Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Impacts of Q-matrix Design on Learning DCM Linear Attribute Hierarchies

Olasunkanmi Kehinde [*,†] Alfonso J. Martinez [‡] and Mubarak O. Mojoyinola [¶]

†Department of Psychology, Norfolk State University, Norfolk, Virginia, USA
‡Department of Psychology, Fordham University, New York, New York, USA
¶Department of Psychological and Quantitative Foundations, University of Iowa, Iowa City, Iowa, USA
*Corresponding author. Email: ojkehinde@ecsu.edu

Abstract

This study examines the classification accuracy of the G-DINA model under three linear hierarchical Q-matrix specifications: Independent, Adjacent, and Reachable. Using a fully crossed simulation design, we evaluated the effects of item–attribute alignment and structural misspecification across varying sample sizes, attribute base rates, and item quality conditions. Results indicate that while linear hierarchies are generally well recovered, Q-matrix specification critically influences model reliability. Independent designs demonstrated robust performance, whereas Reachable configurations were prone to increased classification error. These findings underscore the importance of strategic Q-matrix construction for producing reliable diagnostic feedback in educational assessment.

Keywords: Diagnostic Classification Models, Attribute Hierarchies, Q-matrix Design

1. Introduction

Diagnostic Classification Models (DCMs) have emerged as a pivotal psychometric framework designed to provide fine-grained, skill-level insights into an individual's cognitive strengths and weaknesses (Rupp et al., 2010). Unlike traditional unidimensional models, such as Item Response Theory (IRT), which summarize examinee performance into a single aggregate score, DCMs classify individuals into mastery or non-mastery statuses across multiple discrete latent traits, termed attributes (Madison & Bradshaw, 2015). This multidimensional approach is essential for assessment for learning, as it maps item-level responses to specific cognitive profiles, facilitating the delivery of actionable feedback and tailored instructional interventions (Tang & Zhan, 2021). The theoretical value of DCMs is particularly evident in their capacity to model the sequential and structured nature of educational development. In most academic domains, attributes do not function in isolation; instead, they often form an attribute hierarchy, where the acquisition of a complex skill is predicated on the mastery of more foundational prerequisites (Clements & Sarama, 2012, 2020). For instance, geometric reasoning has been shown to progress through a rigid sequence of stages, such as those described in the van Hiele theory (Chen et al., 2023). A core assumption in hierarchical DCMs (HDCMs) is that a student cannot logically possess a higher-level attribute without first mastering its antecedent. This dependency restricts the potential latent class space to ensure mastery profiles align with established cognitive theories (Templin & Bradshaw, 2014).

The validity of these diagnostic inferences is heavily contingent upon the Q-matrix specification, which serves as the formal mapping between test items and the attributes they intend to measure (de la Torre, 2008). The Q-matrix acts as the bridge between cognitive theory and statistical estimation;

consequently, its design is as critical to model performance as the choice of the statistical estimator itself (Ravand, 2019). If the Q-matrix is poorly aligned with the underlying cognitive structure, the resulting feedback may be contaminated, leading to erroneous conclusions about a student's learning progress and misinformed instructional decisions.

In diagnostic measurement, attribute structures represent the formalization of dependencies within learning trajectories. According to Liu (2018), attribute hierarchies can be defined by their external shape, describing the macro-level relationship between skills. Common configurations include the linear hierarchy, where attributes follow a single sequential chain, and more complex forms such as divergent, convergent, and unstructured hierarchies (Liu et al., 2017). Failing to account for these dependencies can lead to illogical mastery profiles; for example, a student might be classified as a master of advanced calculus while failing basic arithmetic (Tu et al., 2019). Implementing a restricted Q-matrix design forces the model to consider only theoretically plausible patterns, enhancing classification accuracy (Tu et al., 2019).

The effectiveness of a diagnostic test is often determined by how the Q-matrix handles item-attribute alignment. Madison and Bradshaw (2015) found that classification reliability improves significantly when the Q-matrix includes isolated information: items measuring a single attribute in a simple structure format. To integrate hierarchical dependencies into the test design, three primary approaches have been identified: the Independent approach, where items measure one attribute; the Adjacent approach, measuring at most two connected attributes; and the Reachable approach, where items measure a target attribute along with all its direct and indirect prerequisites (Liu et al., 2017). Table 1 illustrates the test formulation for the linear hierarchical Q-matrix designs.

While recent simulation studies have extensively examined the G-DINA model under various conditions (Shi et al., 2021; Tu et al., 2019), there remains a significant gap in the literature regarding its performance under linear hierarchical structures when the hierarchy is misspecified (Liu, 2018). Furthermore, current research has only recently begun to address the complexities of multiple-attempt data in dynamic assessment environments (Kwon et al., 2024). Misspecification of the attribute structure can lead to significant respondent misclassification. Given that human learning is seldom strictly linear, failing to account for these structural complexities may yield diagnostic feedback that is inaccurate and unhelpful for student learning.

This paper addresses this gap by investigating the classification performance of the G-DINA model across varied linear hierarchical Q-matrix designs, including the Independent, Adjacent, and Reachable configurations. By examining the interplay between item-attribute alignment and structural misspecification, this study aims to determine the conditions under which fine-grained diagnostic feedback remains robust and reliable for educational practitioners.

In particular, the goal of this study is to extend the literature by examining the performance of the G-DINA model under different Q-matrix design approaches that reflect linear hierarchies of attribute mastery under varying design factors including sample size, base-rates, Q-matrix.

2. Methods

This study employed a fully-crossed simulation design to investigate the impact of Q-matrix design on classification performance. A total of five factors were specified. In particular, the number of attributes was fixed to three across all conditions. Three levels for sample size were investigated (200, 500, 1000) to reflect small, medium, and large sample sizes. In addition, three Q-matrix structures were considered, including linear independent, linear adjacent, and linear reachable. As stated earlier, these Q-matrices differ in how they encode information about that attribute hierarchy, with the linear independent Q-matrix structure representing the simplest design and the linear reachable representing the most complex design. Attribute mastery base rates were also manipulated to reflect different hierarchy strengths. In particular, a total of six attribute mastery base rate values were specified, three which contained equal base rates across attributes and three where the base rates

Table 1. Q-Matrix Design Specifications for Linear Attribute Hierarchies

	Linear Independent (LI)			**Linear Adjacent (LA)**			**Linear Reachable (LR)**		
Item	A1	A2	A3	A1	A2	A3	A1	A2	A3
Item 1	1	0	0	1	1	0	1	0	0
Item 2	0	1	0	0	1	1	1	1	0
Item 3	0	0	1	0	0	1	1	1	1
Item 4	1	0	0	1	1	0	1	0	0
Item 5	0	1	0	0	1	1	1	1	0
Item 6	0	0	1	0	0	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

decreased sequentially. With respect to the former, the base rates were (0.2, 0.2, 0.2), (0.5, 0.5, 0.5), and (0.7, 0.7, 0.7), respectively. The (0.2, 0.2, 0.2) condition, for instance, represents a linear hierarchy where the three attributes have a relatively low probability of mastery whereas the (0.7, 0.7, 0.7) represents a linear hierarchy where the three attributes have a relatively high probability of mastery. The decreasing base rate conditions were specified to reflect scenarios where attributes further down in the hierarchy are more difficult to master relative to prerequisite attributes. The base rates selected for this condition were (0.3, 0.2, 0.1), (0.7, 0.4, 0.1), and (0.9, 0.8, 0.7) for attributes one through three, respectively. The final manipulated condition was item quality, of which two levels (low/high) where values were chosen following Yamaguchi and Templin (2022). In total, there were 108 conditions in this study.

Data were generated in R using custom data-generation functions and all models were estimated via the GDINA package in R (Ma & de la Torre, 2020). The item parameters were estimated with marginal maximum likelihood (ML) estimation with the expectation-maximization (EM) algorithm and convergence threshold of 0.0001 by default. Each generated dataset was then estimated using the three aforementioned Q-matrices. Each simulation condition was replicated 50 times resulting in $108 \times 50 = 5,400$ unique datasets.

Given that this study is primarily concerned with how different Q-matrix designs impact classification at the attribute level, we limit our results to attribute-level classification rates. For each analysis, the maximum a posteriori (MAP) estimate for each attribute was taken to compute attribute-wise accuracy relative to the true attribute mastery status. In this study, attribute-wise accuracy rates were calculated as follows:

$$\text{AAR}_a = \sum_{i=1}^{n} \frac{I(\hat{\alpha}_{ia,\text{MAP}} = \alpha_{ia})}{n}.$$

where $I(\cdot)$ represents the indicator function that takes on a value of 1 if its argument is true and 0 otherwise, n is the sample size, and $\hat{\alpha}_{ia,\text{MAP}}$ is the MAP estimate of the a-th attribute for the i-th individual for a given simulation replication. Results were then aggregated across replications.

3. Simulation Results

Figure 1, 2, and 3 show the attribute-level classification accuracy as a function of sample size, base-rate condition, item quality, and Q-matrix specification across the three linear hierarchy Q-matrix encoding.

3.1 Linear-Independent Q-matrix

Attribute classification accuracy was uniformly high for all Q-matrix specifications. The minimum observed accuracy was 0.916, occurring at a sample size of 1000 under the equal base-rate condition

(0.2) with low item quality and the linear adjacent Q-matrix specification. The maximum accuracy of 0.999 was observed at a sample size of 1000 under the equal base-rate condition (0.2) with high item quality and the linear independent Q-matrix specification.

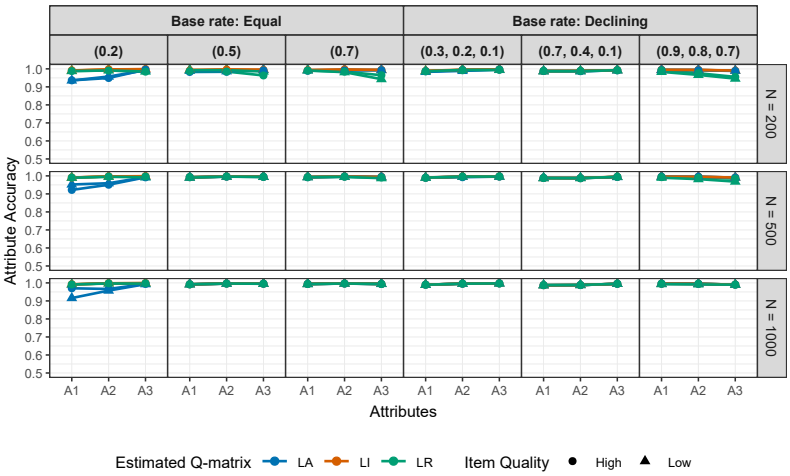

Figure 1. Attribute Classification Accuracy for Linear-Independent Q-matrix

3.2 *Linear-Adjacent Q-matrix*

Attribute classification accuracy exhibited greater variability under the Linear–Adjacent Q-matrix generating condition. Under equal base-rate conditions, classification accuracy was uniformly high across attributes and Q-matrix specifications, with most estimates exceeding 0.85 even at the smallest sample size. The Linear–Independent specification performed comparably to the true Linear–Adjacent Q-matrix. In contrast, the Linear–Reachable specification yielded the lowest attribute classification accuracy, particularly for the third attribute.

Under declining base-rate conditions, attribute classification accuracy was generally lower for both the Linear–Independent and Linear–Reachable Q-matrix specifications. These reductions were consistent across sample size levels. Sample size effects were most evident under the equal base-rate condition (0.20), where classification accuracy increased as sample size exceeded 200.

3.3 *Linear-Reachable Q-matrix*

Under the Linear–Reachable generating condition, the Linear–Independent Q-matrix specification generally performed more closely to the true Linear–Reachable Q-matrix than the Linear–Adjacent specification. Under equal base-rate conditions, classification accuracy was uniformly high across attributes and Q-matrix specifications, with most estimates exceeding 0.85 even at the smallest sample size. An exception was observed for the first attribute under the equal base-rate condition (0.20) when the Linear–Adjacent Q-matrix was used for estimation.

Declining base-rate conditions were associated with reduced attribute classification accuracy, particularly when the rate of decline exceeded 0.10, as in the declining base-rate condition (0.7, 0.4, 0.1). In this setting, item quality and sample size exhibited minimal influence on classification accuracy.

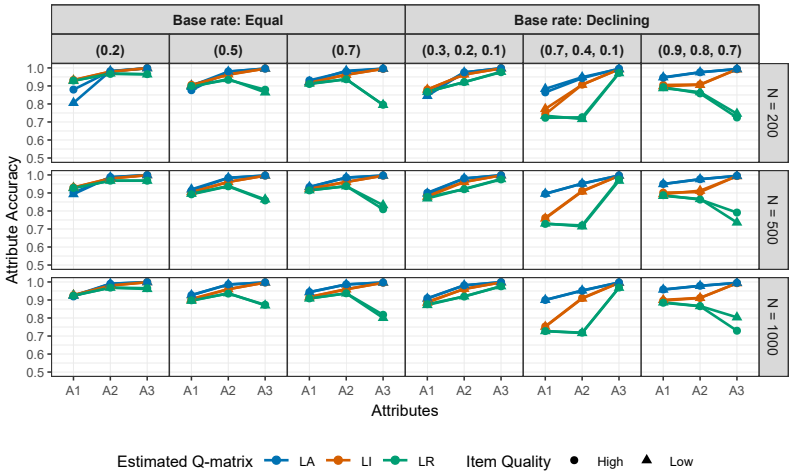

Figure 2. Attribute Classification Accuracy for Linear-Adjacent Q-matrix

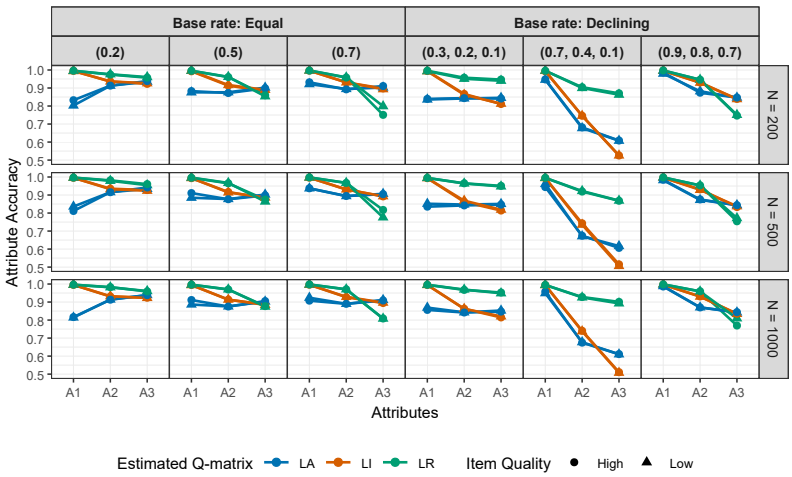

Figure 3. Attribute Classification Accuracy for Linear-Reachable Q-matrix

4. Conclusion

The present study investigated the impact of various Q-matrix designs, specifically Linear-Independent (LI), Linear-Adjacent (LA), and Linear-Reachable (LR), on attribute classification accuracy within the G-DINA framework. The findings reinforce several key principles established in diagnostic classification modeling while highlighting specific challenges associated with structural complexity. Primarily, the results confirm that while linear hierarchies can be effectively modeled, the choice of Q-matrix design is a critical determinant of model integrity.

One key finding concerns the robustness of isolated attribute information within the Q-matrix. Consistent with prior research (Kehinde et al., 2025; Madison & Bradshaw, 2015), classification accuracy remained consistently high when items measured a single attribute. Under equal base-rate conditions, the Linear-Independent specification performed comparably to the true generating models, suggesting that simple structure designs are resilient even when hierarchical dependencies

are present. This alignment with the work of Templin and Bradshaw (2014) and Clements and Sarama (2012, 2020) supports the theoretical value of using hierarchies to restrict the latent class space to plausible mastery profiles, ensuring that diagnostic results match established cognitive theories.

However, the results also revealed notable challenges regarding structural complexity, particularly with the Linear-Reachable design. Contrary to the assumption that more complex designs might enhance diagnostic power, the Reachable specification, which measures a target skill along with all its prerequisites, often yielded the lowest attribute classification accuracy. This was particularly evident for attributes further down the hierarchy, suggesting that overcomplicating item-attribute mappings may introduce noise that outweighs the benefits of modeling prerequisites. These findings extend the work of Liu (2018), who noted that misspecifying attribute structures leads to significant misclassification, by showing that LA and LR specifications are particularly sensitive to misspecification under declining base-rate conditions.

Furthermore, the performance of the Linear-Independent Q-matrix relative to more complex designs provides a nuanced perspective on model selection. In many conditions, the Independent specification performed more closely to the true Linear-Reachable Q-matrix than the Adjacent specification did. This contradicts the intuitive assumption that a "closer" structural match, such as the Adjacent design, is always superior to a simpler design when the true underlying structure is complex. Such findings emphasize that practitioners should be cautious of "contaminated" classification results if they ignore the subtleties of attribute hierarchies and item quality.

In conclusion, these findings suggest that while linear hierarchies are essential for representing learning trajectories, the strategic construction of the Q-matrix is paramount. Practitioners should prioritize validated simple structure designs or carefully calibrated Adjacent designs over complex Reachable configurations to ensure that diagnostic feedback remains reliable for educational interventions. This is especially critical as the field moves toward dynamic assessment environments where multiple-attempt data may further complicate respondent classification.

References

Chen, Y.-H., Hsu, C.-L., Wu, Y.-J., Yi, Z., Wang, Y., & Thompson, D. R. (2023). Exploring attribute hierarchies of the van Hiele theory using diagnostic classification modeling and structural equation modeling. https://ucsmp.uchicago.edu/resources/van-hiele/

Clements, D. H., & Sarama, J. (2012). *Learning and teaching early math: The learning trajectories approach* (2nd). Routledge.

Clements, D. H., & Sarama, J. (2020). *Learning and teaching early math: The learning trajectories approach* (3rd). Routledge.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Applied Psychological Measurement*, *32*(4), 344–362. https://doi.org/10.1177/0146621607304028

Kehinde, O. J., Dai, S., & French, B. (2025). The impact of q-matrix on item parameter and attribute profile estimations in transitional diagnostic classification models [Paper presented at the Annual Meeting of the National Council on Measurement in Education]. *Paper Presentation of the Annual Meeting of the National Council on Measurement in Education*.

Kwon, T. Y., Huggins-Manley, A. C., Templin, J., & Zheng, M. (2024). Modeling hierarchical attribute structures in diagnostic classification models with multiple attempts. *Journal of Educational Measurement*, *61*(2), 198–218. https://doi.org/10.1111/jedm.12387

Liu, R. (2018). Misspecification of attribute structure in diagnostic measurement. *Educational and Psychological Measurement*, *78*(4), 605–634. https://doi.org/10.1177/0013164417702458

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. https://doi.org/10.1177/0013164416645636

Ma, W., & de la Torre, J. (2020). Gdina: An r package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*, 1–26.

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. https://doi.org/10.1177/0013164414539162

Ravand, H. (2019). Q-matrix validation: A comparison of three methods. *Journal of Psychoeducational Assessment*, *37*(2), 192–204. https://doi.org/10.1177/0734282917734891

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

Shi, Q., Tu, D., & Wang, S. (2021). The impact of Q-matrix misspecification on the G-DINA model. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211025562

Tang, F., & Zhan, P. (2021). Does diagnostic feedback promote learning? Evidence from a longitudinal cognitive diagnostic assessment. *AERA Open*, 7(1), 1–15. https://doi.org/10.1177/23328584211060804

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339. https://doi.org/10.1007/s11336-013-9362-0

Tu, D., Wang, S., Cai, Y., Douglas, J., & Chang, H.-H. (2019). Cognitive diagnostic models with attribute hierarchies: Model estimation with a restricted Q-matrix design. *Applied Psychological Measurement*, *43*(4), 255–271. https://doi.org/10.1177/0146621618765721

Yamaguchi, K., & Templin, J. (2022). A gibbs sampling algorithm with monotonicity constraints for diagnostic classification models. *Journal of Classification*, *39*(1), 24–54.