ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Statistical Applications to Cognitive Diagnostic Testing

Susu Zhang,[1] Jingchen Liu,[2] and Zhiliang Ying[2]

[1]Departments of Psychology and Statistics, University of Illinois Urbana-Champaign, Champaign, Illinois, USA

[2]Department of Statistics, Columbia University, New York, New York, USA; email: zy9@columbia.edu

Annu. Rev. Stat. Appl. 2023. 10:651–75

First published as a Review in Advance on November 9, 2022

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-033021-111803

Keywords

cognitive diagnosis, diagnostic classification, educational measurement, psychometrics, latent class analysis, adaptive learning, statistical learning, classification, clustering, supervised learning, unsupervised learning

Abstract

Diagnostic classification tests are designed to assess examinees' discrete mastery status on a set of skills or attributes. Such tests have gained increasing attention in educational and psychological measurement. We review diagnostic classification models and their applications to testing and learning, discuss their statistical and machine learning connections and related challenges, and introduce some contemporary and future extensions.

1. INTRODUCTION

In educational and psychological tests, numbers or labels are assigned to individuals in a coherent manner to represent hypothesized, unobserved (i.e., latent) constructs (Allen & Yen 2001, Cronbach & Meehl 1955). Statistically, this task can be approached with a measurement model that describes the relationship between the theorized construct(s) underlying the test and the observed behavioral data, such as a total score in classical test theory (Spearman 1904) or responses to a series of items (i.e., questions) in item response theory (IRT) (Lord 1980). Early work in psychometrics focused on the measurement of continuous traits in norm-referenced tests (Glaser 1963), whose purpose is to rank individuals on a continuum and to compare individuals with respect to a normative group. An example is an overall math proficiency score on a standardized college admission test, which is informative in college admission decisions but does not differentiate between individuals' fine-grained math skills. Nor do criterion-referenced tests provide diagnostic information on misconceptions or lack of cognitive skills leading to incorrect problem-solving.

In the late 1970s, cognitive diagnostic testing was developed to bridge the gap between psychometric theories of latent trait measurement and cognitive theories for problem-solving (e.g., Macready & Dayton 1977, Tatsuoka 2009). A response to a question is assumed to depend on the discrete mastery status of a specific set of latent skill(s) or knowledge, commonly termed attributes. The aim of cognitive diagnostic testing is hence to classify individuals into discrete latent classes of attribute mastery on the basis of their responses to a series of well-designed assessment items. Such discrete characterization is useful in many ways: to identify mistakes (e.g., misconceptions, lack of skill/knowledge) that explain unsuccessful problem-solving, to create diagnostic profiles of learners' mastery level of specific skills, and to inform learners and teachers as to directions for remediation and improvement. Here, testing is an integral part of the learning process that provides ongoing feedback, also known as formative assessment (Tyler et al. 1972). Cognitive diagnostic testing is most commonly applied to the assessment of mathematics and reading skills but is also applied to other disciplines, including the measurement of psychological disorders and other noncognitive skills and profiles (e.g., de la Torre et al. 2018, Templin & Henson 2006). Sessoms & Henson (2018) review the literature on specific applications of cognitive diagnostic testing.

This article provides an introduction to the application of statistical methods to cognitive diagnostic testing. The task of diagnosing latent attributes on the basis of observed test responses is closely connected to the statistical problems of clustering (e.g., Hartigan 1975, MacQueen 1967), classification (e.g., Bishop 1995, Fisher 1936, Ripley 1996), and latent class analysis (e.g., Langeheine & Rost 1988, Lazarsfeld & Henry 1968). With the practical demand of providing interpretable diagnostic classification, one defining characteristic of cognitive diagnostic testing is the incorporation of substantive expert opinion to define the domain of attributes underlying an assessment, which restricts the space of possible latent classes, as well as the set of attributes measured by each item (Tatsuoka 2009). In the following, we first introduce the general concepts and frameworks of cognitive diagnostic testing, including early work, a selection of parametric models, and nonparametric methods. We discuss the key statistical issues in cognitive diagnostic testing such as parameter estimation, identification, and model/variable selection. Statistical applications for specific scenarios, such as adaptive testing, tracking learning over time, and adaptive learning, are also briefly introduced. We conclude with a discussion of some ongoing and future directions.

2. APPROACHES TO COGNITIVE DIAGNOSTIC TESTING

2.1. Mastery Testing and Mastery Model

IRT models for the relationship between a continuous trait(s) and ordinal item response distribution predated cognitive diagnostic testing (Birnbaum 1968, Lord 1980). Under a dichotomous IRT model, the ith examinee's response to the jth item in a test ($i = 1, \ldots, N, j = 1, \ldots, J$), X_{ij}, is assumed to follow a Bernoulli distribution with success probability

$$P_j(\theta_i) = P(X_{ij} = 1 \mid \theta_i) = f(\theta_i, \boldsymbol{\zeta}_j), \qquad 1.$$

where θ_i is the latent ability of examinee i, $\boldsymbol{\zeta}_j$ denotes the parameters associated with item j (e.g., item difficulty), and $f(\cdot)$ is the item response function (such as the logistic or probit function). One common assumption in IRT is local independence; that is, conditional on the measured latent trait θ_i, responses across questions $(X_{i1}, \ldots, X_{iJ})$ are independent. This enables one to write the joint probability of observing a response vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})'$ given θ_i as the product of item response probabilities for each individual item,

$$P(X_{i1} = x_{i1}, \ldots, X_{iJ} = x_{iJ} \mid \theta_i) = \prod_{j=1}^{J} P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}}. \qquad 2.$$

If one assumes that the responses across N total examinees are independent, the joint likelihood of item parameters, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_J)$, and examinees' ability parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)'$, given the observed response data $\underset{\sim}{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$, can be written as

$$L(\boldsymbol{\zeta}, \boldsymbol{\theta}; \underset{\sim}{\mathbf{x}}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \{P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}}\} = \prod_{i=1}^{N} \prod_{j=1}^{J} \left\{ f(\theta_i, \boldsymbol{\zeta}_j)^{x_{ij}} [1 - f(\theta_i, \boldsymbol{\zeta}_j)]^{1-x_{ij}} \right\}. \qquad 3.$$

The item and examinee parameters can then be estimated on the basis of Equation 3 and the observed response data. The local independence assumption enables one to specify a latent variable model for item responses by specifying an item response function.

Mastery testing based on the IRT framework may be regarded as a predecessor of cognitive diagnostic testing (Hambleton & Novick 1973, Lord 1980). It was proposed to provide decision rules for mastery versus nonmastery on a criterion-referenced test. Different from norm-referenced tests that rank individuals on a continuum, a criterion-referenced test yields measurement on whether an examinee meets the performance standards for a particular domain (Glaser & Nitko 1970). Consider two levels of ability: θ_2, a low ability level that is unsatisfactory for mastering the domain, and θ_1, a high ability level that is satisfactory for the measured domain. The problem of whether an examinee should be classified as a master, on the basis of the observed responses ($\mathbf{x}_i$), can be formulated as a simple-versus-simple hypothesis testing problem (Neyman & Pearson 1933):

$$H_0 : \theta_i = \theta_2, \qquad H_1 : \theta_i = \theta_1. \qquad 4.$$

A decision can be made by comparing the likelihood ratio at θ_2 and θ_1 given the observed responses $\mathbf{x}_i$ to a rejection region at a preset level of type I error.

The mastery model, which explicitly models the relationship between mastery status and the observed responses, was subsequently proposed (e.g., Macready & Dayton 1977). Built upon latent class models (Lazarsfeld & Henry 1968), in the mastery model, examinees are assumed to originate from one of C a priori specified latent classes with a hierarchical structure. One special case is when there are only two latent classes: masters of all items and nonmasters of all items. The assumption of discrete mastery levels, instead of continuous latent traits, may be more consistently aligned

with theories of developmental and educational psychology, in which the acquisition of skills or knowledge occurs as qualitative jumps (e.g., Bloom 1956, Piaget 1950, von Davier & Lee 2019b). Let v_i denote the latent class membership of examinee i. The conditional probability of a response vector $\mathbf{x}_i$ given $v_i = v_c$ ($v_c \in \{v_1, \ldots, v_C\}$) is

$$P(\mathbf{x}_i \mid v_i = v_c) = \prod_{j=1}^{J} \left[g_j^{1-\eta_j(v_c)} (1 - s_j)^{\eta_j(v_c)} \right]^{x_{ij}} \left[(1 - g_j)^{1-\eta_j(v_c)} s_j^{\eta_j(v_c)} \right]^{(1-x_{ij})}, \quad 5.$$

where $\eta_j(\cdot)$ is a predefined function of the latent class membership v_c that is either 1 or 0, indicating whether or not individuals in the cth proficiency class are capable of solving item j, if there was no measurement error. These days, $\eta_{ij} = \eta_j(v_i)$ is commonly referred to as the ideal response on item j by examinee i. Moreover, g_j, commonly termed the guessing parameter, denotes the probability of a correct response by a nonmaster with $\eta_{ij} = 0$, and s_j, the slipping parameter, denotes the probability of an incorrect response (e.g., due to forgetting, a careless mistake) by a master with $\eta_{ij} = 1$. Let $\pi_c = P(v_i = v_c)$ denote the population membership probability of latent class c, i.e., the probability that a randomly chosen examinee comes from the cth proficiency class, with $\sum_{c=1}^{C} \pi_c = 1$. Then, the marginal probability of observing $\mathbf{x}_i$ can be obtained by integrating out the latent class membership,

$$P(\mathbf{x}_i) = \sum_{c=1}^{C} P(\mathbf{x}_i \mid v_c)\pi_c. \quad 6.$$

Macready & Dayton (1977) discussed methods for (*a*) estimating model parameters, $\mathbf{s} = (s_1, \ldots, s_J)'$, $\mathbf{g} = (g_1, \ldots, g_J)'$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{C-1})$; (*b*) evaluating absolute model fit; and (*c*) comparing models (e.g., models with different structural assumptions about latent mastery hierarchy) on the basis of the marginal likelihood. Once a model is selected and its parameters estimated, assigning mastery class labels to examinees amounts to a classification problem (Bishop 1995, Hastie et al. 2009, Ripley 1996), where examinee i's latent class membership v_i is estimated to minimize a loss function that depends on both the likelihood of v_i given the observed responses and the cost of different misclassification errors (e.g., falsely classifying a nonmaster as a master and vice versa).

2.2. Extensions to Multiple Attributes and a Framework for Cognitive Diagnosis

Mastery testing and mastery models allow for statistical decision-making on whether or not an individual should be considered a master on a single domain. Indeed, this approach can sufficiently inform many practical decisions, for example, whether an examinee should pass or fail a licensure test and whether a student is well prepared to proceed to the next level. It is not difficult to imagine, however, test items that are designed to be solved with multiple fine-grained attributes.

The late 1970s and early 1980s saw the advancement of computing power and, as a result, early development of computer-based systems that automatically diagnose learners' use of erroneous rules in elementary arithmetics on the basis of their item responses (e.g., Brown & VanLehn 1980, Tatsuoka et al. 1980). In the example of the signed number subtraction problem $-3 - (-7) = ?$, one example of an erroneous rule is, "The student always subtracts the smaller absolute value from the larger one and takes the sign of the number with larger absolute value in the answer. The conversion of subtraction problems to addition is omitted and the difference between addition and subtraction of two signed numbers seems to be ignored" (Tatsuoka 1983, p. 346). These developments sparked research interest in psychometric methods for measuring misconceptions based on observed arithmetic item responses. Tatsuoka & Tatsuoka (1983) and Tatsuoka (1983, 1984) proposed the rule space method for diagnosing erroneous rules on the basis of observed response vectors. The original rule space method relied largely on proficiency

estimates and person-fit indices under a continuous IRT model to diagnose misconceptions. However, it established a framework for cognitive diagnosis, which is characterized by (*a*) a thorough analysis of the set of specific cognitive skills and knowledge (i.e., attributes) that are required for solving each task and (*b*) the diagnosis of an examinee's error resulting from a lack of a required attribute(s) for a task (Tatsuoka 1990). Another framework that was independently developed at approximately the same time was knowledge space theory (Doignon & Falmagne 1985). A knowledge state is defined as the subset of tasks within a considered domain that an examinee is capable of solving, and whether an examinee is a master of a task may be explained by her latent mastery of a number of skills/knowledge that the task demands (e.g., Falmagne et al. 1990). Heller et al. (2015) provide an analysis of the connection between cognitive diagnosis and knowledge space theory. These statistical frameworks for the measurement of discrete knowledge states are a building block of many well-known computer-based instructional systems, including the PLATO system (Tatsuoka 1983), the ALEKS system (Doignon & Falmagne 2012), and the Cognitive Tutor intelligent tutoring system (Anderson et al. 1995).

Here, we formally define the cognitive diagnosis framework. Performance on a test is assumed to depend on the mastery status, or knowledge state (Tatsuoka 2009), of a collection of K attributes, denoted $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$. 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). 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). The attributes may be either parallel or hierarchical, where one is the prerequisite to mastering another (Gierl et al. 2000, Leighton et al. 2004). In the latter case, the number of permissible attribute patterns can be less than 2^K. Consider a J-item test. The relationship between items and attributes is defined via a Q-matrix, a $J \times K$ 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 $q_{jk} = 1$. As an illustration, consider the fraction subtraction assessment (Tatsuoka 2009, p. 41), a $J = 20$-item test for basic subtraction operations that involved fractions. Experts identified $K = 8$ measured attributes:

1. converting a whole number to a fraction or mixed number,
2. separating a whole-number part from a fraction part,
3. simplifying before getting the common denominator,
4. finding the common denominator,
5. borrowing one from the whole-number part,
6. column borrowing for subtraction of the numerators,
7. reducing the answer to the simplest form, and
8. subtracting numerators.

For this test, the Q-matrix is a 20×8 matrix. As an example, question 2 requires solving $\frac{3}{4} - \frac{3}{8}$. To solve this question, a student would first convert $\frac{3}{4}$ to $\frac{6}{8}$ (attribute 4) and work with $6 - 3$ on the numerator (attribute 8). Correspondingly, $q_{2,4}$ and $q_{2,8} = 1$, and the remaining elements in the second row of the Q-matrix are 0; i.e.,

$$Q = \begin{pmatrix} q_{1,1} & \cdots & q_{1,8} \\ 0 \;\; 0\,0 \;\; 1 \;\; 0\,0\,0 & & 1 \\ \vdots & \ddots & \vdots \\ q_{20,1} & \cdots & q_{20,8} \end{pmatrix}.$$

For an item that requires multiple attributes, a natural question that arises is how different skills work together to influence the correct response probability. For the fraction question above, a master of the question must be able to perform both required steps (attributes 4 and 8). This is known as a conjunctive item, where an ideal response of 1 requires mastery of all requisite skills. In the following section, models assuming other conjunction rules are introduced. The Q-matrix, together with a conjunction rule, allows one to identify the ideal response to a question j by examinees with attribute profile $\boldsymbol{\alpha}$. The ideal response, as seen in the mastery model by Macready & Dayton (1977) (Equation 5), is the response that would have been observed were there no measurement error, and in general, it can be written as a function of $\boldsymbol{\alpha}$ and the jth row of the Q-matrix, $\mathbf{q}_j$, $\eta_j(\boldsymbol{\alpha}, \mathbf{q}_j)$, which is either 0 (nonmaster) or 1 (master). Tatsuoka (1995, 2009) formulated the task of cognitive diagnosis as a pattern recognition problem. She discussed the connection between the task of classifying an examinee's $\boldsymbol{\alpha}$ and the task of statistical pattern recognition and classification (e.g., Ripley 1996). In the latter, on the basis of the observed patterns from a set of feature variables, an observation is grouped into one of C predetermined groups. For cognitive diagnostic testing, the C predetermined groups are knowledge states (i.e., $\boldsymbol{\alpha}$ profiles) that are permissible, given the expert-defined domain of K assessed skills and their hierarchical structures. Evaluating the similarity between an examinee's observed response vector and the ideal response vectors for each proficiency class (derived on the basis of Q and the conjunction rule) enables classification of examinee knowledge state based on the latent feature space formed by attributes.

2.3. Probabilistic Diagnostic Classification Models

Even for a well-designed cognitive diagnostic test with a correctly specified Q-matrix, the observed responses of an examinee can deviate from the ideal response vector for his/her proficiency class. In the development of a Unified Model, DiBello et al. (1995) provided a conceptual analysis of the possible sources of stochastic variation in responses:

- Strategy: The choice of strategy by some examinees can differ from what the expert-defined Q-matrix presumes, and a correct response may arise from other routes.
- Positivity: A master of a skill may fail to apply it correctly to a question, yet a nonmaster may accidentally apply the right strategy.
- Completeness: A correct response may require skills or ability not covered by the K skills; an example is higher-order processing (Samejima 1995), such as reading, gathering and filtering information, and strategizing a solution.
- Slips: Other random errors that cannot be attributed to the aforementioned sources can also occur, resulting in deviations between observed and Q-predicted responses.

Built on the cognitive diagnosis framework, many parametric models were developed to account for the stochastic relationship between the theorized attribute profile and the observed responses. These models are commonly referred to as cognitive diagnosis models (CDMs) or diagnostic classification models (DCMs) (e.g., Rupp et al. 2010). DCMs differ in their complexity, in their assumptions on how different requisite skills are combined to affect the correct response probability for an item, and in how measurement error enters the model. Thorough introductions to various DCMs can be found in DiBello et al. (2006), Hartz (2002), Rupp et al. (2010), and von Davier & Lee (2019a). Here, we survey a few examples.

Before we move on to specific models, it is worth mentioning how proficiency class can be estimated on the basis of a parametric DCM. Similar to IRT, DCMs, which are restricted latent class models, commonly assume local independence of the item responses conditioning on the

measured attributes; i.e.,

$$P(X_{i1}, \ldots, X_{iJ} \mid \boldsymbol{\alpha}_i) = \prod_{j=1}^{J} P(X_{ij} \mid \boldsymbol{\alpha}_i).$$

With the local independence assumption, specifying an item response function (e.g., the probability of a correct response for dichotomous items) amounts to specifying a model, i.e., the likelihood function. In the next section, we provide an overview of DCM parameter estimation. For now, consider the case in which the item and population parameters $(\boldsymbol{\zeta}, \boldsymbol{\pi})$ are known and the only unknown parameters are the $\boldsymbol{\alpha}_i$s for the examinees. Then, the likelihood of $\boldsymbol{\alpha}_i$ given examinee i's observed responses is

$$L(\boldsymbol{\alpha}_i; \mathbf{x}_i) = \prod_{j=1}^{J} P(X_{ij} = x_{ij} \mid \boldsymbol{\alpha}_i). \quad 7.$$

The likelihood in Equation 7 allows for the estimation of $\boldsymbol{\alpha}_i$, the attribute pattern of examinee i, based on the examinee's observed response vector. There are multiple ways to fulfill this task, such as maximum likelihood estimation, where $\hat{\boldsymbol{\alpha}}_i$ is chosen to maximize Equation 7, and Bayesian posterior mean or mode, where the posterior distribution of $\boldsymbol{\alpha}_i$ is proportional to the product of the likelihood and some prior distribution.

One of the simplest and most studied DCMs is the DINA (deterministic input, noisy "and" gate) model (Haertel 1990, Junker & Sijtsma 2001). DINA is a conjunctive DCM, where examinee i needs to master all attributes required by item j to answer correctly with probability $(1 - s_j)$, and missing any required attribute will result in a correct response probability of $g_j < 1 - s_j$. In other words,

$$P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \quad 8.$$

where

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \quad 9.$$

A disjunctive counterpart to the DINA model is the DINO (deterministic input, noisy "or" gate) model (Templin & Henson 2006), where mastery of one requisite skill can completely compensate for the lack of others. Only one of the required attributes for item j is needed for a correct response probability of $(1 - s_j)$, and for those examinees without any of the required attributes, the probability of a correct response is g_j. The DINO model differs from the DINA model in how the ideal response relates to $\mathbf{q}_j$ and $\boldsymbol{\alpha}$:

$$\eta_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}. \quad 10.$$

Both the DINA and DINO models have two parameters for each item, slipping (s_j) and guessing (g_j), which capture the non-Q-explained stochastic fluctuations in responses. Different latent classes can take only one of two possible correct response probabilities, g_j or $1 - s_j$. One should also note a close connection between the DINA/DINO models and the mastery model by Macready & Dayton (1977) in Equation 5, where the DINA ideal response (Equation 9) is 1 for any proficiency class $\boldsymbol{\alpha}_c \succeq \mathbf{q}_j$ (i.e., each element of $\boldsymbol{\alpha}_c$ is greater than or equal to $\mathbf{q}_j$) and 0 otherwise and the DINO ideal response (Equation 10) is 1 for any $\boldsymbol{\alpha}_c' \mathbf{q}_j > 0$. Although one assumes a conjunctive relationship and the other assumes a disjunctive relationship among

requisite skills, Köhn & Chiu (2016) showed that the DINA and DINO models are algebraically equivalent, where one can be reparameterized into the other.

The DINA and DINO models, as suggested by their names (i.e., deterministic input), assume that masters of a required attribute (i.e., input) will have probability 1 of successfully retrieving it on a task and that nonmasters will have probability 0 of retrieving it on a task. The noise of the DINA or DINO model comes at the conjunction stage (i.e., the "and"/"or" gates). Two related models, which assume that the noise comes at the input stage, are the NIDA (noisy input, deterministic "and" gate) (Junker & Sijtsma 2001, Maris 1999) and NIDO (noisy input, deterministic "or" gate) (Templin & Henson 2006) models. The NIDA item response function is

$$P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \prod_{k=1}^{K} [(1 - s_k)^{\alpha_{ik}} g_k^{(1-\alpha_{ik})}]^{q_{jk}}, \tag{11.}$$

where, for each attribute k, s_k and g_k are the probabilities that a master fails to apply it and a nonmaster applies it by chance, respectively, the same across all items. The "and" gate is deterministic; that is, if all requisite skills are successfully retrieved for an item, the correct response probability is 1. The NIDO model, in contrast, requires successful retrieval of any requisite skill by an item.

In reality, an item is neither completely conjunctive nor completely disjunctive but is somewhere in between. An individual who masters more (but not all) requisite skills for an item may have a higher chance of a correct response than someone who masters fewer of the requisite skills. For instance, on a multiple-choice question with four options, an individual with partial mastery of the required skills may have a higher chance of selecting the correct answer, as she is capable of ruling out some distractors. This calls for models with more relaxed assumptions. One model that accommodates this relaxation is the reduced reparameterized unified model (rRUM) (Hartz 2002), where the correct response probability is

$$P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{q_{jk}(1-\alpha_{ik})}.$$

That is, the probability of responding 1 is π_j^* for someone with all attributes required by the item, and a penalty $r_{jk} \in [0, 1]$ is applied for lacking the required attribute k. The rRUM is algebraically equivalent to a more general NIDA model, where attribute-level slipping and guessing can differ across items.

Rather than explicitly defining how skills combine and how measurement error arises in the model, general DCMs have also been proposed. For many of them, the item response function shares connections with generalized linear models (Nelder & Wedderburn 1972), where a transformation (via a link function) of the correct response probability can be written as a linear combination of main effects and (possibly) higher-order interactions of the requisite skills (α_ks with $q_{jk} = 1$). Here, we present three examples:

- The first example is the general diagnostic model (GDM) (von Davier 2005), with item response function

$$P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{1}{1 + \exp[-(\lambda_j + \sum_{k=1}^{K} \lambda_{cj} q_{jk} \alpha_{ik})]}. \tag{12.}$$

  Note that $q_{jk}\alpha_{ik} = 1$ when k is a requisite skill for item j and is mastered by examinee i. The model can extend to include higher-order interactions among requisite skills (von Davier 2019), for example, as seen in the following model.

- The second example is the loglinear CDM (LCDM) (Henson et al. 2009), with item response function

$$\text{logit}P(X_{ij}=1 \mid \boldsymbol{\alpha}_i) = \lambda_j + \sum_{k=1}^{K} \lambda_{jk}(\alpha_{ik}q_{jk}) + \sum_{k=1}^{K}\sum_{k'>k} \lambda_{jkk'}(\alpha_{ik}q_{jk}\alpha_{ik'}q_{jk'}) + \ldots. \quad 13.$$

  Higher-level interactions are permitted for the attributes required by j. For identifiability purposes, Henson et al. (2009) suggested imposing a monotonicity assumption such that mastering any additional skills results in a nondecreasing correct response probability.
- The third example is the generalized DINA (G-DINA) model (de la Torre 2011); here,

$$g[P(X_{ij}=1) \mid \boldsymbol{\alpha}_i] = \lambda_j + \sum_{\forall k:q_{jk}=1} \lambda_{jk}\alpha_{ik} + \sum_{\forall k:q_{jk}=1}\sum_{\forall k':q_{jk'}=1,k'>k} \lambda_{jkk'}\alpha_{ik}\lambda_{ik'} + \cdots + \lambda_{jkk'k''\ldots}\prod_{k:q_{jk}=1}\alpha_{ik}, \quad 14.$$

  where $g(\cdot)$ is a link function such as the identity link or the logit link.

Interpretations of the three models are quite similar: λ_j is the intercept of the linear predictor, which affects the correct response probability of the proficiency class of nonmasters of all attributes; λ_{jk} is the main effect of possessing a requisite skill k on item j; the slope coefficients for higher-order interactions, $\lambda_{jkk'\ldots}$, describe how requisite skills combine to influence correct response probability. Not all main effects or interaction terms need to be active; some λs can be restricted to 0 on the basis of model assumptions. All three models subsume many restricted models as nested cases. For instance, under the G-DINA model (de la Torre 2011), using an identity link and setting all but λ_j and $\lambda_{j\{k:q_{jk}=1\}}$ to 0 yield the DINA model, and similarly for the GDM and the LCDM. This approach allows for model fit comparisons among different item response functions, where an appropriate one can be chosen to balance flexibility and parsimony.

Lastly, we note a number of ways in which a model involves both discrete attributes and a continuous trait or traits. The idea that different (continuous) abilities can combine in a nonadditive manner to affect correct response probability is seen in many continuous item response models; see DiBello et al. (2006) for a review. Continuous traits may also be introduced to model the residual effect on item response unexplained by an expert-specified Q. An early example is the unified model by DiBello et al. (1995), which incorporates the four sources of variation introduced at the beginning of this subsection, with item response function

$$P(X_{ij}=1 \mid \boldsymbol{\alpha}_i, \theta_i) = d_j \prod_{k=1}^{K}(1-s_{jk})^{\alpha_{ik}q_{jk}} g_{jk}^{(1-\alpha_{ik})q_{jk}} P_{c_j}(\theta_i) + (1-d_j)P_{b_j}(\theta_i)$$

that depends on a residual trait θ_i in addition to the attribute profile $\boldsymbol{\alpha}_i$. This person-specific continuous trait is used to explain response variations due to alternative strategies $[P_{b_j}(\theta_i)]$ not specified in Q and those due to the demand for additional skills (i.e., completeness), $P_{c_j}(\theta_i)$. This model, while conceptually meaningful, is statistically unidentifiable (Jiang 1996), leading to simplifications such as the reparameterized unified model (RUM, i.e., the fusion model) (Hartz 2002, Roussos et al. 2007),

$$P\left(X_{ij}=1 \mid \boldsymbol{\alpha}_i, \theta_i\right) = \left[\pi_j^* \prod_{k=1}^{K} r_{jk}^{q_{jk}(1-\alpha_k)}\right] P_c\left(\theta_i\right). \quad 15.$$

The aforementioned rRUM (Hartz 2002), without the completeness term involving continuous θ_i, is the further simplification of the RUM.

Another way in which continuous ($d \geq 1$-dimensional) $\boldsymbol{\theta}$, as a high-order latent trait (de la Torre & Douglas 2004), can be incorporated is through the joint distribution of discrete attributes, specified by

$$P(\boldsymbol{\alpha}_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}_i) = \prod_{k=1}^{K} P(\alpha_{ik} \mid \boldsymbol{\theta}_i . \beta_{k0}, \boldsymbol{\beta}_k) = \prod_{k=1}^{K} \frac{1}{1 + \exp[-(\beta_{k0} + \boldsymbol{\beta}_k' \boldsymbol{\theta}_i)]}. \quad 16.$$

This approach allows for the use of DCM as the measurement model to explain observed item responses but at the same time generate an overall continuous proficiency estimate that is useful for summative evaluations, comparisons, and joint modeling with other continuous traits (e.g., speed) (Wang et al. 2018b, Zhan et al. 2018a). Another by-product of the higher-order model is its computational advantage. With the K attributes assumed conditionally independent given $\boldsymbol{\theta}$, one can circumvent working with the mixture of 2^K discrete latent classes. At the end of this review, we discuss another current development in which continuous random effects are added to DCMs to account for residual dependence between items for tests with testlets.

2.4. Nonparametric Methods for Diagnostic Classification

Applying the aforementioned parametric DCMs to diagnostic classification requires estimation of model parameters. When examinees' $\boldsymbol{\alpha}_i$s are assumed to be random samples from an underlying population, the model parameters often consist of item parameters and structural parameters that characterize the distribution of different proficiency classes in the population. de la Torre & Lee (2010) suggested $N = 500$ to be a sufficient sample size for DINA parameter recovery when the number of attributes is $K = 5$. Reliable parameter estimates for more complex models and larger number of attributes would demand even larger sample sizes. However, such a large sample size is often unattainable for small-scale, classroom assessments, where cognitive diagnostic testing shows its merit as a formative assessment.

Nonparametric methods for clustering and classification provide viable alternatives to parametric models for cognitive diagnosis. Nonparametric clustering methods for cognitive diagnosis aim to infer latent proficiency classes on the basis of observed responses using clustering algorithms (e.g., Chiu et al. 2009). Nonparametric classification methods for cognitive diagnosis (e.g., Chiu & Douglas 2013), in contrast, employ distance-based algorithms to classify individuals into predefined proficiency classes, and they were shown to be not only usable for any sample size but also computationally inexpensive. Chiu & Köhn (2019) thoroughly review nonparametric methods in cognitive diagnostic testing. We give a brief primer to the problem, methods, and theories behind nonparametric clustering and classification for cognitive diagnosis here, although this survey of the literature is by no means complete.

Nonparametric clustering and classification in cognitive diagnosis greatly leverage methods in statistical and machine learning for clustering (unsupervised learning) and classification (supervised learning). In particular, the rule space method (Tatsuoka 1983, 2009) discussed above can be seen as a nonparametric classification approach, where the diagnosis of mastery profiles is formulated as a pattern recognition problem. The key distinctions, again, are that (*a*) cognitive diagnostic testing restricts the latent space to the latent proficiency classes defined by the K assessed attributes and (if applicable) their hierarchical relationships, and (*b*) the relationship between observed features (i.e., item responses) and the underlying latent classes are constrained by the Q-matrix. Classical methods for clustering, such as the k-means algorithm (e.g., Hartigan & Wong 1979), the hierarchical agglomerative clustering algorithm (e.g., Johnson 1967, Ward Jr. 1963), and methods for feature-based supervised classification (e.g., Bishop 1995, Fisher 1936, Ripley 1996), being mostly data driven, are not straightforwardly aligned with theory-driven psychometric practice,

which emphasizes interpretability (e.g., diagnostic reports that can be understood by students and teachers) and reliable and valid measurement of theorized constructs based on psychological or educational theory (Allen & Yen 2001, Cronbach & Meehl 1955). To oversimplify, statistical problems unique to nonparametric cognitive diagnosis are (*a*) how expert inputs (e.g., the Q-matrix) should be incorporated to guide the clustering and classification of individuals into proficiency classes and (*b*) whether and when the true knowledge state structure and examinee proficiency classes can be recovered using data-driven methods for clustering and classification.

Cluster analysis aims to identify maximally homogeneous groups, with observations across groups maximally separated, on the basis of either P-dimensional feature variables summarized in an $N \times P$ matrix (e.g., k-means clustering; Hartigan & Wong 1979) or pairwise distance/dissimilarity between observations, which can be summarized as an $N \times N$ matrix (e.g., hierarchical clustering; Johnson 1967, Ward Jr. 1963). In nonparametric clustering for cognitive diagnosis, raw scores on individual questions are aggregated into attribute-wise sum scores based on the Q-matrix. Specifically, denote the response matrix by $\mathbf{X}_{N\times J}$, where X_{ij} is the score on item j by examinee i. The attribute-wise sum scores, $\mathbf{W} = \mathbf{XQ}$, are an $N \times K$ matrix, with the (i, k)th entry, $\sum_{j=1}^{J} X_{ij}Q_{jk}$, being the sum score of examinee i on all items that involve skill k. $\mathbf{W}$ can be used as the feature variables for k-means clustering or to construct an $N \times N$ pairwise dissimilarity matrix for hierarchical clustering, whose $[i, j]$th entry is the dissimilarity between examinee i's and examinee j's observed responses, computed by the Euclidean distance of their attribute sum score vectors (Chiu et al. 2009).

Cluster analysis, as an unsupervised algorithm, identifies groups that do not come with inherent interpretations. Hence, one important question is when nonparametric clustering for cognitive diagnosis can separate individuals into their true underlying proficiency classes.

Definition 1 (Completeness of a Q-matrix). A Q-matrix is complete if it guarantees the identifiability of all realizable proficiency classes among examinees. Specifically, let $\mathbf{S}(\boldsymbol{\alpha}) = E(\mathbf{X} \mid \boldsymbol{\alpha})$ be the vector of expected scores on a test given latent attribute profile $\boldsymbol{\alpha}$. Then Q is complete if for any two $\boldsymbol{\alpha}, \boldsymbol{\alpha}^*$s from the set of possible attribute profiles, $\mathbf{S}(\boldsymbol{\alpha}) = \mathbf{S}(\boldsymbol{\alpha}^*) \Rightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha}^*$.

Intuitively, a Q-matrix of a test is complete if no two different proficiency classes yield identical expected scores on all J items. An incomplete Q implies that the given test cannot uniquely identify all underlying latent proficiency classes on the basis of the observed response data. Two things are worth noting here. First, whether different latent classes can be uniquely identified depends on the structure of the Q-matrix, i.e., the set of skills that each item measures. Second, because the definition of Q-matrix completeness involves $E(\mathbf{X} \mid \boldsymbol{\alpha})$, the completeness of the Q-matrix of a test is always in reference to the true underlying item response model. In other words, for the same test, a Q-matrix that is complete for rRUM might not necessarily be complete for the DINA model (Köhn & Chiu 2017).

This definition of completeness was introduced by Chiu et al. (2009) in the asymptotic classification theory for cognitive diagnosis. It provides a theoretical justification for the use of the hierarchical agglomerative clustering algorithm for clustering individuals into proficiency classes, when Q is complete. In particular:

- If the data-generating model is DINA, the Q-matrix is complete if and only if it contains an identity submatrix after row permutations. In other words, for each $k \in \{1, \ldots, K\}$, at least one item assesses skill k exclusively. This provides a practical guide to the design of a test that has complete Q under the DINA model.

- Under the DINA model, if the Q-matrix is complete, then the expected attribute-level sum score vector, $\mathbf{T}(\boldsymbol{\alpha}) = E(\mathbf{W} \mid \boldsymbol{\alpha})$, uniquely identifies $\boldsymbol{\alpha}$; that is, $\mathbf{T}(\boldsymbol{\alpha}) \neq \mathbf{T}(\boldsymbol{\alpha}^*)$ for $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$. This allows for the use of attribute-wise sum scores for clustering.
- The hierarchical agglomerative clustering algorithm separates individuals into their true latent classes when the underlying model is a finite mixture model.

These lemmas ensure that the probability that the hierarchical agglomerative clustering algorithm separates individuals into their true latent classes approaches 1 as the number of items, J, approaches infinity. Later work extended this to DCMs other than DINA (e.g., see Chiu & Köhn 2019, Köhn & Chiu 2017), but the conditions for Q-completeness and the separation of proficiency classes by attribute-wise sum scores could differ. Intuitively, this theory justifies the use of hierarchical clustering to identify groups that correspond to the theorized latent proficiency classes. Whether similar theoretical properties can be established for k-means clustering remains an open problem, and diagnosing the proficiency classes requires postclustering labeling, e.g., finding a closest match between within-cluster attribute sum scores and candidate attribute vectors (Chiu et al. 2009).

Nonparametric classification methods for cognitive diagnosis, in contrast, are designed to classify examinees into predefined groups of attribute proficiency. One method proposed by Chiu & Douglas (2013) using distance-based algorithms seeks optimal alignment of an individual's observed response vector ($\mathbf{x}_i$) and the ideal response vector ($\boldsymbol{\eta}_c$) of each candidate proficiency class c. That is,

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

where d is a distance measure, such as the weighted or unweighted Hamming distance, and $\boldsymbol{\eta}_c$ is either the conjunctive (Equation 9) or the disjunctive (Equation 10) ideal response for class $\boldsymbol{\alpha}_c$ given Q. Wang & Douglas (2015) showed the consistency of the nonparametric estimator $\hat{\boldsymbol{\alpha}}$ under certain regularity conditions for a number of conjunctive models (e.g., DINA, NIDA, rRUM) as the true data-generating model. Recognizing that a true data-generating model may not be completely conjunctive or disjunctive, but somewhere in between, Chiu et al. (2018) proposed a generalized nonparametric classification method, where the ideal response to item j for a class l, $\eta_{j,l}$, was taken as a convex combination of the conjunctive $[\eta_{j,l}^{(c)}]$ and disjunctive $[\eta_{j,l}^{(d)}]$ ideal responses; i.e.,

$$\eta_{j,l} = \omega_{lj}\eta_{j,l}^{(c)} + (1 - \omega_{lj})\eta_{j,l}^{(d)}.$$

For an item that measures $K^* \leq K$ attributes, l indexes up to 2^{K^*} proficiency classes that the item can distinguish, termed equivalent classes. Thus, equivalent classes that share the same mastery pattern on the K^* attributes but differ on the remaining $K \backslash K^*$ share the same l. The weight parameters, ω_{lj}, can be estimated from the data on the basis of initial estimates of $\hat{\boldsymbol{\alpha}}$ (e.g., using the conjunctive ideal response), and the final attribute classification can be derived on the basis of the weighted ideal responses.

Comparing model-based and nonparametric classification via simulations, Chiu & Douglas (2013) showed that, when the true data-generating model is known and accurate estimation of a parametric model is feasible, model-based classification provides more efficient and reliable measurement. However, nonparametric classification provides a viable alternative when parametric inference is not feasible, for example, when the sample size is small. Other classifiers may also be used for nonparametric classification; one example is maximizing the cosine similarity between observed and ideal responses (von Davier & Lee 2019b), a commonly adopted distance measure in statistical learning.

3. STATISTICAL INFERENCE

In this section, we discuss some of the issues central to statistical inference with cognitive diagnostic testing. We focus primarily on inferences based on parametric DCMs. Central to the discussion are (*a*) parameter estimation; (*b*) model identifiability; and (*c*) *Q*-matrix learning and refinement, which are closely connected to model fit and model/variable selection. Our hope is to provide a primer on them, but we omit the full technical discussion.

3.1. Parameter Estimation

As the above discussion shows, parametric DCMs can take many different forms, some involving continuous traits and some allowing ordinal observed responses. For clarity of presentation, consider the most common case of a DCM with binary attribute patterns (i.e., mastery versus nonmastery), binary observed responses (correct versus incorrect), and item response function

$$P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i, \boldsymbol{\zeta}_j, \mathbf{q}_j). \tag{17.}$$

An example is the DINA item response function in Equation 8. Here $\boldsymbol{\zeta}_j$ is the item parameter vector for j (e.g., s_j, g_j for DINA). Although we include $\mathbf{q}_j$ to emphasize that the correct response probability depends on the interaction between the examinee's mastery and the item's attribute requirements, $\mathbf{q}_j$ is conventionally expert-specified and hence not a parameter of the model. The joint likelihood of the unknown parameters $\boldsymbol{\alpha}, \boldsymbol{\zeta}$ (i.e., $\boldsymbol{\alpha}_i$ for all is and $\boldsymbol{\zeta}_j$ for all js), given the $N \times J$ response matrix $\underset{\sim}{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$, is

$$L(\boldsymbol{\alpha}, \boldsymbol{\zeta}; \underset{\sim}{\mathbf{x}}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left\{ P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i, \boldsymbol{\zeta}_j, \mathbf{q}_j)^{x_{ij}} [1 - P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i, \boldsymbol{\zeta}_j, \mathbf{q}_j)]^{1-x_{ij}} \right\}. \tag{18.}$$

Several approaches exist for parameter estimation. The first is the joint maximum likelihood estimation via maximizing Equation 18. However, the number of parameters in the joint likelihood in Equation 18 grows with sample size N, and, similar to continuous IRT models, joint maximum likelihood estimation leads to inconsistent parameter estimates (Baker & Kim 2004). Chiu et al. (2016) suggested initializing the examinees' attribute pattern estimates with nonparametric classifications on examinee mastery patterns to improve estimation stability.

More commonly adopted, under a frequentist framework, is marginal maximum likelihood estimation. Here, $\boldsymbol{\alpha}_i$s are assumed to be multinomial, drawn from a distribution of possible attribute patterns in the population, with a population membership parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)'$, where $\pi_c = P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c) = p(\boldsymbol{\alpha}_c)$ and C is the total number of realizable proficiency classes. This allows the marginal likelihood of $\boldsymbol{\zeta}$ to be written by integrating out the $\boldsymbol{\alpha}$s:

$$L(\boldsymbol{\zeta}; \underset{\sim}{\mathbf{x}}) = \prod_{i=1}^{N} \sum_{c=1}^{C} P(\mathbf{x}_i \mid \boldsymbol{\alpha} = \boldsymbol{\alpha}_c, \boldsymbol{\zeta}) p(\boldsymbol{\alpha}_c), \tag{19.}$$

with corresponding log-likelihood

$$l(\boldsymbol{\zeta}) = \log L(\boldsymbol{\zeta}; \underset{\sim}{\mathbf{x}}) = \sum_{i=1}^{N} \log \left[\sum_{c=1}^{C} P(\mathbf{x}_i \mid \boldsymbol{\alpha} = \boldsymbol{\alpha}_c, \boldsymbol{\zeta}) p(\boldsymbol{\alpha}_c) \right]. \tag{20.}$$

Direct maximization of Equation 20 is not straightforward. The expectation-maximization (EM) algorithm provides a way to find $\boldsymbol{\zeta}$ that maximizes Equation 20 iteratively (Dempster et al. 1977). The idea is that the unobserved latent class memberships, just like the unobserved latent abilities in IRT models (Bock & Aitkin 1981), are treated as missing data and the responses $\underset{\sim}{\mathbf{x}}$ as observed data; the two combined are referred to as the complete data. The EM algorithm initializes $\boldsymbol{\zeta}$ with

some starting value and iterates between two steps: an *E*-step, which computes the expectation of the complete data likelihood, under the conditional distribution of the missing data $\boldsymbol{\alpha}$ given the observed responses and current value of $\boldsymbol{\zeta}$, and an *M*-step, where the conditional expectation in the *E*-step is maximized with respect to $\boldsymbol{\zeta}$. de la Torre (2009) adapted the EM algorithm to the DINA model and derived the analytical expressions for the iterative updates:

- Initialize item parameters $\boldsymbol{\zeta}^{(0)}$.
- At iteration $r = 1, 2, \ldots$:
  - *E*-step: For each attribute class $c = 1, \ldots, C$, compute $I_c = \sum_{i=1}^{N} P[\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c \mid \mathbf{x}_i, \boldsymbol{\zeta}^{(r-1)}]$. For each j, c, compute $R_{jc} = \sum_{i=1}^{N} x_{ij} P[\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c \mid \mathbf{x}_i, \boldsymbol{\zeta}^{(r-1)}]$.
  - *M*-step: For each j, set $s_j^{(r)} = \frac{I_j^1 - R_j^1}{I_j^1}$, $g_j^{(r)} = \frac{R_j^0}{I_j^0}$, where, for the classes corresponding to masters of j with DINA ideal response $\eta_{jc} = 1$,

$$I_j^1 = \sum_{c:\eta_{jc}=1} I_c, \quad R_j^1 = \sum_{c:\eta_{jc}=1} R_{jc},$$

    and for the nonmasters classes such that $\eta_{jc} = 0$,

$$I_j^0 = \sum_{c:\eta_{jc}=0} I_c, \quad R_j^0 = \sum_{c:\eta_{jc}=0} R_{jc}.$$

- Stop when the termination criterion is reached (e.g., after M iterations or if change of $\boldsymbol{\zeta}^{(r)}$ from $\boldsymbol{\zeta}^{(r-1)}$ is less than some preset tolerance) and set $\hat{\boldsymbol{\zeta}} = \boldsymbol{\zeta}^{(r)}$.

Standard error estimation for $\hat{\boldsymbol{\zeta}}$ is possible via the asymptotic approximation on the basis of the inverse of the Fisher information matrix; see de la Torre (2009) for details. The EM algorithm for DINA [where $\boldsymbol{\zeta}_j = (s_j, g_j)$] is generalizable to more complex models, including general models such as G-DINA. Software implementation is available through the `G-DINA` R package (Ma & de la Torre 2020).

Bayesian posterior inference is another option. Let $\boldsymbol{\theta}$ denote all parameters, including $\boldsymbol{\alpha}$, $\boldsymbol{\zeta}$, and possibly others (e.g., population membership $\boldsymbol{\pi}$). Then the posterior distribution of the parameters given the observed data is proportional to the product of the prior density and the likelihood:

$$f(\boldsymbol{\theta} \mid \underset{\sim}{\mathbf{x}}) \propto P(\underset{\sim}{\mathbf{x}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{21.}$$

Markov chain Monte Carlo (MCMC) sampling algorithms are often applied: Samples of $\boldsymbol{\theta}$ are sequentially drawn on the basis of some transition kernel (e.g., Gibbs sampler; Albert & Chib 1993) to eventually converge to the (stationary) posterior distribution. Inferences, such as point estimates (and credible intervals) for item parameters and attribute mastery profiles, can then be drawn on the basis of the posterior samples. For DCM parameter estimation, Metropolis-Hastings within-Gibbs (Hastings 1970) sampling algorithms are quite flexible in that they can be adapted to various models and can be readily implemented in many software programs (e.g., BUGS; Lunn et al. 2009), but acceptance-rejection sampling requires tuning and can be computationally costly. Full Gibbs-sampling algorithms have been proposed for a number of well-known DCMs (e.g., Culpepper 2015, Culpepper & Hudson 2018), where carefully chosen priors and augmentation schemes enable sampling in each iteration from well-known distributional families without acceptance-rejection. Software implementation is available through the `edm` R package (Balamuta et al. 2020).

Even flexible algorithms for parameter estimation face computational challenges. For example, MCMC requires sequential sampling with typically tens of thousands of iterations until convergence. Alternatives to MCMC for Bayesian inference with DCMs, such as Hamiltonian Monte

Carlo samplers implemented in STAN (e.g., Carpenter et al. 2017, Jiang & Carter 2019) and variational inference (e.g., Blei et al. 2017, Yamaguchi & Okada 2020), were recently adopted. Another challenge to both frequentist and Bayesian parameter estimation is that, as K increases, the number of latent classes C, which can be as many as 2^K, increases exponentially, making the evaluation of $P(\boldsymbol{\alpha}_c \mid \mathbf{x}, \boldsymbol{\zeta}, \ldots)$ computationally burdensome, if not infeasible, for large Ks. Methods have been proposed in the past few years to leverage advanced computational methods (e.g., reformulating a DCM as a network model and applying noise contrastive estimation; von Davier 2018) and computing power (e.g., parallel computing; Khorramdel et al. 2019) to address the computational challenges of DCMs, but open problems and opportunities for acceleration of algorithms remain.

3.2. Identifiability

Regardless of which parameter estimation scheme from Section 3.1 is adopted, model identifiability is a prerequisite for consistent parameter estimation and valid inference (Gabrielsen 1978). Overparameterization can lead to nonidentifiability. Unrestricted latent class models with binary responses are not identifiable (Gyllenberg et al. 1994); in other words, the parameters do not uniquely determine the corresponding likelihood (Casella & Berger 2002). For DCMs, which are restricted latent class models, it is therefore essential to check whether the appropriate restrictions are in place to meet identifiability conditions, before proceeding to model estimation and inferences.

One may notice a resemblance of the definition of statistical identifiability and the definition for the completeness of the Q-matrix in Definition 1, which examines whether expected responses can uniquely identify $\boldsymbol{\alpha}$. For most discussions about identifiability of parametric DCMs, however, $\boldsymbol{\alpha}_i$s are assumed to be random effects from a population with membership probabilities $\boldsymbol{\pi}$, and the goal instead is to identify the item parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\pi}$, that is, whether for two sets of parameters $(\boldsymbol{\zeta}, \boldsymbol{\pi}) \neq (\bar{\boldsymbol{\zeta}}, \bar{\boldsymbol{\pi}})$,

$$P(\mathbf{X} = \mathbf{x} \mid Q, \boldsymbol{\zeta}, \boldsymbol{\pi}) \neq P(\mathbf{X} = \mathbf{x} \mid Q, \bar{\boldsymbol{\zeta}}, \bar{\boldsymbol{\pi}}). \quad 22.$$

Checking this condition is not straightforward. Many theoretical results for the identifiability conditions of DCMs or restricted latent class models were instead established by checking an equivalent condition involving a T-matrix (e.g., Gu & Xu 2019, 2020; Xu 2017; Xu & Shang 2018; Xu & Zhang 2016). Following Liu et al. (2013), $T(Q, \boldsymbol{\zeta})$ is a $2^J \times 2^K$ matrix whose rows index the possible response patterns $\mathbf{x} \in \{0, 1\}^J$ and whose columns index the possible proficiency classes $\boldsymbol{\alpha} \in \{0, 1\}^K$. Each entry, $t_{\mathbf{x},\boldsymbol{\alpha}}(Q, \boldsymbol{\zeta})$, is the marginal probability that an examinee in proficiency class $\boldsymbol{\alpha}$ answers correctly on all items whose corresponding entry in $\mathbf{x}$ is 1; i.e.,

$$t_{\mathbf{x},\boldsymbol{\alpha}}(Q, \boldsymbol{\zeta}) = P(\mathbf{X} \succeq \mathbf{x} \mid Q, \boldsymbol{\zeta}, \boldsymbol{\alpha}),$$

with $\mathbf{X} \succeq \mathbf{x}$ meaning $X_j \geq x_j$ across all J entries (i.e., items). Then, checking Condition 22 becomes equivalent to checking whether

$$T(Q, \boldsymbol{\zeta})\boldsymbol{\pi} \neq T(Q, \bar{\boldsymbol{\zeta}})\bar{\boldsymbol{\pi}} \quad 23.$$

for $(\boldsymbol{\zeta}, \boldsymbol{\pi}) \neq (\bar{\boldsymbol{\zeta}}, \bar{\boldsymbol{\pi}})$. This is because the dot product between each row of $T(Q, \boldsymbol{\zeta})$ and $\boldsymbol{\pi}$,

$$T_{\mathbf{x},\cdot}(Q, \boldsymbol{\zeta})\boldsymbol{\pi} = \sum_{c=1}^{C} P(\mathbf{X} \succeq \mathbf{x} \mid Q, \boldsymbol{\zeta}, \boldsymbol{\alpha}_c) P(\boldsymbol{\alpha} = \boldsymbol{\alpha}_c) = P(\mathbf{X} \succeq \mathbf{x} \mid Q, \boldsymbol{\zeta}, \boldsymbol{\pi}),$$

gives the marginal probability of $\mathbf{X} \succeq \mathbf{x}$ in the population, which has a one-to-one mapping with $P(\mathbf{X} = \mathbf{x} \mid Q, \boldsymbol{\zeta}, \boldsymbol{\pi})$ in Condition 22.

Easily checkable conditions for DINA model identifiability were proposed in Xu & Zhang (2016), later extended to sufficient and necessary conditions in Gu & Xu (2019). The DINA model parameters $\mathbf{s}, \mathbf{g}, \boldsymbol{\pi}$ are identifiable if and only if

- the Q-matrix is complete—i.e., upon row permutations, it takes the form

$$Q = \begin{pmatrix} \mathcal{I}_K \\ \hline Q^* \end{pmatrix},$$

  where $\mathcal{I}_K$ is a $K \times K$ identity matrix and Q^* corresponds to the remaining $J - K$ items—and
- each attribute k is assessed by at least three items.

The identifiability of DCMs has direct practical implications for the design and administration of cognitive diagnostic testing. For example, for a test that assesses $K = 3$ skills, there needs to be at least 1 item exclusively assessing each skill and a minimum of $J = 9$ total items (3 items for each skill) for drawing valid statistical inference on the basis of the DINA model.

Because DINO is algebraically equivalent to DINA (Köhn & Chiu 2016), the same identifiability results for DINA apply. Beyond specific DCMs like the DINA and DINO models, identifiability conditions have also been established for some general restricted latent class models. Xu (2017) provided sufficient conditions for identifiability of a general restricted latent class model with parameters $\boldsymbol{\pi}$ and Θ, with Θ being a $J \times 2^K$ matrix with $\theta_{j,\boldsymbol{\alpha}} = P(X_j = 1 \mid \boldsymbol{\alpha})$, representing the probability of a correct response on each item by each proficiency class. Here, the parameter space of Θ is restricted by the Q-matrix and the underlying model. Fang et al. (2019) established generic identifiability results for general restricted latent class models allowing for ordinal attributes and responses. Chen et al. (2020) established identifiability conditions for a sparse latent class model. In most cases, these identifiability conditions are easily checkable on the basis of the structure of the Q-matrix or other incidence matrices that characterize the relationship between items and proficiency classes. However, models with less restrictive assumptions generally pose more requirements on test design (e.g., more than one identity matrix in the Q-matrix) for identifiability, and for the more general models, complete characterization of identifiability is not yet available. Xu (2019) provided a detailed technical introduction to the identifiability of restricted latent class models. In addition, the conditions for strict identifiability can be difficult to meet in some scenarios. For example, it might be practically infeasible to design items that assess only a single attribute for some skills. To this end, Gu & Xu (2020) proposed a general framework for checking the strict and partial identifiability of restricted latent class models, where parameters of a restricted latent class model may still be partially identified, even if the strict identifiability conditions are not met.

3.3. Learning and Identifying the *Q*-Matrix

So far in this review, we assume the Q-matrix to be known and provided by content experts and test developers. In reality, an expert-defined Q-matrix is subject to misspecification, which can lead to biased parameter estimates and inaccurate diagnostic classifications (Rupp & Templin 2008). This calls for statistical methods to refine or estimate the Q-matrix.

In Q-matrix refinement, a researcher works with an expert-defined Q and uses objective indices to determine, item by item, whether and how the item's q-vector can be modified to improve fit. As an example, de la Torre (2008) proposed a Q-matrix validation method on the basis of an item discrimination index, ς_j. The intuition is as follows: A correctly specified $\mathbf{q}_j$ vector, which requires $K^* < K$ attributes, should differentiate the 2^{K^*} equivalence classes involving the K^* requisite skills, each containing equivalence classes of $\boldsymbol{\alpha}$s associated with the same correct response probability

for item j. The task of Q-refinement is hence transformed to finding a minimum set of K' skills such that the resulting ς^2, a measure of between-equivalence-class variance in correct probability for j, is adequately close to the maximum achieved by $\mathbf{q}_j = \mathbf{1}$ (with 2^K equivalent classes). An initial expert-defined Q is used to estimate the parameters of the G-DINA model (de la Torre 2011; see Equation 14 in Section 2.3). Then for each item j, ς^2 is computed for each candidate Q-vector $\mathbf{q}'_j$, and the most parsimonious $\mathbf{q}'_j$ that is adequately close to the maximum is chosen to replace the original expert specification, $\mathbf{q}_j$. Chiu (2013) proposed another Q-matrix refinement method based on nonparametric classification: An initial expert-defined Q is used to obtain examinees' preliminary proficiency class classifications. Subsequently, the algorithm iterates between (*a*) updating, one by one, items' q-vectors by minimizing the within-equivalent-class residual sum of squares, $RSS_j = \sum_{m=1}^{M} \sum_{i \in C_m} (X_{ij} - \eta_{mj})^2$, and (*b*) reestimating proficiency classifications until no item's *RSS* can be improved.

In Q-matrix learning, in contrast, the researcher estimates the Q-matrix directly from the observed data. Liu & Kang (2019) provided a thorough introduction to this problem and highlighted two key statistical issues of concern:

- Statistical identifiability: whether the Q-matrix, together with the remaining DCM parameters $\boldsymbol{\zeta}, \boldsymbol{\pi}$, can be identified, and if so, when. The problem of identifiability becomes more challenging when the Q-matrix is an additional unknown parameter, which requires checking whether $(\boldsymbol{\zeta}, \boldsymbol{\pi}, Q) \neq (\bar{\boldsymbol{\zeta}}, \bar{\boldsymbol{\pi}}, \bar{Q})$ yields

$$P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\zeta}, \boldsymbol{\pi}, Q) \neq P(\mathbf{X} = \mathbf{x} \mid \bar{\boldsymbol{\zeta}}, \bar{\boldsymbol{\pi}}, \bar{Q}). \quad 24.$$

  Depending on the underlying true model from a multitude of plausible models, the conditions for Q and parameter identifiability differ. Liu et al. (2013) provided sufficient conditions for consistent estimation of Q, up to column swaps (i.e., permutation of the order of attributes), when the underlying DCM is DINA (or equivalently DINO) and the guessing parameter is known. The sufficient conditions were given by (*a*) nonempty attribute classes in the population ($\pi_c > 0$ for any class c); (*b*) monotonicity, in that the true DINA parameters satisfy $1 - s_j > g_j$ for all items; and (*c*) Q-matrix completeness, in that the true underlying Q contains an identity submatrix after row permutations. Chen et al. (2015) extended the results to unknown guessing parameters under two additional conditions: (*d*) At least three items assess each skill, and (*e*) there are at least two identity matrices in the true Q after row permutations. For general restricted latent class models for both binary and ordinal attributes, Fang et al. (2019) provided sets of conditions for the identifiability of items' partial information structure, that is, the equivalent classes of attribute patterns that an item is able to separate.
- Computational intensity: Estimating Q involves searching among all $2^{J \times K}$ possible incidence matrices, in addition to the model parameters $(\boldsymbol{\zeta}, \boldsymbol{\pi})$. Two plausible estimators are (*a*) the maximum likelihood estimator (Liu et al. 2012),

$$\hat{Q}_{ML} = \arg\max_{Q} \max_{\boldsymbol{\zeta}, \boldsymbol{\pi}} L(\boldsymbol{\zeta}, \boldsymbol{\pi}, Q; \underset{\sim}{\mathbf{x}}),$$

  where $L(\cdot)$ is the marginal likelihood of parameters integrating out $\boldsymbol{\alpha}$ given observed responses $\underset{\sim}{\mathbf{x}}$, and (*b*) an estimator that minimizes the L_2 distance between the model-implied distribution and the empirical distribution $[\hat{P}(\mathbf{x})]$ of the response vector (Liu et al. 2013):

$$\min_{Q, \boldsymbol{\pi}, \boldsymbol{\xi}} \sum_{\forall \mathbf{x} \in \{0,1\}^J} |\hat{P}(\mathbf{x}) - P(\mathbf{x} \mid Q, \boldsymbol{\zeta}, \boldsymbol{\pi})|^2.$$

An exhaustive search over $\mathcal{Q}: \{Q_{J\times K}: q_{jk} \in \{0,1\}\}$, however, is computationally costly. Leveraging methods for regularization and variable selection (e.g., Hastie et al. 2009, 2015), Chen et al. (2015) reformulated the problem of Q estimation as revealing a sparse relationship between attributes and items: Recall that the general DCMs in Section 2.3, such as the GDM (von Davier 2005), the G-DINA model (de la Torre 2011), and the LCDM (Henson et al. 2009), have a linear predictor of the item response model that involves main effects, α_{ik}s, and interactions, $\prod_{kk'\ldots} \alpha_{ik}\alpha_{ak'}\ldots$s. Let Λ denote the matrix of λ parameters of each item for a saturated model with all main effects and interactions among K skills, where columns correspond to the possible main effect and up to order-K interactions of attributes. Chen et al. (2015) proposed the following regularized maximum likelihood estimator,

$$(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\pi}}) = \arg\max_{\boldsymbol{\Lambda},\boldsymbol{\pi}} \left[\log L(\boldsymbol{\Lambda}, \boldsymbol{\pi}; \mathbf{x}) - N \sum_{j=1}^{J} p_{\delta_j}(\boldsymbol{\lambda}_j) \right].$$

Here, $\log L(\cdot)$ is the marginal log-likelihood, and $p_{\delta_j}(\cdot)$ is a penalty function applied to regularize the coefficients for the main effect and interactions of attributes. When certain $\hat{\lambda}$s are shrunk to 0 through the regularization, the corresponding combinations of skills will not affect the correct response probability for j. Here, regularization is applied to select the attribute main effects and higher-order interactions that are active for each item. Recall that these main effects and interactions were also present in the general DCMs in Section 2.3, where, on the basis of an expert-provided Q, λs are restricted to zero for attribute combinations not measured by item j (ks s.t. $q_{jk} = 0$; see Equations 12–14). With specific models, e.g., DINA, as special cases of the general models, this framework allows for Q estimation for both specific and general DCMs. G. Fang, J. Liu, & Z. Ying (unpublished manuscript, available upon request) proposed a group LASSO approach to regularizing the coefficients (e.g., Yuan & Lin 2006), which penalizes all the slope coefficients involving k at the group level. Because the Q-matrix defines the relationship between items and the unobserved latent discrete attributes that arise from the population with membership probabilities $\boldsymbol{\pi}$, the methods proposed by Chen et al. (2015) and Fang et al. work with the marginal likelihood and employ the EM algorithm. Researchers have proposed a number of Bayesian approaches to estimating the Q-matrix, for example, drawing Q-matrices that satisfy identifiability and completeness constraints with efficient Metropolis-Hastings samplers (Chen et al. 2018a), incorporating expert input about Q as prior information (Culpepper 2019), revealing sparse structures via Bayesian variable selection (Chen et al. 2020), and inferring the number of attributes through Bayesian Dirichlet processes (Chen et al. 2021).

4. DISCUSSION: EXTENDED APPLICATIONS AND FUTURE DIRECTIONS

Diagnosing examinees' mastery of discrete attributes is vital for both criterion-referenced testing and formative assessment. This article provides an introduction to cognitive diagnostic testing, a class of restricted latent class models that assess mastery of categorical attributes and the multitude of inferential problems, such as identifiability, estimation, and model selection and validation, that are essential to the design and administration of reliable, valid, and practically useful diagnostic tests.

Beyond a stand-alone test, cognitive diagnostic testing can also be integrated into a wide range of scenarios in the learning setting to promote the attainment of educational goals. We briefly mention a few examples here and refer readers to Chang et al. (2021), who discuss how

statistical and machine learning methods can advance educational measurement to better inform personalized learning.

- Cognitive diagnostic computerized adaptive testing: Unlike linear testing, in which all examinees receive the same test questions, in computerized adaptive testing, test questions are chosen adaptively for examinees on the basis of information collected about their latent proficiency in real time (Lord 1980). This approach allows for tailored selection of test questions that most efficiently inform the researcher about the underlying trait of interest, and computerized adaptive testing can substantially reduce test length without compromising measurement reliability. With the virtue of informing learning through diagnosis of fine-grained skill mastery and misconceptions, cognitive diagnostic testing may be used for interim assessments in the learning process. The integration of cognitive diagnostic testing and computerized adaptive testing enables the use of short diagnostic tests to assess students' attribute mastery (see Cheng 2009, Chiu & Chang 2021, Xu et al. 2003) but at the same time reveals a set of statistical questions, such as algorithms for the adaptive selection of test items to maximize efficiency and statistically identify attribute patterns (e.g., Liu et al. 2015, Xu et al. 2016).
- Longitudinal models for learning: Attribute mastery can be tracked over time in a learning setting through interim, formative assessments, such as exercises and quizzes. Latent transition analysis and hidden Markov models have been developed for this setting (e.g., Kaya & Leite 2017; Li et al. 2016; Wang et al. 2018a,b). DCMs serve as the measurement model that connects unobserved skill mastery profiles at each time point to the observed test responses at that time point; see Zhan (2020) for a review. This allows for not only the measurement of students' learning trajectories over time but also modeling and inference as to how latent skill mastery changes as a result of learners' cognitive and noncognitive traits, the effectiveness and characteristics of treatments (e.g., Zhang & Chang 2020), and other covariates (Chen & Culpepper 2020, Wang et al. 2018a). This information can subsequently be used by educators to tailor and adjust instruction in both classrooms and personalized learning settings.
- Recommendation systems for adaptive learning: The recent surge in online, remote education, at both the K–12 and higher education levels, promotes educational equity by improving the accessibility of good-quality educational materials. At the same time, this trend introduces challenges: Learners' backgrounds and needs become more heterogeneous, and face-to-face interaction between students and teachers becomes less feasible. This calls for recommendation systems for adaptive learning, which can tailor instructional routes and interventions on the basis of real-time information about students' attribute mastery. The task of adaptive content recommendation is closely connected to statistical problems of recommendation systems, Markov decision process, and reinforcement learning. Chen et al. (2018b) proposed a framework based on partially observable Markov decision process, where, at each time point, on the basis of the learner's estimated mastery status, the next intervention can be adaptively recommended to optimize an expected reward, which can be a function of skill mastery trajectories (e.g., expected time to reach mastery). Parameters of the reward function can be defined via a longitudinal model for attribute mastery change, as explained by parameters of particular interventions (e.g., Zhang & Chang 2020). However, more realistically, often no initial data support the estimation of such a complex model. Tang et al. (2019) used reinforcement learning to address the so-called cold start problem (see Sutton & Barto 2018), where limited data are available at the initial implementation of a recommendation system. The expected reward function used for content recommendation is updated in an online manner as more data come in over time.

Statistical applications to cognitive diagnostic testing provide fundamental building blocks for the development of reliable and valid diagnostic assessments, yet many open questions remain as to how statistics can further promote the validity and practical feasibility of cognitive diagnostic testing. An example is the modeling of testlet effects in cognitive diagnostic testing, where sets of questions originate from the same stem. Testlets are highly relevant in classroom and language assessments: For example, reading assessments often contain passage-based questions, where several questions involve the same passage, and math assessments are often designed to contain subquestions under a common setup. This approach introduces local dependence among items in the same testlet, a violation of the fundamental assumption of independence in item response modeling. This development has led to a multitude of DCMs that incorporate testlet effects (Hansen 2013; Zhan et al. 2015, 2018b). Recently, Xu et al. (2022) provided a framework for statistical analysis of testlet effects in cognitive diagnostic testing: the interacted testlet DINA model. Under this general model, the DINA ideal response, a continuous testlet-specific random effect, and the interaction between the ideal response and the testlet-specific effect are allowed to jointly affect the correct response probability. Inferences on the slope coefficient for the testlet effect and the interaction between the ideal response and the testlet effect can support hypothesis testing about the existence of the testlet effect (i.e., local dependence within the testlet), as well as the dependency between discrete attribute mastery and the continuous trait for testlet-induced local dependence. Xu et al. (2022) also provided theoretical results on the identifiability of the interacted testlet DINA model, which informs assessment developers as to the Q-matrix structure, the number of testlets, and the required number of items per testlet for constructing a statistically valid testlet-based diagnostic test.

Real-world implementations of cognitive diagnostic testing often face a difficult trade-off between statistical (or psychometric) soundness and practical feasibility: For instance, reliable measurement of an attribute often requires repeated measures (i.e., multiple items measuring the same skill), and the parameter identifiability and Q-matrix completeness conditions call for single-attribute items. These recommendations by statistical theory may be difficult to implement in practice: In a learning setting, long tests containing a large number of items may be too burdensome, even causing interruptions to the stream of instructions. Single-attribute items that do not require the integration of different concepts may also fail to reflect how well a learner has internalized a concept. We believe that the trade-off might be partially resolved through the incorporation of additional information from problem-solving process data, that is, the ordered sequence of observed events that an examinee executes in pursuit of solving a problem. The growth in computer-based testing has enabled the easy collection of such data, and, compared to a final score, they can provide time-stamped information on the steps that a student took to arrive at a final response. A recent study by Zhan & Qiao (2022) expanded the sequence of actions performed on a single constructed response item on the PISA 2012 problem solving assessment (OECD 2014) to a series of dichotomous indicators, representing the presence or absence of certain subsequences of actions in a student's recorded log. A Q-matrix was constructed to represent the relationship between subsequences of actions and the measured attributes, allowing for the use of the log sequence for building and classifying individuals on the basis of DCMs. By decomposing the observed data on a single item into finer-grained subtasks (Wang et al. 2020), the number of indicators (i.e., items) increases, and each step is more likely to involve just a single skill. Yet, process data are extremely complex and highly variable, and expert-based construction of a Q-matrix, especially on more open-ended problems, can be challenging. Recently proposed methods on automated feature extraction from process data provide data-driven methods for extracting meaningful latent features from observed action sequences (Tang et al. 2020, 2021), and the extracted features have been shown to improve measurement reliability in IRT-based ability

assessment (Zhang et al. 2023). Modern statistical and machine learning has great potential to support the use of process information for diagnostic assessments, which can serve a pivotal role in personalized learning.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank the two reviewers for their very careful reading and numerous helpful comments. This research is supported in part by US National Science Foundation grants SES-1826540, DMS-2015417, and SES-2119938.

LITERATURE CITED

Albert JH, Chib S. 1993. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* 88(422):669–79

Allen MJ, Yen WM. 2001. *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press

Anderson JR, Corbett AT, Koedinger KR, Pelletier R. 1995. Cognitive tutors: lessons learned. *J. Learn. Sci.* 4(2):167–207

Baker FB, Kim SH. 2004. *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker

Balamuta JJ, Culpepper SA, Douglas JA. 2020. *edcm: exploratory cognitive diagnostic models*. Package. **https://github.com/tmsalab/edm**

Birnbaum AL. 1968. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*, ed. FM Lord, JW Tukey, MR Novick, pp. 395–479. Reading, MA: Addison-Wesley

Bishop CM. 1995. *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford Univ. Press

Blei DM, Kucukelbir A, McAuliffe JD. 2017. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112(518):859–77

Bloom BS. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Book 1. Cognitive Domain*. New York: Longman

Bock RD, Aitkin M. 1981. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46(4):443–59

Brown JS, VanLehn K. 1980. Repair theory: a generative theory of bugs in procedural skills. *Cogn. Sci.* 4(4):379–426

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, et al. 2017. Stan: a probabilistic programming language. *J. Stat. Softw.* 76(1):1–32

Casella G, Berger RL. 2002. *Statistical Inference*. Belmont, CA: Cengage Learn.

Chang HH, Wang C, Zhang S. 2021. Statistical applications in educational measurement. *Annu. Rev. Stat. Appl.* 8:439–61

Chen J, de la Torre J. 2013. A general cognitive diagnosis model for expert-defined polytomous attributes. *Appl. Psychol. Meas.* 37(6):419–37

Chen Y, Culpepper S, Liang F. 2020. A sparse latent class model for cognitive diagnosis. *Psychometrika* 85(1):121–53

Chen Y, Culpepper SA. 2020. A multivariate probit model for learning trajectories: a fine-grained evaluation of an educational intervention. *Appl. Psychol. Meas.* 44(7–8):515–30

Chen Y, Culpepper SA, Chen Y, Douglas J. 2018a. Bayesian estimation of the DINA Q matrix. *Psychometrika* 83(1):89–108

Chen Y, Li X, Liu J, Ying Z. 2018b. Recommendation system for adaptive learning. *Appl. Psychol. Meas.* 42(1):24–41

Chen Y, Liu J, Xu G, Ying Z. 2015. Statistical analysis of Q-matrix based diagnostic classification models. *J. Am. Stat. Assoc.* 110(510):850–66
Chen Y, Liu Y, Culpepper SA, Chen Y. 2021. Inferring the number of attributes for the exploratory DINA model. *Psychometrika* 86(1):30–64
Cheng Y. 2009. When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74(4):619–32
Chiu CY. 2013. Statistical refinement of the Q-matrix in cognitive diagnosis. *Appl. Psychol. Meas.* 37(8):598–618
Chiu CY, Chang YP. 2021. Advances in CD-CAT: the general nonparametric item selection method. *Psychometrika* 86(4):1039–57
Chiu CY, Douglas J. 2013. A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *J. Classif.* 30(2):225–50
Chiu CY, Douglas JA, Li X. 2009. Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika* 74(4):633–65
Chiu CY, Köhn HF. 2019. Nonparametric methods in cognitively diagnostic assessment. See von Davier & Lee 2019a, pp. 107–32
Chiu CY, Köhn HF, Zheng Y, Henson R. 2016. Joint maximum likelihood estimation for diagnostic classification models. *Psychometrika* 81(4):1069–92
Chiu CY, Sun Y, Bian Y. 2018. Cognitive diagnosis for small educational programs: the general nonparametric classification method. *Psychometrika* 83(2):355–75
Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. *Psychol. Bull.* 52(4):281–302
Culpepper SA. 2015. Bayesian estimation of the DINA model with Gibbs sampling. *J. Educ. Behav. Stat.* 40(5):454–76
Culpepper SA. 2019. Estimating the cognitive diagnosis Q-matrix with expert knowledge: application to the fraction-subtraction dataset. *Psychometrika* 84(2):333–57
Culpepper SA, Hudson A. 2018. An improved strategy for Bayesian estimation of the reduced reparameterized unified model. *Appl. Psychol. Meas.* 42(2):99–115
de la Torre J. 2008. An empirically based method of Q-matrix validation for the DINA model: development and applications. *J. Educ. Meas.* 45(4):343–62
de la Torre J. 2009. DINA model and parameter estimation: a didactic. *J. Educ. Behav. Stat.* 34(1):115–30
de la Torre J. 2011. The generalized DINA model framework. *Psychometrika* 76(2):179–99
de la Torre J, Douglas JA. 2004. Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69(3):333–53
de la Torre J, Lee YS. 2010. A note on the invariance of the DINA model parameters. *J. Educ. Meas.* 47(1):115–27
de la Torre J, van der Ark LA, Rossi G. 2018. Analysis of clinical data from a cognitive diagnosis modeling framework. *Meas. Eval. Counsel. Dev.* 51(4):281–96
Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* 39(1):1–22
DiBello LV, Roussos LA, Stout W. 2006. Review of cognitively diagnostic assessment and a summary of psychometric models. *Handb. Stat.* 26:979–1030
DiBello LV, Stout WF, Roussos LA. 1995. Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. See Nichols et al. 1995, pp. 361–89
Doignon JP, Falmagne JC. 1985. Spaces for the assessment of knowledge. *Int. J. Man Mach. Stud.* 23(2):175–96
Doignon JP, Falmagne JC. 2012. *Knowledge Spaces*. Berlin: Springer Science & Business Media
Falmagne JC, Koppen M, Villano M, Doignon JP, Johannesen L. 1990. Introduction to knowledge spaces: how to build, test, and search them. *Psychol. Rev.* 97(2):201–24
Fang G, Liu J, Ying Z. 2019. On the identifiability of diagnostic classification models. *Psychometrika* 84(1):19–40
Fisher RA. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7(2):179–88
Gabrielsen A. 1978. Consistency and identifiability. *J. Econom.* 8(2):261–63
Gierl MJ, Leighton JP, Hunka SM. 2000. An NCME instructional module on exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educ. Meas. Issues Pract.* 19(3):34–44
Glaser R. 1963. Instructional technology and the measurement of learning outcomes: some questions. *Am. Psychol.* 18(8):519–21

Glaser R, Nitko AJ. 1970. *Measurement in learning and instruction*. Tech. Rep., Univ. Pittsburgh R&D Cent.
Gu Y, Xu G. 2019. The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika* 84(2):468–83
Gu Y, Xu G. 2020. Partial identifiability of restricted latent class models. *Ann. Stat.* 48(4):2082–107
Gyllenberg M, Koski T, Reilink E, Verlaan M. 1994. Non-uniqueness in probabilistic numerical identification of bacteria. *J. Appl. Probab.* 31(2):542–48
Haertel EH. 1990. Continuous and discrete latent structure models for item response data. *Psychometrika* 55(3):477–94
Hambleton RK, Novick MR. 1973. Toward an integration of theory and method for criterion-referenced tests. *J. Educ. Meas.* 10(3):159–70
Hansen MP. 2013. *Hierarchical item response models for cognitive diagnosis*. Thesis, Univ. Calif., Los Angeles
Hartigan JA. 1975. *Clustering Algorithms*. New York: John Wiley & Sons
Hartigan JA, Wong MA. 1979. Algorithm AS 136: a *K*-means clustering algorithm. *J. R. Stat. Soc. C* 28(1):100–8
Hartz SM. 2002. *A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality*. Thesis, Univ. Illinois Urbana-Champaign
Hastie T, Tibshirani R, Friedman JH. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer
Hastie T, Tibshirani R, Wainwright M. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: CRC Press
Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109
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
Henson RA, Templin JL, Willse JT. 2009. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74(2):191–210
Jiang H. 1996. *Applications of computational statistics in cognitive diagnosis and IRT modeling*. Thesis, Univ. Illinois Urbana-Champaign
Jiang Z, Carter R. 2019. Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. *Behav. Res. Methods* 51(2):651–62
Johnson SC. 1967. Hierarchical clustering schemes. *Psychometrika* 32(3):241–54
Junker BW, Sijtsma K. 2001. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25(3):258–72
Kaya Y, Leite WL. 2017. Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance. *Educ. Psychol. Meas.* 77(3):369–88
Khorramdel L, Shin HJ, von Davier M. 2019. GDM software *mdltm* including parallel EM algorithm. See von Davier & Lee 2019a, pp. 603–28
Köhn HF, Chiu CY. 2016. A proof of the duality of the DINA model and the DINO model. *J. Classif.* 33(2):171–84
Köhn HF, Chiu CY. 2017. A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika* 82(1):112–32
Langeheine RE, Rost JE. 1988. *Latent Trait and Latent Class Models*. Boston, MA: Springer
Lazarsfeld PF, Henry NW. 1968. *Latent Structure Analysis*. New York: Houghton Mifflin
Leighton JP, Gierl MJ, Hunka SM. 2004. The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *J. Educ. Meas.* 41(3):205–37
Li F, Cohen A, Bottge B, Templin J. 2016. A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76(2):181–204
Liu J, Kang HA. 2019. Q-matrix learning via latent variable selection and identifiability. See von Davier & Lee 2019a, pp. 247–63
Liu J, Xu G, Ying Z. 2012. Data-driven learning of Q-matrix. *Appl. Psychol. Meas.* 36(7):548–64
Liu J, Xu G, Ying Z. 2013. Theory of the self-learning *Q*-matrix. *Bernoulli* 19(5A):1790–817
Liu J, Ying Z, Zhang S. 2015. A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika* 80(2):468–90

Lord FM. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, NJ: Lawrence Erlbaum Assoc.
Lunn D, Spiegelhalter D, Thomas A, Best N. 2009. The BUGS project: evolution, critique and future directions. *Stat. Med.* 28(25):3049–67
Ma W, de la Torre J. 2020. GDINA: an R package for cognitive diagnosis modeling. *J. Stat. Softw.* 93:1–26
MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, No. 14, pp. 281–97. Oakland, CA: Univ. Calif. Press
Macready GB, Dayton CM. 1977. The use of probabilistic models in the assessment of mastery. *J. Educ. Stat.* 2(2):99–120
Maris E. 1999. Estimating multiple classification latent class models. *Psychometrika* 64(2):187–212
Nelder JA, Wedderburn RW. 1972. Generalized linear models. *J. R. Stat. Soc. A* 135(3):370–84
Neyman J, Pearson ES. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. A* 231(694–706):289–337
Nichols PD, Chipman SF, Brennan RL, eds. 1995. *Cognitively Diagnostic Assessment*. Mahwah, NJ: Lawrence Erlbaum Assoc.
OECD. 2014. *PISA 2012 Results: Creative Problem Solving: Students' Skills in Tackling Real-Life Problems. Volume V*. Paris: OECD
Piaget J. 1950. *The Psychology of Intelligence*. New York: Harcourt Brace
Ripley BD. 1996. *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge Univ. Press
Roussos LA, DiBello LV, Stout W, Hartz SM, Henson RA, Templin JL. 2007. The Fusion Model skills diagnosis system. In *Cognitive Diagnostic Assessment for Education: Theory and Applications*, ed. J Leighton, M Gierl, pp. 275–318. Cambridge, UK: Cambridge Univ. Press
Rupp AA, Templin J. 2008. The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educ. Psychol. Meas.* 68(1):78–96
Rupp AA, Templin J, Henson RA. 2010. *Diagnostic Measurement: Theory, Methods, and Applications*. New York: Guilford
Samejima F. 1995. A cognitive diagnosis method using latent trait models: competency space approach and its relationship with Dibello and Stout's unified cognitive-psychometric diagnosis model. See Nichols et al. 1995, pp. 391–410
Sessoms J, Henson RA. 2018. Applications of diagnostic classification models: a literature review and critical commentary. *Meas. Interdiscip. Res. Perspect.* 16(1):1–17
Spearman C. 1904. The proof and measurement of association between two things. *Am. J. Psychol.* 15:72–101
Sutton RS, Barto AG. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press
Tang X, Chen Y, Li X, Liu J, Ying Z. 2019. A reinforcement learning approach to personalized learning recommendation systems. *Br. J. Math. Stat. Psychol.* 72(1):108–35
Tang X, Wang Z, He Q, Liu J, Ying Z. 2020. Latent feature extraction for process data via multidimensional scaling. *Psychometrika* 85(2):378–97
Tang X, Wang Z, Liu J, Ying Z. 2021. An exploratory analysis of the latent structure of process data via action sequence autoencoders. *Br. J. Math. Stat. Psychol.* 74(1):1–33
Tatsuoka KK. 1983. Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* 20(4):345–54
Tatsuoka KK. 1984. Caution indices based on item response theory. *Psychometrika* 49(1):95–110
Tatsuoka KK. 1990. Toward an integration of item-response theory and cognitive error analysis. In *Diagnostic Monitoring of Skill and Knowledge Acquisition*, ed. N Frederiksen, pp. 453–88. Hillsdale, NJ: Lawrence Erlbaum Assoc.
Tatsuoka KK. 1995. Architecture of knowledge structures and cognitive diagnosis: a statistical pattern recognition and classification approach. See Nichols et al. 1995, pp. 327–59
Tatsuoka KK. 2009. *Cognitive Assessment: An Introduction to the Rule Space Method*. New York: Routledge
Tatsuoka KK, Birenbaum M, Tatsuoka MM, Baillie R. 1980. *Psychometric approach to error analysis on response patterns of achievement tests*. Tech. Rep., Comput. Based Educ. Res. Lab., Univ. Illinois Urbana-Champaign
Tatsuoka KK, Tatsuoka MM. 1983. Spotting erroneous rules of operation by the individual consistency index. *J. Educ. Meas.* 20(3):221–30

Templin JL, Henson RA. 2006. Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11(3):287–305
Tyler RW, Gagne RM, Scriven M. 1972. *Perspectives of Curriculum Evaluation* (AERA Monograph Series on Curriculum Evaluation, Vol. 1). Chicago: Rand McNally
von Davier M. 2005. A general diagnostic model applied to language testing data. *ETS Res. Rep. Ser.* 2005(2):i–35
von Davier M. 2018. Diagnosing diagnostic models: from von Neumann's elephant to model equivalencies and network psychometrics. *Meas. Interdiscip. Res. Perspect.* 16(1):59–70
von Davier M. 2019. The general diagnostic model. See von Davier & Lee 2019a, pp. 133–53
von Davier M, Lee Y-S. 2019a. *Handbook of Diagnostic Classification Models*. Cham, Switz.: Springer
von Davier M, Lee Y-S. 2019b. Introduction: from latent classes to cognitive diagnostic models. See von Davier & Lee 2019a, pp. 1–17
Wang S, Douglas J. 2015. Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika* 80(1):85–100
Wang S, Yang Y, Culpepper SA, Douglas JA. 2018a. Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden Markov model with covariates. *J. Educ. Behav. Stat.* 43(1):57–87
Wang S, Zhang S, Douglas J, Culpepper S. 2018b. Using response times to assess learning progress: a joint model for responses and response times. *Meas. Interdiscip. Res. Perspect.* 16(1):45–58
Wang Z, Tang X, Liu J, Ying Z. 2020. Subtask analysis of process data through a predictive model. arXiv:2009.00717 [cs.HC]
Ward JH Jr. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58(301):236–44
Xu G. 2017. Identifiability of restricted latent class models with binary responses. *Ann. Stat.* 45(2):675–707
Xu G. 2019. Identifiability and cognitive diagnosis models. See von Davier & Lee 2019a, pp. 333–57
Xu G, Shang Z. 2018. Identifying latent structures in restricted latent class models. *J. Am. Stat. Assoc.* 113(523):1284–95
Xu G, Wang C, Shang Z. 2016. On initial item selection in cognitive diagnostic computerized adaptive testing. *Br. J. Math. Stat. Psychol.* 69(3):291–315
Xu G, Zhang S. 2016. Identifiability of diagnostic classification models. *Psychometrika* 81(3):625–49
Xu X, Chang H, Douglas J. 2003. *A simulation study to compare CAT strategies for cognitive diagnosis*. Presented at Annu. Meet. Am. Educ. Res. Assoc., Chicago, April 21–25
Xu X, Fang G, Guo J, Ying Z, Zhang S. 2022. Modeling interactive testlet effect under diagnostic classification models. In preparation
Yamaguchi K, Okada K. 2020. Variational Bayes inference algorithm for the saturated diagnostic classification model. *Psychometrika* 85(4):973–95
Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* 68(1):49–67
Zhan P. 2020. Longitudinal learning diagnosis: minireview and future research directions. *Front. Psychol.* 11:1185
Zhan P, Jiao H, Liao D. 2018a. Cognitive diagnosis modelling incorporating item response times. *Br. J. Math. Stat. Psychol.* 71(2):262–86
Zhan P, Li X, Wang WC, Bian Y, Wang L. 2015. The multidimensional testlet-effect cognitive diagnostic models. *Acta Psychol. Sin.* 47(5):689–701
Zhan P, Liao M, Bian Y. 2018b. Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Front. Psychol.* 9:607
Zhan P, Qiao X. 2022. DIAGNOSTIC classification analysis of problem-solving competence using process data: an item expansion method. *Psychometrika* 87(4):1529–47
Zhang S, Chang HH. 2020. A multilevel logistic hidden Markov model for learning under cognitive diagnosis. *Behav. Res. Methods* 52(1):408–21
Zhang S, Wang Z, Qi J, Liu J, Ying Z. 2023. Accurate assessment via process data. *Psychometrika*. In press

Annual Review of Statistics and Its Application

Volume 10, 2023

Contents

Fifty Years of the Cox Model
John D. Kalbfleisch and Douglas E. Schaubel 1

High-Dimensional Survival Analysis: Methods and Applications
Stephen Salerno and Yi Li 25

Shared Frailty Methods for Complex Survival Data: A Review of Recent Advances
Malka Gorfine and David M. Zucker 51

Surrogate Endpoints in Clinical Trials
Michael R. Elliott 75

Sustainable Statistical Capacity-Building for Africa: The Biostatistics Case
Tarylee Reddy, Rebecca N. Nsubuga, Tobias Chirwa, Ziv Shkedy, Ann Mwangi, Ayele Tadesse Awoke, Luc Duchateau, and Paul Janssen 97

Confidentiality Protection in the 2020 US Census of Population and Housing
John M. Abowd and Michael B. Hawes 119

The Role of Statistics in Promoting Data Reusability and Research Transparency
Sarah M. Nusser 145

Fair Risk Algorithms
Richard A. Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen 165

Statistical Data Privacy: A Song of Privacy and Utility
Aleksandra Slavković and Jeremy Seeman 189

A Brief Tour of Deep Learning from a Statistical Perspective
Eric Nalisnick, Padhraic Smyth, and Dustin Tran 219

Statistical Deep Learning for Spatial and Spatiotemporal Data
Christopher K. Wikle and Andrew Zammit-Mangion 247

Statistical Machine Learning for Quantitative Finance
M. Ludkovski 271

Models for Integer Data
Dimitris Karlis and Naushad Mamode Khan 297

Generative Models: An Interdisciplinary Perspective
Kris Sankaran and Susan P. Holmes 325

Data Integration in Bayesian Phylogenetics
Gabriel W. Hassler, Andrew F. Magee, Zhenyu Zhang, Guy Baele, Philippe Lemey, Xiang Ji, Mathieu Fourment, and Marc A. Suchard 353

Approximate Methods for Bayesian Computation
Radu V. Craiu and Evgeny Levi 379

Simulation-Based Bayesian Analysis
Martyn Plummer 401

High-Dimensional Data Bootstrap
Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike 427

Innovation Diffusion Processes: Concepts, Models, and Predictions
Mariangela Guidolin and Piero Manfredi 451

Graph-Based Change-Point Analysis
Hao Chen and Lynna Chu 475

A Review of Generalizability and Transportability
Irina Degtiar and Sherri Rose 501

Three-Decision Methods: A Sensible Formulation of Significance Tests—and Much Else
Kenneth M. Rice and Chloe A. Krakauer 525

Second-Generation Functional Data
Salil Koner and Ana-Maria Staicu 547

Model-Based Clustering
Isobel Claire Gormley, Thomas Brendan Murphy, and Adrian E. Raftery 573

Model Diagnostics and Forecast Evaluation for Quantiles
Tilmann Gneiting, Daniel Wolffram, Johannes Resin, Kristof Kraus, Johannes Bracher, Timo Dimitriadis, Veit Hagenmeyer, Alexander I. Jordan, Sebastian Lerch, Kaleb Phipps, and Melanie Schienle 597

Statistical Methods for Exoplanet Detection with Radial Velocities
Nathan C. Hara and Eric B. Ford 623

Statistical Applications to Cognitive Diagnostic Testing
Susu Zhang, Jingchen Liu, and Zhiliang Ying 651

Player Tracking Data in Sports
Stephanie A. Kovalchik 677

Six Statistical Senses
Radu V. Craiu, Ruobin Gong, and Xiao-Li Meng 699

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics