# Methods to Retrofit and Validate Q-Matrices for Cognitive Diagnostic Modeling


**Charles Vincent Hunter, Hongli Li, and Ren Liu**

**Abstract** Cognitive diagnostic models (CDMs) are a family of constrained latent class models that estimate relationships between observed item responses and latent attributes (Rupp and Templin, Educ Psychol Meas 68:78–96, 2008). An important input needed in any CDM is the Q-matrix, an item-by-attribute table that represents a particular hypothesis about which attributes are required to answer each test item successfully. A large number of CDMs have been developed; however, many applications involve retrofitting a CDM to an existing non-diagnostic test. In this study, we conducted a systematic review to describe the current picture of retrofitting Q-matrices to non-diagnostic tests and consequently using the tests for diagnostic purposes.

**Keywords** CDM · Q-matrix · Retrofit

## 1 Introduction

Cognitive diagnostic models (CDMs) are a family of constrained latent class models that estimate relationships between observed item responses and latent traits (Rupp & Templin, 2008). These models assume that the items measure multiple latent traits and the latent traits are categorical (Liu & Shi, 2020). CDMs have been advocated as having the potential to provide rich diagnostic information from tests to aid

C. V. Hunter (✉)
Research, Evaluation, Assessment, and Accountability, Clayton County Public Schools, Jonesboro, GA, USA
e-mail: charles.hunter@clayton.k12.ga.us

H. Li
Georgia State University, Atlanta, GA, USA
e-mail: hli24@gsu.edu

R. Liu
University of California Merced, Merced, CA, USA
e-mail: rliu45@ucmerced.edu



M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 393, https://doi.org/10.1007/978-3-031-04572-1_16

**Table 1** Sample Q-matrix

| | Attribute A | Attribute B | Attribute C |
|---|---|---|---|
| Item 1 | 1 | 0 | 0 |
| Item 2 | 0 | 1 | 1 |
| ... | ... | ... | ... |

instruction and learning, because they provide a profile for each student in regard to whether or not the student has mastered the required skills (a.k.a. attributes) to provide correct responses to the test items. CDMs, therefore, are able to provide useful diagnostic feedback to teachers and students.

As summarized by DiBello et al. (2007), a systematic cognitive diagnostic assessment involves six steps: (i) describing assessment purpose; (ii) describing skill space; (iii) developing assessment tasks; (iv) specifying psychometric model; (v) performing model calibration and evaluation; and (vi) score reporting. However, very few large-scale tests are designed under a cognitive diagnostic modeling framework. Therefore, in most CDM applications, a non-diagnostic preexisting test is analyzed, which is referred to as retrofitting (Liu et al., 2017). A major challenge involved in retrofitting is that constructing the post-hoc Q-matrix is time consuming. In addition, calibrating an existing unidimensional test with a multidimensional CDM may not work or may be inefficient (Haberman & von Davier, 2007).

The Q-matrix that represents a particular hypothesis about which attributes are required to answer each test item successfully (Tatsuoka, 1983). As shown in Table 1, each row represents an item of the test and each column represents an attribute. A Q-matrix can have a simple structure—each item requires only one attribute—or a complex structure—at least one item requires more than one attribute (Rupp et al., 2010). A sound Q-matrix is critical for a successful CDM application (Gorin, 2009).

## 2 Method

Studies that met the following criteria were included in our review. First, the study had to apply CDM(s) to a real dataset. If a study adopted a Q-matrix developed and validated in a previous study, we would only keep the earlier study to avoid duplicates. For example, Jang et al. (2015) used the same Q-matrix that was developed and validated in Jang et al. (2013). We only included Jang et al. (2013). Second, only journal articles in English from 1983 to March 2021 were included.

To begin with, we performed a systematic search of ERIC and APA PsychInfo using keywords "cognitive diagnostic" or "diagnostic classification." The authors read the full-text of each article to decide whether it was eligible. Then, we searched Google scholar using the same keywords "cognitive diagnostic" or "diagnostic classification." The authors went through each entry to decide if any new study would be added to our previous findings. Finally, we consulted the studies included in Sessoms and Henson (2018), who reviewed the application of CDMs from 2008 to 2016.

After selecting the initial set of articles, the authors held several rounds of discussions among themselves to refine the selection criteria. During these discussions, we decided to restrict the study to journal articles published in English, in order to complete the study in a timely manner. We also realized that several articles used a Q-matrix developed in an earlier study, which would have created duplicates. Because of this, we decided to include only the earliest study. In the end, we found 80 articles that met our inclusion criteria. Seven of the articles reported two different CDM studies. For each of these articles, two studies were included for the final coding. Therefore, 87 studies from 80 published journal articles were included in this review.

We first drafted a coding sheet based on existing literature and prior CDM study review (e.g., Sessoms & Henson, 2018), with an emphasis on the Q-matrix development and validation. After several rounds of discussion and training of the coding procedure, one author coded all the studies. Then another author went through the coding of each article multiple times. All discrepancies and questions were resolved through discussion and negotiation until a consensus was reached.

## 3 Results

The earliest study included in our review was published in 1993 (Birenbaum & Tatsuoka, 1993). Fewer than 10 articles per year were published until 2019, when 11 were published. Seventeen were published in 2020. The articles were published in a variety of journals with journals focused on education and psychology being the most frequent categories, such as *Studies in Educational Evaluation* and *Frontiers in Psychology*. The most studies conducted in a single country were 34 in the US, followed by 11 in Iran. Thirteen studies used data collected from multiple countries, primarily because the tests used were the Programme for International Student Assessment (PISA) or the Trends in International Mathematics and Science Study (TIMSS). Many sample sizes were extremely large because of the multinational samples, and the Korean National Assessment of Educational Achievement (NAEA) exam, which had a sample of over 16 million examinees (Table 2). The number of items per assessment ranged from 4 to 216; and the number of attributes ranged from 2 to 27.

**Table 2** Study descriptive statistics

| | Sample size[a] | Sample size[b] | Number of items | Number of attributes |
|---|---|---|---|---|
| Max | 16,928,895 | 120,767 | 216 | 27 |
| Min | 96 | 96 | 4 | 2 |
| Mean | 415,952.8 | 5449.0 | 39.3 | 8.3 |
| Std dev | 2,612,318.7 | 15,749.3 | 34.9 | 5.1 |
| Median | 1454 | 1252 | 33 | 7 |
| Mode | 10,000 | 10,000 | 20 | 5 |

[a]With Kim (2014)
[b]Without Kim (2014)

**Table 3** Content area studied

| Area studied | Frequency |
|---|---|
| Language skills[a] | 30 |
| Mathematics | 30 |
| Psychology (normal and pathological) | 10 |
| Listening | 9 |
| Science and medicine | 2 |
| Civic knowledge | 1 |
| Intrapreneurship | 1 |
| MOOC engagement | 1 |
| Professional competencies | 1 |
| Situational judgment | 1 |
| Social justice advocacy | 1 |

[a]Reading, writing, grammar, foreign language skills, and foreign language arts

In terms of the construct being tested, language skills and mathematics, which appeared in 30 studies each, were the most frequent. Ten studies looked at different areas of psychology, such as personality, and pathological behavior (Table 3).

In all the studies, the attribute classification was dichotomous (i.e., master vs. non-master). Six of the studies reported correlations among the attributes. Although many of the studies discussed how to prepare score reports, none of the studies reported whether such diagnostic results were actually delivered to students or teachers.

The Q-matrices in 74 studies had a complex structure, while eight had a simple structure. In five of the studies, the relationships among the attributes were hierarchical. In 24 of the studies, the authors modified their initial Q-matrices, and in 52 of the studies, the authors provided the final Q-matrix. A variety of CDMs were used in the applications, and it was common for one study to apply multiple CDMs. The deterministic input, noisy "and" gate (DINA) and the generalized deterministic input, noisy "and" gate (GDINA) models were the most frequently used and appeared in 23 studies each, followed by the Reduced Reparameterized Unified Model (RRUM or FUSION) model (20 times; Table 4).

The Q-matrices were developed using combinations of different techniques (Table 5). The most common was using a literature review to determine the attributes (or skills) needed to respond to the items correctly. Review of the assessment items by content experts was the second most common method. Consulting the test specifications was used in eight studies, while asking examinees about how they answered questions was used in seven studies. Thirty-six studies did not report how the Q-matrix was developed.

Checking model fit indices was the most common means of validating the Q-matrix. Indices of both absolute fit (e.g., SRMSR, MAD) and relative fit (e.g., AIC, BIC) were used (Chen et al., 2013). Relative fit indices were used to compare different Q-matrices or different CDMs, and to compare CDM results to results from Classical Test Theory (CTT) or Item Response Theory (IRT) models. The Wald test

**Table 4** CDM used

| CDM | Number of articles |
|---|---|
| ACDM | 6 |
| Attribute hierarchy method (AHM) | 3 |
| DINA[a] | 23 |
| DINO[b] | 13 |
| GDINA[c] | 23 |
| GDM | 4 |
| Hierarchical diagnostic classification model (HDCM) | 1 |
| Log-linear cognitive diagnosis model (LCDM) | 4 |
| LLM | 3 |
| Mixture model | 1 |
| RRUM/FUSION[d] | 20 |
| Rule space method | 10 |

[a]Includes four variant types of DINA
[b]Includes three Bayesian variant types of DINO
[c]Includes one variant type of GDINA
[d]Includes three variant types of RRUM

**Table 5** Q-matrix development and validation

| Development | | Validation | |
|---|---|---|---|
| Task | Frequency | Task | Frequency |
| Literature review | 39 | Model fit indices | 25 |
| Not reported | 36 | Attribute mastery predictions | 13 |
| Content expert review | 29 | Compare with CTT/IRT | 12 |
| Author coding | 15 | Not reported | 12 |
| Test framework – specifications | 8 | Item parameters | 11 |
| Student reports | 7 | Reliability indices | 10 |
| | | Empirical validation algorithms | 8 |
| | | Compare different Q-matrices | 6 |
| | | Cross validation with other criteria | 5 |
| | | Student interviews | 4 |
| | | Factor analysis | 3 |
| | | Regression | 3 |
| | | Discuss with experts | 2 |
| | | Review of misfitting items | 1 |

was also used to compare different models. Checking attribute mastery predictions (both for accuracy and for consistency) was the second most common validation technique. This was used in 13 studies. Other methods include evaluating item parameters and reliability indices, as well as using empirical validation algorithms (e.g., de la Torre and Chiu's (2016) validation method, implemented in R package GDINA (Ma & de la Torre, 2020)). Twelve studies did not report their validation methods.

## 4 Discussion and Significance

The Q-matrix is of vital importance for the proper functioning of a CDM. If the Q-matrix is misspecified, the usefulness of the CDM is impaired (Gorin, 2009). It is, therefore, important to have a Q-matrix that is well-founded theoretically, as well as supported by empirical evidence (Rupp et al., 2010).

Developing a Q-matrix is an iterative process that involves theoretical guidance and content knowledge about the construct being tested. After an initial set of attributes has been developed, the Q-matrix needs to be refined to ensure that there are sufficient high-quality items for each attribute to produce stable results. Also, redundant or closely overlapped attributes need to be identified, combined or removed for parsimony. This is frequently done by removing an attribute that has high correlation with other attributes, or that has low correlation with item difficulty (Buck & Tatsuoka, 1998).

The most common methods of Q-matrix development are literature reviews and ratings by content experts. In addition, consulting test specifications and consulting with examinees via think-aloud or posttest interviews are good ways to understand cognitive processes when examinees respond to the test items. However, test specifications are often not available for a retrofitted test, and consulting examinees, which can provide valuable insights, sometimes may be not be feasible given the development context. It is common to find a CDM study that uses multiple methods discussed above in their Q-matrix development stage (e.g., Li & Suen, 2013). Utilizing evidence from multiple sources greatly strengthens their Q-matrix.

Similarly, high quality CDM application studies tend to adopt multiple procedures to validate their Q-matrices from different perspectives. Our review shows that there are three main types of Q-matrix validation procedures: (a) comparing the CDM model results with results from CTT or IRT models; (b) examination of model fit indices; and (c) examination of attribute mastery predictions. First, it is not always valid to compare CDM results with results from IRT models. CDMs are multidimensional while IRT models are usually unidimensional. Even when multi-dimensional IRT is used, the assumptions are different as the latent variables in CDMs are categorical while the latent variables in IRT models are continuous. Therefore, such comparison does not always lead to meaningful results. Second, model fit indices have played an important role in Q-matrix validation. Both absolute and relative fit indices are utilized. With the lack of well-established criteria for the absolute fit indices for CDMs (Lei & Li, 2016), the relative fit indices (e.g., AIC, BIC) seem to be more useful when results from different models are compared. Some studies (e.g., Ravand et al., 2020) allowed the items to "choose" the best fitting model, but Hemati and Baghaei (2020) found that overall model fit for this procedure was not as good as using the GDINA model for all items.

Attribute mastery predictions usually consist of evaluating both classification accuracy and consistency of the whole latent class pattern for examinee responses. As Park et al. (2020) note, predicting the accuracy and consistency of examinee scores by a model is a measure of reliability. CDM studies in earlier years usually

did not report attribute reliability information, but more recent studies (Min & He, 2021; Wu et al., 2020, 2021) start to report such information.

In addition, in more recent years, a few studies (e.g., Effatpanah, 2019; Javidanmehr & Anani Sarab, 2019; Kilgus et al., 2020) used the Q-matrix empirical validation algorithm (de la Torre & Chiu, 2016) which was further available in the GDINA R package (Ma & de la Torre, 2020). This offers the possibility of a convenient way to validate the Q-matrix empirically. However, this empirical algorithm can only serve as a supplementary information for Q-matrix validation. As recommended by de la Torre (2008), it is always important to combine the Q-matrix empirical validation results with content knowledge.

Suggestions. Our findings suggest several procedures that should be followed when developing a retrofitted Q-matrix, as well as some procedures that should be used only with caution. Our primary recommendation is that researchers and practitioners should consider perspectives from both construct theory and statistical analysis. The process of developing a retrofitted Q-matrix should always include subject matter experts who know both theory and content of the test construct (Rupp et al., 2010). Second, the set of attributes developed should be based on the principle of parsimony where highly correlated attributes may be combined (Buck & Tatsuoka, 1998). Finally, more than one method needs to be used to develop the Q-matrix so that evidence from multiple sources can be combined to strengthen the validity of the Q-matrix (Li & Suen, 2013).

Once the Q-matrix has been developed statistical testing needs to be done to verify the appropriateness of the matrix. This can be done by testing for model fit and reliability using actual data, using both absolute and relative fit indices to compare different models (Lei & Li, 2016) to select the best fitting one. Also, researchers should test for reliability by evaluating both classification accuracy and classification consistency of the whole latent class pattern for examinee responses (Park et al., 2020). An empirical validation algorithm (e.g., de la Torre and Chiu's (2016) empirical validation model for DINA) could also be used.

We recommend against comparing results from CDM models with IRT or CTT models, because they are based on different theory and are not strictly comparable. Results from such comparisons may not be meaningful. For optimal model fit, given sufficient sample size, we recommend starting the analysis with a saturated CDM and examining the significance of the main effects and interaction effects (if any), before considering specific smaller CDMs with particular assumptions on the relationship between items and attributes (Hemati & Baghaei, 2020).

Limitations. A major limitation of this review is that we only included journal articles published in English. Adding dissertations, conference presentations, and articles published in other languages has the potential of opening up more methods of Q-matrix development and validation, as well as insights into the CDM applications. These are areas for continued work. Furthermore, some CDM studies did not provide details about their Q-matrix development and validation procedures so that we were not able to code such information for every study included in the review. We, therefore, call for a detailed report of Q-matrix development and validation procedures in future CDM application studies.

Notwithstanding these limitations, this review contributes to the research into Q-matrix and CDM applications by highlighting the present state of Q-matrix development and validation, some of the possible tools for the process, and the need to use multiple methods in developing and validating Q-matrices.

## References

Birenbaum, M., & Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students' knowledge states in multiplication and division with exponents. *Applied Measurement in Education, 6*, 255–268. https://doi.org/10.1207/s15324818ame0604_1

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119–157. https://doi.org/10.1191/026553298667688289

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273. https://doi.org/10.1007/s11336-015-9467-8

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 979–1030). Elsevier. https://doi.org/10.1016/S0169-7161(06)26031-0

Effatpanah, F. (2019). Application of cognitive diagnostic models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing, 9*, 1–28. https://www.ijlt.ir/?_action=article&au=801055&_au=Effatpanah,%20Farshad

Gorin, J. S. (2009). Diagnostic classification models: Are they necessary? Commentary on Rupp and Templin (2008). *Measurement: Interdisciplinary Research and Perspectives, 7*, 30–33. https://doi.org/10.1080/15366360802715387

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 1031–1038). Elsevier. https://doi.org/10.1016/S0169-7161(06)26040-1

Hemati, S. J., & Baghaei, P. (2020). A cognitive diagnostic modeling analysis of the English Reading Comprehension section of the Iranian National University Entrance Examination. *International Journal of Language Testing, 10*, 11–32. https://eric.ed.gov/?id=EJ1291043

Jang, E. E., Dunlop, M., Wagner, M., Kim, Y. H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning, 63*, 400–436. https://doi.org/10.1111/lang.12016

Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing, 32*, 359–383. https://doi.org/10.1177/0265532215570924

Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high stakes reading comprehension test. *Language Assessment Quarterly, 16*, 294–311. https://doi.org/10.1080/15434303.2019.1654479

Kilgus, S. P., Bonifay, W. E., Eklund, K., von der Embse, N. P., Peet, C., Izumi, J., Shim, H., & Meyer, L. N. (2020). Development and validation of the Intervention Skills Profile–Skills: A brief measure of student social-emotional and academic enabling skills. *Journal of School Psychology, 83*, 66–88. https://doi.org/10.1016/j.jsp.2020.10.001

Kim, H. (2014). Application of cognitive diagnostic model for achievement profile analysis. *KAERA Research Forum, 1*(1), 15–25.

Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement, 40*, 405–417. https://doi.org/10.1177/0146621616647954

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*(1), 1–25. https://doi.org/10.1080/10627197.2013.761522

Liu, R., & Shi, D. (2020). Using diagnostic classification models in psychological rating scales. *The Quantitative Methods for Psychology, 16*, 442–456. https://doi.org/10.20982/tqmp.16.5.p442

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*, 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*(14), 1–26. https://doi.org/10.18637/jss.v093.i14

Min, S., & He, L. (2021). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. *Language Testing., 39*(1), 90–116. https://doi.org/10.18637/jss.v093.i14

Park, Y. S., Morales, A., Ross, L., & Paniagua, M. (2020). Reporting subscore profiles using diagnostic classification models in health professions education. *Evaluation & the Health Professions, 43*, 149–158. https://doi.org/10.1177/0163278719871090

Ravand, H., Baghaei, P., & Doebler, P. (2020). Examining parameter invariance in a general diagnostic classification model. *Frontiers in Psychology, 10*, 2930. https://doi.org/10.3389/fpsyg.2019.02930

Rupp, A. A., & Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96. https://doi.org/10.1177/0013164407301545

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

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*, 1–17. https://doi.org/10.1080/15366367.2018.1435104

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Wu, X., Wu, R., Chang, H.-H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA Mathematics Achievement Test based on cognitive diagnostic models. *Frontiers in Psychology, 11*, 1–13. https://doi.org/10.3389/fpsyg.2020.02230

Wu, X., Zhang, Y., Wu, R., & Chang, H. H. (2021). A comparative study on cognitive diagnostic assessment of mathematical key competencies and learning trajectories. *Current Psychology*, 1–13. https://doi.org/10.1007/s12144-020-01230-0