

Chapter 18

Mastery Classification of Diagnostic Classification Models

Yuehmei Chien, Ning Yan, and Chingwei D. Shin

Abstract The purpose of diagnostic classification models (DCMs) is to determine mastery or non-mastery of a set of attributes or skills. There are two statistics directly obtained from DCMs that can be used for mastery classification—the posterior marginal probabilities for attributes and the posterior probability for attribute profile.

When using the posterior marginal probabilities for mastery classification, a threshold of a probability is required to determine the mastery or non-mastery status for each attribute. It is not uncommon that a 0.5 threshold is adopted in real assessment for binary classification. However, 0.5 might not be the best choice in some cases. Therefore, a simulation-based threshold approach is proposed to evaluate several possible thresholds and even determine the optimal threshold. In addition to non-mastery and mastery, another category called the indifference region, for those probabilities around 0.5, seems justifiable. However, use of the indifference region category should be used with caution because there may not be any response vector falling in the indifference region based on the item parameters of the test.

Another statistic used for mastery classification is the posterior probability for attribute profile, which is more straightforward than the posterior marginal probability. However, it also has an issue—multiple-maximum—when a test is not well designed. The practitioners and the stakeholders of testing programs should be aware of the existence of the two potential issues when the DCMs are used for the mastery classification purpose.

Y. Chien (✉) • C.D. Shin
Pearson, 2510 N Dodge Street, Iowa City, IA 52245, USA
e-mail: yuehmei.chien@pearson.com; david.shin@pearson.com

N. Yan
59-3-406 Southwest Residential Village, Tianjin, China
e-mail: ning.now@gmail.com

18.1 Introduction

The diagnostic classification models (DCM) are latent variable models for cognitive diagnosis, which assumes the latent classes (i.e., mastery or non-mastery of particular skills/attributes/knowledge components) can be represented by binary latent variables. Recently, DCM has drawn much attention of the practitioners because of its promising use in aligning teaching, learning, and assessment. DCMs aim to determine mastery or non-mastery of a set of attributes or skills, or to provide timely diagnostic feedback by knowing students' weaknesses and strengths to guide teaching and learning. In particular, the use of DCM in formative assessments in classroom has been increasing quickly.

The use of DCM is twofold regarding what can be obtained from the model and provided to individual students: the strength and weakness profiles based on estimated attribute mastery probabilities for each attribute, and the classification of mastery or non-mastery based on estimated profile probabilities. For example, a set of estimated attribute mastery probabilities for three skills—0.92, 0.41, and 0.22—indicates the student is strong on Skill 1, but may require some additional learning or practice on the other two skills, especially for Skill 3.

For the mastery classification, there are two statistics obtained from DCM that can be used to determine the mastery or non-mastery status for each attribute. The first statistic is *the posterior probability for attribute profile*. Using DCM, the posterior probabilities for all possible attribute profiles are obtainable and the attribute profile can be the profile with the maximum posterior probability. This estimation method is the maximum likelihood estimation (MLE) or maximum a posteriori (MAP) if a prior applied multiplies the likelihood function. For ease of reference, the method to obtain the mastery classification is referred to as the MLE profile estimation.

Another way to obtain the mastery classification is based on different statistics obtained from DCM, that is *the posterior marginal probabilities for attributes*. To obtain the classification results, a threshold or a cut-off of a probability must be predefined and then used to determine the mastery or non-mastery status. It is not uncommon that a 0.5 threshold is adopted in real assessment. Using 0.5 as a threshold, the previous example has classification [1, 0, 0], where 1 indicates mastery and 0 indicates non-mastery. Similarly, for ease of reference, this method to obtain the mastery classification is referred to as the threshold approach.

In this paper, the focus is on estimation of mastery classification. For classification using the posterior marginal probabilities for attributes, two issues were addressed. First, for binary classification, a simulation-based approach is suggested to evaluate the different thresholds. Second, for the indifference region, in addition to binary classification, evidence demonstrates that examining the values of posterior marginal probabilities for different response patterns or total scores is rational and necessary because there may not have any probability falling in the indifference region. For classification using the posterior probability for attribute profile, the

issue of the multiple maximums on the likelihood in the MLE profile estimation is addressed. Prior to mentioning those focused aspects, DCMs are briefly introduced. Some discussions are also provided at the end of this paper.

18.2 Models

In the literature, there are many cognitive diagnostic models including the rule space model (Tatsuoka 1983), the Bayesian inference network (Mislevy et al. 1999), and the fusion model (Hartz 2002; Hartz et al. 2002), the deterministic inputs, noisy “and” gate (DINA) model (Doignon and Falmagne 1999; Haertel 1989; Junker and Sijtsma 2001), the Deterministic Input, Noisy “Or” Gate (DINO) model (Templin and Henson 2006), the generalized deterministic inputs, noisy “and” gate (G-DINA) model (de la Torre 2011), the log-linear CDM (Henson et al. 2009), and the general diagnostic model (GDM; von Davier 2005). (See more detailed information for various DCMs from Rupp et al. 2010.)

Among those models, DINA and DINO are popular models for educational assessment and for psychological tests, respectively, due to their simplicity. DINA is a noncompensatory model, which assumes the deficiency on one attribute cannot be compensated by the mastery of other attributes. DINA models the probability of a correct response as a function of a slipping parameter for the mastery latent class and as guessing for the non-mastery latent class. On the contrary, the DINO model is a compensatory model, which assumes the deficiency in one attribute can be compensated by the mastery of other attributes.

18.3 The Threshold Approach

To obtain the mastery classification from DCM, the most used approach is the threshold approach (e.g., Hartz 2002; Jang 2005). In practice, the classification of mastery or non-mastery of each attribute is determined by applying cut-offs on the posterior marginal probabilities for attributes. When a binary classification is desired, a convention/intuitive threshold 0.5 is commonly used as the threshold to obtain the mastery (≥ 0.5) and non-mastery states (< 0.5) for each attributes (e.g., DeCarlo 2011). A threshold of 0.5 is statistically sound and a possible optimal threshold in many cases when the classification is binary. However, depending on the Q-matrix structure and the item quality (i.e., the discrimination power), 0.5 might not be the best choice in some cases. Therefore, using a simulation to examine the distribution of the posterior marginal probabilities for attributes and then evaluating several possible thresholds is important for the binary classification.

The simulation-based approach first applies a set of cut-offs, for example, from 0.5 to 0.6 by 0.01. Then the best cut-off for each attribute that results in the largest attribute classification accuracy for each attribute can be obtained. It is possible that different attributes have different cut-offs.

Table 18.1 The classification accuracy using 0.5 vs. the optimal set of cut-offs

	Cut-off = 0.5 for all attributes	Cut-offs = 0.5, 0.51, 0.56, 0.59, 0.6, 0.6
Profile	41.4 %	43.7 %
Attributes	79.4 %, 81.3 %, 84.8 %, 86.4 %, 83.5 %, 85 %	79.4 %, 81.4 %, 85.5 %, 87.1 %, 84.8 %, 87.6 %

Table 18.1 is an example showing the difference of using the convention threshold and using the best cut-offs obtained from the simulation, which is 0.5, 0.51, 0.56, 0.59, 0.6, and 0.6, respectively. The overall profile classification accuracy increased from 41.4 to 43.7 % and the attribute mastery classification accuracy also slightly increased for the attributes 2 to 6.

Note that the largest classification accuracy obtained in the simulation, given a specific cut-off for an attribute, is not population invariant, which means the optimal cut-off obtained might be varied for different populations that are composed of different proportions of student in each of the profiles. Therefore, it is important that the population of simulees can be drawn from an empirical population that represents the real population closely.

A common alternative method to classify students, instead of using binary classification with large uncertainty around the cut-off, is to allow an indifference region aside from the mastery and non-mastery. An indifference region found in the literature is defined between 0.4 and 0.6 (e.g., Hartz 2002; Jang 2005). However, we suggest that the indifference region is defined carefully. Figure 18.1 shows a histogram of the posterior marginal probabilities for an attribute for 2700 students under different test lengths, where n in figure indicates the test length. The original data set contains 8 items measuring one attribute (as shown as $n = 8$ in the figure). The slip parameters are between 0.06 and 0.12 and the guessing parameters are between 0.20 and 0.25 for those 8 items. Because the test length is 8, the possible total scores are 0, 1, 2, 3, 4, 5, 6, 7, or 8. Three hundred students' responses are generated for each of the nine different total scores. To evaluate with shorter test lengths, the posterior marginal probabilities are re-estimated with only the first n items, where $n = 3$ to 7. In total, six different test lengths were evaluated and the results were represented in Fig. 18.1.

For test length = 8, only 21 students fall into the indifference region as defined between 0.4 and 0.6. Note that, in the data set of test length = 8, there are 300 students with a total score 4 and another 300 students with a total score 5, which are the tests with larger measurement error if classification decisions are made. Also, note only test lengths 5 or 8 have some students falling in the indifference region, while other four test lengths have none. Defining an indifference region on the posterior marginal probabilities of attributes and using it to classify students might not obtain the desired results. To further examine the posterior marginal probability for the total scores of 4 and 5, a scatter plot is created, as shown in Fig. 18.2; the total score = 4 all have very low probability values that are obviously classified

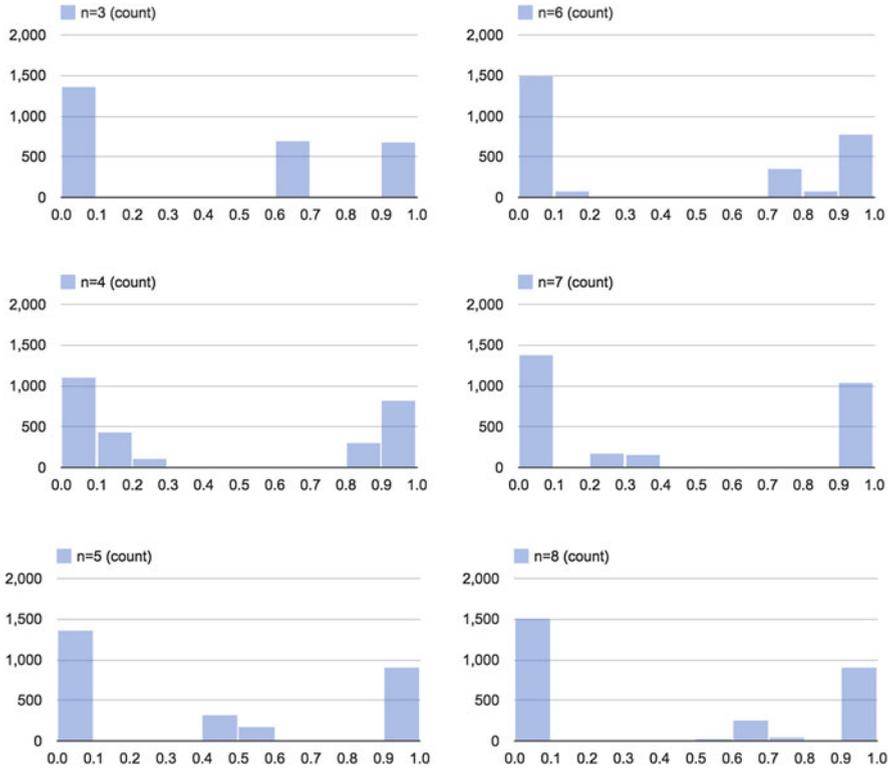


Fig. 18.1 A histogram of the posterior marginal probabilities for an attribute for 2700 students under different test lengths

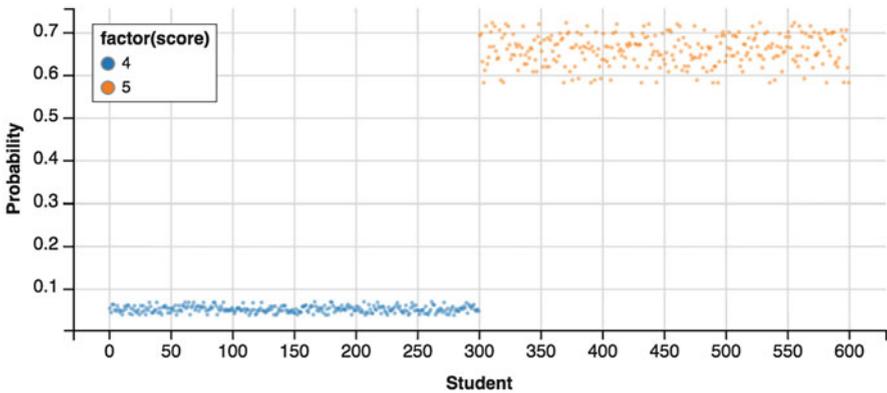


Fig. 18.2 A scatter plot of the posterior marginal probability for the total scores of 4 and 5

as non-mastery, while the total score = 5 have probability values around 0.57 to 0.73 that may be classified as indifference region. This example shows setting up an indifference region might not be straightforward and examining the posterior marginal probability given different responses patterns and different total scores are critical. Indeed, more research is necessary in this area.

18.4 Multiple Maxima (Ties in Posterior Probability)

18.4.1 *The Paradox in the Fraction Subtraction Data*

The well-known fraction-subtraction (FS) data set was collected by Dr. Kikumi Tatsuoka in 1984. Curtis Tatsuoka released the data in 2002 and made it publicly available at [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1467-9876/homepage/fractionsdata.txt](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9876/homepage/fractionsdata.txt). The FS data set contains responses to twenty fraction subtraction test items from 536 middle school students. This test measures eight fine-grained attributes in the domain of fraction subtraction, which includes—(1) convert a whole number to a fraction; (2) separate a whole number from a fraction; (3) simplify before subtracting; (4) find a common denominator; (5) borrow from whole number part; (6) column borrow to subtract the second numerator from the first; (7) subtract numerators; (8) and reduce answers to simplest form. The Q-matrix, which specifies which attributes are measured by each item, is listed in Table 18.2. With eight attributes, the maximum number of latent classes is two hundreds and fifty six, without considering whether some combinations are unlikely such as mastery of “borrow from whole number part” without mastery of “subtract numerators”.

Figure 18.3 shows the likelihood of those 256 latent classes for a student with a total score of 4. It clearly shows there are four latent classes with exactly the same posterior probability. Figure 18.4 demonstrates a more extreme example with a total score of zero, where there are sixty-four latent classes having exactly the same posterior probability. The first latent class and the last latent class among those sixty four are “00000000” and “10111101”, respectively. As mentioned previously, DINA is a conjunctive model that requires all skills measured are mastered to be able to answer an item correctly besides guessing. Therefore, for an incorrect response, depending on the number of attributes measured, DINA may not be able to statistically provide useful information about the state of mastery or non-mastery. In the FS data, items 6, 8, and 9 are simple items, which only measure one attribute, Attribute 7, Attribute 7, and Attribute 2, respectively. The rest of items are complex items measuring more than one attribute. For the all-zero responses in the FS data set, only items 6, 8, and 9 can provide information about the high chance of being non-mastery for attributes 2 and 7; therefore, the mastery status is non-mastery for attributes 2 and 7 while half-half chance for the rest of six attributes.

Table 18.2 The Q-matrix of the FS data

Item\Attribute	1	2	3	4	5	6	7	8
1	0	0	0	1	0	1	1	0
2	0	0	0	1	0	0	1	0
3	0	0	0	1	0	0	1	0
4	0	1	1	0	1	0	1	0
5	0	1	0	1	0	0	1	1
6	0	0	0	0	0	0	1	0
7	1	1	0	0	0	0	1	0
8	0	0	0	0	0	0	1	0
9	0	1	0	0	0	0	0	0
10	0	1	0	0	1	0	1	1
11	0	1	0	0	1	0	1	0
12	0	0	0	0	0	0	1	1
13	0	1	0	1	1	0	1	0
14	0	1	0	0	0	0	1	0
15	1	0	0	0	0	0	1	0
16	0	1	0	0	0	0	1	0
17	0	1	0	0	1	0	1	0
18	0	1	0	0	1	1	1	0
19	1	1	1	0	1	0	1	0
20	0	1	1	0	1	0	1	0

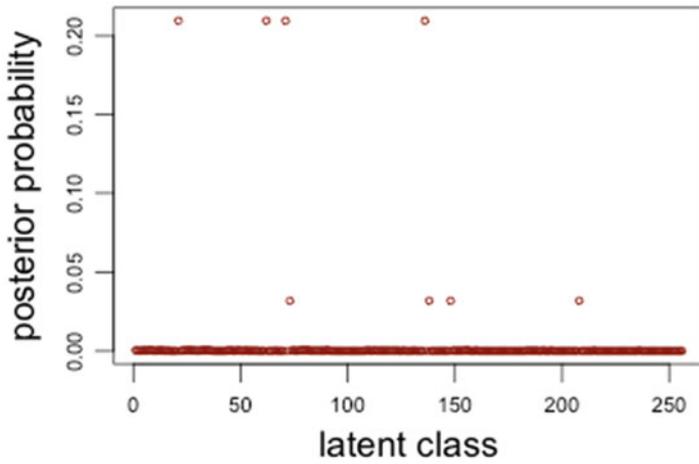


Fig. 18.3 A response vector with a total score of four

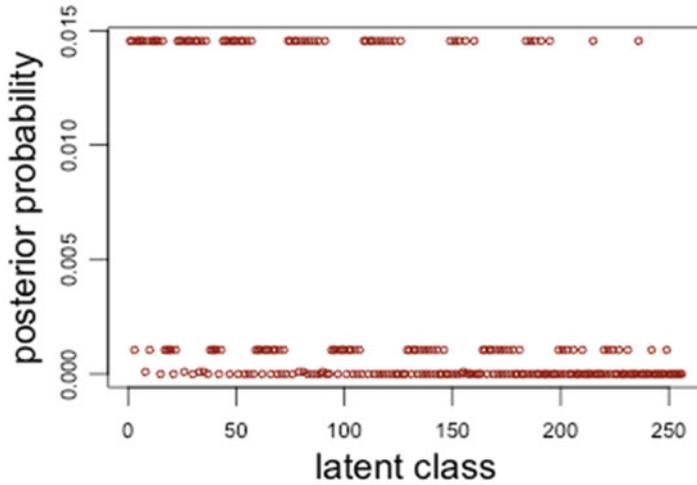


Fig. 18.4 A response vector with a total score of zero

18.4.2 *Q*-score (*The Ideal Response*)

In item response theory, the possible local maximum on likelihood of a continuum latent ability scale is a well-known issue of using the MLE. That means the estimated parameter is not universally best but only in some cases. In other words, the solution found on likelihood is not a real solution. Similarly in DCM, there is a profile estimation issue called multiple-maximum caused by using MLE. That is, given the observed responses on a test, there may be multiple latent classes with exactly the same highest probability. This multiple maxima issue has not been explicitly described in the literature, but is mentioned as a parameter-identification problem (e.g., Zhang 2014). Because of the existence of multiple maximum, the profile estimate using DCMs is not always identifiable for some diagnostic tests when the *Q*-matrix or the test is not well designed.

Depending on the structure of the *Q*-matrix, two different mastery profiles over a set of latent classes could be equivalent; these two mastery profiles generate exactly the same probability distribution of item response patterns, so that they cannot be distinguished on the basis of item response data. To identify this equivalence relationship from the *Q*-matrix, a simple method is proposed. First, a *Q*-score is defined as the most likely observed score on the item for a respondent with the given latent class; i.e., *Q*-score, is the true score for the items given the latent classes. Then, by examining whether there are any two latent classes with the same *Q*-score, the possible existence of a multiple-maximum can be known.

Table 18.3 A Q-matrix of three attributes for four items

Item\Attribute	1	2	3
1	1	1	0
2	1	0	1
3	0	1	1
4	1	1	1

Table 18.4 A Q-score using the condensation rule of the DINA model

	1	2	3	4
000	0	0	0	0
100	0	0	0	0
010	0	0	0	0
001	0	0	0	0
110	1	0	0	0
101	0	1	0	0
011	0	0	1	0
111	1	1	1	1

Table 18.5 A Q-score using the condensation rule of the DINO model

	1	2	3	4
000	0	0	0	0
100	1	1	0	1
010	1	0	1	1
001	0	1	1	1
110	1	1	1	1
101	1	1	1	1
011	1	1	1	1
111	1	1	1	1

The following is a simple example with four items and three attributes (see Table 18.3) to demonstrate the use of the Q-score for finding the possible existence of the multiple-maxima in the mastery profile estimates. Table 18.4 lists the Q-score under the conjunctive assumption of the DINA model. The first four latent classes, or the mastery profiles, all generate exactly the same Q-scores. In other words, a respondent who has an observed total score of zero is equally likely to belong to any of the first four latent classes.

The Q-score rule can be applied to any DCM that has one Q-matrix with a clearly defined condensation rule to specify the relationship between the correct response of each item and the attributes measured by the item. Table 18.5 lists the Q-score of the same four-item test, but using the condensation rule of the DINO model. The last four latent classes, or the mastery profiles, all generate exactly the same Q-scores

under the condensation rule of the DINO model. Therefore, a respondent who has answered the four items correctly is equally likely to belong to any of the last four latent classes.

18.5 Discussion and Future Research

It is statistically sound that the attribute is classified into “mastery” if $p > 0.5$, and “non-mastery” otherwise, where p is the posterior marginal probability of mastery for the attribute. However, because of the complexity of the model used and the structure of Q-matrix, it is suggested that the different threshold values should be used to examine the possible effect on classification. If a sample of population can be obtained and represents the population well, a simulation approach can be used to find a set of optimal thresholds for attributes. To emphasize the importance of this issue, Fig. 18.5 shows the posterior marginal probabilities for those eight attributes, where many posterior marginal probabilities are surround 0.5, and the convention threshold .5 definitely is not a good choice. One might argue that the FS data is not perfect; and yes, the test design regarding whether a complex Q-matrix was employed by the FS test is flawed and therefore, it seriously suffered from uncertainty of the classification and from the multiple-maximum. Therefore, it is even more important to examine the distribution of the posterior marginal probability before a cut-off (for non-mastery and mastery) or two cut-offs (for non-mastery, indifference, and mastery) are applied for classification.

Another importance of this paper is to explicitly call the practitioners’ attention to the multiple-maximum issue. The multiple latent classes might cause a misleading mastery classification for either using the posterior probability for attribute profile or using the posterior marginal probability for attributes. As shown by DeCarlo (2011), the attribute probability for a zero score could be as high as 0.985 for a zero score of the FS data using the DINA model. To avoid this multiple-maximum, simple structure items (solely measuring one attribute) should be added to the test (as suggested by DeCarlo 2011) to make a complete Q-matrix (Chiu et al. 2009) during test construction.

However, DCMs might be used to fit existing items by tagging them with associated attributes (e.g., von Davier 2005). Thus, adding simple structure items into the existing test become cumbersome. Furthermore, with the emergence of cognitive diagnostic computerized adaptive testing (CD-CAT; e.g., Cheng 2009), the interim profile estimates must be calculated based on the items administered so far, and the effect of the multiple-maximum on the CD-CAT is worthy of further research.

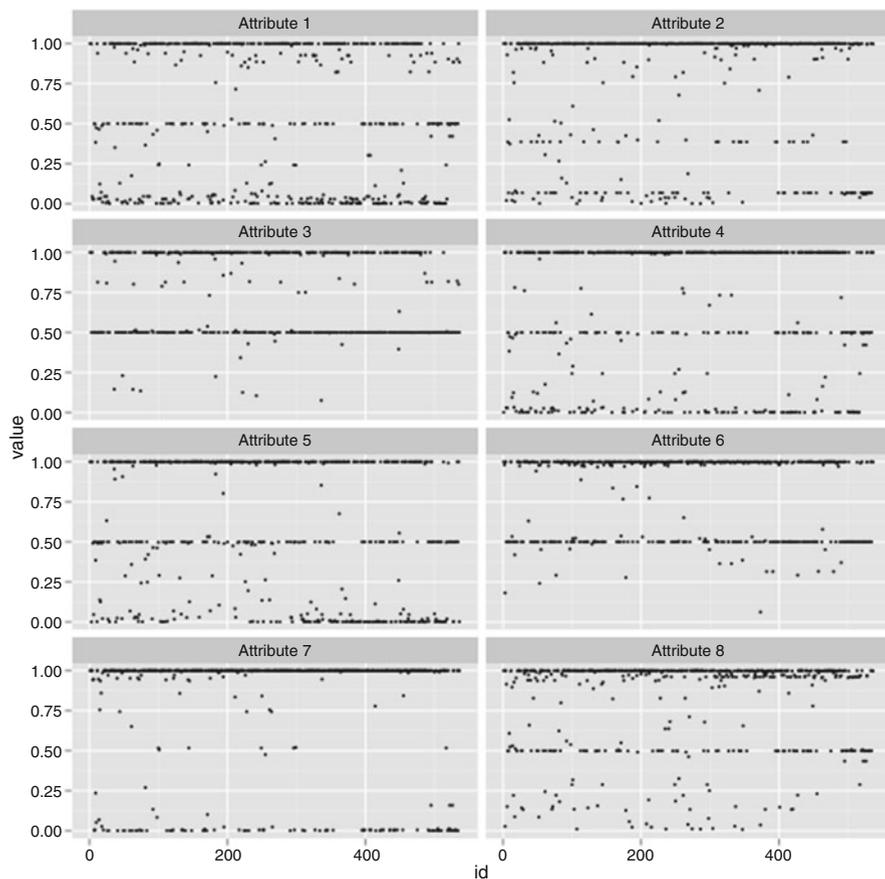


Fig. 18.5 The posterior marginal probabilities for those eight attributes of the FS data

References

- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CDCAT. *Psychometrika*, *74*, 619–632.
- Chiu, C.-Y., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.
- DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.
- Doignon, J. P., & Falmagne, J. C. (1999). *Knowledge spaces*. New York, NY: Springer.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Champaign, IL.

- Hartz, S., Roussos, L., & Stout, W. (2002). *Skills diagnosis: Theory and practice* [User manual for Arpeggio software]. Princeton, NJ: Educational Testing Service.
- 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.
- Jang, E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Doctoral dissertation, University of Illinois at Urbana-Champaign.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.
- Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437–446). San Mateo, CA: Morgan Kaufmann.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and practice*. New York, NY: Guilford.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.
- Templin, J. L., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data*, ETS Research Report RR-05-16. Princeton, NJ: Educational Testing Service. Retrieved from <http://www.ets.org/Media/Research/pdf/RR-05-16.pdf>
- Zhang, S. S. (2014). *Statistical inference and experimental design for Q-matrix based cognitive diagnosis models*. Doctoral dissertation, Columbia University.