment score interpretations. In contrast to traditional psychometric frameworks like item response theory (IRT) that provide scores in the form of examinee ability scalings or rankings, DCMs provide scores in the form of attribute classifications. These attribute classifications can indicate student proficiencies and understandings, as well as areas needing remediation. To achieve their potential in providing accurate and reliable attribute proficiency classifications, a DCM should be paired with a suitable diagnostic assessment with predefined and operationalized attributes.

Assessment specialists often create test blueprints to guide assessment development efforts. These test blueprints can include specifications of content area, levels of rigor or cognitive complexity, item type, and more, depending on the context. When a test closely reflects the intended blueprint, construct and content validity are increased, as well as the validity of score interpretations (Cronbach, 1971; Sireci, 1998). In a classical test theory or an IRT framework, test blueprints can generally be approximated by representing content areas or topics proportionally in the number of items. For example, if mathematics test developers wanted problem solving with fractions to represent 25% of the test, it would generally suffice to allocate 10 out of 40 total items to measure problem solving with fractions. For DCMs, however, this is generally not the case. Jurich and Madison (2023) showed that for a general DCM and a fixed form test, empirical test blueprints cannot be controlled using item representation, only estimated, reported, and potentially adjusted with a revised test form. That is, the proportion of

items measuring a specific content area did not align with the empirical weight of those items on the attribute classifications. If DCM classifications do not reflect prespecified test blueprints, it would pose a significant limitation in the application of DCMs because the validity of model-based classifications would be dubious. To overcome this limitation, Madison et al. (2024) proposed a constrained DCM and with an empirical example, showed that in limited settings, it had the potential to control empirical item weights, thereby approximating prespecified test blueprints. The purpose of this study is to more systematically examine DCMs' ability to approximate prespecified test blueprints. More specifically, we simulate two realistic test scenarios to evaluate the congruency of DCM empirical test blueprints and prespecified test blueprints. First, we describe the two DCMs used in this study: a general DCM, the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) and a constrained DCM, the one-parameter log-linear cognitive diagnosis model (1-PLCDM; Madison, et al., 2024). Then we describe *item influence* (Jurich & Madison, 2023) and how it provides a mechanism to quantify empirical test blueprints and compare them to prespecified test blueprints. Then we move to the simulated scenarios. Finally, we summarize the results of the simulations and discuss the implications of the results.

Diagnostic Classification Models

DCMs can be simply described as siblings to IRT models, with the distinguishing feature being the categorical latent traits. In educational settings, these categorical latent traits, or *attributes*, typically represent skills, domains, or topics to be assessed as a part of the learning process. While reducing the latent trait from continuous to categorical may seem like a small modification, this assumption leads to different interpretations and a different statistical structure. While IRT models support norm-referenced scalings (e.g., Maya is at the 92nd

percentile in reading comprehension), DCMs support criterion-referenced interpretations through attribute proficiency classifications (e.g., Sadie is proficient in identifying numbers 0 to 100 and writing numbers 0 to 20, but is not yet proficient in counting to 100 by tens). Structurally, DCMs are confirmatory and constrained latent class models (Collins & Lanza, 2010). They are confirmatory in two ways: (1) the latent classes are specified a priori as the different patterns of attribute proficiency, or attribute profiles; and (2) the item-attribute alignment is specified a priori in the Q-matrix (Tatsuoka, 1983). DCMs are constrained due to the constraints they place on item parameters; to reflect the assumption that attribute proficiency leads to increased correct response probabilities, DCMs constrain item-attribute effects to be positive. Like IRT, there are several different DCMs that make assumptions about the underlying item response generation process, and these assumptions are reflected in the item response function. Below, we introduce the two DCMs used in this study, the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) and the one-parameter log-linear cognitive diagnosis model (1-PLCDM; Madison, et al., 2024).

Log-linear Cognitive Diagnosis Model

The LCDM is a general DCM that models the probability of a correct response as a function of attribute proficiency status and item-attribute effects, including intercepts, main effects, and interaction terms. To see this more concretely, Equation 1 displays the log-odds of a correct response to an item measuring Attributes 2 and 3:

$$P(X_i = 1) = \lambda_{i,0} + \lambda_{i,1(2)} \cdot \alpha_2 + \lambda_{i,1(3)} \cdot \alpha_3 + \lambda_{i,2(2,3)} \cdot \alpha_2\alpha_3 \quad (1)$$

In Equation 1, the intercept, $\lambda_{i,0}$, represents the log-odds of a correct response for an examinee who is not proficient in Attribute 2 nor Attribute 3. The Attribute 2 and 3 main effects, $\lambda_{i,1(2)}$ and $\lambda_{1(3)}$, represents the increase in log-odds of a correct response for an examinee who is proficient

in Attribute 2 or Attribute 3, respectively. The Attribute 2 and 3 interaction term, $\lambda_{i,2(2,3)}$, represents the additional increase in the log-odds of a correct response for an examinee who is proficient in Attribute 2 and Attribute 3. The LCDM is considered a general model because many other DCMs can be specified by placing constraints on LCDM effects. For example, the popular the deterministic-inputs, noisy-and-gate model (DINA; e.g., Haertel, 1989; Junker& Sijtsma, 2001) can be obtained by constraining LCDM main effects to be zero. General DCMs like the LCDM are preferred because they allow for a flexible and proper model-building process, whereby effects can be reduced on an item-by-item basis if necessary. The potential downside of general DCMs is that they can be difficult to estimate and interpret with highly complex Q-matrices.

One-parameter Log-linear Cognitive Diagnosis Model

The one-parameter log-linear cognitive diagnosis model (1-PLCDM; Madison, et al., 2024) is a special case of the LCDM. In the single attribute case, the 1-PLCDM is analogous to the one-parameter logistic (1-PL) IRT model, estimating a single main effect across all items. In multiattribute settings, the 1-PLCDM estimates a single main effect for each item measuring each respective attribute. The 1-PLCDM can easily be extended to polytomous attributes by constraining each level's effects to be equal (Madison, et al., 2024). The 1-PLCDM has been shown to have properties akin to the 1-PL IRT and Rasch models, including sum-score sufficiency, invariant item and person ordering, and person-free and item-free classifications (Madison, et al., 2023; 2024). We note that in order to achieve the sum-score sufficiency property in the multiattribute setting, attributes are assumed independent and the Q-matrix must be simple structured (i.e., all items measure a single attribute). Studies have shown that the 1-

PLCDM is relatively robust to violations of the assumptions and requirements (Madison, et al., 2023; Mass, et al., 2024)

Item Influence and Diagnostic Test Blueprints

In linear regression, it is common to observe observations that have an undue influence on the estimated regression line and predictions (Belsley, Kuh, & Welsch, 1980). Jurich and Madison (2023) defined an analogous concept, *item influence*, to describe how individual items contribute to DCM classifications. They developed several metrics to quantify item influence. For example, one metric, which they termed item override, is defined as the proportion of examinee classifications that change if an item is omitted from the analysis. The metric we use in this study, *proportion of attribute information*, quantifies how much of an attribute's total information is contained within each item. This metric uses the cognitive diagnostic index (Henson & Douglas, 2005), which quantifies how strongly items distinguish between examinees with different attribute profiles. If a test measures a single attribute with 10 items, and each item is contributing equally, then the proportion of attribute information for each item will be 10%. To the extent that items are differentially providing information about the attribute, we would observe different values for the individual items.

The proportion of attribute information metric provides a mechanism to quantify DCM empirical attribute definitions and blueprints. To define empirical blueprints, we can add the proportion of attribute information for items allocated to measure certain domains or topics. For example, suppose a prespecified blueprint indicates that a domain contains three subdomains with weights of 50%, 30%, and 20%, respectively. Further suppose that to reflect these prespecified weights, test developers wrote Items 1-5 to measure Subdomain 1, Items 6-8 to measure Subdomain 2, and Items 9-10 to measures Subdomain 3. Then we would calibrate the

DCM and sum the proportion of attribute information for Items 1-5, Items 6-8, and Items 9-10. If the DCM empirical blueprint adheres to the prespecified blueprint, then the sums will approximate 50%, 30%, and 20%, respectively. If the empirical blueprint proportions deviate drastically from the prespecified blueprint proportions, construct and content validity come into question, and interpreting classifications with respect to the measured attributes as predefined is challenging.

Diagnostic Assessment Blueprint Demonstration

In this section, we present two realistic diagnostic assessment scenarios where approximating the test blueprint is critical for validity and interpretation. We describe each scenario, detail the simulation study, and summarize results with the goal of comparing a general DCM (the LCDM) and the 1-PLCDM's ability to approximate the desired test blueprint. Each simulation was conducted in R, Version 4.1.1, using the CDM package (George, et al., 2016) for LCDM estimation, TDCM package (Madison, et al., 2025) for 1-PLCDM estimation, and mirt package (Chalmers, 2012) for polytomous LCDM and 1-PLCDM estimation.

Scenario 1: Single Domain with Subdomains

In this first scenario, suppose we are interested in developing a test to be used for summative purposes. For example, the state of North Carolina (NC) in the United States has end-of-course mathematics tests for Grades 3 – 8 (North Carolina Department of Public Instruction, 2022). Within each grade's test, there are blueprints that outline content and rigor specifications. For example, the NC Math 1 Exam covers five domains (Number and Quantity and Algebra; Functions; Geometry; Statistics and Probability) with weight distribution ranges of 36-40%, 32-36%, 8-12%, and 18-20%, respectively. After taking the assessment, each student is classified (using unidimensional IRT cut-scores) into one of four achievement levels (Not Proficient, Level

3, Level 4, Level 5). Here, we map this scenario onto a DCM framework where there is a single attribute with four levels and four sub-attributes.

Simulation Study. Without loss of generality, we approximated the NC Math 1 test and simulated a single attribute (e.g., domain) with four subattributes (e.g., subdomains) represented with prespecified blueprint proportions of 40%, 28%, 12%, and 20%, respectively. To reflect these proportions, we simulated a diagnostic test with a total of 25 items: ten, seven, three, and five items for the four subattributes, respectively. We generated 2000 examinees with base-rates of .50, .15, .15, and .10 for the four attribute levels (e.g., NCDPI, 2022). Items parameters were generated such that Level 1 examinees had correct response probabilities randomly drawn from a uniform distribution ranging from .00 and .25. Each subsequent level (Levels 2, 3, 4) increased the probability of a correct response by between .05 and .30, pulled randomly from a uniform distribution.

Within each replication, we estimated the LCDM and the 1-PLCDM. For each model, we calculated the proportion of attribute information for each item. Then, we summed the proportions of attribute information for all the items within each subattribute to obtain the empirical blueprint. If the model is true to the prespecified blueprint, then the resulting proportions will approximate the blueprint proportions for the subattributes (40/12/28/20). We replicated this scenario and analysis 500 times, recording the empirical blueprint proportions for each replication. To evaluate, for the LCDM and the 1-PLCDM, we summarized the distribution of the subattributes' empirical weights.

Simulation Results. Results show that on average, both models had mean subattribute weights approximately equal to the prespecified blueprint. Figure 1 shows, however, that the 1-PLCDM (red histogram) is much more precise than the LCDM (blue histogram); the blue and

red dashed vertical lines represent the middle 90% of the subattribute weight distributions for the four subattributes for the LCDM and 1-PLCDM, respectively. The LCDM's 90% intervals were, on average, 3.7 times wider than the 1-PLCDM. To make these results more concrete, let us consider Subattribute 1, which had a prespecified blueprint weight of 40%. The 1-PLCDM's worst replications were 35% and 45%, an error of 5%. On the other hand, 35% of the LCDM's replications had more than 5% error for Subattribute 1. The LCDM's worst replications were 25% and 57%, quite a drastic deviation from the desired blueprint weight of 40%. .

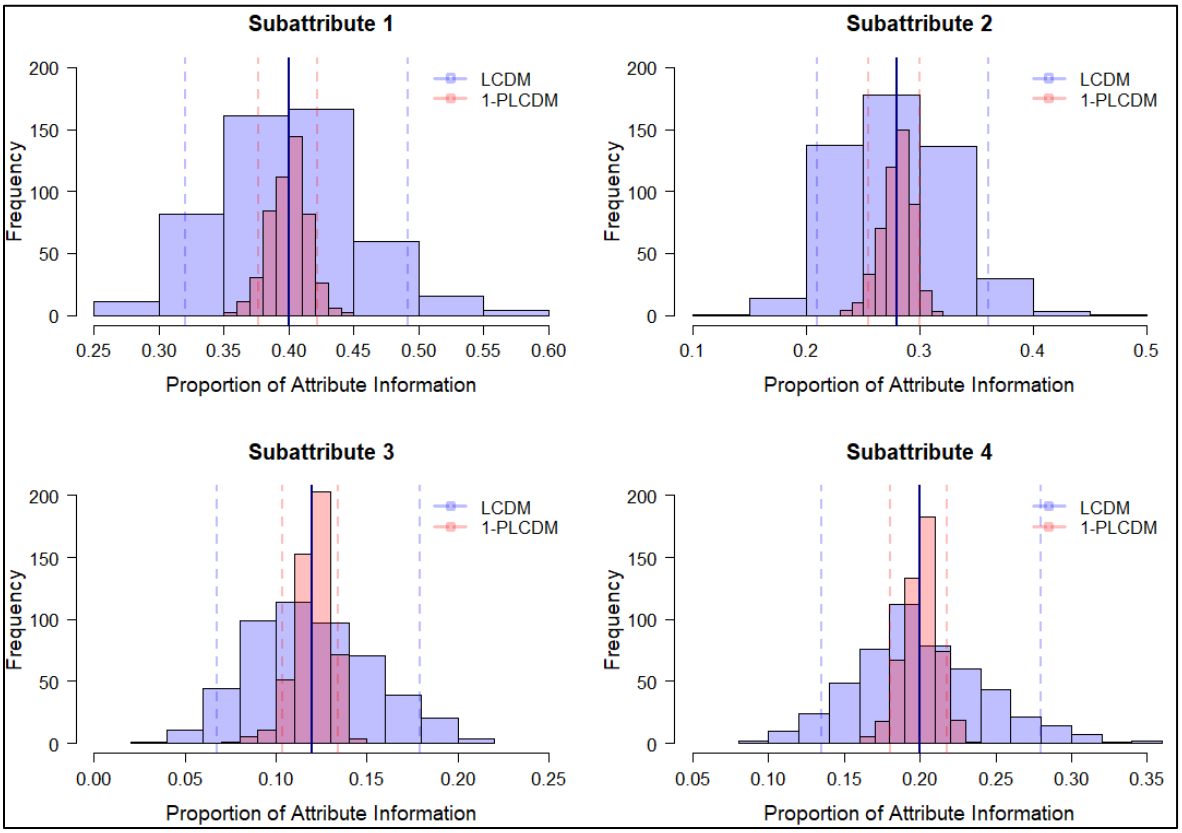

Figure 1. Simulation Study #1 Subattribute Empirical Blueprint Distributions

Another way to quantify error in blueprint estimation is to combine the error across all subattribute weights. We will call this statistic the combined absolute percent deviation (CAPD).

For example, if a model had empirical blueprint weights of 35/33/10/22, and the prespecified blueprint weights were 40/28/12/20, then the CAPD would be 14 ($|40-35|=5; |28-33|=5; |12-10|=2; |20-22|=2; 5+5+2+2=14$). We calculated the CAPD for both models and summarized the distributions. The 1-PLCDM had a mean CAPD of 4%, with a minimum of 0% and a maximum of 10%. The LCDM had a mean CAPD of 14%, with a minimum of 2% and a maximum of 37%. It is clear that in this case, for any given replication, the LCDM cannot be trusted to approximate the prespecified blueprint.

Scenario 2: Multiple Domains, Each with Subdomains

In this second scenario, suppose we are interested in developing a diagnostic test with three attributes, each with their own prespecified blueprint. This scenario can be seen in the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). For example, the Grade 4, Numbers and Operations cluster, *Build fractions from unit fractions* (4.NF.B), has two standards with four and three substandards, respectively. While the Common Core State Standards do not specify weights for any clusters or standards, it is easy to imagine these substandards having different weights. Here, we map this scenario onto a DCM framework with three attributes, each with multiple subattributes and their own prespecified blueprints. This scenario requires application of the multiattribute extension to the 1-PLCDM.

Simulation Study. Without loss of generality, we simulated three attributes. These three attributes had two, three, and two, subattributes, respectively. Attribute 1's two subattributes had prespecified weights of 50/50. Attribute 2's three subattributes had prespecified weights of 40/40/20. And Attribute 3's two subattributes had prespecified weights of 60/40. To reflect these proportions, we simulated a multiattribute diagnostic test. Attribute 1 had a total of six items,

three measuring Subattribute 1.1 and three measuring Subattribute 1.2. Attribute 2 had a total of ten items, four measuring Subattribute 2.1, four measuring Subattribute 2.2, and two measuring Subattribute 2.3. Attribute 3 had a total of five items, three measuring Subattribute 2.1 and two measuring Subattribute 3.2. As the model is limited to simple structure items, all items loaded onto a single attribute and had item parameters generated exactly like Simulation Study #1. We generated 2000 examinees with attribute level base-rates randomly pulled from a uniform distribution ranging from .25 to .75. Attribute correlations were also randomly pulled from a uniform distribution ranging from .25 to .75. We replicated this scenario and analysis 500 times, recording the empirical blueprint proportions for each replication and each model.

Simulation Results. Results followed the trend of Simulation Study #1 with averages for both models approximating the prespecified blueprint and the 1-PLCDM being much more precise. To see this, consider Attribute 1, which had two subattributes with prespecified weights of 50/50. The 90% intervals for the 1-PLCDM ranged from 49% to 51%, while the 90% intervals for the LCDM ranged from 41% to 59%. The worst replication for the 1-PLCDM had subattribute weights of 48% and 52%, while the worst replication for the LCDM had subattribute weights of 100% and 0%. When examining the totality of blueprint weight error across all three attributes, we observed that the 1-PLCDM had a mean CAPD of 1%, with a minimum of 0% and a maximum of 5%. The LCDM had a mean of 16%, with a minimum of 1% and a maximum of 255%. For the multiattribute context, perhaps more pronounced than Simulation Study #1 with the single attribute 1-PLCDM, we see that the LCDM cannot be trusted to approximate the prespecified blueprint; there is far too much variance and volatility in the empirical subattribute weights.

Discussion

This paper examined the ability of DCMs to approximate prespecified test blueprints. Test blueprints are used to guide test development and following blueprints allows for increased construct and content validity. Previous research has indicated that general DCMs are unable to approximate prespecified test blueprints with respect to attribute and subattribute weights. This result follows from the differential influence of individual items on attribute classifications (Jurich & Madison, 2023). A recently developed constrained DCM, however, the one-parameter cognitive diagnosis model (1-PLCDM), was shown in an empirical analysis to have some potential to approximate prespecified blueprints in a single attribute context (Madison, et al., 2024). In this study, we used two simulated scenarios to verify this result and extend to multiattribute assessment situations. In short, in contrast to more general DCMS, our analyses indicated that the 1-PLCDM approximated test blueprints with a high level of accuracy and precision.

In the first scenario, we simulated a summative assessment context where there is one broad domain with multiple subdomains with respective blueprint weights (e.g., 1st grade mathematics with algebra, operations, measurement, and geometry; NGA & CCSSO, 2010). Without loss of generality, we mapped this scenario onto a DCM framework with a single polytomous attribute and four subattributes with prespecified weights. In the second scenario, we simulated an intermediate or formative assessment context with multiple finer-grained domains, each with multiple subdomains. Without loss of generality, we mapped this scenario onto a DCM framework with three dichotomous attributes, each with multiple subattributes with prespecified weights. In both scenarios, we found that the 1-PLCDM and the LCDM were able to, *on average*, approximate the prespecified blueprint. The LCDM, however, had much more variation

in blueprint weight proximity than the 1-PLCDM. For example, for the subattribute with a prespecified weight of 40%, the LCDM empirical weights ranged from 25% to 57%, while the 1-PLCDM empirical weights ranged from 35% to 45%. In terms of total blueprint error in the multiattribute context, the LCDM had a median of 9% error, with many replications having more than 100% error. In the same multiattribute context, the 1-PLCDM had a median of 1% error, with the worst replication having 5% error.

These results have significant implications for diagnostic assessment design and analysis, as well as the interpretation of DCM classifications. In general, because most DCM applications use fewer items than IRT applications, DCMs are susceptible to item influence (Jurich & Madison, 2023). That is, individual items can differentially impact classifications. This poses a problem when items were written to adhere a prespecified test blueprint. We demonstrated that when using a general DCM, it is not feasible to approximate a prespecified blueprint. The extreme variability implies an unstable distribution of subattribute weights, which decreases the validity and reliability of model-based classifications. When using a general DCM, it is critical that classifications are interpreted in conjunction with their associated empirical weights.

If adhering to a prespecified blueprint is required for construct and diagnostic score validity, then using the 1-PLCDM is a preferred modeling option. The blueprint approximation ability of the 1-PLCDM stems from its constraining of item main effects, thereby minimizing differential item influence and allowing subattribute representation be reflected in the empirical weights. For example, in the 1-PLCDM, if Subattribute 1 is measured by 4 out of 10 total items, then Subattribute 1 will have an empirical weight of approximately 40%. This is not the case for general DCMs. The 1-PLCDM, however, is not universally applicable. It requires a simple structure Q-matrix, constrains main effects to be equal, and assumes independent attributes. For

tests designed with complex items, the 1-PLCDM may not provide the best fit. We note, however, that in our simulation study, we generated main effects to be different and generated attributes with correlations ranging from .25 to .75. Therefore, our simulation results indicate that even when the model is misspecified, the 1-PLCDM can approximate a prespecified blueprint. This result is aligned with previous studies indicating the robustness of the 1-PLCDM to model misspecifications (Mass, et al., 2024). Future research could explore the performance of these models under different conditions, as well as apply them to empirical data from diagnostic assessments. Additionally, the exploration of hybrid models or alternative constraints may offer further improvements.

In closing, this study demonstrated that it is possible for DCMs to adhere to prespecified test blueprints. We believe that this resolves a major issue limiting the practical application of DCMs. By sacrificing some model flexibility, the 1-PLCDM was able to reliability approximate prespecified test blueprints, ensuring that empirical attribute definitions align with operational attribute definitions. This alignment is critical for the validity and utility of classifications. This study also highlights the need for prospective assessment design that has proper alignment of assessment objectives and a psychometric model that supports those objectives. As curriculum designs and learning standards continually evolve, and new assessments are developed and applied to assess student knowledge, we hope that this study provides some insights into how DCMs can best be used to support diagnostic assessment efforts.

References

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics; Identifying Influence Data and Source of Collinearity*. Wiley, New York.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley

Cronbach, L. J. (1971). Test Validation. In R. Thorndike (Ed.), *Educational Measurement* (2nd ed., p. 443-507.

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), 1–24.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 333-352.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. Applied Psychological Measurement, 29(4), 262–277.

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

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258-272.

Jurich. D., & Madison, M. J. (2023) Measuring item influence for diagnostic classification models. *Educational Assessment*. https://doi.org/10.1080/10627197.2023.2244411

Madison, M. J., Wind, S. A., Maas, L., Yamaguchi, K., & Haab, S. (2023, August). *A one-parameter diagnostic classification model with familiar measurement properties.* Paper presented at the 2023 Joint Statistical Meetings in Toronto, Ontario, CA.

Madison, M. J., Wind, S. A., Maas, L., Yamaguchi, K., & Haab, S. (2024). A one-parameter diagnostic classification model with familiar properties. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12390.

Madison, M. J., Jeon, M. Cotterell, M. E., Haab, S., & Zor, S. (2025). TDCM: An R package for estimating longitudinal diagnostic classification models. *Multivariate Behavioral Research*.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington, DC.

North Carolina Department of Public Instruction. (2022). *End-of-Course NC Math 1 and NC Math 3 Tests North Carolina Test Specifications*. Raleigh, NC: Office of Accountability and Testing.

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

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83–117.

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