IMPACT OF ASSUMPTION VIOLATIONS IN THE ONE-PARAMETER LOG-LINEAR COGNITIVE DIAGNOSIS MODEL

By

OLUWATOSIN ADELEYE

(Under the Direction of Matthew J. Madison)

ABSTRACT

Diagnostic classification models (DCMs) are psychometric models that enable the categorization of examinees based on their proficiency (proficient or non-proficient) in specific skills or attributes. The information from DCMs can aid educators in identifying specific areas where examinees require additional support and provide actionable feedback. This study examines the robustness of a recently developed DCM, the one-parameter log-linear cognitive diagnosis model (1-PLCDM), that prioritizes interpretability and ease of application. Through a simulation study, I demonstrate that the 1-PLCDM maintains robust classification accuracy and reliability when the assumptions of attribute independence and a simple Q-matrix structure are violated.

INDEX WORDS: diagnostic classification models, cognitive diagnosis model, one-parameter log-linear cognitive diagnosis model, one-parameter logistic, robustness, Q-matrix misspecification, attribute correlation.

IMPACT OF ASSUMPTION VIOLATIONS IN THE ONE-PARAMETER LOG-LINEAR COGNITIVE DIAGNOSIS MODEL

By

OLUWATOSIN ADELEYE

B.S., Obafemi Awolowo University, Nigeria, 2022

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment

of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2025

IMPACT OF ASSUMPTION VIOLATIONS IN THE ONE-PARAMETER LOG-LINEAR COGNITIVE DIAGNOSIS MODEL

by

OLUWATOSIN ADELEYE

Major Professor: Matthew J. Madison
Committee: Zhenqiu Lu
Shiyu Wang

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2025

DEDICATION

To my parents, Gbenga and Roli Adeleye, whose unwavering strength, dedication, and love have been the foundation of our family, and to my sister, Dami Adeleye, for being a constant source of joy and support.

ACKNOWLEDGEMENTS

First and most of all, I want to sincerely thank my advisor, Dr. Matthew Madison. His guidance, patience, encouragement, and deep expertise carried me through every stage of this thesis. I'm especially grateful for him, this would not have been possible without him.

I also want to thank my committee members, Dr. Shiyu Wang and Dr. Laura Lu, for their thoughtful feedback and support.

To my dear friends, thank you for standing by me throughout this journey. Your love, encouragement, and presence have meant more to me than words can say.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Effectively identifying and diagnosing students' skills is essential for ensuring an accurate educational assessment. This serves as a vital tool for teachers in monitoring student progress and ensuring meaningful learning outcomes. Thus, accurate classification of students into different level of proficiency (proficient / non-proficient), enables educators to identify the specific areas where students need support and provide feedback that highlights each student's strengths and areas for improvement. Also, it helps teachers determine if adjustments to their teaching methods are needed to enhance students' understanding of the material.

Diagnostic classification models (DCMs) are psychometric models which classify students based on their mastery of specific skills, classifying them into "proficient" or "non-proficient", "mastery" or "non-mastery", (Rupp et al., 2010). This model provides insights that broaden educators' perspectives on overall student performance, helping them identify the specific skills students have mastered and the areas where additional support is needed. DCMs use criterion-referenced score interpretations that assess a student's mastery of predefined skills, rather than comparing performance to others. This makes it a valuable tool for classification, helping to determine what students know and can do in relation to the specific content of the test. A key feature of DCMs is the use of a Q-matrix, which defines the relationship between test items and attributes. This allows DCMs to estimate whether examinees have mastered specific attributes based on their responses. This diagnosis provides detailed, actionable feedback for both educators and learners (Maas et al., 2022).

In the world of DCMs, there are several models; the log-linear cognitive diagnosis (LCDM) model (Henson et al., 2009), deterministic input noisy or gate (DINO) model (Templin & Henson, 2006), the deterministic inputs noisy and gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001), the compensatory re-parameterized unified model, (C-RUM; Hartz, 2002) and more. These models are designed to be applied in diverse settings based on several reasons such as flexibility, ease of interpretation, statistical fit and so on. The DINA model is the most widely used DCM model in educational assessment due to its estimation simplicity and ease of interpretation, even though it is not necessarily the best in relation to flexibility and statistical fit (36%; Sessoms & Henson, 2018). On the other hand, the LCDM is one of the most flexible discussed models in the literature on DCMs, designed to accommodate a variety of attribute interactions, it is well suited for complex assessments requiring detailed diagnostic feedback. While, complex models, like the LCDM are flexible and capable of fitting data more accurately, parameter estimates can often be difficult to understand and interpret even for experienced psychometricians (Bradshaw and Levy, 2019). This is why the DINA model is the most widely used, because practitioners and researchers tend to prefer models that are easier to understand and interpret. They place a high value on simplicity, especially in educational settings where it's important that the results are clear and the way classifications are made is easily understood. Simpler models often have straightforward interpretations, which is preferred as long as the results are accurate and reliable. In practical applications, ease of estimation and interpretability often outweighs the complexity of achieving the best statistical fit. As a result of these, less complex DCMs continue to be developed prioritizing simplicity and ease of interpretation while maintaining accuracy and reliability. One such model is the one-parameter log-linear cognitive diagnosis model (1-PLCDM) which was designed for this purpose. This model forgoes some flexibility to

enhance understanding and interpretation of parameter estimates. This paper aims to assess the robustness of the 1-PLCDM.

1-PLCDM

The 1-PLCDM is a recently developed DCM that offers a straightforward interpretation of parameter estimates. It imposes item parameter constraints on the LCDM, estimating a singular main effect for all items assessing the same attribute, along with an intercept for each item, similar to the Rasch model which includes a difficulty parameter for each item and a single discrimination parameter across all items (DeMars, 2010).
The measurement properties of the 1-PLCDM are analogous to those of the Rasch model, including sum score sufficiency, item and person free measurement, and invariant item and person ordering. The sum score sufficiency property in the 1-PLCDM means that the total score, or the sum of correct responses across items, is all that is required to ascertain examinee probability of proficiency. Item and person free measurement implies that item parameters remain consistent, regardless of the group of people taking the test. Invariant item and person ordering ensure that if one item is more difficult than another for an individual, this holds true across all examinees, while higher-proficiency individuals are consistently more likely to answer items correctly than those with lower proficiency levels.

While the 1-PLCDM is easy to interpret and implement, it does make certain assumptions which limits the performance of the model. One of those attribute independence (Madison et al., 2024), in other words, it assumes that attributes being tested do not correlate with each other. In many real-world educational situations, multiple attributes often correlate, this is where the model can begin to struggle, as these interactions might reduce its ability to classify students accurately. As a result, it could affect the instructional support a student receives, making it harder to tailor the right interventions. It also assumes that the Q-matrix

has a simple structure, that is an item cannot assess more than one attribute, in many educational settings, attributes often interact, and specific test items may assess multiple attributes, needing a more complex Q-matrix structure. Given these circumstances, the goal of this study is to understand how correlated attributes can affect the model's performance. For example, if two attributes are highly related (e.g., a student's ability to understand both reading comprehension and vocabulary at the same time), how does that influence the accuracy of the model's classification? The second part of this study examines what happens when the Q-matrix, which maps the relationships between test items and student attributes, becomes more complex. When the Q-matrix doesn't follow a simple structure, the assumptions of the 1-PLCDM may not hold, and that could impact its ability to classify students correctly. By understanding these limitations, we can improve how we use these models and make them more effective for everyday educational use.

The study's purposes are to analyze:

1. the degree to which correlation between attributes affects the performance of the 1-PLCDM, and
2. the impact of assuming the Q-matrix is simple structure when it has complex items.

CHAPTER 2

LITERATURE REVIEW

This chapter begins with an overview of DCMs, followed by a review of the current literature on various types of DCMs and DCMs misspecification. It then highlights the focus of this study, the 1-PLCDM. The primary goal of this chapter is to provide a comprehensive overview of the existing literature on DCMs and the 1-PLCDM.

DCMs

DCMs, as described by Rupp et al. (2010), are psychometric models that assign probabilistic classifications to respondents by analyzing their mastery of discrete latent variables, referred to as "attributes". These attributes, when utilized in educational measurement, generally describe the skills essential for solving specific challenges. Students are grouped according to their demonstrated mastery of these skills, resulting in attribute profiles that clearly distinguish mastered attributes from not-mastered attributes.

DCM's are special latent class models (is a constrained and confirmatory latent class model), this means that they constrain force attribute mastery into increased correct response probabilities. A Q-matrix, which defines the relationship between test items and attributes, is used to enforce this structure. This guarantee that people who are proficient in a certain trait are more likely to provide accurate answers to specific questions. DCM is a confirmatory latent class model because with DCMs, the classes are known and predefined and they correspond to the attribute profiles, these profiles are outlined in Q-matrix that indicates which items measure which attributes.

In contrast to the scaled scores derived from Item Response Theory (IRT) models which allow for norm-referenced interpretations, which are used in various assessment contexts where individual performance comparison within a group is needed, such as college admissions testing (e.g., SAT, Graduate Record Examination (GRE)) and extensive international studies (e.g., Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS)), DCMs, makes use of criterion-referenced interpretations which is useful for diagnosis. This makes DCMs particularly valuable in contexts where diagnosing individual strengths and weaknesses is more useful than ranking students.

Types of DCMs

Based on their structure and underlying assumptions, DCMs can be broadly labeled as "general DCMs" or "constrained DCMs".

General DCMs

General DCMs have the ability to better capture the underlying processes that determine how students master different attributes, which influences their item responses. One key advantage of using a general model is its ability to accommodate a wide range of assumptions about both items and attributes. This flexibility enables researchers to test whether certain model constraints are truly necessary based on data provided. Essentially, these general models serve as parent frameworks from which more specific diagnostic models can be derived by enforcing statistical constraint on the parameters. Each specific DCM within these frameworks makes different assumptions about how student skills interact. For example, some models assume that a lack of mastery in one skill can be offset by mastery in another, while others do not allow for such compensatory effects. General DCMs includes,

general diagnostic model (GDM; Von Davier, 2005), the generalized deterministic inputs, noisy "and" gate model (G-DINA; de la Torre, 2011), and LCDM.

The LCDM is a general DCM that has emerged as one of the most flexible models in the literature on DCMs. The LCDM is designed to model the conditional item response probabilities. It accommodates a variety of attribute interactions, the LCDM provides a versatile framework that can model both compensatory and non-compensatory relationships between attributes, making it well-suited for complex assessments requiring detailed diagnostic feedback. Previous studies have shown accuracy of the LCDM model heavily relies on the correct specification of the Q-matrix (Madison & Bradshaw, 2014, Rupp and Templin, 2008).

The LCDM, introduced by Henson, Templin, and Willse (2009), is grounded in a log-linear framework that models the probabilistic relationship between multiple attributes and item responses. Furthermore, the log-linear model is formulated in terms of the log-odds of a correct response for each item, contingent upon the latent variables. For an item measuring two attributes, attribute 2 and attribute 3, the form of LCDM item response function is: Complex structure:

$$P(X_i = 1|\alpha_C) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1(2)}\,(\alpha_2) + \lambda_{i,1\,(3)}\,(\alpha_3) + \lambda_{i,2(2,3)}\,(\alpha_2.\alpha_3))}{1 + \exp(\lambda_{i,0} + \lambda_{i,1(2)}(\alpha_2) + \lambda_{i,1\,(3)}\,(\alpha_3) + \lambda_{i,2(2,3)}\,(\alpha_2.\alpha_3))}$$

$\lambda_{i,0}$, is known as the intercept, it indicates the log-odds that a student who has not mastered neither attribute 2, nor attribute 3 measured by item i will provide a correct response. $\lambda_{i,1\ (2)}$, It indicates the increase in the log-odds of a correct response that a student gets for mastering attribute 2.

$\lambda_{i,1\ (3)}$, It indicates the increase in the log-odds of a correct response that a student gets for mastering attribute 3.

$\lambda_{i,2(2,3)}$ is the sum of all of the products of the two-way interaction parameters, it indicates the change in the log-odds of a correct response that a student gets for mastering both attribute 2 and attribute 3.

In the estimation of the LCDM, main effect parameters are constrained to be greater than zero so that examinees who have mastered more of the measured attributes on an item have increasing predicted correct response probabilities, and to prevent latent class switching (Lao & Templin, 2016).

The LCDM's primary advantage is its flexibility; it can represent both compensatory and non-compensatory interactions between attributes, allowing for a range of diagnostic inferences based on the hypothesized structure of the cognitive processes being assessed. This model can capture complex interactions between attributes, where mastery of some attributes may compensate for deficiencies in others, or where certain combinations of attributes must be mastered together to achieve a correct response (Kunina-Habenicht, 2012). One of the LCDM's significant strengths lies in its ability to serve as a foundational model for other DCMs. Simpler models like the DINA can be represented as special cases of the LCDM or G-DINA by adjusting its parameters.

The LCDM stands out for its flexibility and robustness, especially in handling complex attribute interactions. Simulation studies have demonstrated the robustness of the LCDM under various conditions, including Q-matrix misspecifications and different attribute correlations. Results shows, the accuracy of the model is highly dependent on the correct specification of the Q-matrix, which defines the mapping between items and attributes. A poorly specified Q-matrix can lead to biased estimates and misclassifications, underscoring the need for careful design and validation (Madison & Bradshaw, 2014).

Although the LCDM is powerful and adaptable, its complexity in interpretation can sometimes make it challenging to apply in practice. Building on the flexibility of the LCDM, researchers have worked to create simpler models that are easier to use while still providing accurate and reliable diagnostic information.

The G-DINA model is also a type of general DCM. Where the LCDM use the logit link function the G-DINA uses the identity link function. This model extends the flexibility of the DINA model by allowing for compensatory, conjunctive, and disjunctive interactions among attributes. It provides a general framework for modeling multiple cognitive processes, accommodating scenarios where mastery of one attribute can compensate deficiencies in another. This flexibility makes G-DINA highly adaptable to diverse testing conditions, offering a more nuanced understanding of examinee abilities than the all-or-nothing approach of DINA (Chen & de la Torre, 2013). Like the LCDM, the accuracy of the G-DINA model heavily relies on the correct specification of the Q-matrix.

Constrained DCMs

As highlighted by Henson et al.,(2009), the parameters in the general DCMs can be adjusted in various ways to create different DCMs, each designed to reflect different behavior in estimation. Attribute behavior can range from completely non-compensatory to partially compensatory to fully compensatory. In compensatory models, mastery of one attribute can compensate the lack of mastery in another attribute. In contrast, non-compensatory models require mastery of specific attributes, as the absence of a necessary attribute cannot be compensated for by mastering others. Constrained DCMs includes, DINA, CRUM, 1-PLCDM, DINO model, and so on.

The DINA model, one of the most widely used DCMs (Sessoms & Henson, 2018), strictly non-compensatory model, DINA assumes that all relevant attributes must be mastered

to produce a correct response. It incorporates two item-level parameters: the "guessing" parameter, which estimates the chance of a correct answer from a student lacking mastery, and the "slipping" parameter, which accounts for the possibility that a student with full mastery may still respond incorrectly. While the DINA model's simplicity makes it easy to use and interpret, it has certain limitations, it is not the most flexible model. In the estimation of the DINA model, the main effects are constrained a priori to be zero, only the intercept and interaction are estimated. This constraint forces attributes to behave in a non-compensatory way, meaning that non-mastery of one measured attribute cannot be compensated for by mastery of other measured attributes. Its rigid nature can lead to misclassifications in scenarios where compensatory attributes are important (Bradshaw & Templin, 2014).

Despite this, the DINA model remains attractive due to its straightforward structure and diagnostic insights, especially for educational practitioners seeking actionable feedback making it the most commonly used DCM (36%; Sessoms & Henson, 2018). For an item measuring two attributes, the form of DINA item response function is:

$$P(X_i = 1|\alpha_C) = \frac{\exp(\lambda_{i,0} + \lambda_{i,(23)}(\alpha_2.\alpha_3))}{1 + \exp(\lambda_{i,0} + \lambda_{i,(23)}(\alpha_2.\alpha_3))}$$

In the C-RUM model, also known as the additive CDM (ACDM; de la Torre, 2011), mastering additional attributes increases the probability of a correct response. This reflects the compensatory nature of the model, where the mastery of one attribute can aid the lack of mastery in another. For complex items, the C-RUM only estimates the main effects of the attributes and constrains interaction terms to be zero. This simplification makes the model more straightforward to interpret and apply, but it also limits its ability to capture complex dependencies between attributes. For an item measuring two attributes, the form of CRUM item response function is:

$$P(X_i = 1|\alpha_C) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1(2)} + \lambda_{i,1(3)})}{1 + \exp(\lambda_{i,0} + \lambda_{i,1(2)} + \lambda_{i,1(3)})}$$

THE 1-PLCDM

This inclination towards simplicity continues to be addressed by the advancement of simple DCMs. A recently presented unique and simple DCM which prioritizes interpretability above complexity, the model is referred to as the 1-PLCDM (Madison et al., 2023). This can be accomplished by enforcing item parameter constraints on the LCDM. The 1-PLCDM estimates a singular main effect for all items assessing the same attribute and an intercept for each item. The model is analogous to the unidimensional Rasch model and 1-PL IRT models, both of which utilize a singular discrimination parameter for all items and an individual difficulty parameter for each item (DeMars, 2010). Madison et al. (2023) introduced this model in a single-attribute context and demonstrated its advantages, such as invariant measurement and the sufficiency of sum scores, using real data. They only defined it in the single attribute context, where the form of 1-PLCDM item response function is:

$$P(X_i = 1|\alpha_C) = \frac{\exp(\lambda_{i,0} + \lambda_1\alpha_1)}{1 + \exp(\lambda_{i,0} + \lambda_1\alpha_1)}$$

Notice that the main effect is not subscripted because it is constrained to be equal across all items. Prior studies suggest that even when the equality constraint on main effects is violated, the 1-PLCDM still performs well, maintaining classification accuracy and reliability across different sample sizes and test lengths (Maas et al., 2024).

Single Attribute 1-PLCDM

The single attribute 1-PLCDM centers around on a single latent attribute. Its purpose is to determine whether a respondent has mastered one specific attribute by examining their responses to the items designed to assess a skill. However, the assumption of a single

attribute limits the model's ability in real-world assessments, where tests measure multiple attributes. By narrowing its focus to a single attribute, it is not applicable in multi-dimensional assessment. Expanding this model to handle multiple attributes was suggested as an area for future research (Madison et al., 2023).

Multi-attribute 1-PLCDM

The multi-attribute 1-PLCDM builds upon the foundation of its single-attribute by enabling the assessment of multiple cognitive skills or traits at once (Madison et al., 2023). Through this modification, the model becomes more versatile for multi attribute assessments like many diagnostic tests in educational settings. Despite assessing multiple attributes, the model maintains its structural simplicity by estimating a singular main effect for all items measuring the same attribute. This simplicity allows it to offer clear diagnostic feedback across various attributes without the additional complexity of multiple item parameters. To achieve the same properties (sum score sufficiency, invariant item/person ordering) as the single attribute 1-PLCDM, the model depends on a simple Q-matrix and assumes that the attributes are independent.

DCM Misspecification

DCMs, and other psychometrics, have certain components that need to be specified according to the assessment. In certain situations, these specifications can be incorrect, potentially leading to biased results. For DCMs, there is the structural model, which parameterizes how attribute prevalence and how attributes are related to each other. And there is the measurement model, which parametrizes how attribute mastery corresponds to item responses. Here, I review research that has looked at DCM misspecifications of both model components and their effect on model performance.

Q-matrix Misspecification

The Q-matrix is a binary matrix used in cognitive diagnosis models (CDMs) to map test items to the attributes they assess. Each row represents an item, and each column represents a skill/attribute, with entries indicating whether a skill/attribute is required for an item. The Q-matrix is an item-by-attribute matrix of 0s and 1s indicating which attributes are measured on each item. If item 2 requires attribute 2, then cell 2 in the Q-matrix will be 1, and 0 otherwise.

Kunina-Habenicht et al. (2012) identified two types of Q-matrix misspecification that impact classification accuracy; "*underspecification*" which happens when a skill that is actually needed for a test question is mistakenly left out (changing a "1" to "0"), as a result, the model does not estimate enough parameters for that question, leading to an incomplete or incorrect understanding of what skills are required, while "*overspecification"* occurs when extra, unnecessary skills are assigned to a test question (changing a "0" to "1"). This causes the model to estimate more parameters than needed, adding noise and reducing accuracy (2012). Results from the study revealed adverse effect on classification accuracy and parameter recovery of latent class distributions, correlations, and attribute proportions.

De la Torre (2009) found that while small errors might be tolerated, too many lead to a significant drop in the Q-matrix diagnostic accuracy.

Rupp and Templin (2007) investigated Q-matrix misspecification in the DINA Model, examining its impact on parameter estimates and classification accuracy. Their findings indicate that incorrectly deleting attributes from the Q-matrix results in overestimation of slipping parameters, while unduly adding attributes leads to overestimation of guessing parameters. The study also found high misclassification rates for attribute classes where

attribute combinations were deleted, as well as incorrect dependencies between attributes reducing model accuracy.

Madison and Bradshaw (2014) further emphasize the critical role of Q-matrix design in maximizing classification accuracy in the LCDM. Their study examined different Q-matrix designs and their effects on classification accuracy and reliability and convergence rates. They found that attribute isolation is essential, with each attribute needing to be measured in isolation at least once. Additionally, they warn against conjoined attributes, where two attributes are always measured together, as it causes ambiguity in skill classification, balancing the number of attributes per item is also important, as an excess of attributes per item does not necessarily improve classification accuracy. Together, these findings emphasize the need for precision in Q-matrix development.

Attributes Structure Misspecification

Attribute structures describe how skills relate to each other, they define the relationships between skills or knowledge required to complete specific tasks. There are are two main aspects to attribute structures: the external shape, which shows the visual layout of attribute relationships, and the internal organization, which provides the numerical representation, correctly specifying these structures is important for model fit, item fit, accurate respondent classification, and reliable diagnostic feedback (Leighton et al., 2004).

Misspecifications in attribute structures, such as incorrect relationships between attributes or errors in hierarchical ordering, can impact classification accuracy leading to biased parameter estimates and misclassification of respondents (Templin, Henson, Templin, and Roussos 2008). Research emphasizes the implications of misspecifications of attribute structure which impacts respondent classification, cannot be compensated by other components like item design or correct item-attribute relationships (Kunina-Habenicht et al.,

2012; Rupp & Templin, 2008). Liu and Huggins-Manley (2016) showed that adding more profiles than needed doesn't significantly affect accuracy, but omitting necessary ones does. Liu (2017) categorized attribute misspecifications into two types: sequence misspecification, when the hierarchical order of attributes is incorrectly defined, and design misspecification, which involve changes to the shape or number of connections in the attribute structure. Both types affect the external shape and internal organization of attribute structures, resulting in overfitting, underfitting, or misfitting diagnostic models, all of which impact the accuracy of respondent classifications. Liu's simulations revealed that in three-attribute settings, sequence misspecification caused worse model fit, while in five-attribute settings, design misspecification had greater negative impacts. In terms of internal organization, overfitting models performed better than underfitting and misfitting ones. The results reinforce how important correct structure is to accurately classify student. To reduce error, researchers recommend that attribute structures be supported in theory and empirical validation. Liu and Huggins-Manley (2016) emphasized the need for theory-driven or data-supported structures.

Measurement Model Misspecification

Incorrect measurement model specification poses serious challenges in psychometric modeling, especially for DCMs. When measurement models are wrongly specified, it can lead to flawed parameter estimates and incorrect skill classifications. Kunina-Habenicht, Rupp, and Wilhelm (2012) explored the impact of two key types of measurement model misspecification within LDCM; Q-matrix misspecification and interaction effect misspecification. These types of misspecifications can significantly influence the accuracy of parameter estimates and the resulting classification of respondents. Interaction effects capture relationships between different attributes that influence the probability of a respondent endorsing an item. Misspecification occurs when interaction terms among attributes are

omitted from a diagnosis model. Their research found that Q-matrix misspecification had the biggest negative effect on classification accuracy, while omitting interaction terms especially two-way or three-way interactions mostly reduced parameter precision.

In addition, Madison and Bradshaw (2018) examined the performance of the longitudinal DCMs when measurement invariance is assumed but item parameter drift (IPD) is present. Measurement invariance refers to whether the meaning of the measured construct and its relationship with item responses remain constant over time or across different groups.

Their simulations showed that while the TDCM retained reliable classification accuracy even with high IPD, its ability to recover accurate item parameters dropped as IPD increased. These findings suggest that the model remains reliable for diagnosing skills but not it comes to recovering item parameters. This highlights the importance of accounting for measurement changes when analyzing long term data.

CHAPTER 3

METHODS

This chapter outlines the methods employed to evaluate the robustness of the 1-PLCDM under assumption violations. Two simulation studies were conducted to address the research questions. The first simulation study examines how attribute correlation affects the performance of the 1-PLCDM. The second simulation study explores the impact of assuming a simple Q-matrix structure when the items require a more complex Q-matrix structure.

Simulation Study 1: Effects of Attribute Dependence

The purpose of Simulation Study #1 was to evaluate how attribute dependence affects the classification accuracy and reliability of the 1-PLCDM. Here, I describe the fixed and manipulated factors, then describe the evaluation metrics.

Manipulated factors

This simulation study manipulated one factor: attribute correlation. The attribute correlations were pulled from a random uniform distribution. A compound attribute correlation structure was implemented, ensuring that all pairwise attribute correlations were the same. This made it possible to analyze in detail how the model performs across the entire range of possible attribute correlation values.

Fixed factors

In order to isolate the impact of the correlation of attributes in the study, several parameters were kept fixed across simulation conditions:

Test Length: The number of items used to simulate the test was set at 20 items where four attributes were each assessed by five items. Previous study on DCMs conducted by

(Maas et al., 2024) utilized 5 items per attribute, suggesting that 20 items provide a representative and robust dataset for simulation.

Sample size: The sample size was set at 1000 examinees for each simulation condition. This sample size was chosen to provide sufficient data for reliable estimation of model parameters while reflecting a realistic testing scenario and is within the range of other 1-PLCDM simulation study sample sizes (Maas et al., 2024).

Examinee Attribute Profiles: To simulate item response data, first attribute profiles are generated for each examinee, taking into consideration the base rates of mastery and the attribute correlations. The base rate of mastery was set at .50 for all attributes. A base rate of .50 indicates that out of all examinees, 50% were expected to master each attribute. The attribute profiles were generated using a multivariate binary distribution**,** which simulates mastery states **(**1 = mastered, 0 **=** not mastered**)** while incorporating attribute correlations**.**

Q-matrix design: A simple Q-matrix structure was employed for the study. This design specifies that each item measures a single attribute, simplifying the relationships between test items and attributes. The simplicity of the Q-matrix was intentional to focus on the effects of attribute correlation without introducing additional complexity from the item-attribute mappings.

Item Parameters: To generate item response data, item parameters were assigned fixed values to ensure consistent response probabilities across the simulation. In this study, each item had a fixed intercept of – 1.5 and main effect of 2. These parameter values produced a mean item discrimination of .45 which can be considered average item discrimination (Maas et al, 2024).

Number of Replications: The number of replications was set to 1000 to ensure robust and reliable results in the simulation study. This large number of replications provides

a stable estimate of the model's performance by minimizing the impact of random variability inherent in the data generation process.

Data Generation

Examinees and item responses for each replication were generated with code written within the computational freeware environment R Version 4.4.3 (R Core Team, 2019). I estimated the 1-PLCDM using the TDCM package (Madison et al., 2024). The TDCM package builds on the CDM package (Robitzsch, 2025), which uses marginal maximum likelihood estimation via the expectation maximization algorithm (Dempster, Laird, Rubin 1977). Within the TDCM package, I used summary functions, tdcm.summary and mg.tdcm.summary, for single group and multigroup analyses, respectively to extract item, person, and structural parameters. Finally, results were compiled and summarized using R as well.

Evaluation Metrics

Classification Accuracy. In this study, classification accuracy was defined as the proportion of examinees whose model-estimated mastery status matched their generated mastery status. Classification accuracy was computed for each attribute, or marginally, then averaged across the four attributes.

Classification Reliability. Reliability was evaluated by examining the consistency of classification outcomes across multiple simulation replications. Reliabilities of the classifications were calculated according to the tetrachoric correlation-based metric defined by Templin and Bradshaw (2013). High reliability indicates that the model produces stable and precise classifications.

Simulation Study 2: Effect of Q-matrix Simplification

The purpose of Simulation Study #2 was to explore the impact of reducing to a simple Q-matrix structure when the items require a more complex representation. First, I describe the fixed and manipulated factors, then the evaluation metrics. In this study we define Q-matrix complexity as the number of test items that assess more than one attribute. For example, in the 20% complexity condition below, 20% (4/20) of the items measure two attributes.

Fixed factors

For consistency, most conditions were the same as Simulation Study #1: test length was 20 items (five per attribute); number of attributes was four; sample size was 1000; base-rates were .50 and attribute correlations were fixed at .50. One difference from Simulation Study #1 was item response generation as this study had complex items. For complex items, the responses were modeled to account for interactions between multiple attributes, meaning that a single item could assess more than one attribute simultaneously.

Item Parameters: To estimate the CRUM model, item parameters were pulled from a uniform distribution of intercept (−1.75, -1.25), and the main effects from (1.75, 2.25). We used a distribution to draw the main effects because setting a specific value for the main effects would prevent meaningful comparisons of attribute main effects within the CRUM model.

Manipulated factors

This simulation study manipulated one key factor: Q-matrix design**,** which varied in complexity with four levels (0%, 20%, 40%, 60%, and 80%).

- In the 0% complexity condition, all items were simple, each measuring just one attribute.
- In the 20% complexity condition, four out of the 20 items were complex, with each measuring two attributes. Specifically, items 1, 6, 11, and 16 each measured two attributes.
- In the 40% complexity condition, eight out of the 20 items were complex, with items 1, 2, 6, 7, 11, 12, 16, and 17 each measuring two attributes.
- In the 60% complexity condition, twelve out of the 20 items were complex, with items 1, 2, 3, 6, 7, 8, 11, 12, 13, 16, 17, and 18 each measuring two attributes.
- Finally, in the 80% complexity condition, sixteen out of the 20 items were complex, with items 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, and 19 each measuring two attributes.

Data Generation

In Simulation Study 2, examinees were generated with code written within the computational freeware environment R. I initially estimated the CRUM model to compare main effects, with the aim of simplifying the Q-matrix. In this process, the attribute with the larger main effect was identified as the primary attribute and the second measured attribute was dropped from that item in the updated Q-matrix. After simplifying the Q-matrix, I fitted the 1-PLCDM to the observed responses**.** I estimated 1000 replications per condition, expecting this to be enough replications to provide stable summaries of results.

Evaluation Metrics

Similar to Simulation Study #1, I used classification accuracy and reliability as evaluation metrics.

CHAPTER 4

RESULTS

This chapter presents the results obtained from two simulation studies used to evaluate the robustness of the 1-PLCDM under conditions where its assumptions are violated. Each study assesses classification accuracy and reliability as defined in Chapter 3.

SIMULATION STUDY RESULTS

This chapter has two sections. Section 1 summarizes the findings from Simulation Study 1, which examines how the 1-PLCDM performs when attributes are correlated, violating the independence assumption. Section 2 summarizes the findings from Simulation Study 2, which looks into the impact of simplifying Q-matrix structures by changing the number of items that measure more than one attribute. Both studies look at how assumption violations effect model accuracy.

Section 1: Simulation Study 1

Classification Accuracy

Classification accuracy is the proportion of examinees whose estimated mastery state matched their generated mastery state. Figure 1 shows a scatter plot of the classification accuracy and the corresponding correlation values across 1000 simulation replications. Each point on the plot represents the marginal attribute accuracy of the 1-PLCDM model for a single replication. To better understand the performance of the model in accurately classifying respondents under different levels of attribute correlation, a linear regression was fitted where mean accuracy was the outcome and the attribute correlation was the predictor. Overall, the model's classification accuracy remained stable.

- The intercept ($b_0 = 0.848$) indicates that when the attribute correlation was 0 (i.e. when attributes are independent), the predicted mean classification accuracy is 0.848.
- The slight positive slope ($b_1 = 0.003$) suggests that for every one-unit increase in attribute correlation (i.e., from 0 to 1), the predicted mean accuracy would increase by 0.003.
- The effect size ($R^2 = 0.018$) means that only 1.8% of the variability in the dependent variable is explained by the model.

While the effect was statistically significant ($p < 0.001$), the r-squared value indicates a minimal effect size, the positive effect is not practically significant; when there was no correlation, the model predicts a mean accuracy of 0.848 and with the maximum correlation of 1.0, the model predicts a mean accuracy of 0.851. Even when attributes were highly correlated, the model performed well. In other words, even as the attributes relate with each other, the model's classification accuracy remained consistent. The results show that accuracy did not decline as the correlations increased.

Classification Reliability

In addition to accuracy, we assessed the 1-PLCDM's classification reliability, which measures how consistently the model classifies respondents. In this study, classification reliability was determined by calculating tetrachoric correlations between the model's predicted classifications and the true simulated mastery status (Templin & Bradshaw, 2013). A similar regression was conducted to assess how classification reliability varied across different level of attribute correlation. Here too, the trend was relatively flat, and no significant or meaningful decline in reliability was observed as correlation increased. Reliability stayed within the moderate range, even when attributes were highly correlated.

These results strongly support the robustness of the 1-PLCDM against the violation of the attribute independence assumption.

To better understand the reliability of the model in accurately classifying respondents under different levels of attribute correlation, a linear regression was fitted where mean reliability was the outcome and the attribute correlation was the predictor. Overall, the model's reliability remained stable.

- The intercept ($b_0 = 0.809$) indicates that when the attribute correlation was 0 (i.e. when attributes are independent), the reliability is 0.809.
- The slight positive slope ($b_1 = 0.001$) suggests that for every one-unit increase in attribute correlation (i.e., from 0 to 1), the reliability would increase by 0.001.
- The effect size ($R^2 = 0.0003$) means that only 0.03% of the variability in the dependent variable is explained by the model.

Section 2: Simulation Study 2

Classification Accuracy

The results from Simulation Study 2 reveal how classification accuracy under the 1-PLCDM model varies with increasing levels of Q-matrix complexity. As illustrated in Figure 3, mean classification accuracy declined a little, but still within reasonable range as item complexity increases. At the lowest level of complexity (i.e., 0%), the model demonstrates strong performance, achieving a mean classification accuracy of approximately 0.85. This suggests that when the Q-matrix is simple structure and correctly specified, the model can effectively differentiate between mastery and non-mastery profiles. As the proportion of complex items increases to 20%, accuracy remains relatively stable, indicating that the model maintains robustness in the presence of a small number of multi-attribute items. However, a noticeable decline begins as complexity rises to 40%, with mean accuracy dropping to around

0.83. This trend continues with further increases in complexity. At 60% complexity, the mean accuracy falls to approximately 0.81, and at the highest level of complexity (80%), accuracy declines further to below 0.78. These findings suggest that the 1-PLCDM's ability to classify attribute mastery declines as item complexity increases, but still within an acceptable range.

Thus, as the structure of the Q-matrix became increasingly complex, the model's ability to correctly classify examinees declined somewhat, but still within an acceptable range.

Classification Reliability

Figure 4 shows the average classification reliability across simulation replications for each Q-matrix complexity condition. As shown in the results, reliabilities ranged from .81 to .76, with a mean of .77 and a median of .78**.** These values indicate a generally stable classification performance, even as item complexity increased. Across the five complexity levels (0%, 20%, 40%, 60%, and 80%), reliability began at approximately .81 under the simple Q-matrix structure This high reliability suggests that when items each measure only a single attribute, the 1-PLCDM can classify examinees' mastery statuses with strong consistency. As complexity increased to 20% and 40%, reliability declined only slightly, remaining close to .79 and **.**78, respectively. A more noticeable decrease occurred at 60% complexity, where reliability fell to its lowest point at approximately .76. However, this decline was still relatively modest and did not reflect a breakdown in classification consistency. Interestingly, at the highest complexity level (80%), reliability moved to .79, nearly returning to the levels observed under simpler conditions. Figure 4 shows this trend visually.

These findings suggest that while the model is somewhat sensitive to item complexity, it remains reliable enough for practical use in educational assessments. Assessment designers may therefore consider using the 1-PLCDM, even in complex testing environments.

Summary of Findings:

The 1-PLCDM is highly robust to violations of attribute independence. It maintains strong accuracy and high reliability, even when attributes are correlated. Increasing item complexity shows a slight decline in classification accuracy, while reliability remains consistent.

CHAPTER 5

CONCLUSIONS

This study examined the robustness of a recently developed DCM, the 1-PLCDM, which offers a simple DCM with straightforward interpretation of parameter estimates and preferable measurement properties. It imposes item parameter constraints on the LCDM, estimating a singular main effect for all items assessing the same attribute, along with an intercept for each item. The 1-PLCDM is highly valued for its simplicity and ease of interpretation. To afford these properties, the 1-PLCDM makes strong assumptions about the data and the item response generation process that may not hold in practical educational assessments. The purpose of this study was to examine its performance when these assumptions are violated. Specifically, the assumptions that (1) attributes are independent and (2) each item measures a single attribute, are often violated in real-world testing scenarios.

To conduct this study, I used a simulation study to control attribute relationships and item complexity. The analysis was carried out using the R programming environment, with the TDCM package developed by Madison et al. (2025) utilized for model estimation. The study explored two key variables: attribute correlation and Q-matrix complexity. First, the impact of attribute correlation was investigated by varying the dependencies between attributes using a uniform distribution. The objective was to evaluate how the 1-PLCDM performed when the assumption of attribute independence was violated. The second area of focus was Q-matrix complexity, where test items were designed to assess either one attribute or multiple attributes. The complexity of these items varied from 0% to 80%. A simplified Q-matrix was used to estimate the 1-PLCDM, while the data were generated using a more

complex Q-matrix. This allowed for a direct assessment of how item complexity affected the model's ability to classify respondents accurately.

The results from Simulation Study 1 show that the 1-PLCDM holds up well, even when the attributes being assessed are correlated. This means it can still be trusted to give meaningful diagnostic feedback, even in real-world scenarios where attributes correlate, making it a reliable option for classroom assessments and other educational settings.

The model maintained high classification accuracy and reliability, with regression analysis revealing a statistically significant but small positive relationship between attribute correlation and classification accuracy**.** This suggests that while modest correlations between attributes slightly improved classification accuracy, the effect was minimal. In contrast, as item complexity increased, the model's performance showed a slight reduction in accuracy. However, this decrease was small, and the model still produced consistent, reliable results, with the accuracy remaining within an acceptable range.

These findings highlight that the 1-PLCDM is robust in environments where skills correlate. This makes it a reliable tool for classroom assessments, where attributes are often correlated. Educators can confidently use the 1-PLCDM, knowing that moderate correlations between skills do not undermine its diagnostic capability. The model effectively balances simplicity with accuracy, allowing educators to simplify the assessment process without sacrificing the quality of the results.

For practitioners considering the 1-PLCDM, this study provides the following guidance:

- The model performs well under typical classroom conditions where skill correlations exist.
- Its reliability is stable enough to support data-driven instructional decisions.
- Caution is advised when using the model in assessments with many complex items unless steps are taken to reduce item complexity in the Q-matrix.

In summary, this study validates the 1-PLCDM as a robust and practical tool for diagnostic assessment, particularly in environments where simplicity, clarity, and usability are prioritized. Its overall performance remains stable and interpretable. These findings enhance our understanding of where and how the 1-PLCDM can be applied effectively, thereby contributing to the ongoing refinement of diagnostic measurement modeling for educational use.

Limitations and Future Directions

The simulation study had some limitations. While the study introduced variation in the Q-matrix structure to test different levels of complexity, the most complex condition only included items measuring two attributes. This does not capture scenarios where items assess three or more attributes at once, which is common in real-world diagnostic assessments. In practice, Q-matrices often reflect more diverse and nuanced patterns of relationships between attributes. Additionally, the study fixed factors such as test length**,** sample size**,** and item quality, which may have constrained its ability to fully capture the variability and complexity typically seen in real-world assessment data.

Future research could benefit from testing the model with a broader and more realistic Q-matrix structure, including more complex items that assess multiple attributes. Also applying the model to empirical data will allow for an evaluation of its effectiveness in real-world contexts. Examining model fit across a wider range of conditions can further assess its accuracy and applicability in different settings. Extending the 1-PLCDM to more complex applications, such as longitudinal application or dual-purpose applications (e.g., scaling and classification), would allow for exploration of its potential in more diverse and nuanced settings. By broadening the scope of the study and applying the model to real-world data,

future research can provide deeper insights into the practical utility and limitations of the 1-PLCDM and its ability to handle more complex real-world situations.

REFERENCES

Bradshaw, L., Izsak, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational measurement: Issues and practice*, *33*(1), 2-14.

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

de La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353.

de La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Embretson, S. (1994). Applications of cognitive design systems to test development. In *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). Boston, MA: Springer US.

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, *10*(4), 318-341.

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

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59-81.

Lao, H., & Templin, J. (2016). *Estimation of diagnostic classification models without constraints: Issues with class label switching*. Paper presented at the annual meeting of the National Council on measurement in education in Washington, DC.

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

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

Liu, R., & Huggins-Manley, A. C. (2016, August). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society, Beijing, 2015* (pp. 243-254). Cham: Springer International Publishing.

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and psychological measurement*, *77*(2), 220-240.

Maas, L., Brinkhuis, M. J., Kester, L., & Wijngaards-de Meij, L. (2022). Cognitive diagnostic assessment in university statistics education: Valid and reliable skill measurement for actionable feedback using learning dashboards. *Applied Sciences*, *12*(10), 4809.

Maas, L., Brinkhuis, M. J., Kester, L., & Wijngaards-de Meij, L. (2022, February). Diagnostic classification models for actionable feedback in education: Effects of sample size and assessment length. In *Frontiers in Education* (Vol. 7, p. 802828). Frontiers Media SA.

Maas, L., Madison, M. J., & Brinkhuis, M. J. (2024, January). Properties and performance of the one-parameter log-linear cognitive diagnosis model. In *Frontiers in Education* (Vol. 9, p. 1287279). Frontiers Media SA.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and psychological measurement*, *75*(3), 491-511.

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*(4), 963-990.

Madison, M. J., Wind, S. A., Maas, L., Yamaguchi, K., & Haab, S. (2024). A One-Parameter Diagnostic Classification Model with Familiar Measurement Properties. *Journal of Educational Measurement*, *61*(3), 408-431.

Madison, M. J., Jeon, M., Cotterell, M., Haab, S., & Zor, S. (2025). TDCM: An R package for estimating longitudinal diagnostic classification models. *Multivariate Behavioral Research*, *60*(3), 518-527.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, *1*(1), 3-62.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25-36.

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

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317-339.

Schunk, D. H., & Zimmerman, B. J. (2012). Self-regulation and learning. *Handbook of Psychology, Second Edition*, *7*.

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

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. Item response theory: knowing when to cross the line. *Practical Assessment, Research & Evaluation*, *26*, 11.

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

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*, 317-339.

Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287-307.

Zwitser, R. J., & Maris, G. (2016). Ordering individuals with sum scores: The introduction of the nonparametric rasch model. *psychometrika*, *81*, 39-59.

Table 1

Regression analysis for attribute correlation

Coefficient	Estimate	Std. Error	t-value	p-value
Intercept	0.848	<0.001	2049.8	<.001
Attribute correlation	0.003	0.001	4.3	<.001

Table 2

Regression analysis for Reliability

Coefficient	Estimate	Std. Error	t-value	p-value
Intercept	0.809	0.001	1133.2	<.001
Attribute correlation	0.001	0.001	0.564	0.573

Figure 1

Scatter plot showing classification accuracy for varying levels of attribute correlation

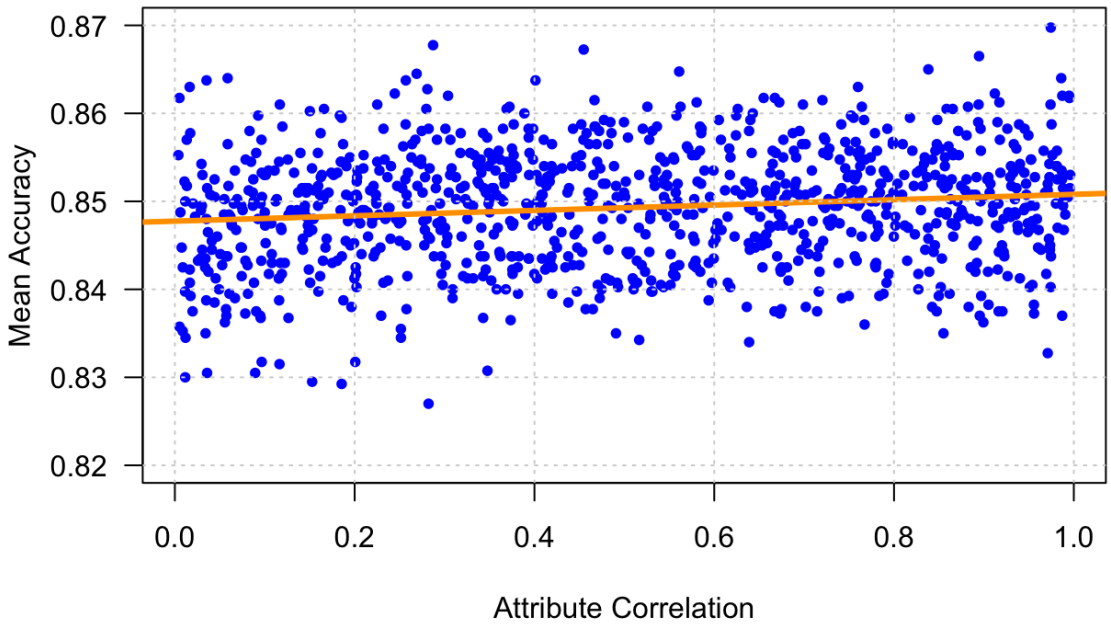

Figure 2

Scatter plot showing classification reliability for attribute correlation

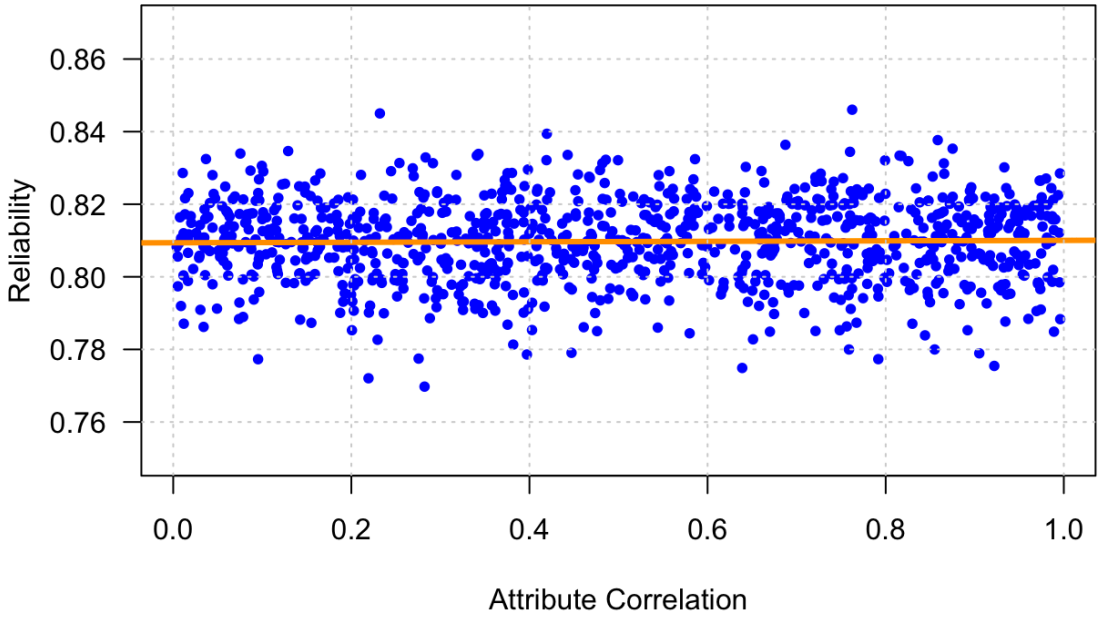

Figure 3

Scatter plot showing classification accuracy for Q-matrix Complexity

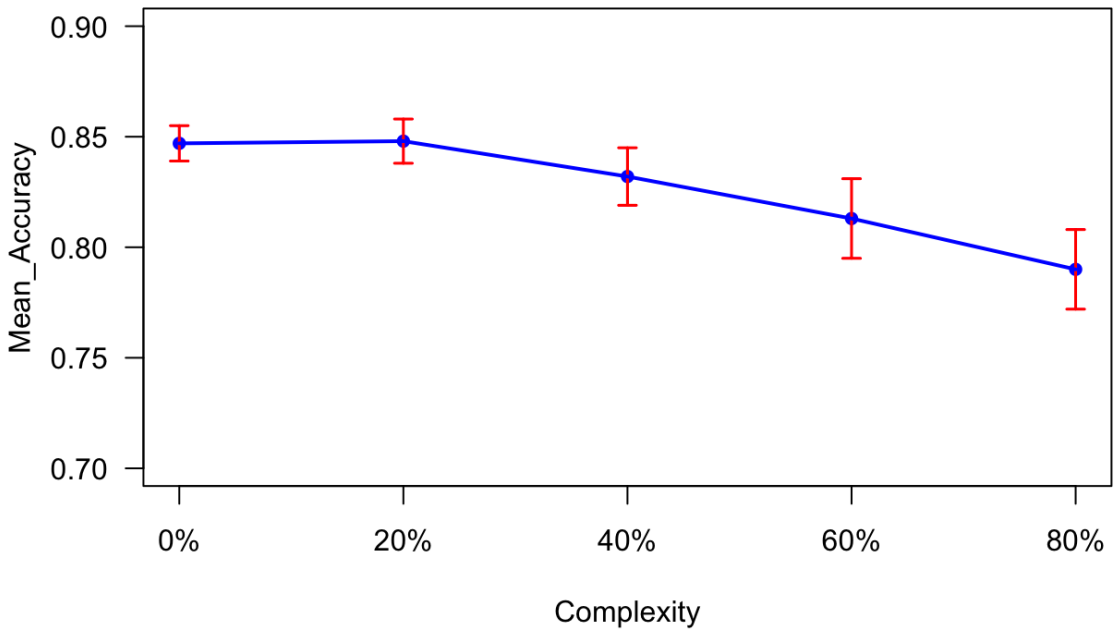

Figure 4

Scatter plot showing classification reliability for Q-matrix Complexity